Machine Learning-Enabled Prediction of Metabolite Response in Genetic Disorders

Christel Sirocchi[1,*,†], Federica Biancucci[2,†], Matteo Donati[1], Nunzio D'Amore[1], Riccardo Benedetti[2], Alessandro Bogliolo[1], Stefano Ferretti[1], Mauro Magnani[2], Michele Menotta[2], Muhammad Suffian[1] and Sara Montagna[1]

[1]*Department of Pure and Applied Sciences, University of Urbino, Piazza della Repubblica 13, 61029, Urbino, Italy*
[2]*Department of Biomolecular Sciences, University of Urbino, Via Saffi 2, 61029 Urbino, Italy*

Abstract

Metabolomics has emerged as a promising discipline in pharmaceuticals and preventive healthcare, holding great potential for disease detection and drug testing. However, analysing large metabolomics datasets remains challenging, with available methods generally relying on limited and incompletely annotated biological pathways. This study introduces a novel approach that leverages machine learning classifiers trained on molecular fingerprints of metabolites, to predict their responses under specific experimental conditions. The model is evaluated on mass spectrometry metabolomic data for a cellular model of the genetic disease Ataxia Telangiectasia. In this study, metabolite structures are encoded using the Morgan fingerprint, a well-established technique widely embraced in drug discovery. The suitability of this fingerprinting method, in generating unique structural encodings for detected metabolites, is analysed, and strategies to mitigate resolution limitations inherent to this fingerprint are introduced. Machine learning classifiers are trained on these fingerprints and exhibit satisfactory performance, providing evidence that the structural encoding holds predictive power over the metabolic response. Feature importance analysis, conducted on the best-performing models, identifies the chemical configurations that have the greatest influence to the classification process, shedding light on affected biological processes. Remarkably, this analysis not only identifies metabolites known to participate in affected pathways but also discovers metabolites not previously associated with the disease, opening up novel opportunities for further exploration. As an initial exploration of the proposed approach, this work lays the foundation for future research that leverages alternative structural encodings, diverse machine learning models, and explainability tools.

Keywords

Ataxia telangiectasia, mass spectrometry, metabolic pathways, metabolomics, machine learning

1. Introduction

Metabolomics, as the quantitative study of small molecule substrates and products of cellular metabolism, occupies a unique position in the -omics landscape due to its proximity to the phenotype [1]. The metabolome, representing the final product of genomic, transcriptomic, and proteomic processes, provides a direct readout of the physiological state of an organism [2].

Second AIxIA Workshop on Artificial Intelligence For Healthcare, November 6, 2023, Rome, Italy
*Corresponding author.
†These authors contributed equally.
c.sirocchi2@campus.uniurb.it (C. Sirocchi); federica.biancucci@uniurb.it (F. Biancucci)
0000-0002-5011-3068 (C. Sirocchi); 0009-0006-2567-5460 (F. Biancucci)

Metabolomic profiling of diseased and healthy tissues can help uncover the disease mechanisms of action and identify metabolic signatures, aiding the identification of potential drug targets [3]. Additionally, metabolomics can help assess the effects of candidate treatments, evaluating the response at the metabolic level [4]. Therefore, metabolomics serves as an indispensable tool in preventive healthcare as well as pharmaceutical research and development, with the potential to enable timely disease diagnosis, early detection, effective drug testing, disease monitoring, and personalised treatment strategies [5].

The potential of metabolomics lies in characterising and quantifying the metabolites present in a particular biological system using a combination of analytical tools [5]. Central to this endeavour are the advancements in mass spectrometry technologies, including in-line chromatographic separation modes, ionisation techniques, mass analysers, and detection methods. Untargeted High-Resolution Mass Spectrometry (HR-MS) has emerged as a powerful tool, capable of simultaneously detecting a vast array of metabolites, thereby facilitating the identification of metabolic alterations and the discovery of novel metabolites [6]. However, the analysis and biological interpretation of the resulting large and complex datasets remains challenging.

In metabolomics studies, a typical approach involves comparing samples from a normal state to those from a perturbed state, often induced by genetic knockout or the administration of specific treatments [6]. Pathway enrichment analysis is the prevalent method for comparing such samples, as it identifies pathways that exhibit a higher degree of overlap with significantly under or over-expressed metabolites than would be expected by chance. This analysis aids in pinpointing the affected pathways and unravelling the underlying metabolic mechanisms. However, this approach poses various challenges, as it heavily relies on the existing knowledge of biological pathways, which is not comprehensive or fully annotated. Furthermore, it should be applied to metabolomic data with caution, as the method was primarily developed for transcriptomic data and guidelines for best practices for metabolic pathway enrichment analysis are still lacking [7]. Lastly, enrichment results were found to be quite sensitive to the pathway definitions used by different metabolomic databases [8]. Consequently, despite significant technological advancements, the full potential of metabolomics remains untapped, as metabolic data analysis often leans on the limited knowledge of known pathways. To achieve a comprehensive understanding of cellular metabolism, there is a pressing need to extend beyond the boundaries of known pathways and consider all detected metabolites in the analyses.

When metabolites lack annotations of their roles in enzymatic reactions and metabolic pathways, chemical similarity emerges as a valuable tool for unravelling potential relationships with other metabolites. This approach leverages the known tendency for chemically similar compounds to be found in close proximity within metabolic pathways [9]. Chemical structures can be analysed to identify enriched chemical features within a specific experimental condition, thereby providing insights into the affected cellular processes. The structure of metabolites can be represented using fingerprints, which are binary vectors that capture the presence or absence of structural properties [10]. Machine Learning (ML) models can then be trained on the structural encoding of metabolites to predict whether the metabolite level significantly differs in the sample under study compared to a control. Relationships between the chemical structures and the metabolic response to given experimental conditions can therefore be explored in a data-driven manner, opening new avenues for understanding metabolic ways and identifying biomarkers.

This study examines the interplay between chemical structure and metabolic response through a novel approach that combines fingerprinting for structure encoding and machine learning for the identification of relevant chemical substructures within a given condition. The approach is evaluated using a cellular model of Ataxia Telangiectasia (AT), a rare neurodegenerative disorder caused by mutations in the Ataxia Telangiectasia Mutated (ATM) gene, known to disrupt numerous metabolic pathways [11]. The study analyses metabolic data obtained through untargeted mass spectrometry, comparing the disease cellular model with a control sample.

In the study, metabolite structures are encoded using the Morgan fingerprint, the most widely embraced and rigorously validated molecular fingerprinting technique in drug discovery [12]. The suitability of this fingerprinting method for providing unique structural encodings for the detected metabolites is carefully evaluated. Resolution limitations, leading to duplicate fingerprints for distinct metabolites, are addressed by proposing extensions to the Morgan fingerprint. ML classifiers are then trained on the fingerprints to predict down-regulated metabolites [13]. The emphasis on down-regulation stems from the disease's well-established tendency to inhibit cellular activities, although up-regulated metabolites can be similarly explored. The trained models achieve satisfactory performance, providing evidence that the structural encoding of a metabolite holds predictive value over its response to a particular condition. Finally, the study analyses feature importance to identify the specific chemical substructures that contribute to the classification process, shedding light on the biological pathways affected by the disease. Remarkably, feature importance computed for one of the best-performing models identifies metabolites known to participate in affected pathways, thereby validating existing knowledge, as well as groups of metabolites not previously associated with AT, opening up novel opportunities for further investigation.

In summary, this article introduces a novel approach that harnesses molecular fingerprinting and machine learning for the analysis of large metabolomic datasets. The evaluation focuses on three key aspects: the suitability of the Morgan fingerprint for representing metabolite structures, the performance of the trained models, and the interpretability of the learned models. As an initial exploration, this work lays the foundation for future research that leverages alternative structural encodings, diverse machine learning models, and explainability tools.

2. Data and Methods

The study used fibroblasts AT GM00648 as a cellular model for AT and AG09429 for the control. Metabolite analysis was conducted in triplicate using the UHPLC Vanquish system with an Accucore 150 amide HILIC column. LC was coupled to an Orbitrap Exploris 240 mass spectrometer equipped with an H-ESI source, operating in positive and negative modes and scanning the 80–800 m/z range. Metabolite identification and quantitation were carried out using Compound Discoverer 3.2 (Thermo Fisher Scientific). Data processing led to the identification of 4643 chemical structures. To enhance the precision of metabolite identification, the mass of each detected compound was compared to the mass of the matched compound recorded in the ChemSpider database and metabolites exhibiting a delta mass exceeding 5 ppm were excluded from the dataset. Duplicate molecules were filtered, retaining the one with

the highest peaks and yielding a set of 2453 distinct metabolites. Within this subset, only 157 metabolites were successfully assigned a KEGG ID that would allow for pathway enrichment analysis. The ratios between the measured quantities in the diseased and healthy conditions, along with the corresponding adjusted p-values, were calculated for each metabolite.

Molecular fingerprints of chemical structures were computed using the RDKit cheminformatics Python library [14]. The chosen encoding uses the Morgan molecular fingerprinting method with a radius of 2 and 1024 bits and accounts for molecule chirality. The target class for the classification task is binary and indicates whether the metabolite is significantly downregulated, i.e. its adjusted p-value is below 0.05 and the ratio of diseased to healthy is less than 1. The data exhibits class imbalance, with the positive class accounting for 17% of the data. T-distributed Stochastic Neighbor Embedding (t-SNE) [15] was employed to map data to a bi-dimensional space. Two oversampling algorithms, Synthetic Minority Over-sampling TEchnique (SMOTE) [16] and ADAptive SYNthetic (ADASYN) [17], were used to balance the data by generating synthetic samples for the minority class.

The metrics to evaluate the models include Accuracy (A), F1-score ($F1$), Recall for class 1 (R_1), Balanced Accuracy (BA), and Matthew's Correlation Coefficient (MCC). A and macro $F1$ provide an assessment of the classifier's performance across both classes but tend to yield overly optimistic results, particularly on imbalanced datasets. BA and MCC serve as a more informative statistical measure for unbalanced datasets, while $F1$ and R_1 place emphasis on the models' ability to correctly classify instances of class 1. Six ML algorithms were used - Decision Tree (DT), gaussian Naive Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), and XGBoost (XGB) - initially without data preprocessing and then applying both t-SNE and either SMOTE or ADASYN. Information gain was used to compute feature importance for the XGB model.

3. Results and Discussion

3.1. Structural encoding

The selected fingerprinting method is evaluated in terms of its resolution power, i.e. the ability to provide molecules with a unique encoding. 196 molecules were found to share identical fingerprints with at least one other metabolite. Further investigation into these groups of molecules with identical fingerprints unveiled a consistent pattern: they consisted of molecules that were identical except for the length of hydrocarbon chains, as exemplified in Figure 1 (a). This is unsurprising given that the selected fingerprint method resolves substructures of diameter 4 and cannot account for longer repetitive structures, which are rare in libraries screened for drug discovery. To enhance the resolution of the structural encoding, two strategies are proposed: (a) to adopt a count fingerprint rather than a binary encoding, capturing not just the presence of substructures but also their quantity, and (b) to incorporate the local information detected by the binary fingerprint with global properties, either measured during data acquisition or computed from the structure. Considering that the length of the hydrocarbon chain influences the molecule's weight, polarity, and interaction with chromatography phases, the molecular weight, partition coefficient (logP), and retention time (RT) are selected as additional features. Both approaches yield unique encodings and are illustrated in Figures 1 (b) and (c).

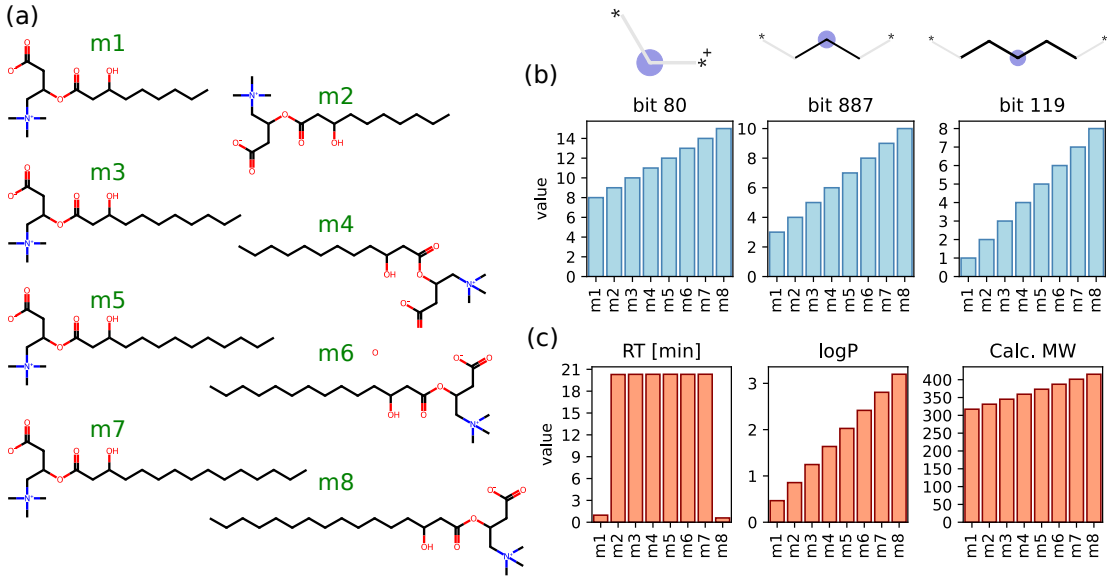

Figure 1: (a) Group of metabolites with identical morgan fingerprint, (b) number of repetitions of given substructures (corresponding to the bits 80, 887, and 119 of the fingerprint) and (c) global physicochemical properties measured or computed for each molecule. Count fingerprints and added global properties can both provide a unique structural encoding for groups of molecules with varying hydrocarbon chain lengths.

3.2. Performance of trained models

The performance metrics for ML models trained on the binary Morgan fingerprint and the two alternative structural encodings are presented in Table 1. Among models trained on binary data, NB and XGB emerge as the top performers, achieving *BA* of 0.73 and 0.68, respectively. For both alternative encodings, the performance of DT, SVM and RF is comparable to that of models trained on binary data, while LR and XGB demonstrate substantial improvements in both R_1 and *BA*. XGB trained on the binary fingerprint with added global features demonstrates good performance, with R_1 of 0.65 and *BA* of 0.75, providing evidence that the structural properties of metabolites hold predictive power over their response to particular conditions.

The original dataset presents notable challenges stemming from class imbalance and a relatively large number of features (1024) in comparison to the sample size (2453). To tackle these issues, the binary Morgan fingerprint data were pre-processed with t-SNE to reduce dimensionality and oversampled with either ADASYN or SMOTE techniques. Table 2 reports the metrics calculated for ML models trained on original and pre-processed data. For models DT, NB, RF, and XGB, an improvement in R_1 is achieved at the expense of *BA*. Conversely, models SVM and LR demonstrate a substantial improvement in both R_1 and *BA*. However, features transformation, while beneficial for model performance, prevents direct feature interpretation.

Table 1
Performance metrics for six ML classifiers trained on three different structural encodings

Oversampling	Binary Morgan FP					Count Morgan FP					Binary FP + global properties				
model	A	F1	R_1	MCC	BA	A	F1	R_1	MCC	BA	A	F1	R_1	MCC	BA
DT	0.83	0.60	0.29	0.21	0.60	0.87	0.62	0.23	0.28	0.60	0.82	0.55	0.18	0.10	0.54
NB	0.82	0.68	**0.61**	0.38	**0.73**	0.59	0.50	0.68	0.17	0.63	0.80	0.62	0.44	0.25	0.64
SVM	0.83	0.60	0.27	0.20	0.59	0.80	0.59	0.35	0.19	0.61	0.80	0.57	0.26	0.13	0.57
LR	0.78	0.54	0.23	0.08	0.54	0.82	0.65	**0.47**	0.31	**0.67**	0.79	0.65	**0.61**	0.34	**0.72**
RF	0.87	0.57	0.15	0.18	0.56	0.87	0.60	0.19	0.25	0.58	0.87	0.56	0.13	0.18	0.55
XGB	0.80	0.64	**0.53**	0.30	**0.68**	0.81	0.66	**0.58**	0.35	**0.71**	0.79	0.65	**0.65**	0.35	**0.75**

Table 2
Performance metrics for six ML classifiers trained on fingerprint data with and without pre-processing

oversampling	None					t-SNE + SMOTE					t-SNE + ADASYN				
model	A	F1	R_1	MCC	BA	A	F1	R_1	MCC	BA	A	F1	R_1	MCC	BA
DT	0.83	0.60	0.29	0.21	0.60	0.72	0.56	0.45	0.16	0.60	0.72	0.56	0.48	0.17	0.62
NB	0.82	0.68	0.61	0.38	0.73	0.64	0.55	0.79	0.28	0.71	0.61	0.53	0.79	0.25	0.69
SVM	0.83	0.60	0.27	0.20	0.59	0.65	0.56	**0.79**	0.28	**0.71**	0.61	0.53	**0.79**	0.25	**0.69**
LR	0.78	0.54	0.23	0.08	0.54	0.64	0.55	**0.76**	0.26	**0.69**	0.63	0.54	**0.77**	0.26	**0.69**
RF	0.87	0.57	0.15	0.18	0.56	0.73	0.57	0.45	0.17	0.61	0.73	0.58	0.52	0.21	0.64
XGB	0.80	0.64	0.53	0.30	0.68	0.60	0.51	0.68	0.18	0.63	0.58	0.51	0.74	0.20	0.65

3.3. Interpretation of trained models

To gain insights into the chemical structures contributing to the classification, feature importance analysis was conducted on the XGB model trained with the non-oversampled Morgan binary fingerprint. As illustrated in Figure 2, the feature importance analysis reveals several key features that influence the model predictions, corresponding to specific chemical configurations within the metabolites. Notably, the influential bits represent chemical configurations contained in saturated (bits 119, 794, and 591) and unsaturated (bit 849) fatty acid chains, phosphate groups (bits 814 and 192), nucleic acids (bits 640 and 932), and amino acids (bits 820 and 573).

Upon closer examination of the affected metabolites containing bits 640 and 932, several nucleotides and nucleotide-containing compounds were identified. These molecules are the building blocks of nucleic acids and play essential roles as coenzymes and signalling molecules, including Acetyl-CoA, Coenzyme A, AMP, and GMP. These observations align with prior knowledge, as the synthesis of nucleotides is known to be promoted by ATM and suppressed in AT [18]. Moreover, the nicotinamide corresponding to bit 984 points to other nucleotide-containing metabolites, NAD+ and NADP+, essential anti-oxidant cofactors whose role in AT as a result of impaired response to reactive oxygen species has been thoroughly characterised [13]. Affected metabolites containing structures represented by bits 119, 794, 591, and 849 encompass

a diverse array of lipids, with a prominent presence of phospholipids. The metabolism of these lipids has been reported to be disrupted in the context of the disease [19]

Notably, bits 573 and 820 were linked to over 100 down-regulated amino acids, dipeptides, and their derivatives. While the role of the tripeptide glutathione in oxidative stress within the context of AT has been characterised [20], the role of other peptides remains unknown, presenting a promising avenue for future investigations.

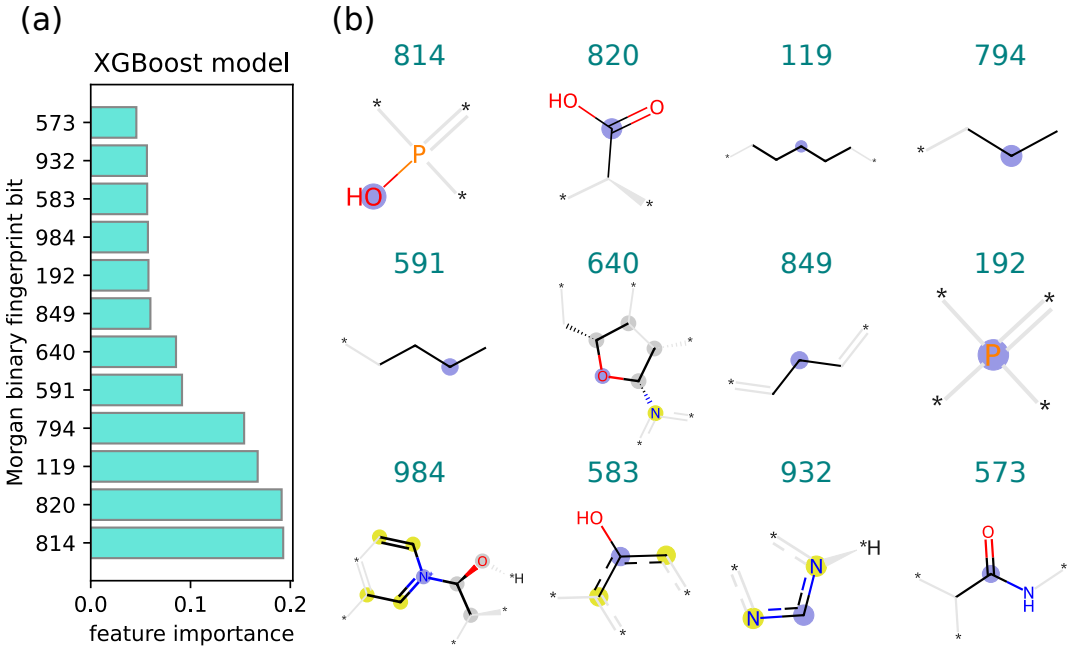

Figure 2: (a) The most important features for the XGB model trained on binary Morgan fingerprint and (b) the corresponding chemical configurations.

3.4. Opportunities for further improvement

Several avenues to further improve model performance can be explored. Alternative hashed fingerprints frequently employed in drug discovery, such as Daylight, Atop Pair, and Topological Torsion, can provide higher encoding resolution. Additionally, fine-tuning parameters such as fingerprint size and radius can positively impact model performance. Expanding the metabolic dataset by integrating measurements from different chromatographic columns, such as HILIC and C18, would effectively double the dataset size and enhance its diversity. The number of features can be reduced by calculating their correlation with the label vector and removing those with coefficients below a certain threshold, thereby reducing the dimensionality of the dataset while preserving the interpretability of features. A feature correlation study can be conducted to eliminate highly correlated features and reduce collinearity. Alternatively, compact molecular fingerprints based on pattern-matching could be explored, such as the MACCS fingerprint with 166 features. Lastly, it should be noted that classification tasks are intrinsically more

challenging for complex conditions with multiple affected pathways and a diverse range of affected metabolites like in AT. Consequently, achieving excellent classification performance in such a complex disease is challenging. Conversely, classification tasks may be relatively easier in experimental conditions where only a few metabolic pathways are affected.

4. Conclusions

This study proposes a novel approach for the comprehensive study of metabolites which does not rely on prior knowledge of metabolic pathways. By representing metabolites using molecular fingerprints and training machine learning classifiers on these structural encodings, the study effectively predicts down-regulated metabolites in the disease under study. Feature importance analysis provides insights into the cellular processes affected by the disease, validating existing knowledge and uncovering novel associations. This study serves as a foundation for future research exploring alternative structural encodings, diverse machine learning models, and advanced explainability techniques. These explorations are essential for the ongoing development of metabolomics as a powerful tool for enhancing the understanding of cellular metabolism and its implications for human health.

Acknowledgments

This work has been funded by the European Union - NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem grant ECS00000041 - VITALITY - CUP H33C22000430006

References

[1] E. Holmes, I. D. Wilson, J. K. Nicholson, Metabolic phenotyping in health and disease, Cell 134 (2008) 714–717.
[2] B. Peng, H. Li, X.-X. Peng, Functional metabolomics: from biomarker discovery to metabolome reprogramming, Protein & cell 6 (2015) 628–637.
[3] G. G. Harrigan, R. Goodacre, Metabolic profiling: its role in biomarker discovery and gene function analysis, Springer Science & Business Media, 2003.
[4] W. J. Griffiths, Metabolomics, metabonomics and metabolite profiling, Royal Society of Chemistry, 2007.
[5] L. Puchades-Carrasco, A. Pineda-Lucena, Metabolomics in pharmaceutical research and development, Current opinion in biotechnology 35 (2015) 73–77.
[6] D. M. Drexler, M. D. Reily, P. A. Shipkova, Advances in mass spectrometry applied to pharmaceutical metabolomics, Analytical and bioanalytical chemistry 399 (2011) 2645–2653.
[7] C. Wieder, C. Frainay, N. Poupin, P. Rodríguez-Mier, F. Vinson, J. Cooke, R. P. Lai, J. G. Bundy, F. Jourdan, T. Ebbels, Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis, PLoS Computational Biology 17 (2021) e1009105.

[8] P. D. Karp, P. E. Midford, R. Caspi, A. Khodursky, Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics, BMC genomics 22 (2021) 1–11.

[9] D. K. Barupal, P. K. Haldiya, G. Wohlgemuth, T. Kind, S. L. Kothari, K. E. Pinkerton, O. Fiehn, Metamapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity, BMC bioinformatics 13 (2012) 1–15.

[10] R. C. Glen, A. Bender, C. H. Arnby, L. Carlsson, S. Boyer, J. Smith, Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme, IDrugs 9 (2006) 199.

[11] M. Menotta, S. Biagiotti, C. Spapperi, S. Orazi, L. Rossi, L. Chessa, V. Leuzzi, D. D'Agnano, A. Soresina, R. Micheli, et al., Atm splicing variants as biomarkers for low dose dexamethasone treatment of at, Orphanet Journal of Rare Diseases 12 (2017) 1–7.

[12] H. L. Morgan, The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service., Journal of chemical documentation 5 (1965) 107–113.

[13] A. Ricci, F. Biancucci, G. Morganti, M. Magnani, M. Menotta, New human atm variants are able to regain atm functions in ataxia telangiectasia disease, Cellular and Molecular Life Sciences 79 (2022) 601.

[14] G. Landrum, Rdkit documentation, Release 1 (2013) 4.

[15] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).

[16] A. Fernández, S. Garcia, F. Herrera, N. V. Chawla, Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, Journal of artificial intelligence research 61 (2018) 863–905.

[17] H. He, Y. Bai, E. A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, 2008, pp. 1322–1328.

[18] C. Cosentino, D. Grieco, V. Costanzo, Atm activates the pentose phosphate pathway promoting anti-oxidant defence and dna repair, The EMBO journal 30 (2011) 546–555.

[19] M. A. Yorek, J. A. Dunlap, A. Manzo-Fontes, R. Bianchi, G. T. Berry, J. Eichberg, Abnormal myo-inositol and phospholipid metabolism in cultured fibroblasts from patients with ataxia telangiectasia, Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids 1437 (1999) 287–300.

[20] P. Degan, M. d'Ischia, F. V. Pallardó, A. Zatterale, A. Brusco, R. Calzone, S. Cavalieri, K. Kavaklı, A. Lloret, P. Manini, et al., Glutathione levels in blood from ataxia telangiectasia patients suggest in vivo adaptive mechanisms to oxidative stress, Clinical Biochemistry 40 (2007) 666–670.