UPPSALA
UNIVERSITET

Automated Prediction of the Endocrine Disruptive Potency of Chemicals detected with LC/ESI/HRMS based on Mass Spectral Networks

Master's thesis in Analytical Chemistry

Leonardo Soto

Uppsala 2023

Degree project E in chemistry, 1KB053, 30 credits
Uppsala University
Supervisors: Anneli Kruve, Jeffrey Hawkes, Riin Rebane
Subject Specialist: Daniel Globisch
Examiner: Per Sjöberg

Abstract

The widespread exposure to chemicals has raised concerns about their toxicity impact on public health and the environment. Identifying and quantifying these chemicals in complex samples is not always possible, making the assessment of their toxicities difficult. In an effort to quickly screen chemicals for potential risks to human health, this study aims to predict toxicities based on tandem mass spectrometry MS^2 data. To achieve this goal, endocrine-disrupting activity data and other relevant human endpoints from the Tox21 Challenge were collected and combined with mass spectra from Mass Bank Europe. A *k*-nearest neighbors (*k*-NN) and a spectra network-based algorithm were implemented to predict the activity from MS^2 mass spectra. For *k*-NN, 5-fold cross-validation, the highest recall and precision were 47.1% and 44.4% (both for NR.AR), respectively. The implementation of a spectral similarity network enhanced the overall prediction power, leading to recall and precision of 81.8 % and 75.0% (both for NR.AR), respectively. The spectral networks showed clustering tendencies for the endpoints NR.AR, NR.ER, NR.AR.LBD, and NR.ER.LBD. The approach was applied to retrospective analysis of MS^2 mass spectra of a wastewater sample, showing potential for toxicity alerts. To refine the predictive capabilities of the model, further investigations should focus on feature selection techniques, network optimization, and integration with other domain datasets.

Popular summary

The brain has a great ability to classify information. It organizes and categorizes a vast amount of sensory data based on shared characteristics and similarities. When examining a new object, the brain quickly identifies common features like shape, color, smell, texture, and size and then associates these features with known categories.

At the same time, there is a vast universe of chemicals that is expanding. Some of them have adverse effects to humans and have been categorized as toxic. Among these chemicals, the endocrine disruptors are of particular concern because they interfere with the normal functioning of hormones in the body, potentially leading to health issues.

The identification of these chemicals in samples can be time-consuming and challenging. One of the most sensitive analytical techniques used for this purpose is tandem mass spectrometry. Tandem mass spectrometry, like solving a complex puzzle from small pieces, weights molecules and uses their fragments to infer its identity. Even though the fragments are detected, it is not always possible to fully identify all the chemicals in a sample.

To predict the endocrine disruptive activity of chemicals in a sample, this study mimics the brain classification ability and applies it to chemical toxicity. A digital brain builds a network of known chemical toxicity based on a vast amount of tandem mass spectrometry and toxicity data. Then, this network is used to interpret the fragments of unknowns and to infer their toxicity.

In this way, this study advances the analytical capabilities in chemical toxicity by combining mass spectrometry and a network-based approach. Ultimately, it can aid in the assessment of chemical risk and provide valuable insights for environmental monitoring and public health protection.

Contents

Abstract ii

Popular summary iii

Contents iv

Abbreviations vi

1 Introduction 1

2 Background 3

3 Aims of the project 9

4 Methodology 10

5 Results 13
Toxicity dataset 13
Spectra in MassBank library 14
Mass spectra similarity 15
Spectral similarity networks 18
k-NN algorithm 20
k-NN algorithm in locally connected spectra 21
Retrospective analysis of MS^2 features 21

6 Discussion 26
Toxicity dataset 26
Spectra in MassBank library 26
Spectra similarity 27
Mapping cosine 27
Pairwise comparison 28
Cosine budget 29
Deep learning-based similarity metric 35
Spectral similarity networks 36

k-NN algorithm 43
k-NN algorithm in locally connected spectra 44
Factors influencing the predictions 45

7 Conclusions and future perspectives 47

References 48

Appendix 55
A Distribution of endpoint labels in the toxicity dataset 55
B Cosine similarity and MS2DeepScore 56
C *k*-NN cross-validation 58
D Sample MS^2 features 59
E Spectral similarity networks with labels for all the endpoints . 61

Abbreviations

DDA	Data Dependent Acquisition
DIA	Data Independent Acquisition
ESI	Electrospray Ionization
GC	Gas Chromatography
GNPS	Global Natural Products Social Molecular Networking
HRMS	High Resolution Mass Spectrometry
LC	Liquid Chromatography
MS	Mass Spectrometry
MS^2	Tandem Mass Spectrometry
QSAR	Quantitative Structure-Activity Relationship
REACH	Regulation for Registration, Evaluation, Authorization and Restriction of Chemicals
SAR	Structure Activity Relationship

1 Introduction

The increasing production of chemicals has brought numerous benefits to society, from life-saving medicines to essential materials and components. However, the production and commercialization of new chemicals have raised concerns about their potential harmful effects on human health and the environment [1]. Therefore, regulations have been put in place to ensure their safe use. For instance, the European Union's Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) aims to ensure the safe production, commercialization, and use of chemicals by assessing their risks to human health and the environment. The database have listed thousands of chemicals with their properties and toxicity information [2]. Nevertheless, the understanding of the effects of these chemicals is limited, and research efforts are ongoing to uncover their distribution and potential risks [3][4][5].

Efforts to characterize the toxicity of chemicals have traditionally relied on animal testing and epidemiological studies, which are time-consuming, expensive, and often raise ethical concerns. Emerging technologies, such as high-throughput in-vitro testing, high-content data analysis (omics, image analysis), and bioinformatics, are transforming the traditional approach towards a mechanistic understanding of toxicity [6]. However, these approaches mostly rely on the identified compounds and toxicity in complex mixtures are more difficult to assess.

Mass Spectrometry (MS) is an analytical technique that can identify and quantify chemicals in complex matrices. Regarding toxic chemicals, MS can provide a more comprehensive understanding of their distribution and fate in the environment [7]. However, identifying and quantifying all toxic chemicals in a sample is indeed a challenging task [8] due to the complexity of sample matrices, the wide range of chemical substances involved, and the need for suitable chemical standards.

At the same time, several large mass spectra databases have been expanding their repositories, primarily through collaborative efforts, over the past decade (e.g., MassBank Europe[9], MoNA[10], HMDB[11], GNPS[12]). These databases offer opportunities for data mining and the exploration of prediction models. Very recently, it has been proposed that toxicity relevant information

can also be gained from the mass spectra collected in chemicals analysis of complex samples [13][14].

The lack of comprehensive toxicological data for all chemicals compounds and the difficulty to identify all chemicals in environmental samples challenges the assessing of potential risks [15][16]. To address these issues, this study seeks to directly predict the endocrine activity of detected chemicals using LC-ESI-HRMS data. By utilizing this approach, it is hoped that in the future, prioritizing toxic samples for further analysis will become more efficient, while extensive screening of contaminated environmental areas or commercialized products will become more feasible. Ultimately, such efforts will help to identify and mitigate potential risks to human health and the environment.

2 Background

Toxicity prediction

Computational models and data analysis techniques can be used to predict the toxicity of chemicals based on their structural properties, activity profiles, and other relevant factors. Some approaches include Quantitative Structure-Activity Relationship (QSAR) models and machine learning algorithms. These models use data on chemical structures, biological activity, physicochemical properties, and toxicological endpoints to establish relationships and make predictions about the potential toxicity of new or untested chemicals.

QSAR models, for example, employ statistical and mathematical techniques to correlate chemical descriptors (molecular features) with toxicological data. They assume that chemicals with similar structural and physicochemical properties are likely to exhibit similar toxicological effects. Machine learning algorithms, on the other hand, learn patterns and relationships from training data to make predictions on new, unseen chemicals. These algorithms can handle large datasets and capture complex relationships between chemical features and toxicity. Also deep learning techniques, such as neural networks, have shown promise in capturing intricate patterns in chemical data and achieving high predictive performance [17][18].

Despite these advancements, challenges remain in toxicity prediction. Some major challenges are the availability and quality of comprehensive toxicity data [19] and the multi-factorial nature of toxicity that involves interactions between the chemical, biological systems, and environmental factors. Capturing the complex relationships and underlying mechanisms challenges predictions. Furthermore, incorporation of toxicity prediction methods into regulatory frameworks is an ongoing process [20][21] that requires interpretability, and regulatory relevance for their adoption.

Recently, there is a growing focus on the integration of diverse data sources, such as omics data (genomics, proteomics, metabolomics) and high-throughput screening data, to enhance the prediction of toxicity [22]. This multi-dimensional data integration and systems biology can potentially provide a more comprehensive understanding of the molecular mechanisms of toxicity and enable the development of more holistic predictive models.

Read-Across

Testing all the universe of chemicals for toxicity is not feasible. In some cases, there is sufficient evidence to infer toxicity. Read-across is a gap filling technique used for the prediction of toxicity of chemicals based on the available toxicity data of related compounds. This relation might be based on similar structure, properties and/or activities [23]. It leverages the principle that chemicals with similar structural features are likely to exhibit similar biological activities or toxicological properties. In read-across, known toxicity information of a reference compound is extrapolated or "read across" to predict the toxicity of a target compound with similar chemical structure but without available experimental data.

Read-across offers a cost-effective and time-efficient strategy for toxicity prediction, especially when experimental data for a large number of chemicals is limited or unavailable. However, it is important to consider the reliability and relevance of the predictions as in some cases structural similarity alone may not be sufficient to conglomerate chemicals with the same potency [24].

Tox21 Challenge

The Tox21 Challenge was launched in 2014 to advance the toxicology prediction methods. This challenge is a collaborative research initiative aimed at advancing toxicology testing methods and predicting the potential toxicity of chemical compounds using high-throughput screening approaches. The challenge involves the evaluation of around ten thousand chemicals for their response of twelve toxicity endpoints for the panels of nuclear receptor (NR) and stress response (SR) pathways. All the endpoints are summarized in Table 2.1. Research teams develop and apply computational models, such as machine learning algorithms, to predict the activity of these chemicals on various toxicity endpoints. The high-performing models achieved prediction accuracies comparable to experimental errors, highlighting the potential of these models as screening tools for chemical prioritization [25]. The Tox21 Challenge serves as a platform for the development and validation of innovative computational methods for toxicity prediction.

Tandem mass spectra

Tandem mass spectrometry (MS^2) can provide insights into the structural information of unknown compounds in a sample by analyzing their fragmentation patterns. The annotation process consists of comparing the experimental MS^2 spectra with reference databases or spectral libraries. Spectral matches are based on the similarity of the fragmentation patterns by examining the mass-to-charge ratios and relative intensities of the fragment ions.

Table 2.1: Endpoints of the Tox21 Challenge

Panel	Abbreviation	Description
Nuclear Receptor	NR.AhR	Activation of the aryl hydrocarbon receptor
	NR.AR	Activation of the androgen receptor
	NR.AR.LBD	Binding to the ligand-binding domain of the androgen receptor
	NR.Aromatase	Inhibition of aromatase activity
	NR.ER	Activation of the estrogen receptor
	NR.ER.LBD	Binding to the ligand-binding domain of the estrogen receptor
	NR.PPAR.gamma	Peroxisome Proliferator-Activated Receptor Gamma
Stress Response	SR.ARE	Activation of the antioxidant response element
	SR.ATAD5	Inhibition of ATPase family AAA domain-containing protein 5
	SR.HSE	Activation of the heat shock response element
	SR.MMP	Changing the Mitochondrial Membrane Potential
	SR.p53	Activation of the p53 tumor suppressor pathway

Some of challenges in annotation rely on the spectra complexity and the data availability. For example, the presence of adducts, isotopic effects, and other sources of fragmentation can make it difficult to confidently assign a specific molecular formula or compound to a feature. Even with the availability of databases and spectral matching algorithms, accurate annotation can still be challenging, especially for compounds that have not been well-characterized or are not present in the databases. This is the case of the analysis of complex mixtures such as environmental samples or biological fluids. These samples often contain numerous compounds with varying levels of abundance and structural diversity. In such cases, the presence of overlapping peaks and co-elution of compounds can further complicate the annotation process, making it difficult to distinguish and assign individual features to specific compounds.

To improve the accuracy of feature annotation in MS^2 analysis, efforts are being made to enhance spectral libraries [26][27], assign classes [28], develop better algorithms for spectral matching [29][30], and integrate complementary information from other analytical techniques [31]. Recent approaches involves the utilization of multi-layer networks [32], which aim to incorporate knowledge-based networks into mass spectra networks. This integration enhances the annotation and understanding of chemicals within their context, such as metabolic pathways. These advancements aim to overcome the challenges associated with feature annotation in tandem mass spectrometry and enable more reliable and comprehensive compound identification in complex samples. However, it is not always possible to fully annotate all the mass spectra and only a comparison can be made base on their similarity.

Similarity scores

The similarity between mass spectra is the degree to which two or more mass spectra are alike. This is achieved by comparing the patterns and intensities of peaks in the spectra, the presence or absence ions or neutral losses. Some similarity metrics for mass spectra include cosine similarity, modified cosine similarity, and neutral loss similarity.

Similarity scores play an crucial role for library matching and allow the annotation of small molecules in MS studies, e.g., metabolomics [33], environment [34][35]. Several approaches to similarity can be found in literature. Some common similarity measures are cosine [36], modified cosine, and neutral loss [37].

The cosine similarity measures the cosine of the angle between two vectors and provides a similarity score ranging from -1 to 1. The cosine similarity, $\cos(\theta)$, is shown in Equation (2.1), where $\mathbf{A}$ and $\mathbf{B}$ represent the vectors corresponding to two spectra forming a θ angle.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{2.1}$$

Additionally, machine learning approaches have been employed to predict spectral similarities and aid in the annotation process. Neural networks and deep learning approaches have been studied, e.g. GLEAMS [38], DLEAMSE [39], MS2DeepScore [29], MS2Query [30]. Huber et al. (2021) [29] proposed a Siamese neural network trained with spectra from Global Natural Products Social Molecular Networking (GNPS) that outperforms the classical spectral similarity measures, with root mean squared error about 0.15. This score has the potential to be used in library matching and for clustering similar spectra. Combination of scores for the same pairs of spectra and removing outliers for IQR led to an improvement on the Tanimoto score predictions [29].

However, some limitations of deep learning approaches include their dependency on training data and the lack of explainability [19]. Deep learning models typically require extensive datasets for training, and the quality and diversity of the data can greatly impact their effectiveness. Due to their complex architectures and intricate computations, it can also be challenging to understand and interpret the reasoning behind the resulting scores.

Toxicity prediction from MS data

Some approaches have been described to predict toxicity from mass spectrometry data for GC-MS and LC-HRMS. GC-MS data benefits from technique standardization, such as common electron ionization (EI) energies and extensive mass spectra libraries. At the same time, collaborative projects, such as

GNPS (Wang et al., 2016) [12], are now providing a rich source of mass spectra from LC-MS.

Zushi (2022) [40] utilized GC-MS data to predict physicochemical properties and toxicities employing random forest, deep neural networks, and XGBoost-based models. The input variables comprised retention times and mass spectra obtained through electron ionization. The random forest and XGBoost-based models exhibited high accuracy, with root mean square errors (RMSEs) of 1.1 and 0.74, respectively, for log(LD50, mouse, oral) predictions. However, the neural network demonstrated comparatively lower accuracy and could be further optimized to enhance performance. In a different study, Peets et al. (2022) [14] employed LC-HRMS data to predict lethal concentration (LC50) and effective concentration (EC50) values for fish in static and flow-through exposures and for water flea and algae toxicity. With extreme gradient boosting Dropouts Additive Regression Trees (xgbDART) a promising performance with a root mean square error (RMSE) below 0.89 was achieved.

These efforts showcase the potential and promising opportunities for toxicity prediction. While significant progress has been made, it is important to note that certain endpoints, particularly those related to endocrine activity, have not yet been thoroughly investigated. Exploring these more human-specific endpoints holds tremendous value and contributes to research on the application of mass spectrometry to toxicity prediction.

Spectral similarity networking

Spectral similarity networking is an approach that allows to cluster mass spectra and find similar structural chemicals [41] [42]. It is an important tool for the study of mass spectra in non-targeted analysis. Several studies have used spectral similarity networking for the annotation of mass spectra in areas including metabolomics [43] and environmental screening [44][45]. Spectral similarity networking shows a great potential for a better understanding of complex samples, and it can also be explored for its application in toxicity prediction from MS data.

In a spectral similarity network, normally the nodes represent mass spectra and the edges the intensity of the similarity. The construction of a network can be influenced by several factors, which includes those described below. Additional attributes can also be depicted in the network, e.g. number of peaks, chemical class.

Similarity metric: It quantifies the similarity between different mass spectra to assess the structural similarity or relatedness of compounds. A common metric for this purpose is cosine similarity. Cosine similarity measures the cosine of the angle between two vectors and ranges from -1 to 1.

Threshold for Minimum Similarity: It determines the minimum level of

similarity required for two spectra to be considered connected in the network. Spectra with cosine similarity scores below this threshold were not included as edges in the network. This threshold helps in controlling the density of the network and ensures that only sufficiently similar spectra are connected.

Minimum Number of Matched Peaks: It ensures that there is sufficient overlap in the peaks between spectra to establish a meaningful connection.

Maximum Number of Edges from a Node: This parameter restricts the number of connections from a node, preventing excessive clustering and improving the readability of the network.

3 Aims of the project

The production and commercialization of chemicals have raised concerns regarding their potential impact on public health and the environment. Endocrine disruptors have been identified as a significant concern due to their ability to interfere with the normal functioning of the endocrine system, which regulates hormones in the body. To assess their toxicity, traditional and in-vitro high-throughput assays have been applied to pure substances. However, these approaches can be time-consuming and difficult to implement in mixtures and complex samples for which the constituents are not determined, such as environmental samples.

Mass spectrometry can capture the fragmentation patterns of molecules, providing valuable analytical information that can potentially reveal the toxicity potency of these chemicals within samples. However, the use of MS^2 data for predicting endocrine activity has not been explored to date. This study proposes two main hypotheses: (1) chemicals that are closely connected within a spectral similarity network exhibit comparable endocrine disruptive activity, and (2) the similarity of MS^2 spectra can be utilized to predict the endocrine disruptive activity of unknown chemicals.

The primary objective of this study was to predict the endocrine disruptive potency of chemicals detected with LC-ESI-HRMS, based on the similarity of their MS^2 spectra. To achieve this goal, the following specific objectives have been outlined:

I. To obtain, combine and process MS^2 spectra for the chemicals included in the Tox21 Challenge.

II. To build spectral similarity networks and analyze their ability to connect mass spectra with similar endocrine activity.

III. To predict endocrine activity from MS^2 spectra using k-nearest neighbors and spectral similarity networks.

By achieving these objectives, this study explores the prediction of toxicity from MS^2 spectra based on mass spectral networks, thereby contributing to the understanding of the application of mass spectra in toxicity prediction.

4 Methodology

Data collection and pre-processing

Toxicity

The toxicity dataset was obtained in tabular format from http://bioinf.jku.at/ research/DeepTox/tox21.html Mayr et al. (2016) [18][46]. The dataset included 8975 unique InChIKeys for the twelve endpoints of nuclear receptor and stress responses shown in Table 2.1. If an endpoint had at least one active label for an InChIKey, the active label was retained.

MS^2 spectra

Mass spectra were collected from MassBank.eu (Feb 2023, v. 2022.12.1) [47]. The MS framework *matchms* (v.0.18.0) [48] was used for pre-processing the metadata and peaks. The collected mass spectra was filtered for MS^2, LC-ESI, $[M+H]^+$, and InChIKeys of the Tox21 dataset. Mass spectra with the same InChIKey were combined into a list of all peaks and intensities, called here as combined spectra. The intensities were normalized, and those below 5% were removed. Additionally, the intensities were square rooted to increase the importance of low intensity peaks.

k-NN algorithm

To predict the toxicity labels based solely on the most similar spectra on the entire dataset, a *k*-NN algorithm was evaluated. The results would indicate the closeness of compounds having the same labels.

The *k*-NN algorithm was tested using *scikit-learn* (v.1.2.0), a module for machine learning. Vector representations of the mass spectra was created with bin width of 0.1 *m/z* from 10 *m/z* to 1000 *m/z*. The mass-to-charge ratio was averaged, and the highest peak intensity of each bin was kept. The dataset was split into train (80%) and test (20%) sets with stratified sampling. The stratified split ensured that the train and test sets had roughly similar ratios of active/inactive compounds. Weighted voting was employed to assign labels to

the unknown compounds based on the labels of the *k* most similar neighbors. The influence of closer nodes was given more weight than nodes that were further away. The weights were calculated as the inverse of the distance between the unknown compound and its neighbors. The *k*-NN classifier using cosine weights is represented in Equation (4.1),

$$\hat{y} = \text{argmax} \left(\sum_{i=1}^{k} \frac{1}{1 - \cos(\theta)_i} \cdot \mathbf{y}_i \right) \tag{4.1}$$

where, $\hat{y}$ is the predicted class label, k is the number of nearest neighbors, $\mathbf{y}_i$ is the class label of the ith training vector, and $\cos(\theta)_i$ is the cosine similarity score between the input vector and the ith training vector.

To determine the number of neighbors, *k*, a 5-fold stratified cross-validation was performed, with recall used as the evaluation metric. Recall measures the proportion of true positive predictions (correctly classified active compounds) out of all actual active compounds, which can be beneficial for the application in toxicological alerts.

The performance of the *k*-NN algorithm was assessed based on its ability to accurately assign endpoint labels to the MS^2 in the test set. Considering the objective of correctly identifying positive results and enabling the generation of toxicological alerts, recall was selected as the primary criterion.

k-NN in spectral similarity networks

Spectral similarity

Cosine similarity between pairs of combined spectra was computed using *matchms* for a tolerance of 0.1 *m/z*. Each peaks in the combined spectra could be matched only once. Additionally, MS2DeepScore [29], a deep learning-based score, was calculated and compared with cosine for its ability to form clusters within a network.

Spectral similarity networking

A spectral network was built based on the cosine similarity between pairs. Mass spectra were represented as nodes, and cosine similarity was represented as edge connecting two nodes. Nodes were included and connected with an edge according to a minimum similarity of 0.6, at least 3 matched peaks (bins), and a maximum of 10 edges from a node. If there were more than 10 similar spectra with a similarity over 0.6 for a node, the additional edges were not included for better readability. The resulting network was loaded to Cytoscape (v.3.9.1), an open-source software for network visualization. Table 6.2 shows

the different parameters tested before the selection of a network for further processing.

Endpoint activity prediction

To predict the toxicity labels based on the spectral similarity networks a voting scheme of the directly connected nodes was applied. If the number of active neighbors was greater or equal than the inactive ones, then an active label was assigned to the node. Recall and precision were calculated based on leave-one-out over the entire network as the number of spectra were significant lower than the initial sample.

Retrospective analysis of wastewater samples

To test the approach, MS^2 features from an effluent sample of a wastewater treatment plant in Stockholm were analyzed. The MS^2 features were obtained and provided by Kruve Lab, Stockholm University. The HRMS analysis of the selected sample was carried out on a Q Exactive Orbitrap (Thermo Fisher Scientific, USA) equiped with Electrospray Ionization (ESI) source in positive mode. The acquisition modes were Data Dependent Acquisition (DDA) and Data Independent Acquisition (DIA). The collision energies were 20 V and 70 V. The MS data was processed in MS-DIAL (v. 4.80). The mass spectra features were examined using the spectral similarity network.

Code and data availability

The workflow code as a Jupyter Notebook, the network, and mass spectra data are available on the GitHub platform http://https://github.com/LeoSotoJ/MNTox.

5 Results

Toxicity dataset

The dataset showed a greater proportion of inactive labels as illustrated in Figure 5.2. For the whole dataset, the average ratio of active/inactive labels was 8.6% and average missing labels was 16.6%. After filtering for the available mass spectra, these values changed to 12.1% and 13.9%, respectively. The frequency of labels for each endpoint is shown in Table A.1 and Table A.2.

The association of binary classification between endpoints was investigated. The phi coefficient ranged from 0.013 (NR.AR, NR.PPAR.gamma) to 0.756 (NR.AR, NR.AR.LBD). An additional high phi coefficient was 0.7405 (NR.ER, NR.ER.LBD). The phi coefficients for all the pairs are illustrated in Figure 5.1.

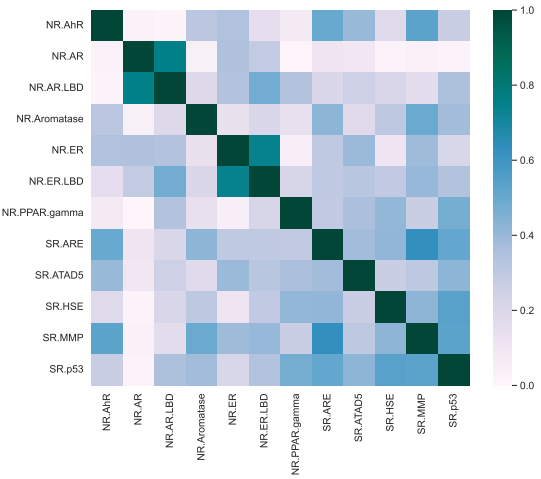


Figure 5.1: Phi coefficient matrix. Phi coefficient measures the association of two binary variables, analogue to the Pearson correlation. It ranges from -1 for high negative association to 1 high positive association. Zero corresponds to no association between the variables. In this matrix, all endpoints had a positive association.

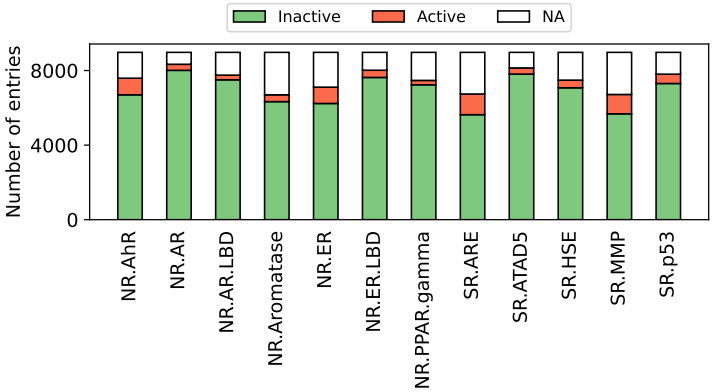


Figure 5.2: Distribution of endpoint labels in the toxicity dataset. Total number of unique InChIKeys: 8975.

Spectra in MassBank library

The mass spectra available in MassBank was filtered for the compounds with endpoint labels. The resulting collection included 15808 spectra and 1350 unique InChIKeys, which represents 15% of the unique InChIKeys in the toxicity dataset. The distribution of the precursors and the number of peaks per spectra are shown in Figure 5.3. The average precursor and number of peaks per spectra were 283 *m/z* and 5, respectively.

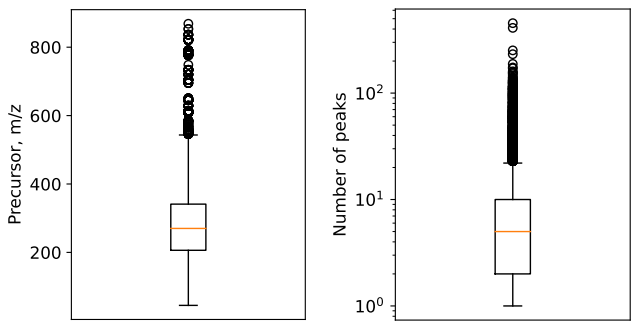


Figure 5.3: (a) Precursor *m/z* and (b) number of peaks per spectra. The spectra was filtered for the Tox21 Challenge compounds, MS^2, and LC-ESI. Number of filtered spectra: 15808. Unique InChIKeys: 1350.

Mass spectra similarity

The cosine similarity was calculated to compare the combined mass spectra. Very few chemicals scored above 0.7 (139, representing 0.02% of all pairs) as shown in Figure 5.6 and Table A.3. The distribution of cosine values across the spectra is illustrated in Figure 5.7. The spectra were ordered using hierarchical clustering that minimizes the variance between all groups at each step of the process. In this way, it is possible to notice regions of high and low similarity on the heatmap.

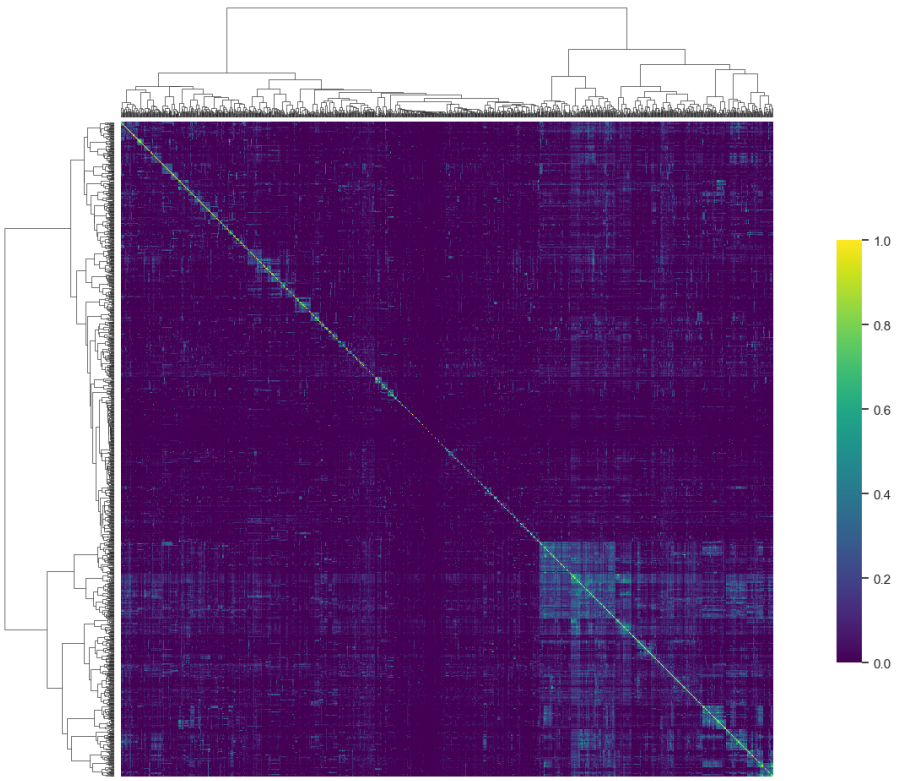


Figure 5.4: Cosine similarity heatmap. The x-axis and y-axis of the heatmap includes the combined mass spectra for unique InChIKeys ordered by hierarchical clustering. The hierarchical distribution allows to visualize groups with similar spectra. Green zones of high similarity (lower right corner and small groups across the diagonal) suggest groups of chemical classes, while purple areas indicate low similarity and no groups. The diagonal is the cosine similarity for the same spectra and it is 1.

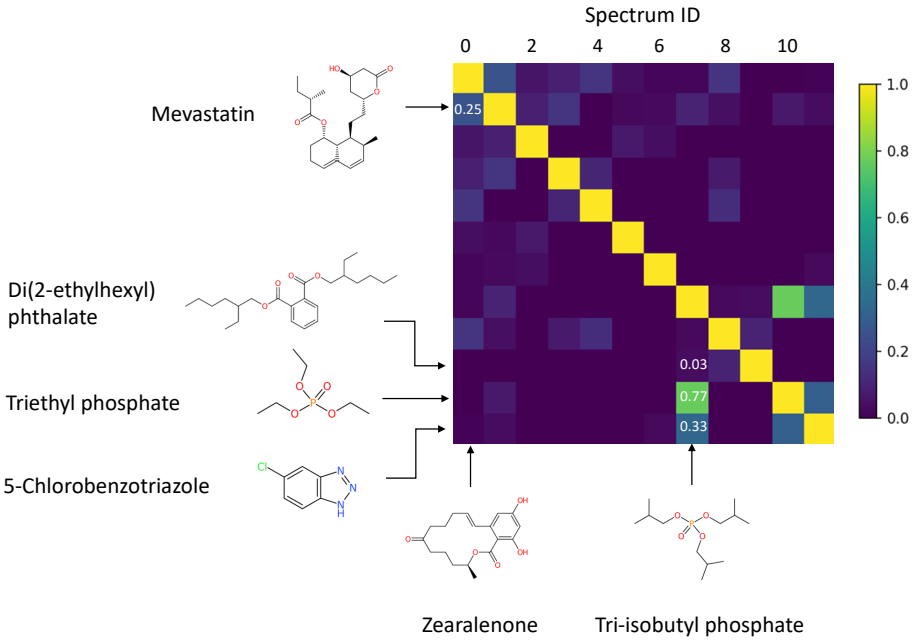
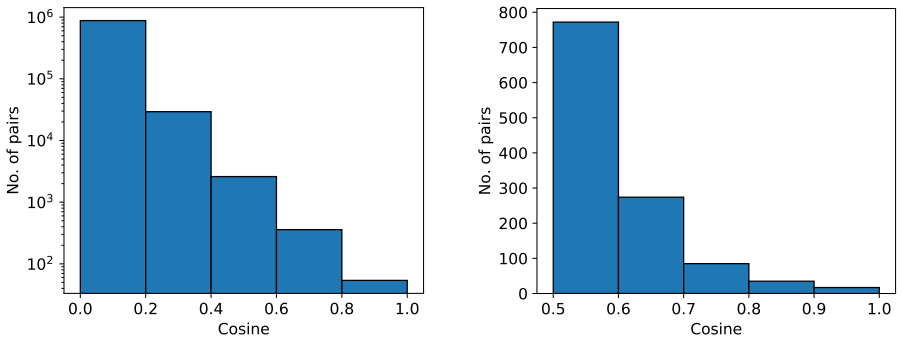


Figure 5.5: Cosine similarity matrix for a subset of mass spectra. The cosine score is high for the phosphate pair, while low for zearalenone in the subset indicating favorable selectivity of the cosine score. Additional pairs with cosine score greater than 0.6 are illustrated with their chemical structures in Figure 6.2. The diagonal is the cosine similarity for the same spectra.

The cosine similarity of a subset of mass spectra is shown on Figure 5.5. It can be seen that the cosine score is high for the pair of phosphates, correctly reflecting their structural similarity. Conversely, zearalenone (first column from the left) exhibits low cosine similarities, indicating a favorable selectivity of the cosine score for these pairs.

The relationship between the cosine and the number of matched peaks in the combined spectra is shown in Figure 5.7. The number of matched peaks corresponds to the number of common peaks within a tolerance of 0.1 *m/z*. The figure shows a low density in the high cosine range over 0.7.



(a) (b)

Figure 5.6: (a) Absolute frequency for all cosine scores and (b) for cosine scores greater than 0.5. Cosine is calculated for all pairs in the processed dataset with tolerance 0.1 *m/z*.

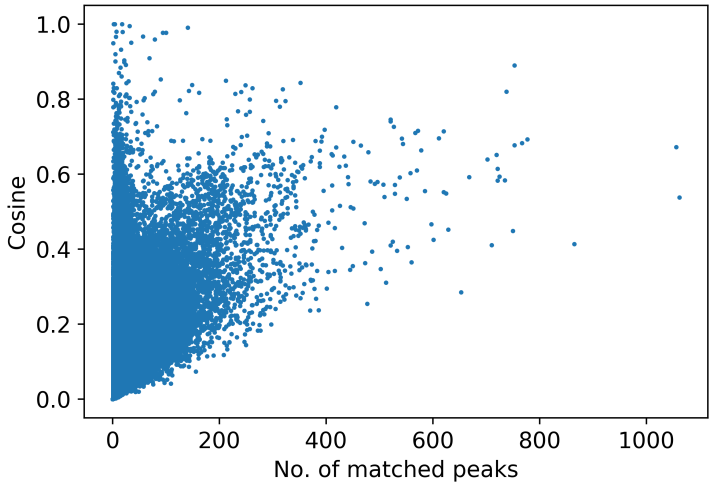


Figure 5.7: Relationship between the cosine score and the number of matched peaks in the combined spectra. A combined spectra contains all the peaks from library spectra having the same InChIKey. A peak from the combined spectra A is matched with a peak from the combined spectra B if the second lies within a tolerance of 0.1 *m/z*. Each peak can only be matched once. The figure shows a high-density area in the lower-left part and low density in the upper part. The latter depicts pairs of spectra that can form the spectral networks. High and atypical values are further discussed.

Spectral similarity networks

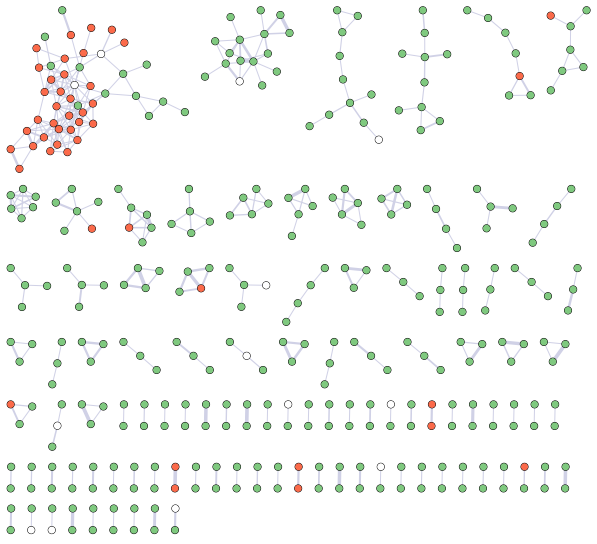
The cosine similarity was used to build the network. Several parameters were tested including the tolerance for the cosine calculation, minimum score, and maximum edges from a node, as described in Table 6.2. Then, a network was selected considering the number of nodes, edges, other network statistics, and visual inspection. The final network had 363 nodes, 409 edges and 105 connected components. The statistics summary for the selected network is shown in column "cosine" in Table 5.1. Additionally, the same table includes the statistics for an experimental network based on a machine learning similarity metric. These two networks share similar clustering coefficient, network density, and connected components. Not remarkable differences were observed for the distribution of the endpoint labels compared to cosine similarity for the studied collection of spectra. The MS2DeepScore shows great promise for spectra clustering, a network is examplified for the endpoints NR.AR and SR.ARE in Figure A.1.

The distribution of endpoints for the endpoint NR.AR and SR.ARE is shown in Figure 5.8. The NR.AR is a representative case of an endpoint that showed clustering tendency, while the SR.ARE represents low clustering tendency with active labels distributed across different clusters. All endpoint labels distribution is illustrated in the figures of Appendix E.

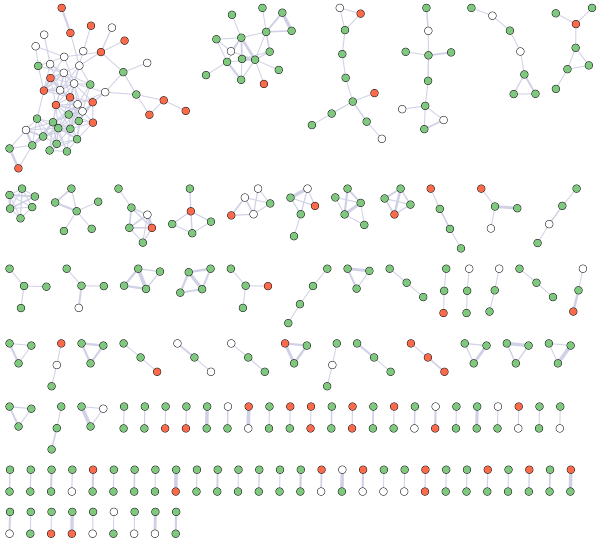
Table 5.1: Summary of conditions and statistics of the cosine similarity network and a network based on a deep learning-based score.

Description	Cosine	MS2DeepScore
Minimum score	0.6	0.85
Max connections from node	10	10
Number of nodes	363	475
Number of edges	409	745
Avg. number of neighbors	6.043	8.898
Network diameter	7	7
Network radius	4	4
Characteristic path length	2.970	2.807
Clustering coefficient	0.452	0.576
Network density	0.131	0.185
Network heterogeneity	0.683	0.548
Network centralization	0.226	0.198
Connected components	105	108

Note: The second column is the selected network for further processing and predictions. The third column is an example of a network built with MS2DeepScore. The statistics were obtained in Cytoscape.



(a)



(b)

Figure 5.8: Spectral similarity network for the endpoint (a) NR.AR and (b) SR.ARE. Minimum cosine score: 0.6. Active endpoints are in red, inactive in green, and blank if the a label was not available. The nodes represent the combined mass spectra of unique chemicals. The edge width indicates the intensity of the cosine similarity. The spectral similarity networks for all the endpoint labels are shown in Appendix E.

k-NN algorithm

Two approaches for the prediction of the endpoints were tested: *k*-NN and spectral similarity networks. k-NN algorithm explored the entire dataset and assigned a label based on the *k* nearest neighbors according to their cosine similarity. On the other hand, the spectral similarity network narrowed the chemical space and pre-defined the neighbors for a chemical based on, among other factors, a cosine similarity threshold.

The k-NN algorithm requires the definition of two hyperparameters, the distance metric and *k*. The distance parameter was the cosine distance. The best *k* was determined by cross-validation. The change of recall for different values of *k* for NR.AR and NR.AhR is shown Figure A.2. In the case of NR.AR, the recall was stable for the explored values of *k*, but for the rest of endpoints the recall decreased with the increase of *k* as illustrated with NR.AhR.

Table 5.2: Recall and precision for *k*-NN

Endpoint	Active	Inactive	Active/Inactive %	Recall %	Precision %
NR.AhR	48	191	25.1	31.3	29.4
NR.AR	17	243	7.0	47.1	44.4
NR.AR.LBD	13	230	5.7	15.4	33.3
NR.Aromatase	21	180	11.7	4.86	42.9
NR.ER	38	180	21.1	29.0	22.9
NR.ER.LBD	17	233	7.3	11.8	9.5
NR.PPAR.gamma	7	217	3.2	0.0	0.0
SR.ARE	40	176	22.7	25.0	27.8
SR.ATAD5	12	244	4.9	0.0	0.0
SR.HSE	13	228	5.7	7.7	8.3
SR.MMP	38	165	23.0	26.3	26.3
SR.p53	18	225	8.0	5.56	5.56

Note: The recall and precision were calculated based on a stratified test set (20%) of the entire dataset. Number of samples: 1350. The active and inactive columns are the number of compounds with the respective labels in the toxicity data.

k-NN algorithm in locally connected mass spectra

After building the network, its prediction ability was tested. A representation of the chemical class of the nodes, the predicted and true labels are illustrated in Figure 5.9. Recall and precision were estimated by leave-one-out and they are shown in Table 5.3. The highest recall and precision were obtained for NR.AR, being 81.8% and 75.0%, respectively; and the lowest for SR.HSE, being 15.8% and 13.0%. The same table also shows the number of labels for each endpoint. The active/inactive ratio range was between 27.3% for SR.ARE and 2.9% for NR.PPAR.gamma.

The network was used for the retrospective analysis of a wastewater sample from an effluent of a treatment plant. Some features of the real sample show similarity with the nodes in the reference network. The reference network and the MS^2 features of the sample are displayed in Figure 5.10.

Table 5.3: Recall and precision for the *k*-NN based on locally connected nodes within a spectral network

Endpoint	Active	Inactive	Active/Inactive %	Recall %	Precision %
NR.AhR	55	257	21.4	54.5	52.6
NR.AR	44	294	15.0	81.8	75.0
NR.AR.LBD	33	284	11.6	66.7	57.9
NR.Aromatase	29	209	13.9	41.4	42.9
NR.ER	59	216	27.3	49.2	39.7
NR.ER.LBD	29	299	9.7	27.6	29.6
NR.PPAR.gamma	8	274	2.9	50.0	66.7
SR.ARE	49	225	21.8	40.8	39.2
SR.ATAD5	12	318	3.8	33.3	36.4
SR.HSE	19	307	6.2	15.8	13.0
SR.MMP	42	204	20.6	42.9	46.2
SR.p53	19	293	6.5	36.8	43.8

Note: The recall and precision were estimated based on leave-one-out over the entire network. Number of nodes: 363. The active and inactive columns are the number of compounds with the respective labels in the toxicity data.

Retrospective analysis of MS^2 features

The MS^2 features of a wastewater sample were analyzed using the spectra similarity network for the prediction of NR.AR activity. The sample exhibited multiple dense clusters of spectra with limited connections to the toxicity network, as depicted in Figure 5.10.

Mass spectra features were screened based on their connectivity to active

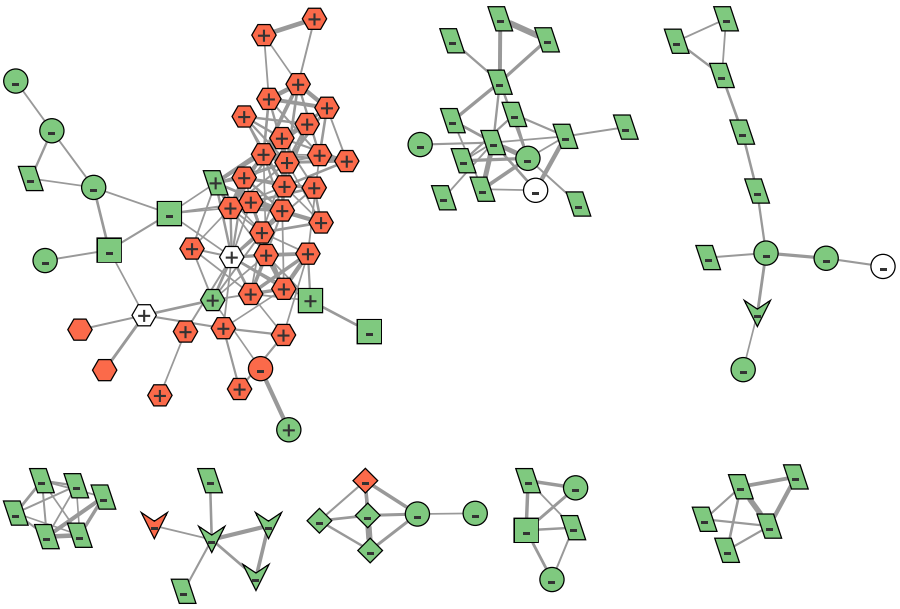


Figure 5.9: Predicted and true labels for the NR.AR endpoint on a subset of nodes. The nodes represent mass spectra, the fill color indicates the true labels as active (in red), inactive (in green) or not available (in white). The symbols inside the nodes show the predicted labels as active (+) or inactive (-). The edge width illustrates the intensity of the spectral similarity. The shape of the node corresponds to the chemical class according to ChemOnt obtained from ClassyFire [49]. Hexagon: steroids and steroid derivatives. Diamond: phenols. Parallelogram: benzene and substituted derivatives. Rectangle: prenol lipids. V: Organic phosphoric acids and derivatives. Circle: other classes.

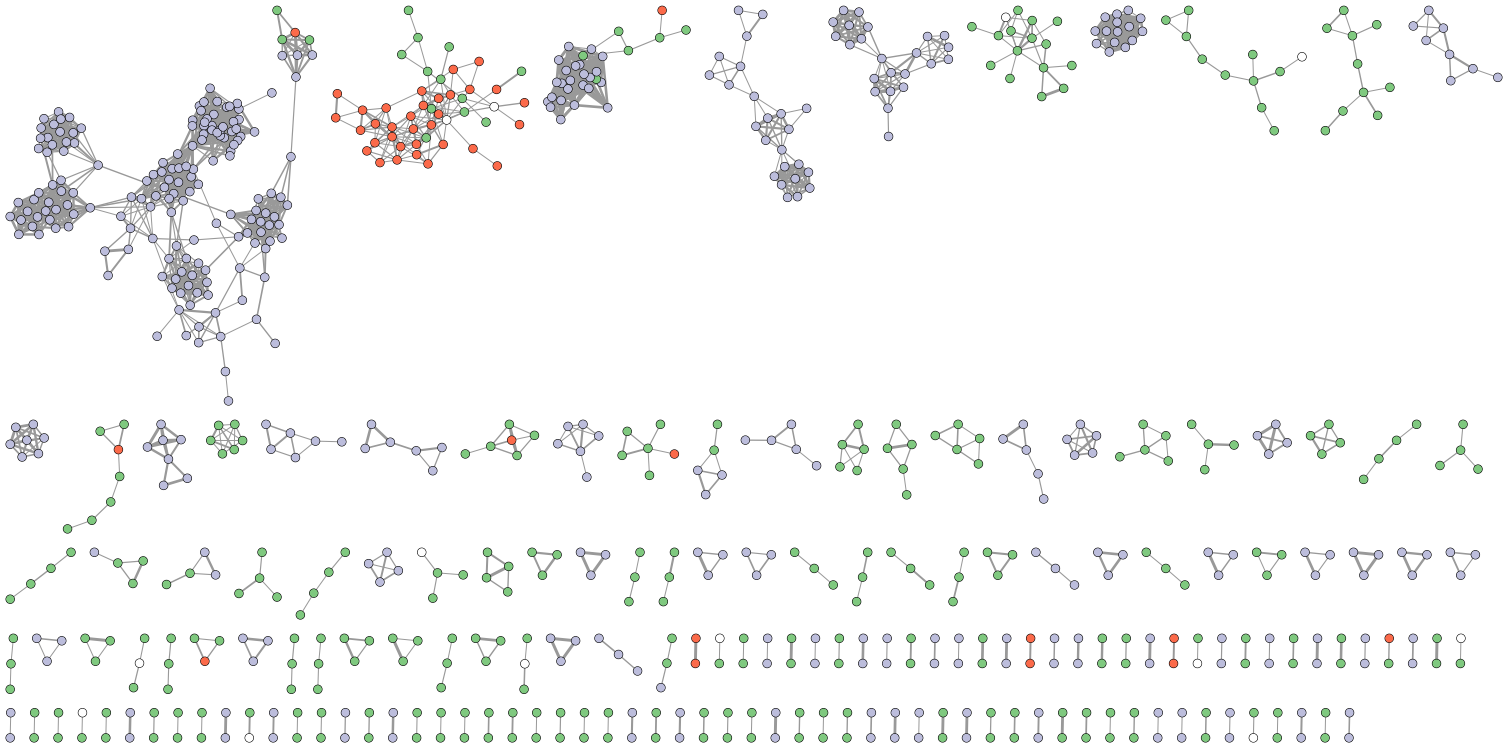


Figure 5.10: Toxicity network for the NR.AR endpoint and unknown chromatographic MS^2 features from a wastewater sample. The active compounds are in red, inactive in green, and detected MS^2 features in light purple. Nodes without color do not have a NR.AR label in the toxicity dataset. The graph shows independent clusters suggesting, in most cases, low similarity between the reference and sample spectra, and reduced overlap with the chemical space of the toxicity dataset.

spectra in the spectral network. A few mass spectra features were found to have connections with the reference network. When compared to the reference network, the sample features displayed low cosine similarities. Examples of spectra comparison for features and nodes are shown below. Certain sample spectra exhibited low number of matched peaks, as illustrated in Figure 5.11 (3 peaks) and 5.12 (3 peaks), while others had a higher number of matched peaks, as exemplified in Figure 5.13 (82 peaks). Even with high number of matched peaks, the cosine similarity score can be low for reference spectra containing a high number of peaks.

To establish node connections, an alternative approach involves considering the common peaks, such as directly by the number of matched peaks or by incorporating weights. The number of matched peaks can provide useful information to describe the connections. By considering the number of matched peaks and lowering the minimum cosine score, an edge can be established between each sample feature and the network. Subsequently, the endocrine disruptive label is assigned based on the votes from connected nodes. For example, the feature of Figure 5.13 is connected to 8 active and 1 inactive nodes, then this feature can be predicted as active for NR.AR. This method allows for determining the relatedness of each feature to the nodes in the reference network and identifying those with a higher number of common peaks with active nodes. Consequently, the identified compounds with positive predicted activity can be prioritized for further investigations.

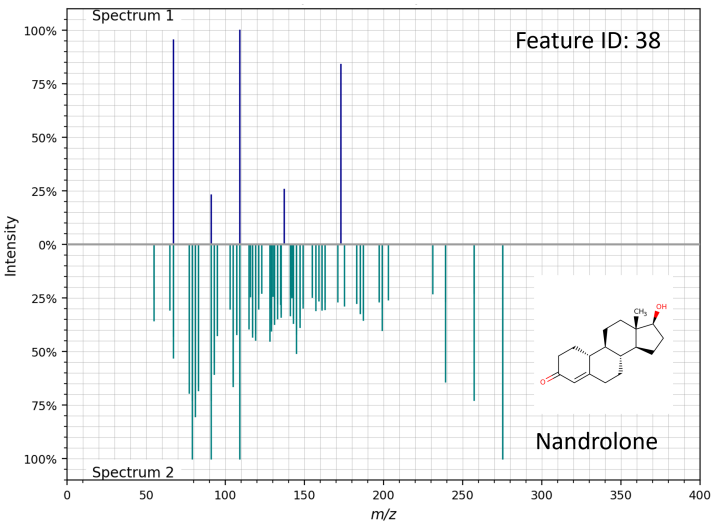


Figure 5.11: Comparison of a sample feature with a reference spectra (nandrolone: active for NR.AR). The figure exemplifies the cases in which there are a high number of peaks for the reference node.

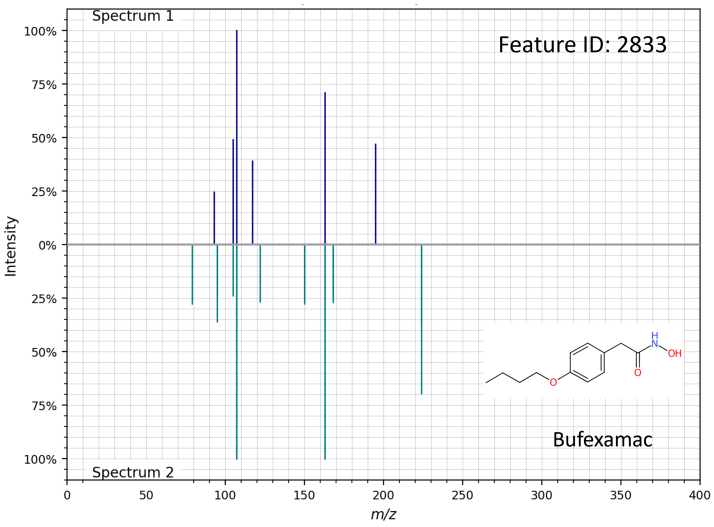


Figure 5.12: Comparison of a sample feature with a reference spectra (bufexamac: inactive for NR.AR). The figure exemplifies the cases where the number of peaks in the sample feature and the reference node are comparable leading to high cosine scores.

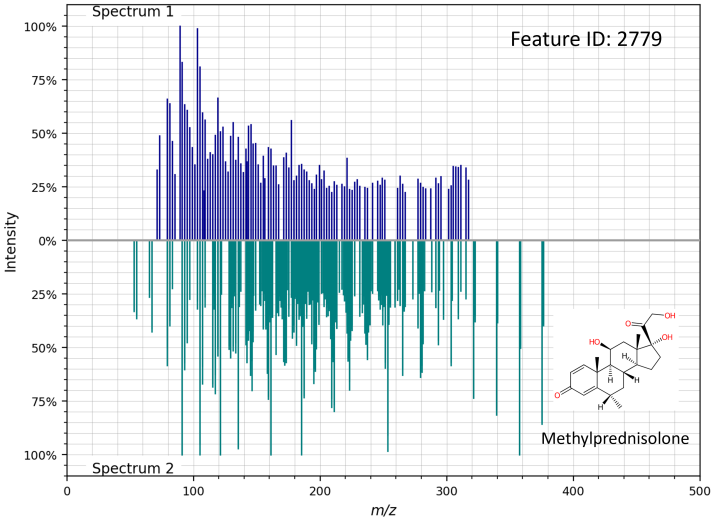


Figure 5.13: Comparison of a sample feature with a reference spectra (methylprednisolone: active for NR.AR). This pair exemplifies the cases where the sample features and reference node have a high number of peaks. This could result in a high number of matched peaks but low to medium cosine scores. The number of matched peaks could be then used to associate the detected compounds with the reference network.

6 Discussion

Toxicity dataset

The dataset exhibited an imbalance towards the inactive compounds as shown in Figure 5.2. This imbalance has significant implications for the clustering of active labels, as they are less prevalent. This observation is consistent with the occurrence of negative results in toxicity assays, where most of the tested compounds tend to be inactive. The endpoint with the highest ratio of active to inactive labels was SR.ARE (0.197), indicating a relatively higher response for this endpoint. Conversely, the lowest ratio was observed for NR.PPAR.gamma (0.032), suggesting a lower response to the test. Nevertheless, SR.ARE is also one of the endpoints having the highest missing values ratio. The scarcity of active chemicals and missing values can impact the overall performance of the subsequent analysis and should be considered while interpreting the results.

High correlation was found for some endpoints. The phi coefficient is a measure of association between two binary variables, indicating the strength and direction of the relationship. In Figure 5.1, the phi coefficients reveal the extent of similarity between various pairs of endpoints. The high phi coefficients for NR.AR - NR.AR.LBD (0.76) and NR.ER - NR.ER.LBD (0.74) indicate a strong positive association between these pairs, suggesting a close relationship and similarity in their binary classifications and it could explain their similar predictions. These two pairs of assays are usually correlated because of their similar objectives, the first pair are related to the androgen receptor and the second one to the estrogen receptor. For example, if a chemical can bind to the ligand binding domain (LBD) of the androgen receptor (AR) (labelled as active for the NR.AR.LBD endpoint), then is very likely that the binding can affect the conformation and activity of the androgen receptor, which lead to a positive result for the NR.AR endpoint.

Spectra in MassBank library

In the dataset, small molecules are predominant as shown in Figure 5.3 (a). This observation aligns with the general understanding that smaller molecules

have a higher potential to penetrate biological barriers and interact with cellular processes, which can impact their bioactivity and toxicity. Small molecules are also predominant in toxicity datasets, e.g., drugs and pesticides. However, it is worth noting that the relationship between molecular mass and toxicity is much more complex involving various mechanisms. For instance, the big carbohydrate-binding protein ricin (64-65 kDa) can also penetrate biological barriers and disrupt protein synthesis within cells leading to toxicity [50].

This collection of spectra constitues the chemical space that is explored in the scope of this study. The understanding of these boundaries is important to interpret the prediction capabilities of this approach.

Spectra similarity

Mapping cosine

From all pairwise comparisons, 1185 (0.13%) pairs exhibited a cosine similarity greater than 0.5, while only 139 pairs had a cosine greater than 0.7, as detailed in Table A.3. Some possible factors leading to a small number of highly similar pairs could be the intrinsic structural similarity of the dataset and the similarity metric for the combined spectra .

The similarity across the dataset shows certain characteristics highlighted in Figure 5.4. Some spectra exhibited broad similarities with the rest of the dataset, as can be seen by the transversal marks at the bottom and right parts. This could be in part due to the greater amount of analytical information. As multiple collision energies are available in the library, the number of total peaks and matched peaks increases. For example, cortisol has the highest mean similarity with all the compounds (0.127, the global mean is 0.0428) and has 26 different mass spectra (the average is 11.7 spectra per compound). Certain molecules were more similar within specific groups, forming clusters, as seen in the lower right corner of Figure 5.4. The upper part of the diagonal indicates the presence of multiple smaller groups. In contrast, the dark cross region, originating from the center of the axis, indicates subsets with low similarity. In general, the heatmap shows groups of spectra with shared similarity where the network can be built, as it be will further explored later.

The cosine similarity is a widely utilized metric for assessing the similarity between spectra. In principle, having a greater amount of empirical information can be beneficial for associating structures and predicting bioactivity, such as shared toxophores between compounds. Having more peaks, obtained from different ionization energies, can assist in identifying coincidental fragments. However, the cosine similarity can be susceptible to the influence of isotopic peaks, noise, or background interference. When confronted with a large number of peaks, including noise or isotopic peaks, the cosine score may decrease

as the vector lengths increase. In such cases, the cosine similarity may yield low scores, potentially leading to erroneous assessments of structural similarity.

Some alternatives for calculating similarity between spectra with high density of peaks would be computing a consensus spectra per InChIKey and machine learning-based scores. A consensus spectra would cluster the peaks across *m/z* and assign an intensity based on their frequency on the library. This approach would significantly reduce the number of peaks per spectra and have the potential of improving the performance of cosine similarity. Additionally, a machine learning-based score can select the fragments that contain certain structural information by recognizing patterns across the spectra dataset. These two alternatives are further discussed later in the networking section.

In general, the cosine scores alone for the combined spectra provided a satisfactory number of pairs for networking and demonstrated sufficient structural similarity upon examination of the pairs. This study specifically focused on the application of cosine similarity, which will be discussed in more detail in the following sections.

Pairwise comparison

After analyzing the cosine map, the pairwise comparison were explored in more detail. For this, the number of matched peaks was examined for its relationship with the cosine score in Figure 5.7. From this graph, it can be seen the density of pairs sharing certain levels of similarity, this would be helpful later to interpret the networks.

The scores are concentrated on the lower left part of Figure 5.7. Conversely, the upper part of the graph exhibits a low-density area containing pairs with sufficient similarity to establish nodes in the spectral networks. On the left side of the graph, low number of matched peaks can exhibit also high similarities; while on the right side of the graph, large number of matched peaks (and consequently total peaks) are less likely to have very low cosine scores.

The right upper region, sparsely populated, represents a highly similar mass spectra where the number of matched peaks is close to the total number of peaks. For instance, the right farthest pair has a cosine value of 0.54 and 1062 matched peaks. This pair consists of colchicine and prednisone, both medications. These substances are well-documented in the MassBank library, with 39 and 40 spectra, respectively. Moreover, their complex fragmentation patterns are evident from the number of peaks per spectrum, surpassing 100 peaks in some cases. The presence of several fused rings and functional groups in both molecules results in multiple fragmentation pathways, contributing to the high number of matched peaks. However, their cosine similarity is relatively low partly because of the increase in the lenght of the vectors. This is an

example that even at high number of matched peaks, the abundance of peaks can be detrimental for the cosine score. Most of the high cosine values are, in fact, found at a low number of matched peaks.

Cosine budget

Some spectra share similarities across the dataset more than others. The cosine budget is here presented as the sum of all the cosine values of a spectra. A high cosine budget indicates that similarities are shared over a wide range of spectra, while a low cosine budget represents narrowed similarity and specificity. Some factors that are related to the cosine budget are number of available spectra, total peaks per spectra and total matched peaks, which are described below.

Number of spectra

The increase of the number of spectra in the library is more likely to yield high cosine budgets until a certain point, as shown in Figure 6.1a. This could be explained by the increase in empirical analytical information (e.g., more fragments as result of more collision energies in the library) that results in the identification of additional common peaks. However, the graph also shows a negative effect at high number of available spectra. High abundance of multiple mass spectra for a compound shows less frequency of high cosine budgets. This could be explained by the negative effect of noise features in the cosine similarity measure. Moreover, it could correspond to a cosine budget threshold, for which the increment of spectra do not provide additional fragment information that can be shared within the dataset.

Total number of peaks

Some mass spectra can have a high number of peaks, in some cases more than a hundred. Therefore, the total number of peaks per InChIKey across the dataset and its relationship with cosine was investigated. The combined spectra contain all the peaks associated with an InChIKey. A greater amount of peak information tends to have a greater cosine budget as it is shown in Figure 6.1c. The graph shows high density of compounds in the lower matched peaks and cosine budget suggesting low similarities within the dataset and therefore low clustering tendencies for those chemicals. On the upper right part, there are fewer and dispersed compounds with high similarity across the dataset more likely to be in clusters.

The dispersion of cosine budget is higher as the number of peaks increases. High cosine similarity budget could reflect structural features that are common in the chemical space of the dataset. A high number of peaks can be due to complex fragmentation patterns and a greater availability of mass spectra in the

library. The ten highest cosine budgets are for the chemicals described in Table 6.1. Steroids and other cyclic compounds are predominant chemical structures. Steroids are structural complex and have multiple rings and functional groups. Hydroxyl (-OH), carbonyl (C=O) groups, and double bonds provide different sites for bond cleavage, leading to diverse fragmentation pathways and resulting in multiple peaks. Additionally, the fused ring system can be fragmented in one or more rings and produce different fragment ions. Other factor is that the isomeric structures (e.g., positional and stereoisomers) can contribute to the generation of multiple peaks. The complex complex nature of steroids could also explain their high cosine budgets and their location in the spectral network. The dispersion of cosine budget is higher as the number of peaks increases. High cosine similarity budget could reflect structural features that are common in the chemical space of the dataset. A high number of peaks can be due to complex fragmentation patterns and a greater availability of mass spectra in the library. The ten highest cosine budgets are for the chemicals described in Table 6.1. Steroids and other cyclic compounds are predominant chemical structures. Steroids are structural complex and have multiple rings and functional groups. Hydroxyl (-OH), carbonyl (C=O) groups, and double bonds provide different sites for bond cleavage, leading to diverse fragmentation pathways and resulting in multiple peaks. Additionally, the fused ring system can be fragmented in one or more rings and produce different fragment ions. Other factor is that the isomeric structures (e.g., positional and stereoisomers) can contribute to the generation of multiple peaks. The complex complex nature of steroids could also explain their high cosine budgets and their location in the spectral network.

Total number of matched peaks

The number of total matched peaks across the dataset is also an indication of clustering tendency. Figure 6.1b illustrates the relation of matched peaks with tolerance 0.1 *m/z* and the cosine budget. The x-axis shows the total number of peaks from all spectrums per unique InChIKey in mass spectra dataset after filtering. The y-axis represents the sum of all the cosine similarity of a chemical against the others in the dataset. The number of peaks represents the amount of analytical information available for a chemical. A higher dispersion of the cosine budget can be noticed at higher number of peaks. The cosine budget indicates the total similarity of a chemical with the dataset. Most of the spectra are concentrated in the lower left part of the scatter plot. This region constitutes the set of compounds that have low similarity with the chemical space, therefore will not be forming clusters in spectral networks. On the other side, the upper right corner of the graph captures the compounds that are more likely to form clusters as they have high cosine budget and high number of matched

Table 6.1: Compounds with the highest cosine budgets

Compound	Structure	Formula	Exact mass (Da)	No. of spectra	No. total of peaks	Cosine budget	Normalized cosine budget
Danazol		$C_{22}H_{27}NO_2$	337.2042	6	302	200.9	1.00
Resibufogenin		$C_{24}H_{32}O_4$	384.5160	10	439	188.18	0.94
Betamethasone valerate		$C_{27}H_{37}FO_6$	476.2574	6	354	186.8	0.93
Corticosterone		$C_{21}H_{30}O_4$	346.2144	19	881	178.2	0.89
Etofenprox		$C_{25}H_{28}O_3$	376.2038	6	309	177.6	0.89
Norethindrone		$C_{20}H_{26}O_2$	298.1933	21	608	175.5	0.88
Gestoden		$C_{21}H_{26}O_2$	310.1933	6	421	175.3	0.88
Glycyrrhetinic Acid		$C_{30}H_{46}O_4$	470.3396	34	379	175.2	0.88
Cyproterone acetate		$C_{24}H_{29}ClO_4$	416.1754	6	463	174.6	0.88
Betamethasone		$C_{22}H_{29}FO_5$	392.1999	8	319	173. 5	0.86

peaks across the dataset. However, the density in this region is low and this will be observed in the networks where only a small fraction of the dataset will establish clusters.

The combined effect of the total number of peaks and matched peaks is illustrated in Figure 6.1d. The graph shows a random sample of 10 % for better visibility. The x axis shows the total number of peaks that are available in the MassBank.eu for each InChIKey. The number of peaks can be read as an indication of the available peak information for a chemical. The y axis shows the total number of matched peaks of the chemical across all the dataset. The latter is an indication of shared features. Additionally, the circle size illustrates the cosine budget as the sum of all the cosine similarities of a chemical across the dataset. The trend in Figure 6.1d shows that those chemicals that have a greater number of peaks in the library result in a greater cosine budget, which could be due to small cosine score and fragments shared across the dataset.

Chemical class

Another aspect that could be responsible for the high cosine budget is the presence of chemical classes. Some predominant classes could explain the broad similarities that some chemicals share across the dataset. One indication of predominant classes is the clustering of steroids as shown Figure 5.9 and Figure 6.2. This aspect is discussed in the section of spectral networking. Steroids have a greater number of fragments compared to other chemical classes. The abundance of peaks increases the number of total matched peaks and cosine budget. Additionally, some compounds, e.g., as carboxylic acids and indoles, might receive lower scores using cosine similarity [37].

Combination of several spectra

The combination of several mass spectra can influence the cosine similarity. In this case, all the peaks from different mass spectra were combined. Then, the cosine function paired the matched peaks within a tolerance of 0.1 *m/z*. Each peak can be paired only once. The effect of the matching tolerance was also explored and it is described in Table 6.2. A tolerance of 0.001 *m/z* reduced the number of nodes to 235 and edges to 193, the resulting network is illustrated in Figure 6.5.

Neutral losses

The occurrence of neutral losses can influence the cosine similarity by modifying the intensity and abundance of peaks. The measurement conditions, sample preparation, and the nature of the compound make it difficult to correctly

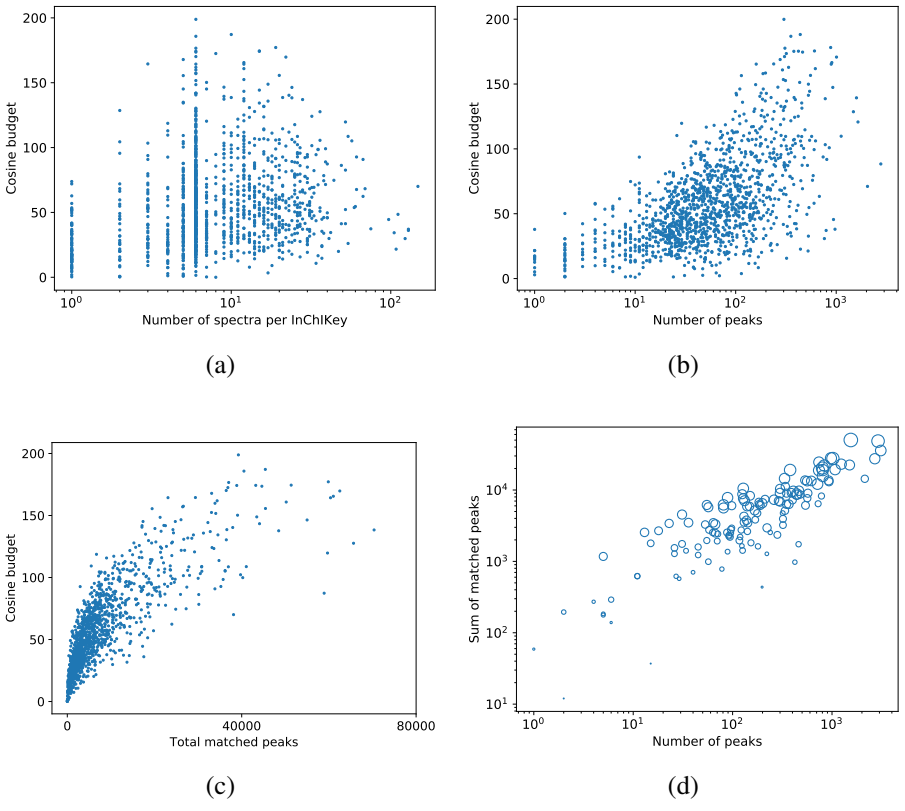
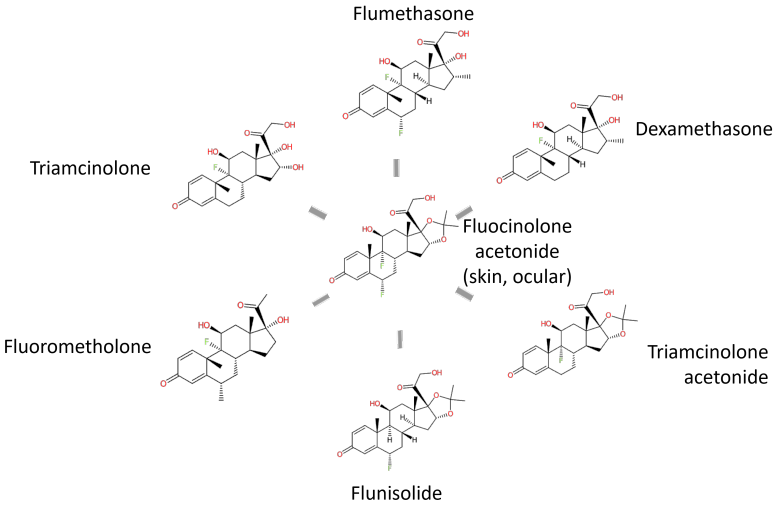
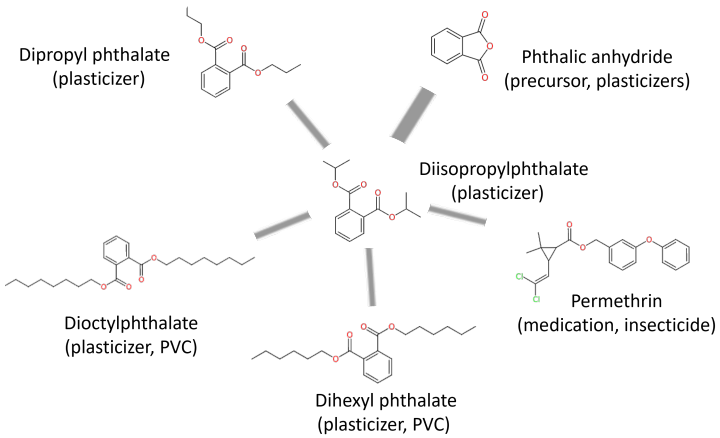


Figure 6.1: (a) Relationship between the cosine budget and the number of spectra per InChIKey; (b) number of peaks and the cosine budget; (c) cosine budget and the sum of total matched peaks within a tolerance of 0.1 *m/z*; (d) matched peaks, total number peaks, and the cosine budget.



(a) Steroids



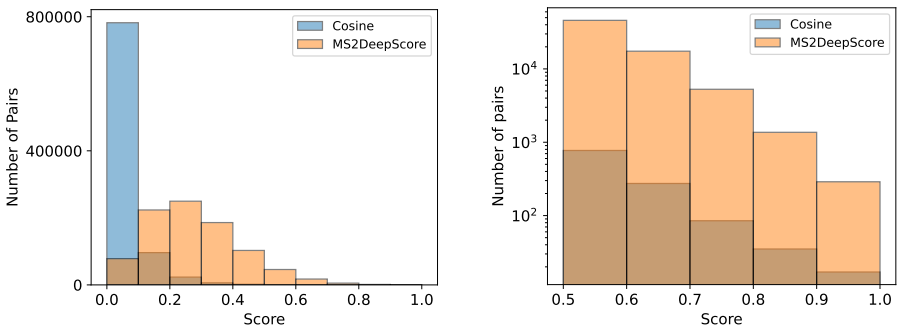
(b) Phthalates

Figure 6.2: Molecular structures of a subset of pairs with cosine similarity greater than 0.6. A wider edge represents a higher cosine score. The figure shows that the connected nodes have similar molecular structures. In the case of (a) steroids, the majority of the group shows activity, whereas (b) phthalates mostly are inactive for the NR.AR endpoint.

identify all losses during fragmentation. The abundance of peaks is particularly increased in large molecules with several functional groups. For instance, steroids can exhibit neutral losses due to the hydroxyl attached to their fused rings. These losses lead to an increase in the number of peaks and it can result in lower cosine scores. Further investigation on accounting for neutral losses could have a positive effect on the accuracy of the similarity metrics.

Deep learning-based similarity metric

Additionally, to the cosine similarity, a deep learning-based metric was examined for its ability to cluster similar molecules within a network. MS2DeepScore [29] is a deep learning-based similarity score that uses convolutional neural networks to predict structural similarity scores. This score exhibited a more balanced distribution in comparison to the cosine similarity, as depicted in Figure 6.3. This can be attributed to MS2DeepScore's ability to capture the complex relationships between mass spectra peaks.



(a) (b)

Figure 6.3: Comparison of pairwise scores between cosine and MS2DeepScore for the whole range (a) and higher half (b).

In the higher range of the similarity scale (0.5 to 1), MS2DeepScore exhibited approximately 10 times more pairs compared to the cosine score as shown in Figure 6.3. However, it is important to note that this disparity in pair counts does not directly imply a greater number of structurally similar pairs being identified. The calculation and interpretation of these two metrics are distinct. While MS2DeepScore may generate more pairs in this range, the interpretation and significance of those pairs may differ from those obtained using the cosine similarity.

Given the better explainability and interpretability of the cosine similarity, this study primarily focused on its exploration and analysis in the subsequent sections. The use of cosine similarity allows for a clearer understanding of the

structural similarities between mass spectra, facilitating the interpretation of the results.

Spectral similarity networks

The cosine similarity, calculated in the previous section, was used to construct networks in the form of a graph. On the network, the nodes represent the combined spectra, each representing a unique compound identified by its InChIKey, and the edges represent the cosine similarity between the spectra. The topology of the spectral network is influenced by various factors, such as the minimum similarity required to form an edge between two nodes and the maximum number of connections a node can establish with other nodes.

The selection of the optimal network was based on a combination of statistical summaries and visual inspection. Key factors considered included the total number of nodes and edges, the average number of neighbors, characteristic path length, and clustering coefficient. These measures provide insights into the local connectivity patterns among mass spectra that are structurally or functionally similar. By examining the network statistics, it is possible gain a more objective description of the relationships in the spectra dataset.

For a detailed overview of the graph statistics under different conditions, please refer to Table 6.2. The table provides a summary of the network properties for various sets of conditions, allowing for a comparative analysis of the different network configurations. The effect of some of those conditions are further described below.

Effect of tolerance, minimum similarity and maximum edges

When applying a matching tolerance of 0.1 *m/z*, the change from minimum cosine from 0.5 to 0.6 led to a change in the characteristic path length from 7.44 to 2.97 (-60%). The improvement in the characteristic path length is relevant and denotes a more similar mass spectra in the network and interconnected structure with nodes closer to each other. This is slightly enhanced for cosine 0.7, giving a characteristic path length of 2.68 (-10%). However, the number of nodes and edges are reduced by 50% and 60%, respectively. In similar way, the average number of neighbors went down from 6 to 3. The clustering coefficient slightly increased and suggested neighboring nodes densely interconnected in local clusters.

For matching peaks with a tolerance of 0.001 *m/z*, there were fewer nodes and edges compared to the 0.1 *m/z* tolerance at the respective cosine, cut-offs as shown in Table 6.2 and in Figure 6.5. This reduction of nodes could be benefitial as it increased the clustering coefficient and peak matching pairing accuracy; however, the chemical space is also reduced which could be detri-

mental for connecting unseen spectra with the network. The characteristic path length was slightly higher (+11%) for the cosine threshold 0.6. The increase of the minimum cosine similarity required to form a node negatively consequently reduced the number of nodes and edges.

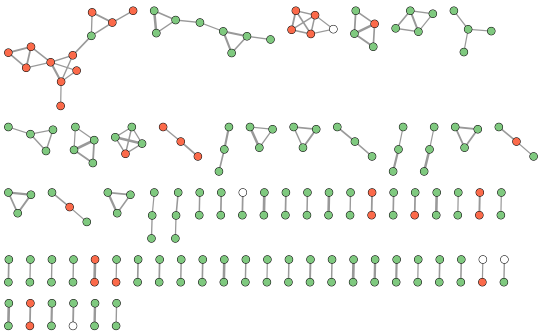
Increasing the maximum edges from a node from 10 to 20 added 4 (1%) new edges and did not affect the topology of the graph when analyzing the network with cosine tolerance 0.6. As the number of nodes sharing high similarity is limited, an increase of this parameter did not substantially change the topology of the networks.

Additionally, a reduction of the number of peaks in the combined spectra was explored for its effect on the clustering tendencies of the endpoints. The reduction was based on bins of 0.001 *m/z* and merging of peaks. The merging kept the average *m/z* and the highest intensity of the peaks. After visual examination of the distribution of active labels across the network, not relevant changes occurred compared to using the combined spectra alone. Figure 6.6 shows the resulting network from this reduction of peaks for NR.AR.

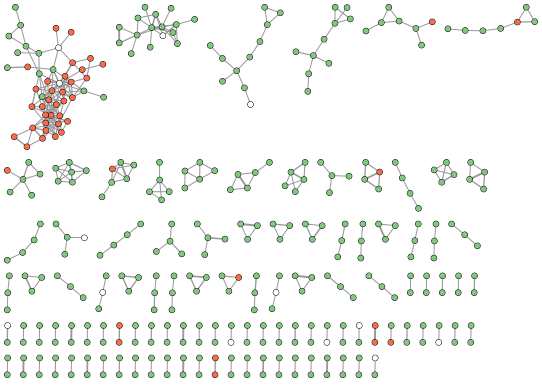
Table 6.2: Effect of several parameters on the construction of spectral similarity networks

Parameter	A	B	C	D	E	F	G	H	I	J	K	L	M
Score	cos	cos	cos	cos	cos	cos	cos	cos	cos	cos	MS2DS	MS2DS	MS2DS
Tolerance	0.1	0.1	0.1	0.1	0.05	0.005	0.0015	0.001	0.001	0.001	NA	NA	NA
Minimum score	0.5	0.6	0.6	0.7	0.5	0.6	0.5	0.5	0.6	0.7	0.8	0.85	0.9
Max connections from node	10	10	20	10	10	10	10	10	10	10	10	10	10
Number of nodes	647	363	363	182	406	260	395	385	235	128	783	474	236
Number of edges	1053	409	413	139	569	239	516	480	193	92	1499	745	288
Avg. number of neighbors	5.736	6.043	6.213	3	6.393	3.92	5.057	4.68	2.769	2	4.558	8.898	8.357
Network diameter	17	7	7	6	9	9	11	11	7	3	18	7	6
Network radius	9	4	4	3	5	5	6	6	4	2	9	4	3
Characteristic path length	7.439	2.970	2.938	2.667	3.45	3.563	3.89	4.066	3.295	1.7	6.432	2.807	2.246
Clustering coefficient	0.438	0.452	0.46	0.497	0.544	0.578	0.55	0.587	0.492	0.467	0.434	0.576	0.673
Network density	0.027	0.131	0.135	0.242	0.116	0.163	0.097	0.096	0.231	0.5	0.024	0.185	0.31
Network heterogeneity	0.923	0.683	0.715	0.415	0.618	0.505	0.615	0.593	0.379	0.316	0.744	0.548	0.364
Network centralization	0.121	0.226	0.245	0.255	0.181	0.185	0.159	0.156	0.22	0.417	0.093	0.198	0.145
Connected components	102	105	105	69	97	85	95	94	84	51	103	107	76

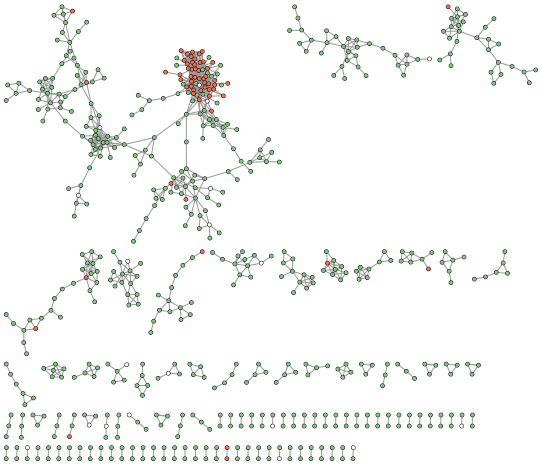
Note: The columns show several spectral networks obtained from a filtered similarity matrix. B is selected as the network for the prediction algorithm. The first four rows indicate the parameters used to construct the network; the other rows contain the network statistics. Tolerance is the maximum difference between peaks to be considered a match between spectra. Columns A-D examine the effect of the minimum score and number of maximum connections from a node. Columns E-J show the effect of reducing the bin width for the vector representations of the mass spectra, and the score cut-off. Additionally, columns K-M show the exploratory application of MS2DeepScore (MS2DS). The statistics were calculated with the built-in feature in Cytoscape.



(a) $\cos(\theta) \geq 0.7$



(b) $\cos(\theta) \geq 0.6$



(c) $\cos(\theta) \geq 0.5$

Figure 6.4: Spectral networks showing the active (red) and inactive (green) NR.AR compounds at different cut-offs of cosine similarity.

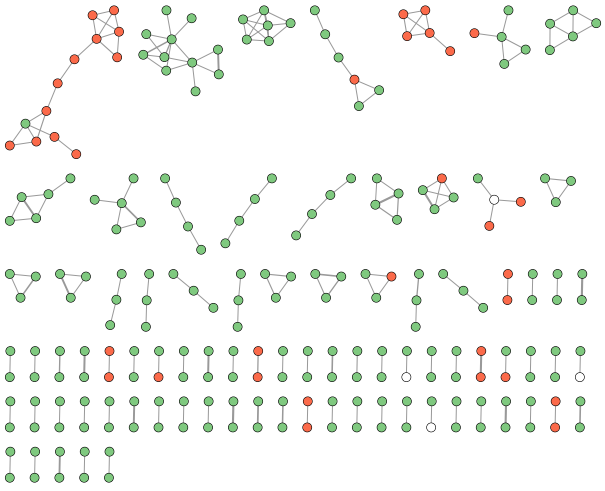


Figure 6.5: Spectral network I Table 6.2 showing the active (red) and inactive (green) NR.AR compounds with minimum cosine similarity of 0.6. For the construction of this network, the tolerance for the cosine calculation was changed from 0.1 *m/z* as in Figure 6.4 (b) to 0.001 *m/z*.

Selected network

All the networks were visually examined for their ability to form clusters of active compounds across the endpoints, taking into consideration the previously mentioned statistics. Then a specific network was selected for further analysis.

The chosen network had a cosine tolerance of 0.1 *m/z*, a minimum cosine similarity of 0.6, and maximum number of 10 connections from each node. This network demonstrated balanced clustering, as indicated by its low characteristic path length and the presence of predominantly active compounds clustered together for the NR.AR endpoint. Additionally, it exhibited a relatively high average number of neighbors, which was among the highest for the tested options. The network comprised 363 connected nodes, 409 edges, and an average of 6 neighbors per node.

These characteristics make the selected network a suitable candidate for further analysis and investigation, as it showcases a significant portion (27%) of the dataset and exhibits favorable clustering patterns for certain chemical classes. The network with labelled endpoints for (a) NR.AR and (b) SR.ARE can be seen in Figure 5.8 and for with all endpoints labels in the Appendix E.

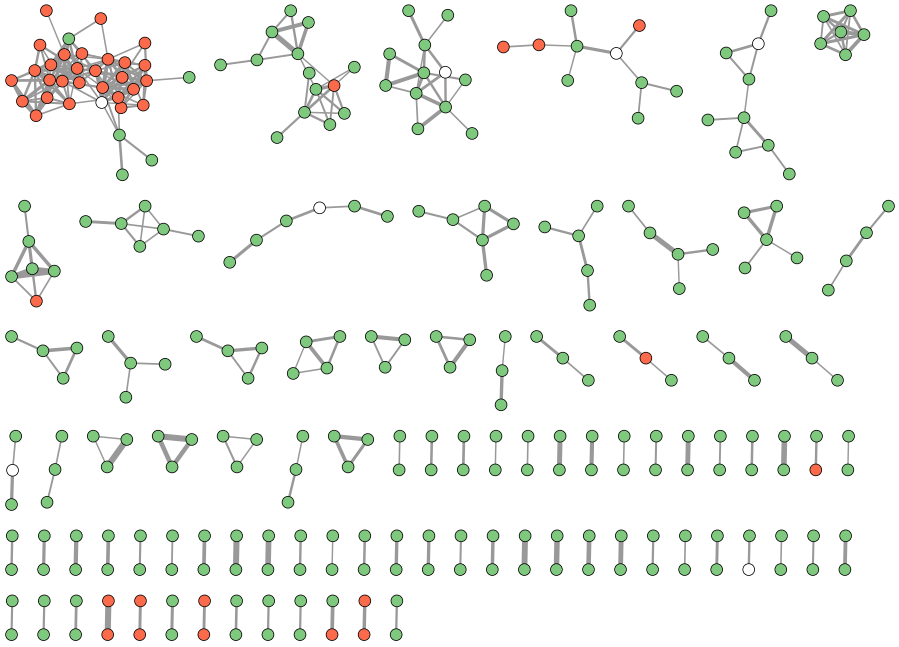


Figure 6.6: Cosine similarity network for the NR.AR endpoint with merged peaks. This network illustrates the effect of merging peaks of the combined spectra before the cosine calculation on the final distribution of active (in red) and inactive (in green) labels. From the combined spectra, peaks were merged if they were within a tolerance of 0.001 *m/z*. This merging had the effect of reducing the total number of peaks per spectra prior to the cosine calculation. After visual comparison of the network, not relevant changes in the distribution of the active labels across of endpoints was found compared to the combined spectra.

Clustering tendency

To verify the structural features responsible for spectral similarity, a visual examination of the closest pairs was conducted. It was observed that cyclic structures were one of the most prevalent structural features that led to spectral similarity. These compounds displayed high levels of spectral similarity, indicating similar fragmentation patterns. For instance, in the case of steroids, structural analogues such as norgestrel and dydrogesterone showed common peaks that resulted in higher cosine scores, over 0.80.

The driving factors behind these clustering patterns can vary significantly, with some clusters being composed of structural analogues while others contain chemicals with common fragments despite lacking close structural sim-

ilarities (e.g., due to the peaks matching not all matched peaks might correspond to the same fragment across molecules). For instance, triamcinolone, triamcinolone acetonide, and flunisolide are clustered together. They exhibit shared peaks (e.g., 147.08 *m/z* $C_{10}H_{11}O^+$ and 225.13 *m/z* $C_{16}H_{17}O^+$) that can be attributed to cyclic fragments from the steroid core structure. Steroids are known to be bioactive and interact with the androgen receptor[51] and they were shown to be clustered together for the NR.AR endpoint. In contrast, inactive clusters (yellow nodes) contain chemicals with shared fragments related to inactive substances, such as tributyl phosphate and triethyl phosphate, which both exhibit a peak for phosphate ion (98.98 *m/z*), which is known to not be bioactive.

Additionally, the clustering tendency is influenced by the parameters selected during the data preprocessing stage, such as noise reduction, scaling, and feature selection, as well as the construction of the network itself, including the minimum number of edges per node and the minimum similarity threshold.

Endpoints distribution

Some of the active endpoints, such as NR.AR, NR.AR.LBD, NR.ER, and NR.ER.LBD, were observed to form distinct clusters within the network. The case of clustered NR.AR and dispersed SR.ARE active endpoints are shown in Figure 5.8. Please refer to Appendix E for a detailed visualization of all endpoints. In the case of NR.AR, active compounds were found to cluster into a central group, characterized by common steroid core structures. The upper part predominantly consisted of compounds with alkyl sidechains, while the lower part comprised compounds without sidechains.

In contrast, the clustering tendency for the SR.ARE endpoint appeared to be more dispersed across the network. This could be attributed to different bioactivities that are not solely dependent on structural similarity. These clustering patterns provide valuable insights for identifying chemical classes and specific toxicity labels. For instance, if a mass spectrum exhibits high similarity to a cluster of molecules showing activity for NR.AR, it suggests a potential for similar activity.

In some cases, even structurally similar molecules may not always exhibit similar toxicity. For instance, 2-sulfanilamidoquinoxalin (active) and its neighbor sulfachlorpyridazine (inactive) share a similar chemical structure, but differ in activity potentially due to the presence of chloride substitutes that can alter the bioactivity of the molecule. Bioisosteres, which are structurally similar except for some group substituents, can also have different toxicity and bioavailability [52][53]. Similarly, stereoisomerism can also result in different biological outcomes [54].

It is important, however, to note that this approach may not be universally applicable to all endpoints, as later the predictions will show. The presence or absence of active labels in the clusters may vary depending on the specific toxicity endpoint.

Deep learning and similarity network

In addition, the application of MS2DeepScore [29] produced a similar distribution of active compounds within the network. Figure A.1 illustrates the network with the NR.AR endpoint labels, where it is evident that the active compounds are primarily clustered together, and to some extent also SR.ARE. Based on visual inspection of cluster formation and examination of pairs, a minimum score of 0.85 was found to be an appropriate threshold for visualization. Similar clustering tendencies were observed for the remaining endpoints, comparable to the cosine metric.

However, for the purpose of toxicity prediction, the cosine similarity metric was chosen due to its superior explainability and interpretability. A summary of statistics comparing the cosine-based network and the MS2DeepScore network can be found in Table 5.1. The constructed network demonstrates the potential of MS2DeepScore for spectra clustering, as it can capture compound similarities and identify structurally related compounds. Further investigations could focus on refining and optimizing the implementation of MS2DeepScore, potentially incorporating chemical classes (e.g., CANOPUS [28]) to enhance the accuracy of toxicity predictions.

k-NN algorithm

Several values of the hyperparameter k, ranging from 1 to 30, were tested on the training set using 5-fold stratified cross-validation to determine the optimal k based on recall. The effect of k on recall is presented in Figure A.2. The best k value obtained through cross-validation for all cases was 1. This could be attributed to the low or non-existent clustering tendency observed for most endpoints.

Figure A.2 (a) demonstrates a consistent recall value for NR.AR across different k values, suggesting that this endpoint is more likely to form a homogeneous cluster. A similar tendency was also observed in the molecular network, Figure 5.8 (a). The homogeneity of the NR.AR cluster might be influenced by the chemical space of the dataset, which could be unbalanced towards certain classes, such as steroids.

Conversely, for NR.AhR (b), the recall reaches its maximum at $k = 1$ and subsequently decreases. This indicates that few or no active compounds were found near the test nodes, in the local area. As k increases, the resulting value

becomes "diffused" throughout the regional area. Furthermore, due to the imbalanced labels in the training samples, the unknown class is more likely to receive votes from inactive compounds, which are more abundant. This can occur in despite of low similarities between their mass spectra.

The results of applying *k*-NN on the test set are summarized in Table 5.2. The highest recall and precision values were achieved for NR.AR, with values of 47.1 % and 44.4 %, respectively. On the other hand, the endpoints NR.PPAR.gamma and SR.ATAD5 had zero values for recall and precision. This result could respond to the scarcity of active labelled for some endpoints as shown in Table A.2, where NR.PPAR.gamma and SR.ATAD5 had the lowest active/inactive ratios. In some cases, the number of active compounds in the test set was less than 20, with only 7 active compounds for NR.PPAR.gamma. This potentially impact the predictions as there may be insufficient information available to support accurate predictions.

k-NN algorithm in locally connected mass spectra

The *k*-NN algorithm considers the votes of all neighbors based solely on their cosine similarity, which directly affects the recall values. As previously illustrated in Figure 5.7, spectra can exhibit very low similarity scores, leading to scenarios where no structurally related compounds can contribute to the endpoint prediction. While this approach gives insights about the distribution of endpoint labels on the entire chemical space, its prediction power is diminished when certain pairs lack the minimum spectral similarity required for read-across. Using the voting of only the locally connected nodes can improve the prediction capabilities as shown by the recall and precision in Table 5.3. In this way, the voting of the *k*-NN is restricted to the requirements set by the network.

The higher recall value for NR.AR shown in 5.3 could be attributed to the strong clustering tendency observed among spectra that share the steroid core structure. Chemical classification based on InChIKeys, obtained from ClassyFire [49], allowed the verification of clustering of spectra sharing the same chemical class. Steroids and steroid derivatives are represented as nodes with hexagonal shapes in Figure 5.9 and form part of the main cluster. The higher abundance of active labels and certain chemical classes could explain the higher recall for some endpoints. Additionally, a similarly high recall of NR.AR.LBD can be explained by their phi coefficient against NR.AR, as previously shown in the phi matrix of Figure 5.1.

Factors influencing the predictions

The predictions can be influenced by multiple factors, ranging from available mass spectra and toxicity data, similarity metric, and network topology. These aspects have been separately discussed in the previous sections. Here, applicability domain, regional and local similarity, and Structure Activity Relationship (SAR) are brought into further discussion.

Applicability domain

The ideal case for the prediction would involve clustering of mass spectra in a group with high similarity and the same endpoint activity; however, this is a rare case and may only apply to certain endpoints, e.g., NR.AR. The applicability domain significantly determines the reliability of predictions. By reducing the scope of the prediction with the spectral network, a minimum level of structural similarity is set as a criterion for the prediction. However, this has the drawback of narrowing the chemical space, which could increase the false negatives rate. In general, the predictions obtained by the spectral network are more reliable than those obtained from the *k*-NN algorithm alone. Predicting unknowns that fall outside the chemical space covered by the network is inherently less reliable compared to those within the domain.

Local and regional similarity

The distribution of compounds for most endpoints does not exhibit clear clusters with active labels, which poses challenges in accurately annotating toxicity and leads to low recall rates. It is important to note that most chemicals are non-toxic, not all new chemicals can be automatically classified as non-toxic. There were small non-active clusters where the local similarity would suggest non-activity, but they also included active compounds. The regional similarity can be considered as an indication of the prediction uncertainty. For instance, if the regional similarity is consistent (e.g., most spectra have active labels with no nearby inactive ones), it provides greater confidence in the accuracy of the predicted value within the local area. This regional similarity analysis can aid in assessing the reliability of the predictions.

SAR *paradox*

When predicting toxicity, it is crucial to consider whether non-property assignments, such as labeling a certain endpoint as inactive, can be universally applied to chemicals lacking this property. Structural factors alone may not provide a complete understanding of toxicity, which may explain lower accuracy in some cases. For instance, 2-acetyl amino fluorene and 4-acetyl amino

fluorene are structurally similar, but they differ in toxicity, with the former labeled as a liver carcinogen while the latter is not. Therefore, in some cases structural similarity alone may not be enough to achieve the desired prediction accuracy. Alternatives include classification based on mode of action analogues, for example those sharing the same receptor activation properties or common metabolites and degradation products. Integrating systems biology could complement the chemical toxicity evaluation with a biological context and potentially improve the overall accuracy.

7 Conclusions and future perspectives

In this study, the application of a mass spectra similarity networking was studied for predicting toxicity endpoints based on tandem mass spectra. First a *k*-NN classification was explored to determine the ability of predicting active labels in the whole dataset. The classification results showed low recall and precision, being the highest for NR.AR at 47.1 % and 44.4 %, respectively. The implementation of mass spectra networks showed a better predictive ability for the same endpoint, with recall and precision of 81.8 % and 75.0 %, respectively. The definition of the applicability domain using the networks narrows the chemical space and improves the recall and precision. The higher tendency of clustering within the NR.AR endpoint can be due to the prevalence of the steroid core structure in the dataset. The results showed promising outcomes in terms of distinguishing between active and inactive compounds for NR.AR, NR.AR.LBD, NR.ER and NR.ER.LBD, suggesting that the spectral similarity networks can capture meaningful relationships among chemicals based on their mass spectra that can be used to predict activity of toxicity endpoints.

To improve future predictions, several aspects can be considered. First, expanding the dataset to include a larger number of diverse compounds could potentially enhance the model's performance by capturing a broader range of chemical activity and classes. Additionally, incorporating feature selection techniques, such as machine learning algorithms or ensemble methods, could help identifying the most informative fragments related to specific endpoints. Moreover, exploring different similarity metrics and optimizing the spectral networking parameters may further enhance the performance of the predictions. Lastly, adopting systems biology approaches can provide valuable insights into the mechanisms underlying toxicity that is not related to chemical structure alone. By integrating data from multiple domains, a more comprehensive understanding of the molecular basis of toxicity can enhance future predictions.

References

[1] Philip J Landrigan, Richard Fuller, Nereus J R Acosta, Olusoji Adeyi, and et. al. The lancet commission on pollution and health. *The Lancet*, 391(10119):462–512, February 2018.

[2] European Chemicals Agency. Registered substances - european chemicals agency. Website: https://echa.europa.eu/information-on-chemicals/registered-substances/.

[3] Beate I. Escher, Heather M. Stapleton, and Emma L. Schymanski. Tracking complex mixtures of chemicals in our changing environment. *Science*, 367(6476):388–392, 2020.

[4] Anekwe Jennifer Ebele, Mohamed Abou-Elwafa Abdallah, and Stuart Harrad. Pharmaceuticals and personal care products (PPCPs) in the freshwater aquatic environment. *Emerging Contaminants*, 3(1):1–16, March 2017.

[5] Fenna Sillé. The exposome – a new approach for risk assessment. *ALTEX*, pages 3–23, 2020.

[6] Daniel Krewski, Daniel Acosta, Melvin Andersen, Henry Anderson, John C. Bailar, Kim Boekelheide, Robert Brent, Gail Charnley, Vivian G. Cheung, Sidney Green, Karl T. Kelsey, Nancy I. Kerkvliet, Abby A. Li, Lawrence McCray, Otto Meyer, Reid D. Patterson, William Pennie, Robert A. Scala, Gina M. Solomon, Martin Stephens, James Yager, Lauren Zeise, and Staff of Committee on Toxicity Test. Toxicity testing in the 21st century: A vision and a strategy. *Journal of Toxicology and Environmental Health, Part B*, 13(2-4):51–138, June 2010.

[7] Susan D. Richardson. Environmental mass spectrometry: Emerging contaminants and current issues. *Analytical Chemistry*, 84(2):747–778, December 2011.

[8] Julia E. Rager, Mark J. Strynar, Shuang Liang, Rebecca L. McMahen, Ann M. Richard, Christopher M. Grulke, John F. Wambaugh, Kristin K.

Isaacs, Richard Judson, Antony J. Williams, and Jon R. Sobus. Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring. *Environment International*, 88:269–280, March 2016.

[9] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, and et.al. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, July 2010.

[10] MoNA - massbank of north america. Online: https://mona.fiehnlab.ucdavis.edu/.

[11] David S Wishart, Yannick Djoumbou Feunang, Ana Marcu, An Chi Guo, and et. al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, 46(D1):D608–D617, November 2017.

[12] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, and et. al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology*, 34(8):828–837, August 2016.

[13] Yasuyuki Zushi. Direct prediction of physicochemical properties and toxicities of chemicals from analytical descriptors by GC–MS. *Analytical Chemistry*, 94(25):9149–9157, June 2022.

[14] Pilleriin Peets, Wei-Chieh Wang, Matthew MacLeod, Magnus Breitholtz, Jonathan W. Martin, and Anneli Kruve. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. *Environ. Sci. Technol.*, 56(22):15508–15517, November 2022.

[15] Philippe Grandjean and Philip J Landrigan. Neurobehavioural effects of developmental toxicity. *The Lancet Neurology*, 13(3):330–338, March 2014.

[16] Laura N. Vandenberg, Marlene Ågerstrand, Anna Beronius, Claire Beausoleil, and et. al. A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. *Environmental Health*, 15(1), July 2016.

[17] Thomas Unterthiner, Andreas Mayr, Günter Klambauer, and Sepp Hochreiter. Toxicity prediction using deep learning, 2015.

[18] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.*, 3, February 2016.

[19] Gabriel Idakwo, Joseph Luttrell, Minjun Chen, Huixiao Hong, Zhaoxian Zhou, Ping Gong, and Chaoyang Zhang. A review on machine learning methods for in silico toxicity prediction. *Journal of Environmental Science and Health, Part C*, 36(4):169–191, October 2018.

[20] A. Paini, J.A. Leonard, E. Joossens, J.G.M. Bessems, and et. al. Next generation physiologically based kinetic (NG-PBK) models in support of regulatory decision making. *Computational Toxicology*, 9:61–72, February 2019.

[21] Heather L. Ciallella and Hao Zhu. Advancing computational toxicology in the big data era by artificial intelligence: Data-driven and mechanism-driven modeling for chemical toxicity. *Chemical Research in Toxicology*, 32(4):536–547, March 2019.

[22] Hao Zhu, Mounir Bouhifd, Elizabeth Donley, Laura Egnash, Nicole Kleinstreuer, E. Dinant Kroese, Zhichao Liu, Thomas Luechtefeld, Jessica Palmer, David Pamies, Jie Shen, Volker Strauss, Shengde Wu, and Thomas Hartung. Supporting read-across using biological data. *ALTEX - Alternatives to animal experimentation*, 33(2):167–182, May 2016.

[23] OECD. *Guidance on Grouping of Chemicals, Second Edition*. 2017.

[24] Thomas Luechtefeld, Alexandra Maertens, James M. McKim, Thomas Hartung, Andre Kleensang, and Vanessa Sá-Rocha. Probabilistic hazard assessment for skin sensitization potency by dose-response modeling using feature elimination instead of quantitative structure-activity relationships. *Journal of Applied Toxicology*, 35(11):1361–1371, June 2015.

[25] Ruili Huang and Menghang Xia. Editorial: Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. *Frontiers in Environmental Science*, 5, January 2017.

[26] Yi Yang, Xiaohui Liu, Chengpin Shen, Yu Lin, Pengyuan Yang, and Liang Qiao. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications*, 11(1), January 2020.

[27] Mingxun Wang, Alan K. Jarmusch, Fernando Vargas, Alexander A. Aksenov, and et.al. Mass spectrometry searches using MASST. *Nature Biotechnology*, 38(1):23–26, January 2020.

[28] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A. Hoffmann, Daniel Petras, William H. Ger-

wick, Juho Rousu, Pieter C. Dorrestein, and Sebastian Böcker. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology*, 39(4):462–471, November 2020.

[29] Florian Huber, Sven van der Burg, Justin J. J. van der Hooft, and Lars Ridder. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J Cheminform*, 13(1):84, December 2021.

[30] Niek F. de Jonge, Joris J. R. Louwen, Elena Chekmeneva, Stephane Camuzeaux, Femke J. Vermeir, Robert S. Jansen, Florian Huber, and Justin J. J. van der Hooft. MS2query: reliable and scalable MS2 mass spectra-based analogue search. *Nature Communications*, 14(1), March 2023.

[31] Danijel Djukovic, Daniel Raftery, and Nagana Gowda. Mass spectrometry and NMR spectroscopy based quantitative metabolomics. In *Proteomic and Metabolomic Approaches to Biomarker Discovery*, pages 289–311. Elsevier, 2020.

[32] Adam Amara, Clément Frainay, Fabien Jourdan, Thomas Naake, Steffen Neumann, Elva María Novoa-del Toro, Reza M Salek, Liesa Salzer, Sarah Scharfenberg, and Michael Witting. Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Front. Mol. Biosci.*, 9:841373, March 2022.

[33] Yuping Cai, Zhiwei Zhou, and Zheng-Jiang Zhu. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. *TrAC Trends in Analytical Chemistry*, 158:116903, January 2023.

[34] Caiming Tang, Guangshi Chen, Bin Jiang, Jianhua Tan, Yutao Liang, Hui Lin, Yanhong Zeng, Xiaojun Luo, Bixian Mai, and Xianzhi Peng. High-performance nontarget analysis of halogenated organic compounds in tap water, fly ash, soil and sediment using ultrahigh resolution mass spectrometry and scripting approaches based on cl/br-specific search algorithms. *Analytica Chimica Acta*, 1204:339618, April 2022.

[35] Francois Lestremau, Alexandre Levesque, Abdelmoughit Lahssini, Tanguy Magnan de Bornier, Romain Laurans, Azziz Assoumani, and Hugues Biaudet. Development and implementation of automated qualification processes for the identification of pollutants in an aquatic environment from high-resolution mass spectrometric nontarget screening data. *ACS*, 3(3):765–772, February 2023.

[36] Thomas Naake and Emmanuel Gaquerel. MetCirc: navigating mass spectral similarity in high-resolution MS/MS metabolomics data. *Bioinformatics*, 33(15):2419–2420, August 2017.

[37] Wout Bittremieux, Robin Schmid, Florian Huber, Justin JJ van der Hooft, Mingxun Wang, and Pieter C Dorrestein. Comparison of Cosine, Modified Cosine, and Neutral Loss Based Spectrum Alignment For Discovery of Structurally Related Molecules. preprint, Bioinformatics, June 2022.

[38] Wout Bittremieux, Damon H. May, Jeffrey Bilmes, and William Stafford Noble. A learned embedding for efficient joint analysis of millions of mass spectra. preprint, Bioinformatics, November 2018.

[39] Chunyuan Qin, Xiyang Luo, Chuan Deng, Kunxian Shu, Weimin Zhu, Johannes Griss, Henning Hermjakob, Mingze Bai, and Yasset Perez-Riverol. Deep learning embedder method and tool for mass spectra similarity search. *Journal of Proteomics*, 232:104070, February 2021.

[40] Yasuyuki Zushi. Direct Prediction of Physicochemical Properties and Toxicities of Chemicals from Analytical Descriptors by GC–MS. *Anal. Chem.*, 94(25):9149–9157, June 2022.

[41] Ari M. Frank, Nuno Bandeira, Zhouxin Shen, Stephen Tanner, Steven P. Briggs, Richard D. Smith, and Pavel A. Pevzner. Clustering millions of tandem mass spectra. *Journal of Proteome Research*, 7(1):113–122, December 2007.

[42] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, and et. al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol*, 34(8):828–837, August 2016.

[43] Leonardo Perez De Souza, Saleh Alseekh, Yariv Brotman, and Alisdair R Fernie. Network-based strategies in metabolomics data analysis and interpretation: from molecular networking to biological interpretation. *Expert Review of Proteomics*, 17(4):243–255, April 2020.

[44] Daniela Oberleitner, Robin Schmid, Wolfgang Schulz, Axel Bergmann, and Christine Achten. Feature-based molecular networking for identification of organic micropollutants including metabolites by non-target analysis applied to riverbank filtration. *Analytical and Bioanalytical Chemistry*, 413(21):5291–5300, July 2021.

[45] Gang Wu, Xuebing Wang, Xuxiang Zhang, Hongqiang Ren, Yanru Wang, Qingmiao Yu, Si Wei, and Jinju Geng. Nontarget screening based

on molecular networking strategy to identify transformation products of citalopram and sertraline in wastewater. *Water Research*, 232:119509, April 2023.

[46] Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A. Shahane, Anna Rossoshek, and Anton Simeonov. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3, January 2016.

[47] MassBank - a public repository of reference spectra for identification of chemical compounds. `https://massbank.eu/MassBank/`. Accessed in February 2023.

[48] Florian Huber, Stefan Verhoeven, Christiaan Meijer, Hanno Spreeuw, Efraín Manuel Villanueva Castilla, Cunliang Geng, Justin JJ van der Hooft, Simon Rogers, Adam Belloum, Faruk Diblen, et al. matchms-processing and similarity evaluation of mass spectrometry data. *bioRxiv*, pages 2020–08, 2020.

[49] Yannick Djoumbou Feunang, Roman Eisner, Craig Knox, Leonid Chepelev, Janna Hastings, Gareth Owen, Eoin Fahy, Christoph Steinbeck, Shankar Subramanian, Evan Bolton, Russell Greiner, and David S. Wishart. Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics*, 8(1), November 2016.

[50] K. Sandvig, S. Grimmer, T.G. Iversen, K. Rodal, M.L. Torgersen, P. Nicoziani, and B. van Deurs. Ricin transport into cells: studies of endocytosis and intracellular transport. *International Journal of Medical Microbiology*, 290(4):415–420, 2000.

[51] Wenqing Gao, Casey E. Bohl, and James T. Dalton. Chemistry and structural biology of androgen receptor. *Chemical Reviews*, 105(9):3352–3370, August 2005.

[52] Nicholas A. Meanwell. Erratum to: The influence of bioisosteres in drug design: Tactical applications to address developability problems. In *Topics in Medicinal Chemistry*, pages 389–390. Springer Berlin Heidelberg, 2014.

[53] Miloš Svirčev, Mirjana Popsavin, Aleksandar Pavić, Branka Vasiljević, Marko V. Rodić, Sanja Djokić, Jelena Kesić, Bojana Srećo Zelenović, Velimir Popsavin, and Vesna Kojić. Design, synthesis, and biological evalu-

ation of thiazole bioisosteres of goniofufurone through in vitro antiproliferative activity and in vivo toxicity. *Bioorganic Chemistry*, 121:105691, April 2022.

[54] S. FABRO, R. L. SMITH, and R. T. WILLIAMS. Toxicity and teratogenicity of optical isomers of thalidomide. *Nature*, 215(5098):296–296, July 1967.

A Distribution of endpoint labels in the toxicity dataset

Table A.1: Distribution of endpoint labels in the entire toxicity dataset

Endpoint	Active	Inactive	NA	Active/Inactive %	NA/Total %
NR.AhR	890	6699	1386	13.3	15.4
NR.AR	325	8011	639	4.1	7.1
NR.AR.LBD	250	7504	1221	3.3	13.6
NR.Aromatase	366	6333	2276	5.8	25.4
NR.ER	876	6233	1866	14.1	20.8
NR.ER.LBD	386	7630	959	5.1	10.7
NR.PPAR.gamma	235	7236	1504	3.2	16.8
SR.ARE	1110	5632	2233	19.7	24.9
SR.ATAD5	330	7810	835	4.2	9.3
SR.HSE	411	7075	1489	5.8	16.6
SR.MMP	1039	5679	2257	18.3	25.1
SR.p53	500	7305	1170	6.8	13.0
Average:				8.6	16.6

Note: Labels after processing the toxicity tabular dataset. NA, not available in the toxicity dataset. Total compounds: 8975

Table A.2: Distribution of endpoint labels for the mass spectra collection

Endpoint	Active	Inactive	NA	Active/Inactive %	NA/Total %
NR.AhR	238	956	156	24.9	11.6
NR.AR	87	1213	50	7.2	3.7
NR.AR.LBD	64	1151	135	5.6	10.0
NR.Aromatase	103	902	345	11.4	25.6
NR.ER	188	899	263	20.9	19.5
NR.ER.LBD	86	1162	102	7.4	7.6
NR.PPAR.gamma	37	1081	232	3.4	17.2
SR.ARE	202	876	272	23.1	20.1
SR.ATAD5	59	1219	72	4.8	5.3
SR.HSE	63	1141	146	5.5	10.8
SR.MMP	191	822	337	23.2	25.0
SR.p53	89	1125	136	7.9	10.1
Average:				12.1	13.9

Note: NA, not available in the toxicity dataset. Total compounds: 1350

B Cosine similarity and MS2DeepScore

Table A.3: Frequency distribution of cosine score

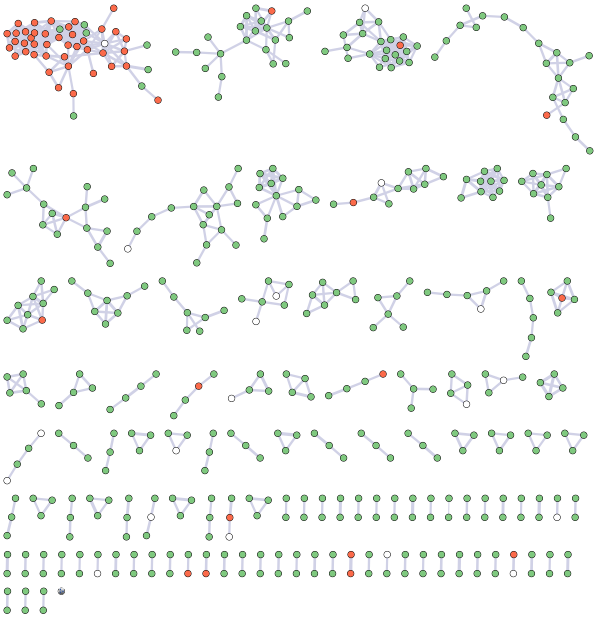
x0	x1	f	F	fr	Fr	f'r	F'r
0.0	0.1	782031	782031	0.8588	0.8588	910575	1.0000
0.1	0.2	96231	878262	0.1057	0.9645	128544	0.1412
0.2	0.3	23226	901488	0.0255	0.9900	32313	0.0355
0.3	0.4	6073	907561	0.0067	0.9967	9087	0.0100
0.4	0.5	1829	909390	0.0020	0.9987	3014	0.0033
0.5	0.6	772	910162	0.0008	0.9995	1185	0.0013
0.6	0.7	274	910436	0.0003	0.9998	413	0.0005
0.7	0.8	85	910521	0.0001	0.9999	139	0.0002
0.8	0.9	35	910556	0.0000	1.0000	54	0.0001
0.9	1.0	19	910575	0.0000	1.0000	19	0.0000

Note: Cosine scores of all unique pairs of the combined spectra. Tolerance cosine: 0.1 *m/z*. x0 lower limit, x1 upper limit, f frequency, F cumulative frequency, rf, relative frequency, rF cumulative relative frequency, rf' inverse cumulative frequency, rF' inverse cumulative relative frequency.

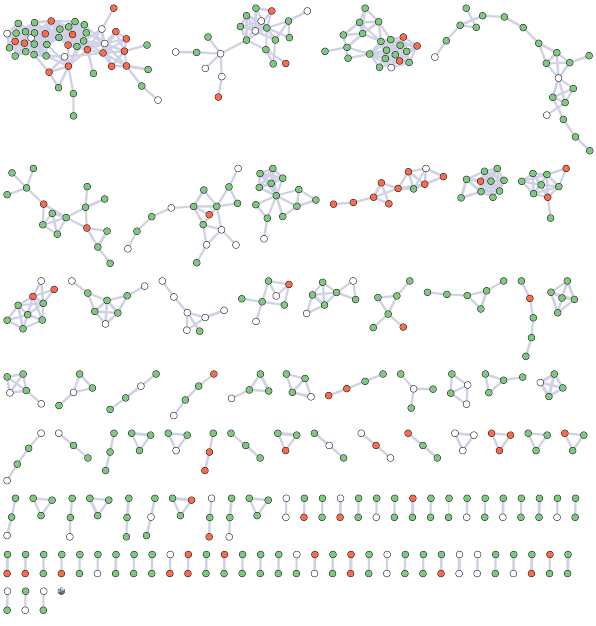
Table A.4: Frequency distribution of MS2DeepScore

x0	x1	f	F	fr	Fr	f'r	F'r
0.0	0.1	78392	78392	0.0861	0.0861	910574	1.0000
0.1	0.2	223489	301881	0.2454	0.3315	832182	0.9139
0.2	0.3	249682	551563	0.2742	0.6057	608693	0.6685
0.3	0.4	185945	737508	0.2042	0.8099	359011	0.3943
0.4	0.5	102789	840297	0.1129	0.9228	173066	0.1901
0.5	0.6	45914	886211	0.0504	0.9732	70277	0.0772
0.6	0.7	17433	903644	0.0191	0.9924	24363	0.0268
0.7	0.8	5274	908918	0.0058	0.9982	6930	0.0076
0.8	0.9	1366	910284	0.0015	0.9997	1656	0.0018
0.9	1.0	290	910574	0.0003	1.0000	290	0.0003

Note: MS2DeepScore similarity score for all unique pairs of the combined spectra, 0.1 *m/z* bin width. x0 lower limit, x1 upper limit, f frequency, F cumulative frequency, rf, relative frequency, rF cumulative relative frequency, rf' inverse cumulative frequency, rF' inverse cumulative relative frequency.



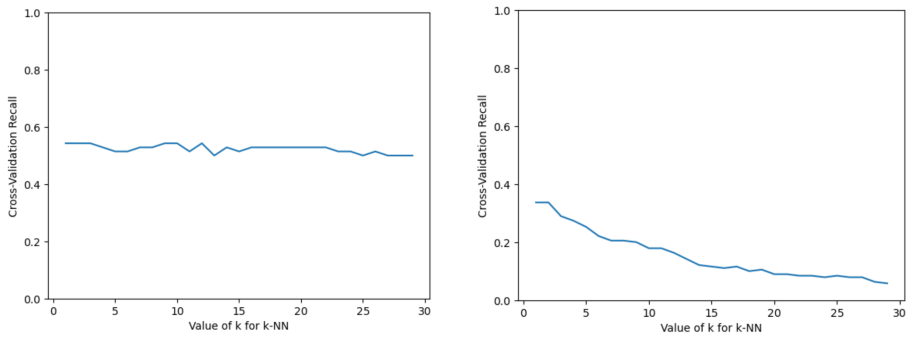
(a)



(b)

Figure A.1: Spectral similarity network based on the MS2DeepScore similarity showing the (a) NR.AR and (b) SR.ARE endpoints. Nodes represent active (red), inactive (green) labels for the endpoints. Nodes in blank did not have a label. The edge width indicates the intensity of the score.

C *k*-NN cross-validation



(a) (b)

Figure A.2: (a) NR.AR and (b) NR.AhR recall for different values of *k* for the *k*-NN algorithm with 5-fold cross-validation.

D Sample MS^2 features

The tandem mass spectra from a wastewater sample was used to test the spectral network. The MS^2 features were provided by Kruve Lab, Stockholm University. The sample has been collected from an effluent of a wastewater treatment plant in Stockholm in March 2022. Prior to the analysis, the sample has been stored at -20°C and filtered through a 0.45 μm filter. The measurement conditions are summarized in Table A.5 and A.6. The data processing parameters for MS-DIAL (v. 4.80) are summarized in Table A.7.

Table A.5: Liquid chromatography conditions

Parameter	Description
LC system	Dionex UltimateTM 3000 UHPLC (Thermo Fischer Scientific, USA)
Column	Kinetex (Phenomenex, Germany) C18 reversed phase, 150 x 3.0 mm, 2.6 um
Column temperature	40 °C
Mobile phase	Solution A: 0.1% formic acid in HPLC grade water Solution B: 0.1% formic acid in acetonitrile:water (95:5)
Gradient	Linear increase of 5% B to 100% B with 20.0 min Held for 5 min. Lowered back to 5% B with 0.1 min
Equilibration time	4.9 min

Source: Kruve Lab

Table A.6: HRMS measurement conditions

Parameter	Description
HRMS system	Q Exactive Orbitrap (Thermo Fisher Scientific, USA) with electrospray ionization (ESI)
Ionization mode	Positive
Acquisition mode	DDA and DIA
Collision energy	20 V and 70 V
Resolution	30000
ESI spray voltage	3.5 kV
Capillary temperature	320 °C
Maximum spray current	100 uA
S-lens RF level	50%
Gas	50 arbitrary units (AU) for the sheath gas 3.0 AU for aux gas and 0.0 AU for spare gas

Source: Kruve Lab

Table A.7: MS-DIAL processing parameters

Parameter	Description
Data collection	
MS1 tolerance	0.001 Da
MS2 tolerance	0.003 Da
RT begin	2 min
RT end	25 min
MS1 mass range begin	90 Da
MS1 mass range end	1050 Da
MS/MS mass range begin	90 Da
MS/MS mass range end	1050 Da
Maximum charged number	2
Consider Cl and Br elements	Yes
Peak detection	
Min peak height	10000 amplitude
Mass slice width	0.07 Da
Smoothing method	Linear weighted moving average
Smoothing level	5 scan
Min peak width	8 scan
MS2Dec	
Sigma window value	1
MS/MS abundance cut off	0 amplitude
Exclude after precursor ion	Yes
Keep the isotopic ions until	5 Da
Keep the isotopic ions w/o MS2Dec	Yes
Adduct	
Positive ionization mode	[M+H]+, [M+Na]+, [M+NH4]+
Alignment	
RT tolerance	0.75 min
MS1 tolerance	0.003 Da
RT factor	0.5
MS1 factor	0.5
N% detected in at least one group	100%
Remove features based on blank information	Sample max/blank average
Fold change	5
Keep ref.matched features	Yes
Keep suggested w/o MS2 features	No
Keep removable features and assign the tag	Yes
Gap filling by compulsion	Yes

Source: Kruve Lab

E Spectral similarity networks with labels for all the endpoints

The selected spectral network for all endpoint labels is shown below. The spectral network uses a minimum cosine similarity of 0.6 and maximum number of edged from a node of 10. The active compounds are represented as nodes in red and the inactive ones in green. Nodes without color do not have activity information for the endpoint.

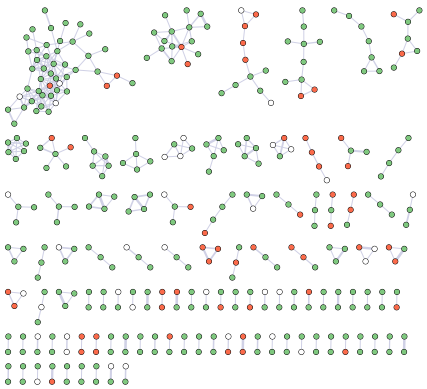


Figure A.3: NR.AhR.pdf

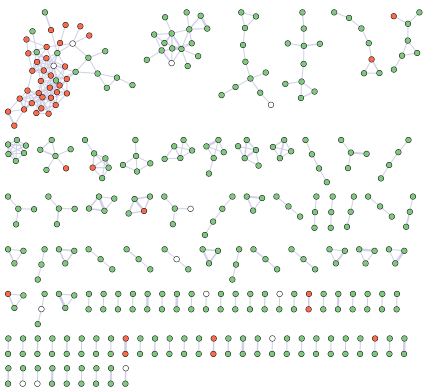


Figure A.4: NR.AR.pdf

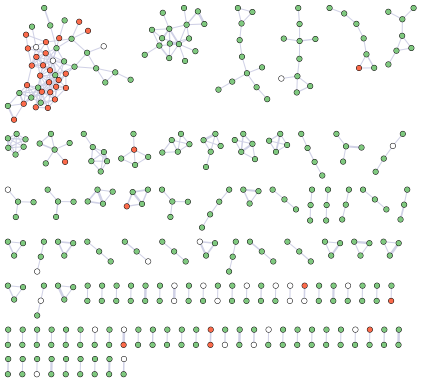


Figure A.5: NR.AR.LBD.pdf

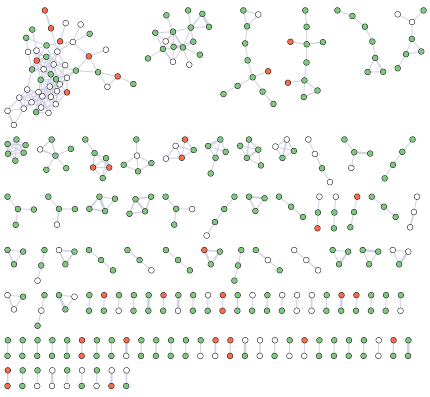


Figure A.6: NR.Aromatase.pdf

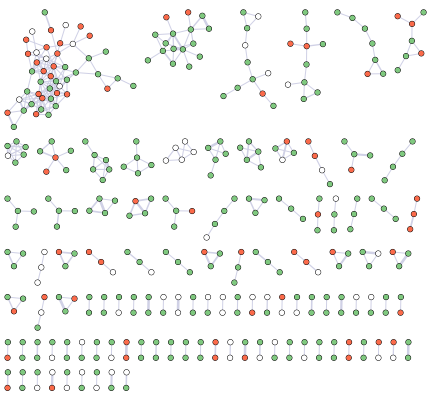


Figure A.7: NR.ER.pdf

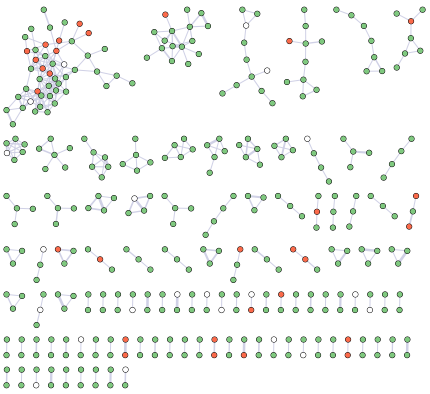


Figure A.8: NR.ER.LBD.pdf

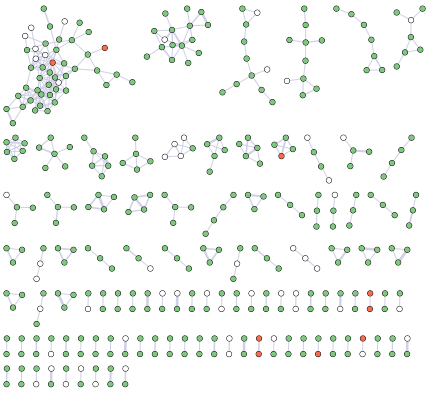


Figure A.9: NR.PPAR.gamma.pdf

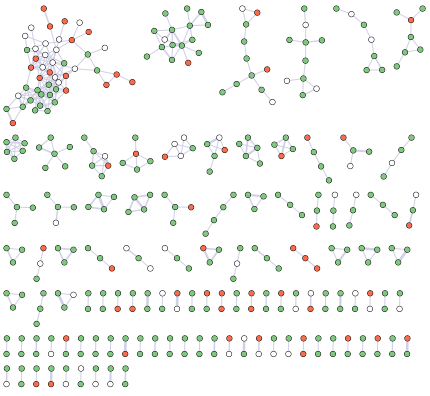


Figure A.10: SR.ARE.pdf

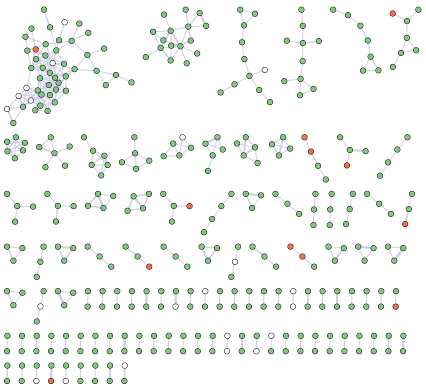


Figure A.11: SR.ATAD5.pdf

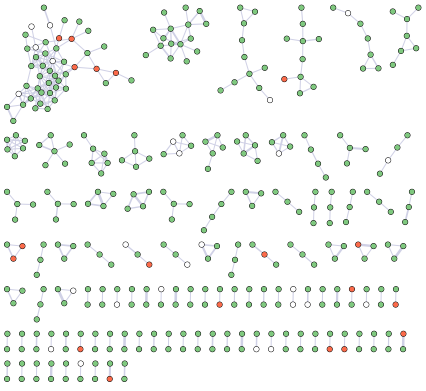


Figure A.12: SR.HSE.pdf

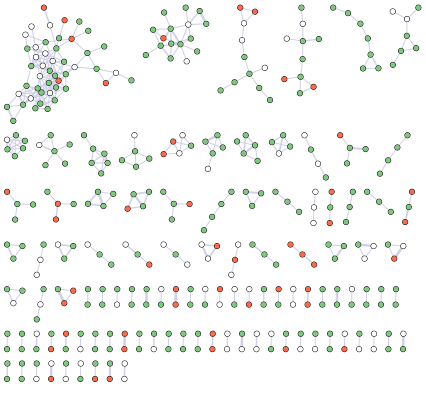


Figure A.13: SR.MMP.pdf

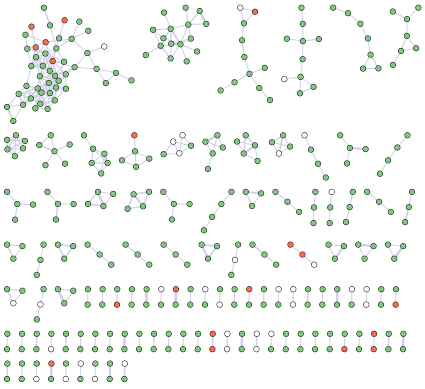


Figure A.14: SR.p53.pdf