Computer Methods and Programs in Biomedicine 250 (2024) 108163

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Exploring machine learning for untargeted metabolomics using molecular fingerprints

Christel Sirocchi [a,*,1], Federica Biancucci [b,1], Matteo Donati [a], Alessandro Bogliolo [a], Mauro Magnani [b], Michele Menotta [b], Sara Montagna [a]

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

ARTICLE INFO

Keywords:
Ataxia telangiectasia
Mass spectrometry
Molecular fingerprinting
Untargeted metabolomics
Machine learning

ABSTRACT

Background: Metabolomics, the study of substrates and products of cellular metabolism, offers valuable insights into an organism's state under specific conditions and has the potential to revolutionise preventive healthcare and pharmaceutical research. However, analysing large metabolomics datasets remains challenging, with available methods relying on limited and incompletely annotated metabolic pathways.

Methods: This study, inspired by well-established methods in drug discovery, employs machine learning on metabolite fingerprints to explore the relationship of their structure with responses in experimental conditions beyond known pathways, shedding light on metabolic processes. It evaluates fingerprinting effectiveness in representing metabolites, addressing challenges like class imbalance, data sparsity, high dimensionality, duplicate structural encoding, and interpretable features. Feature importance analysis is then applied to reveal key chemical configurations affecting classification, identifying related metabolite groups.

Results: The approach is tested on two datasets: one on Ataxia Telangiectasia and another on endothelial cells under low oxygen. Machine learning on molecular fingerprints predicts metabolite responses effectively, and feature importance analysis aligns with known metabolic pathways, unveiling new affected metabolite groups for further study.

Conclusion: In conclusion, the presented approach leverages the strengths of drug discovery to address critical issues in metabolomics research and aims to bridge the gap between these two disciplines. This work lays the foundation for future research in this direction, possibly exploring alternative structural encodings and machine learning models.

1. Introduction

Metabolomics is the study of small molecule substrates and products of cellular metabolism and provides valuable insights into the state of an organism under specific conditions [1]. Metabolomic profiling of diseased and healthy tissues is instrumental in discovering distinctive metabolic signatures and biomarkers, thereby aiding the development of screening tests and identification of potential drug targets [2]. Additionally, metabolomics can help to assess the effects of candidate treatments, evaluating the response at the metabolic level [3]. Therefore, metabolomics serves as an indispensable tool in preventive healthcare as well as pharmaceutical research and development, with the potential to enable timely disease detection and facilitate drug testing [4].

Beyond its clinical applications, metabolomics shows significant potential in basic research for unravelling the mechanisms of action behind diseases and treatments. In this context, one of the prominent methods for analysing metabolomic data is pathway enrichment analysis, which identifies metabolic pathways with a higher-than-expected abundance of affected metabolites, offering insights into disrupted cellular processes. However, this method presents several challenges, as it heavily relies on existing knowledge of metabolic pathways. Such knowledge is neither comprehensive nor consistently annotated, and it can vary across different metabolomic databases [5]. By focusing on metabolites mapped onto known metabolic pathways, which represent only a fraction of all annotated metabolites, this approach overlooks a substantial portion of metabolomic data.

* Corresponding author.
E-mail address: c.sirocchi2@campus.uniurb.it (C. Sirocchi).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2024.108163
Received 18 December 2023; Received in revised form 15 March 2024; Accepted 3 April 2024
Available online 8 April 2024

Recognising the pressing need for novel methods in metabolomic data analysis, that extend beyond the boundaries of known metabolic pathways by exploiting all detected metabolites, this study explores an approach leveraging Machine Learning (ML) to get insights into the interplay between the chemical structure of metabolites and their involvement in a perturbed state under study. Drawing inspiration from the well-established practices in drug discovery and drug design, this study employs molecular fingerprinting to encode chemical structures and ML models trained on such fingerprints to predict functional properties from structural characteristics, extending this paradigm to metabolomics [6]. The analysis uncovers enriched chemical configurations within the perturbed sample, thereby identifying groups of affected metabolites sharing structural similarities. Considering that chemically similar compounds are generally found in metabolic proximity, the identification of structurally related affected metabolites can provide valuable insights into the underlying biological processes and assist researchers in formulating hypotheses [7].

Initially, the study undertakes a comparative analysis of molecular fingerprinting techniques to evaluate their suitability for metabolite representation, shedding light on their limitations and providing recommendations for improving fingerprint resolution. Additionally, it explores challenges associated with training ML models on molecular fingerprints of cellular metabolites, addressing issues such as dataset imbalance and high dimensionality. Subsequently, a variety of ML classifiers are trained on molecular fingerprints to predict perturbations in metabolite levels under experimental conditions. Finally, feature importance is analysed to identify the chemical configurations most influential in the classification process. The above described pipeline is exemplified and tested on metabolomics datasets acquired through two case studies: *(i)* an investigation into a cellular model for Ataxia Telangiectasia (AT), a rare neurodegenerative disorder caused by mutations in the Ataxia Telangiectasia Mutated (ATM) gene, disrupting numerous metabolic pathways [8]; *(ii)* a study on the effects of redox regulation on endothelial cells exposed to low oxygen levels, simulating conditions in the bone marrow niche, with implications for hematopoietic stem cell maintenance [9].

In both case studies, the satisfactory performance of the trained ML models suggests that the structural properties of metabolites hold predictive value over their response to a particular condition and confirms that ML can be effectively applied to predict function from structure in the realm of cell metabolism. Remarkably, feature importance computed on 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 the conditions under study, opening up novel opportunities for further investigation.

The proposed approach leverages the strengths of drug discovery to address critical issues in metabolomics research, aiming to bridge the gap between these two disciplines and lay the foundations for future research. Moreover, the integrated approach adopted in this paper provides a comprehensive data analysis methodology along with an open-access tool, which can be readily utilised for further research endeavours.

2. Background

2.1. Metabolomics in healthcare and drug discovery

Metabolomics occupies a unique position in the omics landscape for its close relationship with phenotype. The metabolome, representing the final product of genomic, transcriptomic, and proteomic processes, serves as a direct readout of the physiological state of an organism. The inherent sensitivity of metabolomics allows for the detection of subtle alterations in biological pathways, shedding light on the underlying mechanisms governing various physiological conditions and the complex interplay between genetics, environmental factors, and the physiological state of the organism [10].

In healthcare, metabolomics has played a crucial role in the discovery of biomarkers and the understanding of diseases. Its impact encompasses early disease diagnosis, disease monitoring, and personalised treatment strategies. In clinical laboratories worldwide, metabolomics accounts for over 95% of the workload. Notably, certain targeted metabolic profiles are now routinely employed in newborn and genetic screenings [11]. Metabolomics also finds application in the field of personalised medicine, which considers individual genomic variations, biochemical profiles, environmental factors, and lifestyles. Identifying metabolic biomarkers is central to this paradigm shift in healthcare [12].

Furthermore, metabolomics plays a pivotal role in drug discovery. By deciphering the metabolic changes associated with diseases, it guides the identification of novel drug targets and informs the design of drugs that specifically target these pathways. Metabolomics also assesses the safety and efficacy of pharmaceutical compounds, streamlining the drug development pipeline from early pre-clinical studies to clinical trials [13].

As advances in chemoinformatics and bioinformatics empower metabolomics, the field has transitioned from merely identifying metabolites as biomarkers to investigating their role in metabolic networks and effect on phenotypic outcomes [10]. In the future, the evolution of existing statistical methods is expected to provide even deeper insights into metabolic processes from the larger perspective of systems biology [14].

2.2. Traditional methods in metabolomics

The potential of metabolomics lies in characterising and quantifying all the metabolites present in a particular biological system using a combination of analytical tools [4]. Central to this endeavour is the remarkable advancements in mass spectrometry, which has emerged as a preeminent analytical platform for metabolomic analysis due to its high sensitivity, specificity, reproducibility and versatility [15].

The study of metabolome encompasses two fundamental approaches. Untargeted metabolomics aims to comprehensively measure a wide spectrum of metabolites without prior knowledge of the metabolome, while targeted metabolomics concentrates on specific metabolites and pathways guided by prior information. The latter offers higher sensitivity and selectivity and plays a crucial role in validating and extending results from untargeted analysis [10].

In practice, metabolomics experiments often involve selecting metabolites based on statistical significance and fold change. Typically, comparative analyses employ statistical tests such as ANOVA and t-tests to compare different conditions. Metabolites exhibiting statistical significance, usually defined as having a p-value below 0.05, and fold changes greater than two, are considered for further investigation [16].

Metabolomics studies typically compare samples from a normal state to those from a perturbed state, induced by genetic modifications or treatment administration [15]. Pathway enrichment analysis is the prevalent method for comparing such samples, as it identifies metabolic pathways that exhibit a higher degree of overlap with significantly under or over-expressed metabolites than would be expected by chance. However, this approach has its shortcomings, as it relies on existing knowledge of biological pathways, which can be incomplete and not fully annotated, and is sensitive to the pathway definitions used by different metabolomic databases [5].

2.3. Machine learning in metabolomics

Metabolomics is a challenging field due to the complex nature of metabolite interactions and the rapid and dynamic nature of metabolite changes in response to the (patho) physiological context. Metabolites partake in multiple pathways, acting as the product or substrate

of biological processes shared among several competing biochemical reactions, so pinpointing their flux direction can be difficult [17]. Metabolomic studies are relatively fewer compared to other omics, domain knowledge is more scattered, and analytical methods tailored to metabolic data are limited [18]. Consequently, metabolomics has turned to ML to navigate the complexities of its domain, although its use is still in its early stages and limited to certain applications [19].

The use of ML is fairly established at the level of data pre-processing, in tasks such as baseline correction, noise filtering, peak detection and alignment, data normalisation and scaling, retention time prediction, and handling of missing data [18]. The use of ML has recently gained attention in clinical metabolomics for patient classification based on metabolic profiles. This application proves valuable in predicting clinical outcomes, assisting disease diagnosis, prognosis, and risk assessment, as well as guiding personalised treatment interventions. It is also instrumental in identifying biomarkers associated with specific conditions, streamlining the development of screening tests [14].

Despite the contributions of ML to metabolomic data analysis, it is crucial to acknowledge that the biological insights derived from these studies are limited. As of our current knowledge, there are no studies applying ML to metabolite chemical structures to offer insights into affected cellular processes. Such applications require representing metabolite structures in a machine-readable format, a process that can draw upon methodologies from drug discovery, where the encoding of structures for ML training is well established.

2.4. The problem of molecular representation

ML has attained a level of maturity in the field of drug discovery, with optimised ML pipelines tailored to predict functional attributes from the structural features of vast and diverse biochemical libraries [20]. In this context, ML has not only demonstrated impressive predictive capabilities but has also contributed to a deeper understanding of the mechanisms of target interaction [21]. In conjunction with molecular fingerprinting, which provides a concise representation of chemical structures, ML has revolutionised drug discovery and design by enabling the prediction of drug activity and crucial drug properties such as bioavailability, solubility, and toxicity [22]. Additionally, the precise prediction of interactions between a potential drug and its designated target enables investigating the mechanism by which a drug operates, providing valuable insights into the underlying pathological processes [23].

Despite the continuous development of increasingly sophisticated ML models for drug discovery, representing molecular chemical data in a concise and machine-readable manner to effectively train these models is not a trivial task and remains one of the main challenges in ML-powered drug discovery. Frequent issues are non-canonical encoding, where multiple different representations may describe the same molecule, non-unique or clashing encoding, where different molecules are encoded into the same representation, erroneous assumptions about the number of implicit hydrogen atoms, or the failure to adequately capture tautomerism [24].

There exist four primary approaches to rendering molecules in a machine-readable format: as a text string, using a connection table, as a set of features - such as fingerprints or a series of physical descriptors -, and most recently, harnessing a ML-learned molecular representation. The choice among these methods depends on various considerations, such as the need for human readability, compatibility with other software or algorithms, space limitations, and more [24]. Numerous molecular representations have been developed and refined within the domain of drug discovery to effectively portray extensive libraries of prospective pharmaceutical compounds. However, the extent to which these representations can be seamlessly extended to other fields, such as the study of functional properties of metabolites, remains an open question.

3. Materials and methods

This section elaborates on the methodologies utilised in the proposed approach, which leverages a diverse array of data pre-processing techniques and ML algorithms to investigate the relationship between chemical structure and metabolic response. Fig. 1 illustrates the phases of this analysis. This section details how structural encoding and data pre-processing techniques are used to tackle challenges like class imbalance, high dimensionality, and duplicate encoding. It also outlines the ML models adopted, discusses the process of hyperparameter fine-tuning, introduces evaluation metrics for assessing model performance, and illustrates feature importance analysis.

3.1. Sample preparation

3.1.1. Hypoxia case study

Human umbilical endothelial cells (HUVEC) from Lonza (Switzerland) were cultured in endothelial growth medium 2 (EGM-2™) supplemented with EGM-2 BulletKit™ (Lonza). The experiment involved two groups: normoxia (21% O2), hypoxia (3%O2). Untreated cells were set in the Hypoxia Chamber (StemCell Technologies) and flushed with 3% O2, 92% N2, and 5%CO2 gas mixture for 15 min before being placed in a humidified incubator (37_C, 5% CO2) for 1, 6, and 24 h. HUVEC were seeded at 5×10^6 cells per T75 flask in technical duplicate for each condition: normoxia, hypoxia. After 6 and 24 hours, cells were washed with ice-cold PBS and harvested in cold 80/10/10 LC/MS-grade methanol/acetonitrile/water (Carl Roth, Germany). Insoluble material was pelleted by centrifugation at 20,000 g for 20 minutes. Extraction buffers followed established protocols to preserve redox-sensitive metabolites. The resulting supernatants were evaporated, and pellets were dissolved in 350 µL of 50/30/20 LC/MS-grade methanol/acetonitrile/water containing 0.1% formic acid. For more detail on sample preparation refer to the original manuscript [9].

3.1.2. Ataxia case study

Fibroblasts WT AG09429 (Atm+/+) and AT GM00648 (Atm-/-) sourced from the Coriell Institute (Camden, NJ, USA) were utilized as the cellular model for metabolomics analysis, following the approach established in previous studies on AT [25–28] The cells were cultured in MEM (Eagle formulation) supplemented with 10 mM glucose, 2 mM l-glutamine, 100 U/mL penicillin, 0.1 mg/mL streptomycin (Sigma Aldrich), and 10% fetal bovine serum (Thermo Fisher Scientific). A total of $2x10^5$ WT AG09429 (Atm+/+) and AT GM00648 (Atm-/-) cells were seeded into ten 90 mm Petri dishes (Thermo Fisher Scientific), five dishes for each cell type. Cell growth was monitored until reaching 70-80% confluence. Upon confluence, the growth medium was aspirated, and the cells were washed twice with pre-chilled PBS. Subsequently, the cells were quenched with 200 µL of pre-chilled (-80 °C) extraction Buffer (per dish), composed of a CH3OH/CAN/H2O solution 2 (78:20:2, v/v/v). To preserve the integrity of redox metabolites, H2O solution 2 containing 250 mM ammonium acetate and 2.5 mM sodium ascorbate was used [29]. The obtained cell pellets were vortexed for 35 minutes and kept at -80 °C for 1 hour for complete cell lysis. Following this, cells were stored at -20 °C overnight to facilitate protein precipitation. The resulting cell extracts were centrifuged at 16,000 x g for 20 min at 4 °C, and the supernatants were collected and dried using a SpeedVac centrifuge for 10 hours (Savant-SPD121P). The dried metabolite pellets were then resuspended in 350 µL of pre-chilled CH3OH/H2O (50:50, v/v) containing 0.1% formic acid and transferred to LC-vials for injection into the LC-MS/MS system. Quality Control samples were prepared by mixing an equal volume (100 µL) of each sample to assess the reproducibility and reliability of the LC-MS/MS system.

3.2. Data acquisition

In both experiments, supernatants were analysed using a Vanquish ultra-high-performance liquid chromatography (UHPLC) system

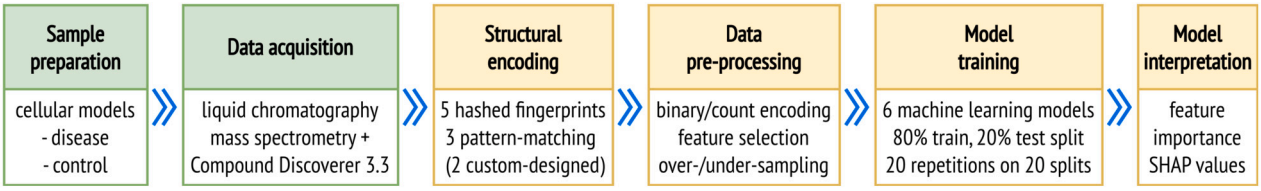


Fig. 1. Key phases of the conducted analysis.

(Thermo Fisher Scientific), coupled with high-resolution mass spectrometry (Exploris 240, Thermo Fisher Scientific). Chromatographic separations were performed using a reversed-phase C18 Hypersyl GOLD column (150 × 2.1 mm, 1.9 µm, Thermo Fisher Scientific) maintained at 40 °C. The mobile phase consisted of water (Solvent A) and acetonitrile (Solvent B), both containing 0.1% formic acid, delivered at a flow rate of 0.3 mL/min. Compounds were also resolved using an AccucoreTM-150-amide-HILIC column (100 × 2.1 mm, 2.6 µm, Thermo Fisher Scientific) maintained at 60 °C with a flow rate of 0.5 mL/min. The mobile phase for HILIC chromatography consisted of water (Solvent A) and acetonitrile (Solvent B), both with 10 mM ammonium formate and 0.1% formic acid. Acquisitions were performed in positive and negative ion polarity modes. Calibration was conducted before each analysis sequence, and internal calibrants were used in each run. The untargeted metabolomics acquisition workflow was conducted in positive and negative ion polarity modes using the deep scan AcquireX software (Xcalibur 4.2, Thermo Fisher Scientific), with 5 ID (Identification Only), 5 QC (Quality Control for system normalisation), and 3 replicates of each sample for statistical analysis. Raw data were processed using Compound Discoverer software Version 3.3 (Thermo Fisher Scientific).

Raw data files were imported into Compound Discoverer 3.3 and processed using a customised workflow within the software. This workflow encompassed peak alignment, peak detection, feature filtering, and metabolite putative annotation. Peak alignment parameters, primarily retention time (RT) and mass tolerances were set at 0.5 min and 5 ppm, respectively; normalisation was based on quality control (QC). The interference from the chemical background was eliminated by employing a procedural blank sample composed exclusively of extraction solvent. Metabolites were putatively annotated by comparing theoretical and detected MS data with the ChemSpider database and by matching MS2 spectra with reference spectra from the mzCloud database[2] (Thermo Fisher Scientific). Additionally, MS2 spectra were also compared using m/zVault search on a customised fragmentation library in Compound Discoverer 3.3. Statistical analysis was conducted using Compound Discoverer 3.3 by a one-way ANOVA model with Tukey as post-hoc test and p-values adjusted by the Benjamini-Hochberg algorithm, as specified in the Compound Discoverer 3.3 Manual.[3]

Compounds for which a full match could not be obtained with spectra in ChemSpider, mzCloud, or mzVault (where a full match indicates that the current formula and structure annotations match the best available item from the particular source) and for which the fragmentation spectra MS2 were not detected were excluded from the analysis. To further improve the precision of metabolite annotation, each detected molecule's mass was compared to the corresponding mass in the ChemSpider database and metabolites with a mass difference exceeding 5 ppm were excluded. Duplicates within the dataset were addressed by retaining only the instances of each metabolite with the highest peak intensity. In the Ataxia experiment, this process yielded a final set of 3999 putatively annotated metabolites. Among these metabolites, only 840 were successfully assigned a KEGG ID, enabling enrichment analysis. The target for classification was defined as metabolites significantly reduced in the diseased sample. Thus, the target takes the value 1 if the adjusted p-value is less than 0.05 and the ratio of diseased to healthy quantities is less than 1, and zero otherwise. The focus on down-regulation stems from the disease's known tendency to inhibit cellular activities [27]. In the Hypoxia experiment, 1331 metabolites were annotated, of which 340 were assigned a KEGG ID. The target for classification reflects metabolites significantly increased after 6 h in hypoxic conditions. Thus, the target takes the value 1 if the adjusted p-value is less than 0.05, and the ratio of hypoxic to healthy quantities is greater than 1, and zero otherwise. Here, the focus on up-regulation reflects the primary interest of the study to identify metabolites activated by hypoxia [9].

3.3. Structural encoding

The chemical structure of metabolites was encoded using hashed and pattern-matching fingerprinting techniques. This section provides a detailed account of the fingerprint generation process.

3.3.1. Hashed fingerprints

Hashed fingerprints include Morgan, Topological Torsion, Atom Pair, and Daylight fingerprints. The Morgan fingerprint, also known as the extended-connectivity fingerprint, captures the molecular structure by considering substructures within a predefined radius surrounding each atom. This versatile method proves suitable for extracting both local and global structural features [30]. The Topological Torsion fingerprint quantifies topological torsion angles between pairs of atoms, effectively revealing subtle structural nuances influencing molecular properties [31]. The Atom Pair fingerprint quantifies the occurrences of atoms within a molecule, commonly employed in tasks like similarity searching and virtual screening [32]. The Daylight fingerprint represents specific molecular fragments and finds common usage in virtual screening [33]. All fingerprints were generated using the RDKit library [34], encoded into a varying number of bits (from 256 to 4096) and accounting for chirality when applicable. The Morgan fingerprint was generated with both radii 2 and 3. For each hashed fingerprint, count-based fingerprints are generated in addition to the standard binary form, allowing not only to detect the presence of substructures but also to quantify their occurrences. Count fingerprints were normalised to the interval [0,1].

3.3.2. Pattern-matching fingerprints

Pattern-matching fingerprints involve the matching of substructures within a molecule to reference structural patterns. This category includes the widely used MACCS keys, along with two additional custom-designed fingerprints within this study. The MACCS fingerprint encodes the presence of 166 predefined chemical substructures within molecules, making it a popular choice for pattern recognition and substructure-based similarity assessments [35]. MACCS keys were generated using the RDKit library [34]. Out of the 166 keys, 152 are used to determine the presence of specific substructures, while the remaining 14 keys indicate whether a particular substructure appears at least a specified number of times. Therefore, distinct binary and count forms cannot be defined for this fingerprint. A custom-designed fingerprint, referred to as the *fragment fingerprint*, was constructed using the 85 molecular fragments defined in the rdkit.Chem.Fragments module of the RDKit library. A second custom-designed fingerprint, here termed

[2] http://www.mzcloud.org.

[3] https://assets.thermofisher.com/TFS-Assets/CMD/manuals/XCALI-98478-Compound-Discoverer-User-Guide-LC-Studies-XCALI98478-en.pdf.

functional fingerprints, was constructed based on 195 functional groups listed in the Daylight manual.[4] These two custom-designed fingerprints are generated in binary and count form, akin to hashed fingerprints, with count fingerprints normalised to the interval [0,1].

3.4. Data pre-processing

Both datasets exhibit class imbalance, with the positive class accounting for 15% of the data in the Ataxia dataset and 6% in the Hypoxia dataset. Balanced training datasets were obtained either by undersampling the majority class or by oversampling the under-represented class, using the RandomUnderSampler and RandomOverSampler functions provided by the imbalance-learn Python library [36]. The generation of synthetic data, as accomplished through techniques such as SMOTE [37] and ADASYN [38], is not applicable, as none of the available variations of these algorithms can effectively handle datasets exclusively composed of categorical features.

The datasets generated using hashed fingerprints display high dimensionality, comprising 1024 features for 1-4000 samples, alongside data sparsity, as only 3% of the data contains non-zero values. To preserve feature interpretability while reducing dimensionality, univariate feature selection techniques are applied. Three distinct statistical tests, namely Chi-squared (χ^2) test, mutual information, and ANOVA are employed to identify the most promising subset of n features, with n values ranging from 100 to 300. The first two tests are known to be particularly effective with sparse data.

3.5. Model training

The models adopted to classify affected metabolites include Decision Tree (DT), known for its interpretability and suitability for binary datasets [39]; Bernoulli Naive Bayes (BNB), recognised for its efficiency and compatibility with high-dimensional data [40]; Logistic Regression (LR), providing interpretability and insights into feature importance [41]; Random Forest (RF), an ensemble method robust to outliers and noise [42]; XGBoost (XGB), offering high predictive accuracy and robustness in handling imbalanced datasets [43]; and MultiLayer Perceptron (MLP), capable of capturing complex feature interactions [44]. All experiments were conducted using the scikit-learn[5] and PyTorch[6] Python libraries.

The chosen metrics for evaluating the models prioritise the accurate classification of instances belonging to the minority class, which in the problem at hand often represent affected metabolites, of greater interest. Metrics such as accuracy, which tend to favour the majority class and can yield artificially high scores even if the model performs poorly on the minority class, were not utilised. The selected metrics include F1-score (F1), recall for the positive class (R), specificity (SP), and the area under the curve of the receiver operating characteristic (ROC AUC). Each metric serves a distinct purpose in evaluating the model's performance. ROC AUC evaluates the model's ability to distinguish between the positive and negative classes, the F1-score balances precision and recall, while recall and specificity focus on correctly identifying the positive and negative classes, respectively.

For each model, a range of hyper-parameters was defined based on prior knowledge and model-specific considerations. Grid-search was utilised to explore various combinations of hyper-parameters and 5-fold cross-validation was used to assess the model's generalisation. The hyperparameter combination that yielded the highest recall score was selected as the optimal configuration. For each dataset, models were trained and evaluated on 20 distinct random splits of the data, consisting of an 80% training set and a 20% testing set, with different models trained on the same 20 splits for consistency. Class imbalance was addressed by implementing a stratified data split, which preserves the percentage of positive class samples in each split, and by assigning class weights depending on the class size during model training.

3.6. Model interpretation

The influence of each feature on predictive outcomes was estimated by calculating feature importance, employing distinctive methodologies contingent on the classifier category and using specific methods available in the scikit-learn library. For DT, the method employed for feature importance estimation is the mean decrease in impurity which quantifies how much each feature contributes to reducing the impurity (i.e., uncertainty) of the target variable within the tree nodes. For ensembles of trees, as in RF and XGB, variants of mean decrease in impurity are used. In the case of BNB, feature importance is determined through the analysis of model log probabilities. LR and MLP models rely on model coefficients, which indicate the magnitude and direction of the influence that each feature has on the model's predictions. In addition, SHAP (SHapley Additive exPlanations) values are computed for each sample within the test dataset and then averaged across all samples [45], shedding light on the collective impact of features on model predictions.

Feature importance analysis was primarily conducted to verify whether the substructures corresponding to important features align with prior knowledge on the experimental condition under study (namely, ATM mutation and hypoxia), and to provide novel insight for further exploration. In an additional experiment, feature importance scores were also utilised to select the top features, offering an alternative method to statistical tests like χ^2.

The entire pipeline for computing molecular fingerprints, training ML models and computing feature importance with SHAP values is available on Github https://github.com/ChristelSirocchi/metabo-ML

4. Results

4.1. Resolution power of fingerprints

To evaluate the ability of the selected fingerprints to distinguish between the diverse structures found in the two datasets under study, the fraction of metabolites that are not assigned a unique fingerprint is counted. Since all metabolites in the dataset have distinct structural configurations, designated by a unique SMILE, a suitable representation should also be unique. The effect of size on the fingerprint resolution is evaluated by varying the bit number between 256 and 4096, and computing the percentage of non-unique encodings, illustrated in Table 1 for the two datasets under study. As the number of bits increases, the molecules with non-unique encoding decrease. However, for most fingerprints, the value stabilises after 2048, indicating that it is not able to discriminate certain configurations independently from the number of bits it hashed into. The fraction of encodings that are resolved passing from 516 to 1024 bits is in some cases substantial (as in the Daylight fingerprint), while that from 1024 to 2048 bits is always very small if not null. Consequently, this standard fingerprint size of 1024 bits is retained for all subsequent experiments, achieving a compromise between resolution and dimensionality.

The groups of molecules sharing identical encodings range from 2 to 25 molecules, with frequency decreasing exponentially with size. For Morgan fingerprints with radii 2 and 3, unresolved metabolite groups differ in the length of one or more hydrocarbon chains, exemplified in Fig. 2 (a). Daylight fingerprint exhibits a similar resolution pattern, as it is unable to resolve metabolites with varying hydrocarbon chain lengths. In the case of the Topological Torsion fingerprint, the largest group of metabolites sharing the same encoding, containing 25 molecules, predominantly comprises small halogen-containing metabolites, such as phosphates, sulphates, and chloric acids. Similar to Morgan and Daylight fingerprints, the remaining groups differ in their

[4] https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html.
[5] https://scikit-learn.org/.
[6] https://pytorch.org/.

Table 1
Percentage of metabolites with non-unique encodings in the Ataxia and Hypoxia datasets for various fingerprints (fp). Hashed fingerprints, including Morgan of radius 2 (MG2) and 3 (MG3), Atom Pair (AP), Topological Torsion (TT), and Daylight (DL), are compared. The analysis includes binary chiral fingerprints of standard size 1024 (highlighted in bold), binary chiral fingerprints of various sizes (from 256 to 4096), as well as non-chiral binary and chiral count fingerprints of size 1024. For pattern-matching fingerprints, including MACCS and custom fingerprints based on functional groups (FUNC) and molecular fragments (FRAG), binary and count fingerprints are compared.

	Ataxia dataset - % non-unique encodings							Hypoxia dataset - % non-unique encodings						
	Hashed fingerprints													
config.	binary					binary	count	binary					binary	count
chirality	chiral					non-chiral	chiral	chiral					non-chiral	chiral
fp size	256	512	**1024**	2048	4096	1024	1024	256	512	**1024**	2048	4096	1024	1024
MG2	8.0	7.0	**7.0**	7.0	7.0	10.7	0.0	8.9	8.3	**8.2**	8.2	8.2	13.2	0.0
MG3	5.1	5.0	**5.0**	5.0	5.0	8.5	0.0	6.8	6.5	**6.5**	6.4	6.4	11.6	0.0
AP	1.6	1.1	**0.9**	0.8	0.8	4.1	0.6	1.7	1.1	**0.9**	0.9	0.8	5.9	0.8
TT	7.8	7.0	**6.8**	6.7	6.7	10.2	1.1	8.8	7.9	**7.7**	7.7	7.7	13.0	1.3
DL	24.8	12.7	**11.9**	11.8	11.8	11.9	4.0	24.3	16.0	**14.9**	14.9	14.9	14.9	5.9
	Pattern-matching fingerprints													
config.	binary					count		binary					count	
FUNC	31.5					3.9		29.8					5.5	
FRAG	53.4					18.8		51.0					18.0	
MACCS	17.3					-		21.3					-	

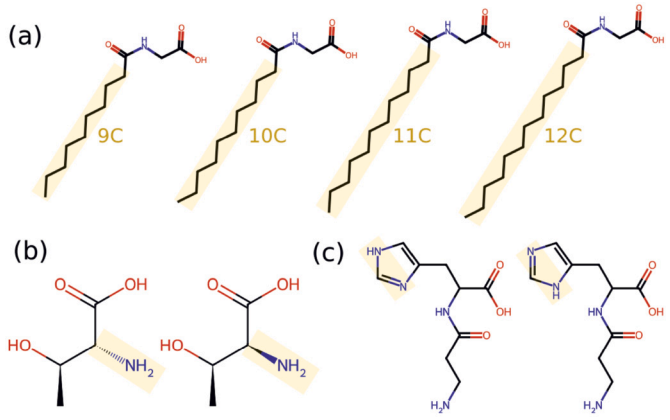


Fig. 2. Groups of metabolites sharing identical encodings in hashed molecular fingerprints. (a) Topological Torsion, Daylight, and Morgan fingerprints are unable to differentiate metabolites with varying hydrocarbon chain lengths. Atom Pair fingerprints cannot resolve (b) certain chiral structures and (c) oxygen and nitrogen-containing groups losing or gaining an hydrogen atom.

ability to distinguish hydrocarbon chain lengths. Atom Pair fingerprint can resolve hydrocarbon chain lengths, and for this reason, holds the highest resolution power among all considered fingerprints, but fails to resolve some chiral structures, as shown in Fig. 2 (b), which are instead effectively distinguished by Morgan and Topological Torsion. Furthermore, it struggles to differentiate oxygen and nitrogen-containing groups that have either lost or gained a hydrogen atom, seen in Fig. 2 (c), a distinction handled effectively by Morgan fingerprints.

The influence of chirality on fingerprint resolution is also explored. All hashed fingerprints, except Daylight, incorporate a parameter to consider molecular chirality. Fig. 2 (b) exemplifies two metabolites that are assigned unique Morgan fingerprints with a radius of 3 when chirality is factored into the analysis. Results in Table 1 show that a considerable fraction of metabolites becomes resolved when accounting for chirality, indicating the presence of a substantial proportion of chiral molecules within the datasets.

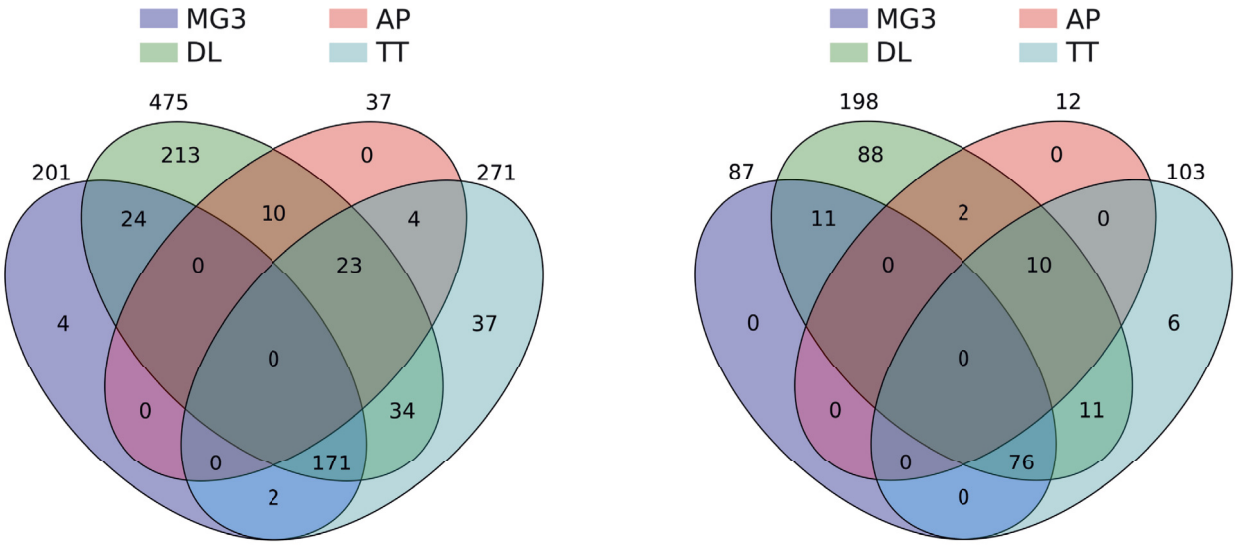
Additionally, the resolution power of fingerprints that count the occurrences of each substructure rather than only detecting its presence is explored. In this regard, all hashed fingerprints have a corresponding *count* version. As presented in Table 1, the count Both Morgan fingerprints of radii 2 and 3 successfully assign a unique representation to all the detected metabolites. While the resolution is notably enhanced for the other fingerprints, there are still instances where different molecules share identical encodings.

The number of metabolites having non-unique encodings for one or more fingerprints is illustrated in Fig. 3 for each analysed dataset. Numbers at the intersections between fingerprints indicate metabolites that cannot be resolved by either. The intersection of all fingerprints is empty for both analysed datasets, indicating that for each metabolite, there is at least one fingerprint that generates a unique vector. Therefore, even though no fingerprint has maximum resolution power, it is possible to engineer such a fingerprint by combining existing ones.

The fraction of metabolites with non-unique representations obtained from any of the pattern-matching fingerprints exceeds that of any hashed fingerprint, often by multiple folds, as evident from Table 1. Among the pattern-matching fingerprints, MACCS exhibits the highest resolution power by assigning unique encoding to approximately 80% of metabolites, due to some keys that detect the presence of up to 3 repetitions of a given functional group. Conversely, functional and fragment fingerprints in their binary form demonstrate relatively lower resolution, providing unique encoding to about 70% and 50% of metabolites, respectively. All pattern-matching fingerprints prove ineffective in distinguishing metabolites differing in hydrocarbon chain length, the number of carbon double bonds, or molecule chirality, as visually depicted in Fig. 4 (a) and (b). These fingerprints also fail to discriminate metabolites differing in the number and arrangement of functional groups within the molecule. Furthermore, the fragment-based fingerprint cannot adequately detect the presence of rings, for instance, failing to differentiate between the linear and ring forms of the glucose molecule, or the presence of phosphate or sulfur-containing groups, as demonstrated in Fig. 4 (c). In their count form, both the functional and fragment fingerprints provide unique encoding to a greater fraction of metabolites as they can discriminate in the length of hydrocarbon chains. However, both of these fingerprints still remain ineffective in discriminating chiral molecules. Moreover, the count-based fragment fingerprint struggles to differentiate molecules that differ in phosphate or sulfur-containing groups.

4.2. Predictive power of fingerprints

The performance of ML models trained on binary fingerprints in their original form (*Binary FP*), after oversampling the positive class (*Binary Oversampled*), undersampling the negative class (*Binary Undersampled*), and performing feature selection (*Binary Selected*), as well as their count form (*Count FP*) is evaluated across all defined fingerprints (5 hashed, 3 pattern-matching) for the two considered datasets. The average performance values over 20 random splits of the datasets for mod-



(a) Ataxia dataset. (b) Hypoxia dataset.

Fig. 3. Number of metabolites with non-unique encoding in one or more hashed fingerprints, including Morgan of radius 3 (MG3), Atom Pair (AP), Topological Torsion (TT), and Daylight (DL), in the Ataxia and Hypoxia datasets. The intersections between fingerprints display the number of metabolites that cannot be distinguished by either fingerprint.

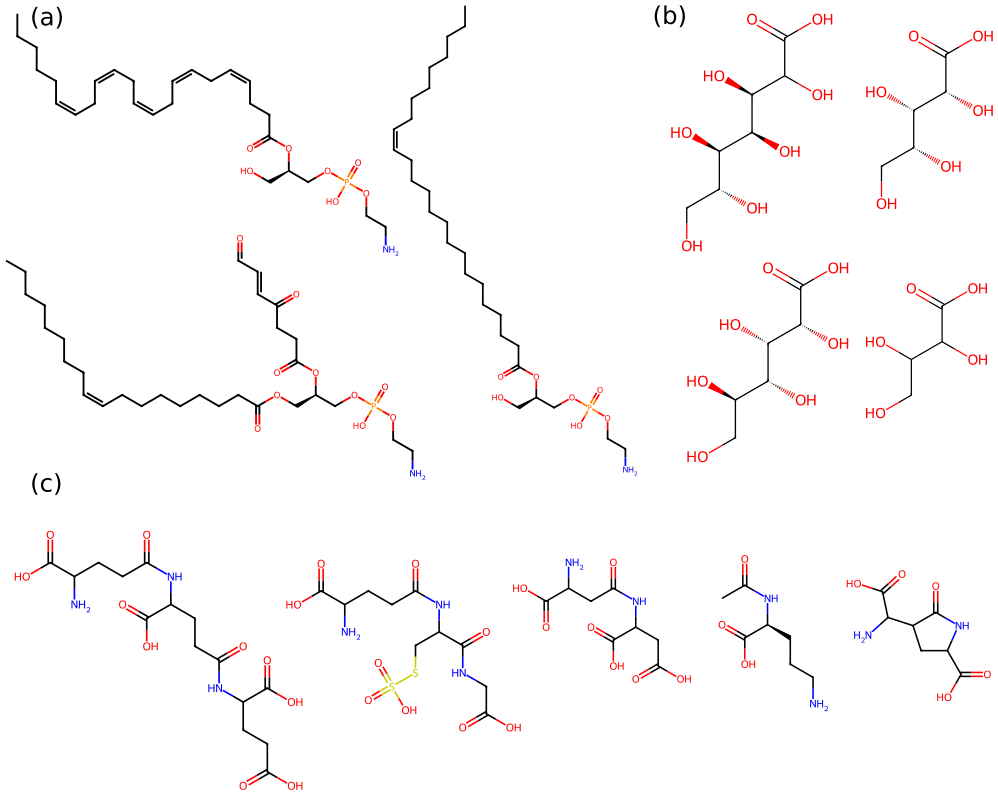


Fig. 4. Groups of metabolites sharing identical encodings in pattern-matching molecular fingerprints. All pattern-matching fingerprints are ineffective in distinguishing metabolites that differ in (a) the number of carbon double bond, (b) chirality, as well as the number and arrangement of functional groups within the molecule. (c) The fragment-based fingerprint also fails to discriminate the presence of rings and phosphate/sulfur-containing groups.

els trained on Ataxia and Hypoxia datasets are presented in Tables 2 and 3, respectively. In these tables, indicate that the model trained on a specific variation of the binary fingerprint (oversampled, undersampled, selected, and count) significantly outperformed the model trained on the original binary fingerprint, based on a t-test. On the other hand, italicised and underlined values denote that the model significantly outperformed all others regarding a given performance metric, within a specific fingerprint and across all fingerprints, respectively.

Across both datasets, it is consistently observed that models trained on count fingerprints do not outperform their binary counterparts. Specifically, in the Hypoxia dataset, improvements are only seen in terms of specificity for two models across all metrics and fingerprinting methods. In the Ataxia dataset, a few performance enhancements are

noted across various metrics, particularly for Atom Pair and Topological Torsion, but never for the Morgan fingerprints. Overall, it appears that the count representation fails to provide a significant performance boost. Additional tests conducted on a transformed count fingerprint, which accounts for multiple occurrences of the same substructure while reducing the contribution of highly frequent substructures, yield similar results: there is no consistent improvement, with some metrics showing improvement while others degrade. These findings suggest that, although the count-based fingerprints have greater resolution and can provide a unique representation for metabolites, such representation does not necessarily enhance the ability of ML classifiers to predict the metabolite response. It underscores the notion that the presence of specific functional groups within the molecule holds greater relevance than their number and that a unique encoding is indeed crucial but not sufficient to ensure the comprehensive representation of structural characteristics. Being the binary representation the standard in cheminformatics, all subsequent experiments are conducted using binary fingerprints.

Strategies to address class imbalance were systematically evaluated across both datasets, revealing discernible patterns. Models trained on oversampled data often exhibit an improvement in their ability to correctly identify the negative class, as evidenced by a significant increase in specificity. This trend is particularly pronounced for pattern-matching fingerprints, observed in an average of four out of five models. However, this increase in specificity is seldom accompanied by an improvement in overall performance metrics, suggesting a trade-off with recall. There are isolated instances (not consistently observed across models or fingerprints) where an improvement in F1 score is also noted. Models trained on undersampled data demonstrate an opposite trend. They frequently demonstrate an improvement in their ability to correctly identify the positive class, marked by a significant increase in recall. However, this improvement is rarely associated with an increase in overall metrics (except twice as an increase in ROC), indicating that an increase in recall comes at the expense of specificity. Therefore, while these techniques do not unequivocally enhance the overall quality of the models, they may offer better performance in certain metrics, which could prove useful in specific applications. Subsequent experiments were conducted without employing any sampling techniques.

In addressing high dimensionality, the effect of feature selection on model performance presents a more nuanced pattern. In hashed fingerprints, models were trained with varying numbers of features selected via the χ^2 test, ranging from 100 to 300. The results showed that performance tends to decline when the feature count falls below 200 and does not improve above 200. Thus, 200 emerges as the minimum number of features that avoids a drop in performance for hashed fingerprints, effectively encoding the structural properties of a diverse range of metabolites. Additionally, the selection of the top 200 features in hashed fingerprints was carried out using various statistical tests, including χ^2, mutual information, and ANOVA F-value. Among these tests, the χ^2 test emerged as the most effective in selecting features for training better-performing models. Additionally, feature importance analysis was conducted on each model to identify the top 200 features for that model, but it did not yield a better feature set compared to that identified by the χ^2 test. Models trained on hashed fingerprints with selected features demonstrated improved performance in terms of recall, often accompanied by increased ROC and occasionally by enhanced F1 scores. This improvement was particularly noticeable in Morgan fingerprints and for models like Naive Bayes and Logistic Regression, but it was less pronounced in other tree-based models, which can inherently identify relevant features. The consistent improvement of a global metric alongside recalls underscores the ability of feature selection to improve the overall quality of the model, particularly for those models lacking internal mechanisms to balance feature contributions. In an attempt to enhance models trained on pattern-matching fingerprints, the top 100 features of MACCS and functional fingerprints were selected using the χ^2 test, and models were retrained on the reduced dataset. However, feature selection methods did not yield noticeable improvements in model performance. Specifically, no improvement was observed in the Ataxia dataset, and only twice in the Hypoxia dataset was there an improvement in terms of recall, without a corresponding enhancement in global metrics. It seems that for pattern-matching fingerprints, which are less affected by high dimensionality, further reducing the number of features does not confer any advantage. Therefore, while feature importance can be effective for high-dimensional hashed fingerprints, particularly on non-tree-based models, it does not present obvious advantages for pattern-matching fingerprints.

Across both datasets, comparing model performances on a given variation of a fingerprint (where fingerprint variations are denoted in the tables are *binary*, *binary oversampled*, *binary undersampled*, *binary selected*, and *count*) reveals that certain models significantly outperform others in terms of a particular metric. However, such occurrences are rare, and it is uncommon for a model to significantly outperform all others across multiple metrics, suggesting that superiority is limited to one aspect of the model. Furthermore, across all experiments, no single model type consistently and significantly outperforms all others. Although Naive Bayes and Logistic regression consistently exhibit higher metrics, these differences are seldom statistically significant, suggesting that no model can be conclusively preferred. Interestingly, there is no discernible correlation between the complexity of the model and its performance. In fact, the neural network model generally exhibits poorer performance compared to the simpler linear logistic model, underscoring the effectiveness of simple, interpretable models. When comparing model performances across fingerprints (a specific variation across all fingerprints), similar results are observed. While Morgan-based fingerprints typically yield models with higher metrics, these differences are rarely significant, and a model that significantly outperforms all others across fingerprints never does so across more than one metric. Once again, no fingerprint emerges as conclusively superior.

4.3. Explaining predictions

After exhaustively exploring various pre-processing techniques and learning strategies, no model emerges as decisively superior to all others. However, both BNB and LR consistently exhibit strong performance, often yielding the highest recall and ROC metrics, especially when feature selection is applied. While the performance differences between BNB and LR are frequently not statistically significant, both models consistently outperform others, achieving satisfactory ROC scores exceeding 0.65 in the Ataxia dataset and 0.70 in the Hypoxia dataset. Among these two models, LR stands out as a simpler and more interpretable option due to its linear nature. LR's coefficients directly reveal the influence of each feature on the classification process, enhancing interpretability. Consequently, LR was chosen for its satisfactory performance and interpretability as the model for further investigation into feature contributions to the classification output. Model coefficients were used to evaluate feature importance and identify the most relevant corresponding substructures.

Feature importance in LR was also evaluated using SHAP values, computed for each test sample and averaged for each feature. This analysis revealed that features with the highest SHAP values identified most of the same chemical configurations as model coefficients, but SHAP value calculation is computationally more intensive. Moreover, while SHAP values provide a more detailed explanation of individual feature contributions to model predictions, they are generally less interpretable than model coefficients. While SHAP values remain an available option for estimating the relevance of features, in this context, the use of model coefficients was preferred.

4.3.1. Relevant substructures in AT

For the Ataxia dataset, feature importance analysis was conducted on binary Morgan of radius 2 and binary functional fingerprints, to exemplify feature importance analysis on hashed and pattern-matching

Table 2
Ataxia experiment: performance analysis of Decision Tree (DT), Bernoulli Naive Bayes (BNB), Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), and MultiLayer Perceptron (MLP) trained on fingerprints. Performance metrics include area under the curve of the receiver operating characteristic (ROC), F1-score (F1), recall (R), and specificity (SP). Bold values indicate that the model trained on a specific variation of the binary fingerprint (oversampled, undersampled, top features selected by χ^2, and count) significantly outperformed the model trained on the original binary fingerprint, based on a t-test. Italicised and underlined values denote that the model significantly outperformed all others regarding a given performance metric, within a specific fingerprint and across all fingerprints, respectively.

	Binary FP				Binary Oversampled				Binary Undersampled				Binary Selected				Count FP			
	ROC	F1	R	SP	ROC	F1	R	SP	ROC	F1	R	SP	ROC	F1	R	SP	ROC	F1	R	SP
Morgan radius 2																				
BNB	0.624	0.576	0.504	0.744	0.623	0.585	*0.474*	**0.773**	0.615	0.551	**0.544**	0.686	**0.657**	**0.609**	**0.535**	0.778	0.624	0.576	0.504	0.744
LR	0.636	0.584	0.526	0.746	0.553	0.538	0.324	**0.783**	0.631	0.568	0.555	0.707	**0.652**	0.590	**0.569**	0.735	0.631	0.585	0.502	0.759
DT	0.611	0.597	0.385	0.837	0.592	0.568	0.392	0.791	0.596	0.556	**0.451**	0.742	0.608	0.592	0.387	0.829	0.592	0.566	0.397	0.786
RF	0.639	0.623	0.429	0.850	0.623	0.625	0.350	**0.895**	0.631	0.574	**0.539**	0.724	0.644	*0.626*	0.437	0.850	0.644	*0.625*	0.445	0.844
XGB	0.639	0.610	0.465	0.813	0.619	0.609	0.384	**0.855**	0.606	0.535	**0.559**	0.654	0.650	0.611	**0.508**	0.793	0.636	0.607	0.461	0.812
MLP	0.560	0.565	0.216	*0.904*	0.561	0.562	0.247	0.876	0.565	0.461	**0.607**	0.523	**0.589**	**0.587**	**0.308**	*0.870*	**0.575**	0.578	**0.259**	*0.890*
Morgan radius 3																				
BNB	0.633	0.574	0.544	0.722	0.633	0.577	*0.537*	0.729	0.620	0.546	0.574	0.665	**0.669**	**0.597**	**0.612**	0.726	0.633	0.574	0.544	0.722
LR	0.638	0.585	0.529	0.746	0.550	0.536	0.317	**0.783**	0.634	0.573	0.549	0.719	**0.659**	0.590	**0.597**	0.722	0.632	0.584	0.509	0.754
DT	0.599	0.583	0.374	0.824	0.585	0.549	0.425	0.744	0.591	0.556	**0.432**	0.750	0.610	0.595	0.384	0.836	0.593	0.575	0.370	0.816
RF	0.647	*0.628*	0.447	0.846	0.617	*0.621*	0.333	***0.900***	0.633	0.565	**0.567**	0.698	0.647	*0.628*	0.447	0.847	0.647	*0.624*	0.460	0.835
XGB	0.630	0.599	0.459	0.801	0.608	0.605	0.344	**0.872**	0.605	0.527	**0.580**	0.630	**0.658**	0.609	**0.540**	0.776	0.628	0.600	0.449	0.808
MLP	0.555	0.561	0.190	*0.921*	0.562	0.559	**0.258**	0.865	**0.572**	0.475	**0.614**	0.531	**0.593**	**0.590**	**0.316**	*0.870*	0.565	0.570	0.218	*0.911*
TopologicalTorsion																				
BNB	0.614	0.545	0.560	0.669	0.616	*0.550*	0.550	0.682	0.601	0.518	0.589	0.613	**0.632**	**0.560**	0.582	0.683	0.611	**0.563**	0.490	**0.731**
LR	0.624	0.566	0.532	0.715	0.576	0.547	0.400	**0.753**	0.609	0.540	0.553	0.665	0.632	0.571	0.550	0.714	0.629	**0.582**	0.504	**0.754**
DT	0.573	0.534	0.412	0.733	0.566	0.544	0.361	0.771	0.563	0.478	**0.554**	0.572	0.576	0.537	0.427	0.725	0.587	**0.571**	0.353	0.821
RF	0.614	0.592	0.410	0.817	0.602	0.605	0.315	**0.890**	0.604	0.528	**0.571**	0.636	0.623	*0.590*	**0.458**	0.788	0.622	***0.616***	0.372	**0.872**
XGB	0.614	0.578	0.456	0.772	0.609	**0.598**	0.370	**0.848**	0.595	0.511	**0.589**	0.601	0.614	0.566	**0.495**	0.733	0.621	0.583	0.471	0.771
MLP	0.576	0.581	0.256	*0.896*	0.577	0.580	0.262	0.892	0.576	0.495	**0.568**	0.584	0.569	0.568	0.274	*0.865*	0.583	0.579	**0.305**	0.861
AtomPair																				
BNB	0.577	0.444	*0.705*	0.448	0.575	0.444	*0.699*	0.452	0.577	0.441	*0.712*	0.442	0.588	0.443	***0.739***	0.436	**0.611**	**0.495**	*0.687*	**0.536**
LR	0.650	0.584	0.579	0.722	0.543	0.528	0.315	**0.771**	0.632	0.556	0.592	0.672	0.645	0.579	0.575	0.715	0.657	**0.597**	0.569	**0.745**
DT	0.600	0.547	0.499	0.702	0.561	**0.559**	0.266	**0.856**	0.580	0.506	0.547	0.612	0.599	0.547	0.495	0.704	0.602	0.557	0.472	0.732
RF	0.650	0.618	0.486	0.813	0.577	0.593	0.196	***0.958***	0.633	0.561	**0.580**	0.686	0.647	0.610	0.497	0.797	0.640	0.609	0.469	0.811
XGB	0.629	0.616	0.400	0.857	*0.593*	0.606	0.259	**0.928**	0.619	0.541	**0.589**	0.649	0.638	0.612	**0.453**	0.822	0.640	0.617	**0.448**	0.833
MLP	0.568	0.574	0.224	*0.912*	0.572	0.573	0.258	0.886	**0.596**	0.508	**0.600**	0.592	0.579	0.578	**0.279**	*0.879*	0.567	0.571	0.241	*0.894*
Daylight																				
BNB	0.598	0.548	0.491	0.705	0.592	0.542	0.486	0.698	0.567	0.499	0.526	0.608	**0.624**	**0.576**	0.507	**0.742**	0.598	0.548	0.491	0.705
LR	0.637	0.587	0.524	0.751	0.555	0.537	0.333	**0.776**	0.615	0.544	**0.563**	0.666	0.639	0.582	0.544	0.734	0.635	0.575	0.550	0.721
DT	0.585	0.544	0.448	0.721	0.573	**0.567**	0.302	**0.844**	0.566	0.492	**0.542**	0.591	0.586	0.538	0.472	0.701	0.597	0.541	**0.506**	0.688
RF	0.629	0.620	0.390	0.867	0.579	0.589	0.234	***0.924***	0.620	0.557	**0.544**	0.696	0.632	0.621	0.401	0.863	0.643	0.621	**0.448**	0.837
XGB	0.621	0.614	0.374	0.868	0.584	0.592	0.258	**0.910**	0.595	0.520	**0.563**	0.628	0.624	0.610	0.400	0.848	0.628	0.617	**0.399**	0.858
MLP	0.570	0.567	0.278	0.862	0.565	0.517	**0.457**	0.672	0.560	0.480	**0.524**	0.595	0.567	0.565	0.276	0.859	0.586	0.579	0.315	0.856
Functional																				
BNB	0.616	0.549	0.555	0.677	0.615	0.549	0.551	0.679	0.610	0.540	0.561	0.660	0.617	0.551	0.553	0.682	0.616	0.548	0.554	0.677
LR	0.625	0.563	0.545	0.705	0.617	0.556	0.536	0.698	0.617	0.546	0.566	0.667	0.628	0.566	0.548	0.708	0.639	**0.581**	0.549	**0.729**
DT	0.616	0.572	0.484	0.747	0.571	0.557	0.334	**0.808**	0.604	0.555	0.484	0.723	0.609	0.565	0.477	0.740	0.604	0.555	0.487	0.720
RF	0.626	*0.613*	0.401	0.851	0.589	0.589	0.302	***0.877***	0.623	0.575	**0.500**	0.745	0.624	*0.611*	0.396	0.852	**0.645**	0.615	**0.474**	0.815
XGB	0.628	0.598	0.456	0.801	0.618	0.594	0.423	**0.812**	0.608	0.541	**0.547**	0.669	0.626	0.593	0.463	0.789	0.637	0.602	0.479	0.795
MLP	0.575	0.566	0.314	0.836	**0.591**	0.543	**0.478**	0.705	0.575	0.491	**0.566**	0.585	0.587	0.576	0.340	0.834	0.581	0.579	0.289	***0.872***
Fragments																				
BNB	0.608	0.531	0.575	0.640	0.601	0.528	0.561	0.641	0.601	0.521	0.582	0.620					0.608	0.531	0.575	0.640
LR	0.623	0.549	0.577	0.668	0.614	0.540	0.574	0.654	0.611	0.530	0.590	0.632					0.619	0.559	0.537	**0.701**
DT	0.593	0.558	0.437	0.750	0.598	0.559	0.450	0.745	0.595	0.530	**0.528**	0.662					0.600	0.556	0.468	0.731
RF	0.625	*0.601*	0.430	*0.820*	0.608	0.586	0.405	*0.811*	0.619	*0.555*	**0.546**	0.691					0.634	*0.606*	0.456	0.812
XGB	0.621	0.574	0.501	0.742	0.618	0.578	0.477	**0.760**	0.613	0.532	**0.594**	0.632					0.625	0.584	0.482	**0.768**
MLP	0.602	0.573	0.421	0.782	0.601	0.555	**0.479**	0.724	0.590	0.505	**0.589**	0.591					0.587	0.576	0.340	**0.835**
MACCS																				
BNB	0.618	0.546	0.572	0.664	0.618	0.545	0.570	0.665	0.610	0.535	0.576	0.645	0.623	0.549	0.581	0.665				
LR	0.623	0.553	0.567	0.678	0.614	0.545	0.558	0.670	0.617	0.537	0.595	0.640	0.622	0.550	0.570	0.673				
DT	0.595	0.540	0.502	0.688	0.579	**0.572**	0.313	**0.844**	0.599	0.524	**0.564**	0.634	0.602	0.540	0.529	0.675				
RF	0.644	0.612	0.479	0.810	0.598	0.606	0.283	***0.913***	0.630	*0.566*	**0.558**	*0.703*	0.645	*0.614*	0.478	0.812				
XGB	0.634	0.599	0.473	0.794	0.612	0.602	0.372	**0.851**	0.623	0.544	**0.594**	0.651	0.638	0.599	0.490	0.786				
MLP	0.582	0.582	0.291	*0.873*	0.593	0.571	**0.387**	0.799	0.593	0.516	**0.571**	0.616	0.583	0.581	0.299	*0.868*				

fingerprints. Fig. 5 (a) presents the feature importance, computed as model coefficients, for the top 25 bits of the Morgan binary fingerprint of radius 2 and illustrates the chemical configuration corresponding to each bit. Structures enclosed in rectangles represent instances of fingerprint clashing, i.e. distinct configurations mapped onto the same bit. The analysis of the most influential bits reveals chemical configurations that are most associated with the disease and point to specific classes of biochemical compounds. Bit 192, the most predictive, along with bits 486 and 814, identifies phosphate groups, while bits 692, 295, 795, and the second most frequent configuration of bit 573 correspond to phosphate group is involved in one or two phosphodiester bonds. These bits isolate a variety of phosphorylated compounds affected in the disease, including sugar phosphates found in the pentose phosphate metabolic pathway and glycolysis, nucleotides and their derivatives, as well as phospholipids. Bit 640 identifies a pentose sugar, while the most common configuration of bit 363 represents a phosphate group bound to a pentose, pointing to metabolites of the pentose phosphate pathway, as well as phosphorylated nucleosides, known to be affected in AT [46]. The second most prevalent configuration of bit 786 identifies adenine-containing nucleotides, encompassing criti-

Table 3
Hypoxia experiment - performance analysis of machine learning models trained on fingerprints as described in Table 2.

	Binary FP				Binary Oversampled				Binary Undersampled				Binary Selected				Count FP			
	ROC	F1	R	SP	ROC	F1	R	SP	ROC	F1	R	SP	ROC	F1	R	SP	ROC	F1	R	SP
Morgan radius 2																				
BNB	0.532	0.487	0.288	0.776	0.518	0.517	0.100	**0.935**	0.528	0.396	**0.532**	0.524	**0.721**	**0.591**	**0.621**	0.821	0.532	0.487	0.288	0.776
LR	0.538	0.482	0.318	0.759	0.519	0.516	0.124	**0.915**	0.544	0.392	**0.585**	0.504	**0.683**	0.522	**0.656**	0.709	0.541	0.468	0.376	0.705
DT	0.563	0.461	0.453	0.673	0.527	**0.503**	0.215	**0.840**	0.540	0.408	0.518	0.563	0.587	0.447	0.538	0.635	0.541	0.441	0.421	0.661
RF	0.535	0.540	0.106	0.965	0.547	0.554	0.126	*0.968*	0.563	0.417	**0.576**	0.550	**0.573**	0.567	**0.218**	0.929	0.540	0.546	0.112	0.967
XGB	0.573	0.554	0.238	0.907	0.552	0.555	0.153	**0.952**	0.566	0.421	**0.576**	0.555	0.599	0.569	0.306	0.892	0.570	0.556	0.224	0.917
MLP	0.519	0.521	0.088	0.949	0.525	0.527	0.097	0.953	0.550	0.433	**0.503**	0.596	**0.551**	**0.555**	**0.144**	*0.959*	0.522	0.526	0.091	0.953
Morgan radius 3																				
BNB	0.507	0.484	0.209	0.805	0.511	0.510	0.094	**0.928**	0.504	0.410	**0.409**	0.600	**0.752**	**0.612**	**0.671**	0.833	0.507	0.484	0.209	0.805
LR	0.532	0.471	0.326	0.737	0.540	**0.537**	0.144	**0.935**	0.524	0.379	**0.562**	0.487	**0.708**	**0.551**	**0.665**	0.752	0.531	0.508	0.212	0.851
DT	0.546	0.486	0.329	0.762	0.549	0.521	*0.244*	**0.855**	0.525	0.408	**0.488**	0.562	0.561	0.499	0.341	0.781	0.549	0.498	0.306	0.792
RF	0.530	0.534	0.094	0.967	0.552	**0.563**	0.132	0.972	0.537	0.415	**0.512**	0.562	**0.565**	0.562	**0.191**	0.939	0.531	0.536	0.091	0.972
XGB	0.560	0.547	0.200	0.920	0.550	0.554	0.141	**0.958**	0.551	0.426	**0.524**	0.579	0.588	0.578	0.250	0.925	0.552	0.548	0.168	0.936
MLP	0.527	0.533	0.091	0.963	0.533	0.538	0.106	0.961	0.518	0.418	**0.453**	0.582	**0.582**	**0.585**	**0.209**	0.955	0.524	0.528	0.088	0.960
TopologicalTorsion																				
BNB	0.533	0.483	0.309	0.757	0.531	0.501	0.235	**0.826**	0.515	0.369	**0.547**	0.483	**0.680**	**0.558**	**0.565**	0.794	0.524	0.492	0.238	0.810
LR	0.529	0.510	0.197	0.860	0.528	0.519	0.156	**0.900**	0.512	0.380	**0.532**	0.491	**0.648**	0.525	**0.550**	0.745	0.559	0.522	0.279	0.838
DT	0.516	0.444	0.326	0.705	0.522	**0.509**	0.165	**0.880**	0.515	0.407	0.462	0.568	0.518	0.444	0.329	0.706	0.511	0.494	0.165	0.857
RF	0.523	0.523	0.082	0.964	0.539	0.549	0.103	*0.976*	0.513	0.402	**0.482**	0.543	0.547	0.543	0.147	0.947	0.511	0.508	0.035	***0.986***
XGB	0.536	0.535	0.144	0.929	0.542	0.546	0.126	**0.957**	0.513	0.402	**0.485**	0.540	0.565	0.546	**0.229**	0.901	0.539	0.535	0.144	0.934
MLP	0.524	0.526	0.088	0.960	0.529	0.517	0.138	0.919	0.526	0.417	**0.479**	0.573	**0.563**	0.555	**0.197**	0.928	0.532	0.535	0.112	0.953
AtomPair																				
BNB	0.529	0.403	*0.506*	0.551	0.517	0.418	*0.450*	0.585	0.520	0.358	0.565	0.476	**0.637**	0.425	***0.750***	0.524	0.561	0.450	0.488	0.634
LR	0.553	0.504	0.312	0.795	0.533	0.527	0.138	**0.927**	0.552	0.397	**0.594**	0.510	**0.664**	0.509	**0.629**	0.698	0.587	0.518	0.391	0.784
DT	0.540	0.478	0.335	0.744	0.502	0.501	0.085	**0.918**	0.554	0.426	**0.532**	0.576	0.561	0.499	0.347	0.775	0.545	0.497	0.297	0.792
RF	0.525	0.528	0.082	0.967	0.504	0.497	0.021	***0.987***	**0.564**	0.436	**0.532**	0.595	**0.577**	0.561	**0.241**	0.913	0.523	0.526	0.079	0.966
XGB	0.514	0.515	0.059	0.970	0.506	0.504	0.035	0.978	**0.558**	0.425	**0.547**	0.570	**0.576**	**0.568**	**0.212**	0.939	0.517	0.517	0.071	0.963
MLP	0.511	0.508	0.047	0.974	0.506	0.503	0.038	0.975	**0.556**	0.435	**0.515**	0.597	**0.541**	**0.547**	**0.115**	*0.967*	0.515	0.517	0.062	0.968
Daylight																				
BNB	0.553	0.396	*0.603*	0.503	0.555	0.400	*0.600*	0.511	0.546	0.342	***0.691***	0.401	**0.596**	0.404	***0.694***	0.498	0.553	0.396	*0.603*	0.503
LR	0.539	0.490	0.300	0.778	0.535	**0.525**	0.162	**0.907**	0.535	0.414	**0.503**	0.566	**0.632**	0.515	**0.526**	0.738	0.556	0.505	0.312	0.801
DT	0.530	0.483	0.285	0.775	0.524	0.520	0.129	**0.918**	0.505	0.396	**0.474**	0.536	0.543	0.478	0.335	0.751	0.516	0.462	0.300	0.731
RF	0.549	0.553	0.132	*0.965*	0.544	0.550	0.124	0.964	0.537	0.415	**0.518**	0.557	0.562	0.545	**0.221**	0.903	0.551	0.553	0.153	0.949
XGB	0.553	0.554	0.159	0.948	0.540	0.543	0.124	0.956	0.522	0.412	**0.482**	0.561	0.563	0.553	0.200	0.925	0.542	0.545	0.132	0.952
MLP	0.542	0.539	0.138	0.946	0.549	0.547	0.159	0.940	0.541	0.426	**0.497**	0.584	0.538	0.534	0.147	0.929	0.533	0.533	0.115	0.952
Functional																				
BNB	0.508	0.434	0.338	0.678	0.514	0.433	0.403	0.625	0.515	0.398	0.447	0.583	0.553	0.422	**0.538**	0.567	0.508	0.434	0.338	0.677
LR	0.550	0.475	0.385	0.714	0.556	0.487	0.371	0.741	0.537	0.378	**0.597**	0.476	0.536	0.413	0.506	0.567	0.548	0.463	0.415	0.682
DT	0.547	0.444	0.441	0.652	0.551	**0.532**	0.215	**0.888**	0.539	0.398	0.556	0.522	0.529	0.453	0.368	0.691	0.536	0.472	0.341	0.731
RF	0.552	0.541	0.188	0.915	0.554	0.552	0.168	*0.941*	0.564	0.404	**0.612**	0.516	0.530	0.501	0.238	0.822	0.533	0.512	0.206	0.860
XGB	0.570	0.531	0.291	0.849	0.554	0.539	0.206	**0.903**	0.576	0.420	**0.606**	0.547	0.553	0.500	0.321	0.785	0.520	0.513	0.135	**0.905**
MLP	0.550	0.539	0.182	0.917	0.537	0.526	0.168	0.907	0.568	0.438	**0.541**	0.595	0.549	0.535	0.194	*0.903*	0.555	*0.552*	0.174	*0.936*
Fragments																				
BNB	0.525	0.393	0.526	0.524	0.533	0.402	*0.538*	0.528	0.524	0.371	0.553	0.495					0.525	0.393	0.526	0.524
LR	0.509	0.427	0.403	0.615	0.525	0.446	0.397	0.653	0.508	0.381	**0.521**	0.496					0.528	0.441	0.421	0.636
DT	0.570	0.477	0.432	0.708	0.582	**0.537**	0.312	**0.851**	0.534	0.397	0.547	0.521					0.530	0.419	0.444	0.616
RF	0.567	0.558	0.218	0.917	0.598	0.568	0.297	0.900	0.567	0.412	**0.600**	0.533					0.528	0.525	0.121	0.935
XGB	0.584	0.528	0.344	0.823	0.596	**0.554**	0.321	**0.872**	0.591	0.431	**0.621**	0.562					0.531	0.511	0.203	0.860
MLP	0.587	0.558	0.279	0.894	0.582	0.549	0.285	0.880	0.568	0.431	**0.559**	0.577					0.545	0.536	0.174	0.917
MACCS																				
BNB	0.530	0.405	0.512	0.549	0.543	0.412	*0.538*	0.547	0.526	0.371	0.576	0.475	0.558	0.405	0.594	0.522				
LR	0.562	0.477	0.415	0.709	0.574	**0.508**	0.368	**0.780**	0.528	0.388	**0.553**	0.502	0.585	0.467	**0.512**	0.658				
DT	0.575	0.504	0.379	0.770	0.546	0.541	0.171	**0.922**	0.553	0.406	**0.576**	0.530	0.552	0.483	0.359	0.745				
RF	0.538	0.530	0.165	0.912	0.539	0.542	0.124	*0.954*	0.569	0.422	**0.582**	0.556	0.536	0.510	0.229	0.842				
XGB	0.579	0.558	0.253	0.905	0.550	0.548	0.162	**0.938**	0.573	0.424	**0.588**	0.558	0.571	0.539	0.268	0.874				
MLP	0.536	0.534	0.135	0.936	0.542	0.534	0.162	0.922	0.569	0.435	**0.550**	0.587	0.542	0.533	0.159	*0.925*				

cal molecules such as adenosine triphosphate (ATP) and cofactors like flavin adenine dinucleotide (FAD), Coenzyme A, nicotinamide adenine dinucleotide (NAD), and nicotinamide adenine dinucleotide phosphate (NADP). Research has revealed that cells lacking ATM exhibit a compromised ability to replenish ATP in response to increased energy demands, leading to chronic ATP insufficiency [47]. Moreover, perturbations in the levels of pyridine nucleotides, - small molecules comprised of adenosine monophosphate and nicotinamide mononucleotide - have been observed in AT cells, marked by a significant reduction in both the reduced and oxidized forms of NAD [48]. Further investigations have unveiled the role of ATM in stimulating NADPH production by inducing the activity of glucose-6-phosphate dehydrogenase, the limiting enzyme of the pentose phosphate pathway responsible for the production of NADPH. ATM-mediated response to reactive oxygen species also promotes nucleotide synthesis, facilitating the repair of double-strand breaks [46]. Consequently, the suppression of ATM has been associated with diminished levels of nucleotides and their derivatives on multiple occasions. Bit 820, the second most predictive bit, identifies carboxylic acids and is found in two primary groups of affected metabolites. The first comprises standard and modified (often acetylated) amino acids and dipeptides, while the second includes phosphatidylserines. Bit 389 also identifies carboxylic groups found in (modified) amino acids and small peptides. One notable example is the tripeptide glutathione, known for its crucial role in preventing damage to important cellular components caused by reactive oxygen species and free radicals. While the impairment of glutathione levels in AT cells has been

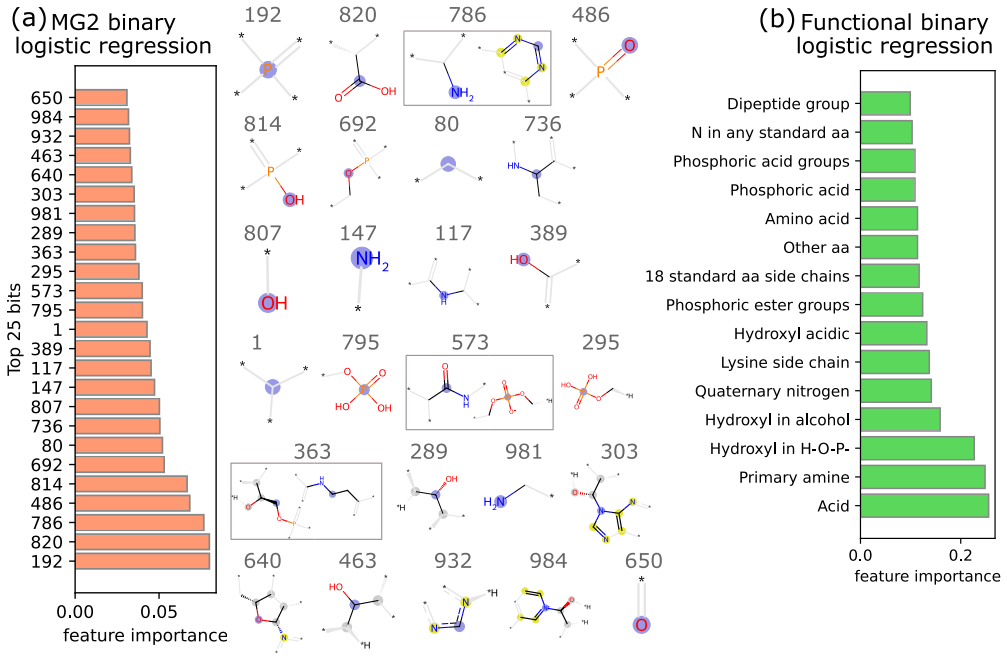


Fig. 5. Feature importance analysis of logistic regression models trained on the Ataxia dataset. (a) Top 25 bits of the Morgan binary fingerprint of radius 2, with visual representations of the most frequent substructures associated with each bit. Structures enclosed in rectangles represent instances of fingerprint clashing, i.e. distinct configurations mapped onto the same bit. (b) Top 15 features of the functional binary fingerprint.

characterised [49], the involvement of other peptides remains unclear. Bits 147, 981, and 786 identify primary amines in various configurations, which are also found in (modified) amino acids and dipeptides, as well as in phosphatidylserines. Bit 981, in particular, identifies primary amines at the end of a carbon chain, also present in several affected polyamines such as putrescine, spermine, and spermidine. This bit also identifies several affected phosphoethanolamines. Phosphatidylserines and phosphoethanolamines identified by several important bits, constitute integral components of cell membranes, with roles in apoptosis and cell signalling. Their metabolic pathways are intricately linked, and their synthesis predominantly occurs within mitochondria-associated membranes [50]. ATM loss in AT has been linked to mitochondrial abnormalities, including elevated reactive oxygen species, increased aberrant mitochondria, high cellular respiratory capacity, and decreased mitophagy [51] and could be responsible for the lower level of phospholipids synthesised within this organelle. Bits 736 and 117 identify secondary amines found in amino acids, small peptides and their derivatives, often as part of the peptide bond. These bits also identify ceramides and sphingolipids, which serve as secondary messengers in activating the apoptotic cascade [52]. The second most common configuration in bit 363 is a secondary amine in a configuration found in pantothenic acid, which is essential for the synthesis of Coenzyme A. While several groups of affected metabolites belong to pathways with well-established roles in the disease, other groups, such as peptides and polyamines, offer opportunities for further investigation, as their exact impact on the disease remains unclear.

A parallel analysis was conducted on the binary functional fingerprint. Fig. 5 (a) illustrates the top 15 features with the highest model coefficients. Notably, the most predictive configuration within the functional fingerprint corresponds to *Acid*, also detected by Morgan bit 820, primarily associated with affected amino acids, dipeptides, their derivatives, and pentose phosphates. Patterns such as *Phosphoric acids*, *Phosphoric acid groups*, *Phosphoric ester groups*, and *Hydroxyl in H-O-P* identify the same groups of affected phosphorylated compounds as detected by Morgan bits, i.e. phospho-sugars from the pentose phosphate pathway, early glycolytic intermediates, nucleotides, derivatives thereof, and phospholipids. Several patterns associated with amino acids are highly predictive, including *Amino acid*, *Other aa*, *18 standard aa side chain*, *N in any standard aa*, *Dipeptide group*. In this fingerprint, amino acid and dipeptide configurations are distinctly identified as predictive, while in the Morgan fingerprint, the predictive configurations identifying amino acids and dipeptides are amines and acids. *Primary amine* identifies the same metabolites as its counterpart Morgan bits, encompassing amino acids, dipeptides, phosphatidylserines, and adenine-containing nucleotides. Finally, *Lysine side chain* and *Two primary or secondary amines* identify affected polyamine.

4.3.2. Relevant substructures in hypoxia

Feature importance analysis of the logistic regression model, trained on the top 200 bits of Morgan fingerprints of radius 3 selected via the χ^2 test, uncovers patterns consistent with documented shifts in metabolic pathways induced by hypoxia. The fingerprint bits with higher model coefficients are depicted in Fig. 6 (a). Notably, bit 509, linked to dipeptides containing aromatic amino acids, aligns with known alterations in amino acid pathways during hypoxia. While pathways like alanine-aspartate-glutamate and arginine synthesis are well-documented in hypoxic conditions, the synthesis of aromatic amino acids remains underexplored, presenting an avenue for further investigation. Bits 273 and 873 represent the guanine nitrogen base, aligning observed alterations in purine metabolism under hypoxia. Previous studies have noted an increase in guanine-containing nucleotides like GDP and GTP, alongside nucleotide depletion and precursor accumulation during hypoxia [53]. The most frequent configuration in bit 953 highlights sulfur bonds in cysteine derivatives, reflecting disruptions in cysteine-methionine metabolism and changes in levels of cysteine-derived antioxidants such as glutathione and taurine [54]. Conversely, the second most frequent configuration indicates a peptide bond, further emphasising the role of amino acid metabolism. Bits 674 and 547 identify polyamines such as spermine, putrescine, and acetylated spermine, whose role in endothelial cell survival during hypoxia has been previously studied [55]. Bit 482 is found in sugars of the pentose phosphate pathway and sphingosine, including sphingosine-1-phosphate, known for its protective role in response to hypoxia [56]. Bit 730 identifies a benzene group found in aromatic amino acids, overlapping with the metabolites identified by bit 509. Similarly, bit 752, representing a primary amine, identifies

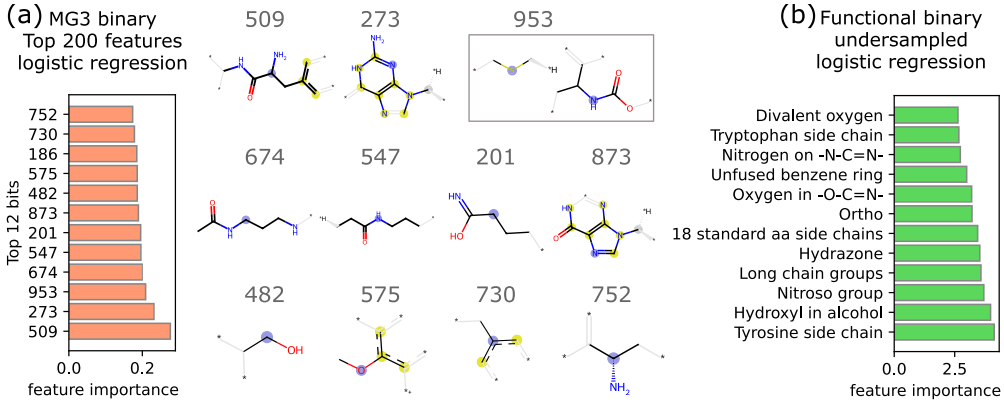


Fig. 6. Feature importance analysis of logistic regression models trained on the Hypoxia dataset. (a) Top 12 bits of the Morgan binary fingerprint of radius 3 with feature selection, where the top 200 features were selected using the χ^2 test, with visual representations of the most frequent substructures associated with each bit. Structures enclosed in rectangles represent fingerprint clashing, i.e. distinct configurations mapped onto the same bit. (b) Top 12 features of the undersampled functional binary fingerprint.

metabolite groups previously detected by other bits such as polyamines, sphingosine, and derivatives of amino acids and peptides.

The key structures identified in the undersampled binary functional fingerprint, shown in Fig. 6 (b), align with observations from the Morgan fingerprint of radius 3. For instance, both *Tyrosine side chain* and *Tryptophan side chain* were associated with peptides containing aromatic amino acids, mirroring the findings from the Morgan fingerprint. Similarly, the presence of *Unfused benzene ring* indicates the same metabolites in both fingerprints. Additionally, the presence of *18 standard aa side chain* further underscores the significance of peptides in the dataset. Finally, the detection of *Long chains* points to various groups of already detected metabolites, including polyamines, sphingosine, and certain amino acids/peptides.

5. Discussion

This study explores a novel approach integrating molecular fingerprinting and ML to analyse metabolic data and offer insights into affected pathways, drawing inspiration from well-established practices in drug discovery. Introducing this approach to a new context requires careful consideration. The study aims to unveil both the challenges and opportunities inherent in deploying this method, thereby laying the groundwork for future research. Specifically, the investigation focuses on four key areas: encoding structural information, data pre-processing, model training, and model interpretation.

1) Structural encoding. The analysis comprehensively investigates seven fingerprints, five widely adopted in cheminformatics and drug discovery (Morgan, Atom Pair, Topological Torsion, Daylight, and MACCS), and two introduced within this study (named fragment and functional). The examination delves deeply into the merits and limitations of these diverse fingerprints, with a specific emphasis on their resolution power.

Hashed fingerprints offer flexible and effective methods for mapping a wide range of substructures without requiring prior knowledge of the structures under study. This approach grants flexibility through user-defined size and radius parameters which determine the quality of the structural representation, and consequently, the interpretability of the model. A larger radius enables the identification of more complex configurations, providing more detailed and potentially more useful insights. Nevertheless, it also results in a greater number of potential configurations. Without an increase in bit number, this can lead to multiple substructures being mapped to the same bit, diminishing the interpretability of that bit and the structural uniformity of molecules associated with it. However, increasing the bit number increases data dimensionality, potentially impacting the performance of models trained on such data. Additionally, as dimensionality increases, also increases the number of bits identifying distinct but very similar configurations and providing overlapping information. To mitigate this, evaluating mutual information for each pair of bits can help filter out bits that do not improve resolution, reducing dimensionality and potentially enhancing model performance. The resolution analysis reveals that none of the hashed fingerprints in their binary form assigns a unique structural encoding to each metabolite. While the resolution power generally increases with the radius, it stabilises between 1024 and 2048, indicating that the inability to discriminate molecules is not attributed to the hashing process but to inherent limitations in the structural encoding. Resolution varies considerably among the selected hashed fingerprints, from Daylight, failing to resolve 12-15% of the metabolites, to Atom Pair, effectively encoding over 99% of metabolites. For almost all hashed and pattern-matching fingerprints, molecules sharing identical fingerprints are often metabolites that differ in the length of one or more hydrocarbon chains. This aspect does not pose significant challenges during the screening of compound libraries for drug discovery, as molecules with extended hydrocarbon chains are typically deemed unsuitable for drug candidates and rarely found in such libraries. However, it becomes a more serious concern in metabolomics, where many such molecules are found. Exploring alternative fingerprints or modifying existing ones to address the issue might be necessary for an accurate fingerprint-based encoding of the metabolome. The performance of models trained on different fingerprints indicates that there is no discernible correlation between resolution power and model performance. The comparable performance observed between models trained on the two Morgan fingerprints, with differences generally lacking statistical significance, along with the higher resolution capability of the Morgan fingerprint with a radius of 3, suggests that the more complex configurations identified by Morgan of radius 3 are not inherently more predictive than the simpler structures captured by bits in Morgan of radius 2. Similarly, the higher resolution of count fingerprints over binary counterparts (which in the case of Morgan offer a unique encoding for each metabolite) does not consistently yield significant performance improvements.

Pattern-matching fingerprints efficiently capture predefined chemical structures while maintaining a reasonable dimensionality. These fingerprints offer a high degree of flexibility, enabling researchers to focus on specific structures of interest to suit the unique research questions and datasets at hand. Moreover, these fingerprints are not affected by substructure clashes; the substructures corresponding to individual bits consistently identify the same chemical structures and offer readily interpretable insights. Nonetheless, pattern-matching fingerprints inherently encompass a finite set of substructures, which do not comprehensively represent the full spectrum of possible variations in chemical

structures. They require prior knowledge of discriminating substructures, which can be limited, particularly in the study of less known biological systems. Comparatively, the pattern-matching fingerprints exhibit lower resolution power compared to their hashed counterparts. Nevertheless, in some cases, models trained on pattern-matching fingerprints demonstrated comparable performance levels and identified similar predictive substructures.

In light of the distinctive advantages and limitations of both fingerprint types, there exists a promising opportunity to design novel fingerprints precisely tailored to the demands of metabolomics. Such fingerprints may potentially merge the strengths of pattern-matching and hashed approaches.

2) Data pre-processing. The datasets present several challenges, encompassing high dimensionality, sparsity, class imbalance, and the binary nature of features. Class imbalance, a common challenge in metabolic datasets where only a small portion of metabolites is generally affected, cannot be handled with data augmentation techniques like SMOTE and ADASYN due to the absence of continuous variables. Instead, random undersampling and oversampling methods are employed. Generally, oversampling the positive class consistently enhances specificity, while undersampling the negative class improves recall across all fingerprints and models, except for MLP, where in some cases both sampling techniques enhanced recall. Therefore, either technique could be employed to improve the desired performance metric.

The problem of dimensionality, determined by both the feature number and sample number, is most relevant in pattern-matching fingerprints and smaller metabolomics datasets (which can range from tens of thousands of metabolites to just a few hundred in untargeted metabolomics studies). To address high dimensionality while retaining feature interpretability, conventional techniques like principal component analysis are not suitable. Instead, feature selection methods guided by statistical tests such as χ^2, ANOVA, and mutual information are often employed. In the Ataxia dataset, which comprises 3999 samples and 1024 features, models trained on the top 200 features did not consistently enhance performance across all fingerprints, Improvements in model performance, in terms of both recall and ROC, were predominantly observed in the two Morgan fingerprints. Conversely, the Hypoxia dataset, with a third of the metabolites compared to the Ataxia dataset, exhibits a more pronounced dimensionality issue. Here, enhancements in recall and ROC were noticeable across all hashed fingerprints, particularly in logistic regression and Naive Bayes models. In both datasets, while feature selection doesn't definitively enhance performance, it also does not degrade it, and the performance remains comparable to models trained on the original data. This indicates that some features contain redundant or overlapping information, and their removal does not impact the predictive power of the model. In both datasets, no performance improvements were observed in pattern-matching fingerprints with feature selection, except for two instances, indicating that the majority of features hold predictive value for the target variable, ruling out concerns of over-fitting.

3) Model performance. The obtained results, while not outstanding, are in line with expectations established by other omics studies. Considering the intrinsic complexity of the metabolomics domain, achieving a ROC above 0.65 with recall above 0.60 represents a respectable predictive value, especially for noisy and unbalanced datasets. Among the various models explored, logistic regression emerges as the most consistent performer across all fingerprints, striking a good balance between recall and sensitivity. Surprisingly, extensive experimentation with diverse ML models, including linear, non-linear, probabilistic, tree-based models, and neural networks, coupled with thorough parameter optimisations, did not yield any substantial performance gains. This suggests that the core limitations may not be primarily rooted in the model choices. One avenue for potential improvement lies in the exploration of novel data representations, such as graph-based models like graph neural networks. However, initial experiments in this direction failed to deliver the anticipated performance boosts, reaffirming Morgan as the most effective structure encoding strategy.

To achieve better predictive modelling in the context of metabolomics, it is imperative to address data quality. Metabolites annotated by software like Compound Discoverer often come with varying degrees of confidence. These tools seek a match between the detected spectrum and that of molecules catalogued in databases such as ChemSpider. However, this matching process is far from flawless. The dataset of known structures is continually expanding but remains incomplete, and the software is more prone to making incorrect matches when the number of spectra available is limited. In this regard, besides enriching available datasets, effort should be devoted to evaluating the matching accuracy and filtering out molecules that are not assigned structures with sufficient confidence. The consequent improvement in data quality could be instrumental in enhancing the performance of ML models in metabolomics. An alternative strategy bypasses the conventional process of mapping molecules to known metabolites by training ML models directly on spectral data and extracting valuable features directly from these spectra. This approach offers the advantage of leveraging the entirety of available spectra, even those lacking direct matches to known structures. Furthermore, it holds the potential to aid the annotation of previously unknown metabolites if the informative features extracted from the spectra can be mapped onto specific chemical configurations.

To enhance model performance, integration of multiple data sources and domain knowledge could also be considered. While traditional metabolomics methods rely on established knowledge and only account for known metabolites, the approach undertaken in this study is purely data-driven, avoiding potential biases stemming from incomplete or inaccurate metabolic pathway annotations. A promising path forward is to explore the integration of these two approaches, leveraging established knowledge while harnessing the insights provided by data-driven techniques. Additionally, integrating diverse data sources, encompassing various omics data types, can offer a more comprehensive and detailed snapshot of the condition under study and potentially improve the predictive power of ML models.

4) Model interpretation. Feature importance plays a pivotal role in quantifying the contribution of each feature to the model prediction. This study demonstrates its application to both hashed and pattern-matching fingerprints. For hashed fingerprints, this was exemplified on Morgan, but the approach can be readily extended to all analysed fingerprints by tracking the bonds and atoms involved in setting each fingerprint bit. For pattern-matching fingerprints, the process was illustrated for the functional fingerprint, although a similar approach can be applied to any pattern-matching fingerprint.

Hashed fingerprints present a distinct advantage in representing a wide range of structures. However, not all configurations within these fingerprints contribute equally informative details. For instance, some configurations, such as those capturing a peptide bond, can precisely identify specific groups of metabolites and their associated cellular processes. In contrast, configurations like the presence of a chiral carbon atom may be too generic to provide actionable insights. Screening fingerprint bits before model training, retaining only those that can be effectively interpreted, might prove beneficial. Pattern-matching fingerprints, on the other hand, offer inherent interpretability as they are explicitly designed to identify predefined patterns but often have limited resolution. Additionally, the output of such an investigation is inevitably biased towards detecting groups of metabolites better characterised by the fingerprint. For example, the functional fingerprint was designed to include many substructures mapped in amino acids but lacks substructures tailored to nucleotides. Consequently, in both case studies, while Morgan successfully identified both nucleotide and amino acid groups as affected, the functional fingerprint only detected the latter. Overall, in the examined case studies, Morgan provided a more varied set of relevant substructures and richer insights.

Remarkably, the substructures found important for pattern-matching fingerprints correspond to chemical configurations also found important in hashed fingerprints, confirming partial alignment in these two outputs. To strike a balance between dimensionality, resolution, and interpretability, a hybrid approach might prove effective. This approach combines the utilisation of hashed fingerprints for a comprehensive exploration of substructures within the fingerprint, followed by the creation of custom pattern-matching fingerprints composed of informative and interpretable features, so to minimise feature count while maximising resolution power.

Metabolites sharing common bits in their structural representation identify groups of molecules characterised by particular substructures, potentially indicating functional relationships and insights into cellular processes. Bits most relevant for the classification process detect groups of structurally related metabolites affected in the condition under study. These groups of affected metabolites can then be presented to domain experts for in-depth examination and hypothesis formulation. For instance, the interpretation of a substructure enriched among affected metabolites, such as the peptide bond, may lead to the hypothesis of perturbations in amino acid metabolism. Subsequently, these hypotheses can be rigorously tested using established methodologies.

6. Conclusion

This research introduces an approach combining molecular fingerprinting and ML to analyse metabolic data and gain insights into affected pathways. Inspired by established practices in drug discovery, these techniques are adapted to the field of metabolomics. The study operates on the premise that structurally similar molecules, often found in close metabolic proximity, share similar metabolic responses. The investigation aims to determine if ML models trained on structural features can predict metabolic functions from structure. Feature importance is used to identify key features influencing predictions and to reveal clusters of structurally related metabolites associated with specific diseases. Key areas of focus include structural encoding, data pre-processing, model training, and model interpretation. The exploration of structural encoding assesses the advantages and limitations of various fingerprinting techniques. Pattern-matching and hashed fingerprints are both examined, with a suggestion that a hybrid approach combining their strengths may be promising. Data pre-processing addresses challenges such as high dimensionality, sparsity, and class imbalance, with a focus on feature selection. While substantial performance improvements were not achieved, further exploration is recommended for optimal pre-processing techniques. Model training results are consistent with the complexity of metabolomics studies, with room for improvement through alternative data representations. Data quality and integration of multiple data sources and domain knowledge are also recommended to enhance model performance. Model interpretation is facilitated through feature importance analysis in both hashed and pattern-matching fingerprints, providing insights into structural configurations and affected metabolite groups. This approach lays the foundation for future research, offering the potential to advance metabolomics through ML.

CRediT authorship contribution statement

Christel Sirocchi: Writing – original draft, Conceptualization, Methodology, Software. **Federica Biancucci:** Writing – original draft, Data curation, Conceptualization. **Matteo Donati:** Software. **Alessandro Bogliolo:** Supervision. **Mauro Magnani:** Supervision. **Michele Menotta:** Writing – review & editing, Data curation, Conceptualization. **Sara Montagna:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Michela Bruschi and Alessandra Fraternale for kindly providing the Hypoxia dataset used in this study. This work has been funded by the European Union - NextGenerationEU - National Recovery and Resilience Plan M4-C2-I1.5 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 (5) (2008) 714–717.
[2] G.G. Harrigan, R. Goodacre, Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis: Its Role in Biomarker Discovery and Gene Function Analysis, Springer Science & Business Media, 2003.
[3] W.J. Griffiths, Metabolomics, Metabonomics and Metabolite Profiling, Royal Society of Chemistry, 2007.
[4] L. Puchades-Carrasco, A. Pineda-Lucena, Metabolomics in pharmaceutical research and development, Curr. Opin. Biotechnol. 35 (2015) 73–77.
[5] 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.
[6] C. Sirocchi, F. Biancucci, M. Donati, N. D'Amore, R. Benedetti, A. Bogliolo, S. Ferretti, M. Magnani, M. Menotta, M. Suffian, et al., Machine learning-enabled prediction of metabolite response in genetic disorders, in: CEUR Workshop Proceedings, vol. 3578, 2023, pp. 1–9.
[7] 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 Bioinform. 13 (1) (2012) 1–15.
[8] 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 J. Rare Dis. 12 (1) (2017) 1–7.
[9] M. Bruschi, F. Biancucci, S. Masini, F. Piacente, D. Ligi, F. Bartoccini, A. Antonelli, F. Mannello, S. Bruzzone, M. Menotta, et al., The influence of redox modulation on hypoxic endothelial cell metabolic and proteomic profiles through a small thiol-based compound tuning glutathione and thioredoxin systems, BioFactors 49 (6) (2023) 1205–1222.
[10] C.H. Johnson, J. Ivanisevic, G. Siuzdak, Metabolomics: beyond biomarkers and towards mechanisms, Nat. Rev. Mol. Cell Biol. 17 (7) (2016) 451–459.
[11] M. Jacob, A.L. Lopata, M. Dasouki, A.M. Abdel Rahman, Metabolomics toward personalized medicine, Mass Spectrom. Rev. 38 (3) (2019) 221–238.
[12] D.K. Trivedi, K.A. Hollywood, R. Goodacre, Metabolomics for the masses: the future of metabolomics in a personalized world, New Horiz. Transl. Med. 3 (6) (2017) 294–305.
[13] A. Palermo, Metabolomics- and systems-biology-guided discovery of metabolite lead compounds and druggable targets, Drug Discov. Today (2022) 103460.
[14] N.R. Anwardeen, I. Diboun, Y. Mokrab, A.A. Althani, M.A. Elrayess, Statistical methods and resources for biomarker discovery using metabolomics, BMC Bioinform. 24 (1) (2023) 1–18.
[15] D.M. Drexler, M.D. Reily, P.A. Shipkova, Advances in mass spectrometry applied to pharmaceutical metabolomics, Anal. Bioanal. Chem. 399 (2011) 2645–2653.
[16] C. Guijas, J.R. Montenegro-Burke, B. Warth, M.E. Spilker, G. Siuzdak, Metabolomics activity screening for identifying metabolites that modulate phenotype, Nat. Biotechnol. 36 (4) (2018) 316–320.
[17] L. Cui, H. Lu, Y.H. Lee, Challenges and emergent solutions for lc-ms/ms based untargeted metabolomics in diseases, Mass Spectrom. Rev. 37 (6) (2018) 772–792.
[18] U.W. Liebal, A.N. Phan, M. Sudhakar, K. Raman, L.M. Blank, Machine learning applications for mass spectrometry-based metabolomics, Metabolites 10 (6) (2020) 243.
[19] A. Galal, M. Talal, A. Moustafa, Applications of machine learning in metabolomics: disease modeling and classification, Front. Genet. 13 (2022) 1017340.
[20] A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications, Drug Discov. Today 20 (3) (2015) 318–331.
[21] P. Carracedo-Reboredo, J. Liñares-Blanco, N. Rodríguez-Fernández, F. Cedrón, F.J. Novoa, A. Carballal, V. Maojo, A. Pazos, C. Fernandez-Lozano, A review on machine learning approaches and trends in drug discovery, Comput. Struct. Biotechnol. J. 19 (2021) 4538–4558.
[22] Y.-C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discov. Today 23 (8) (2018) 1538–1546.

[23] M. Staszak, K. Staszak, K. Wieszczycka, A. Bajek, K. Roszkowski, B. Tylkowski, Machine learning in drug design: use of artificial intelligence to explore the chemical structure–biological activity relationship, Wiley Interdiscip. Rev. Comput. Mol. Sci. 12 (2) (2022) e1568.
[24] D.S. Wigh, J.M. Goodman, A.A. Lapkin, A review of molecular representation in the age of machine learning, Wiley Interdiscip. Rev. Comput. Mol. Sci. 12 (5) (2022) e1603.
[25] A. Ricci, L. Galluzzi, M. Magnani, M. Menotta, Ddit4 gene expression is switched on by a new hdac4 function in ataxia telangiectasia, FASEB J. 34 (1) (2020) 1802–1818.
[26] A. Ricci, S. Orazi, F. Biancucci, M. Magnani, M. Menotta, The nucleoplasmic interactions among lamin a/c-prb-lap2α-e2f1 are modulated by dexamethasone, Sci. Rep. 11 (1) (2021) 10099.
[27] A. Ricci, F. Biancucci, G. Morganti, M. Magnani, M. Menotta, New human atm variants are able to regain atm functions in ataxia telangiectasia disease, Cell. Mol. Life Sci. 79 (12) (2022) 601.
[28] A. Ricci, F. Biancucci, G. Morganti, M. Magnani, M. Menotta, Dexamethasone induces p21cip1/waf1 expression via foxo3a independently of the lamin a/c-hdac2 interaction in ataxia telangiectasia, FEBS Open Bio. 13 (8) (2023) 1459–1468.
[29] B. Petrova, A. Warren, N.Y. Vital, A.J. Culhane, A.G. Maynard, A. Wong, N. Kanarek, Redox metabolism measurement in mammalian cells and tissues by lc-ms, Metabolites 11 (5) (2021) 313.
[30] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (5) (2010) 742–754.
[31] R. Nilakantan, N. Bauman, J.S. Dixon, R. Venkataraghavan, Topological torsion: a new molecular descriptor for sar applications. Comparison with other descriptors, J. Chem. Inf. Comput. Sci. 27 (2) (1987) 82–85.
[32] R.E. Carhart, D.H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure-activity studies: definition and applications, J. Chem. Inf. Comput. Sci. 25 (2) (1985) 64–73.
[33] C.A. James, Daylight theory manual, http://www.daylight.com/dayhtml/doc/theory/theory.toc.html, 2004.
[34] G. Landrum, Rdkit documentation, Release 1 (1–79) (2013) 4.
[35] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of mdl keys for use in drug discovery, J. Chem. Inf. Comput. Sci. 42 (6) (2002) 1273–1280.
[36] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (17) (2017) 1–5, http://jmlr.org/papers/v18/16-365.html.
[37] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[38] 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.
[39] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[40] I. Rish, et al., An empirical study of the naive Bayes classifier, in: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, 2001, pp. 41–46.
[41] R. Koenker, Quantile Regression, vol. 38, Cambridge University Press, 2005.
[42] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[43] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[44] D.P. Kingma, J. Ba Adam, A method for stochastic optimization, arXiv preprint, arXiv:1412.6980, 2014.
[45] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[46] C. Cosentino, D. Grieco, V. Costanzo, Atm activates the pentose phosphate pathway promoting anti-oxidant defence and dna repair, EMBO J. 30 (3) (2011) 546–555.
[47] H.-M. Chow, A. Cheng, X. Song, M.R. Swerdel, R.P. Hart, K. Herrup, Atm is activated by atp depletion and modulates mitochondrial function through nrf1, J. Cell Biol. 218 (3) (2019) 909–928.
[48] N. Stern, A. Hochman, N. Zemach, N. Weizman, I. Hammel, Y. Shiloh, G. Rotman, A. Barzilai, Accumulation of dna damage and reduced levels of nicotine adenine dinucleotide in the brains of atm-deficient mice, J. Biol. Chem. 277 (1) (2002) 602–608.
[49] M.J. Meredith, M.L. Dodson, Impaired glutathione biosynthesis in cultured human ataxia-telangiectasia cells, Cancer Res. 47 (17) (1987) 4576–4581.
[50] J.E. Vance, Phosphatidylserine and phosphatidylethanolamine in mammalian cells: two metabolically related aminophospholipids, J. Lipid Res. 49 (7) (2008) 1377–1387.
[51] Y.A. Valentin-Vega, K.H. MacLean, J. Tait-Mulder, S. Milasta, M. Steeves, F.C. Dorsey, J.L. Cleveland, D.R. Green, M.B. Kastan, Mitochondrial dysfunction in ataxia-telangiectasia, Blood, J. Am. Soc. Hematol. 119 (6) (2012) 1490–1500.
[52] A. Haimovitz-Friedman, R.N. Kolesnick, Z. Fuks, Ceramide signaling in apoptosis, Br. Med. Bull. 53 (3) (1997) 539–553.
[53] E.B. Cohen, R.C. Geck, A. Toker, Metabolic pathway alterations in microvascular endothelial cells in response to hypoxia, PLoS ONE 15 (7) (2020) e0232072.
[54] J. Marcinkiewicz, E. Kontny, Taurine and inflammatory diseases, Amino Acids 46 (2014) 7–20.
[55] P. Kucharzewska, J.E. Welch, K.J. Svensson, M. Belting, The polyamines regulate endothelial cell survival during hypoxic stress through pi3k/akt and mcl-1, Biochem. Biophys. Res. Commun. 380 (2) (2009) 413–418.
[56] F.-c. Yu, C.-x. Yuan, J.-y. Tong, G.-h. Zhang, F.-p. Zhou, F. Yang, Protective effect of sphingosine-1-phosphate for chronic intermittent hypoxia-induced endothelial cell injury, Biochem. Biophys. Res. Commun. 498 (4) (2018) 1016–1021.