Progress in Artificial Intelligence
https://doi.org/10.1007/s13748-024-00338-9

REGULAR PAPER

Predicting metabolic responses in genetic disorders via structural representation in machine learning

Christel Sirocchi[1] · Federica Biancucci[2] · Muhammad Suffian[1] · Matteo Donati[1] · Stefano Ferretti[1] · Alessandro Bogliolo[1] · Mauro Magnani[2] · Michele Menotta[2] · Sara Montagna[1]

Received: 29 February 2024 / Accepted: 13 August 2024

Abstract

Metabolomics has emerged as a promising discipline in pharmaceuticals and preventive healthcare. However, analysing large metabolomics datasets remains challenging due to limited and incompletely annotated biological pathways. To address this limitation, we recently proposed training machine learning classifiers on molecular fingerprints of metabolites to predict their responses under specific conditions and analysing feature importance to identify key chemical configurations, providing insights into the affected biological processes. This study extends our previous research by evaluating various metabolite structural representations, including Morgan fingerprint and its variants, graph-based structural encodings and proposing novel representations to improve resolution and interpretability of the state-of-the-art approaches. These structural encodings were evaluated on mass spectrometry metabolomic data for a cellular model of the genetic disease Ataxia Telangiectasia. The study found that machine learning classifiers trained on the new representations improved in classification accuracy and interpretability. Notably, models trained on graph-based encoding do not exhibit performance gains, not even with pre-training on a larger metabolite dataset, underlining the efficacy of our proposed representations. Finally, feature importance analysis across different encoding methods consistently identifies similar structures as relevant for classification, underscoring the robustness of our approach across diverse structural representations.

Keywords Ataxia telangiectasia · Mass spectrometry · Metabolic pathways · Metabolomics · Machine learning

1 Introduction

Metabolomics, by which the metabolites present in a particular biological system are characterised and quantified, offers insights into an organism's physiological state, aiding in disease mechanism understanding, drug target identification, and treatment assessment [1]. In this context, the advancements in mass spectrometry enable to detect and characterise a vast array of metabolites, producing large and complex datasets. Relying on these data, typically derived from normal and perturbed states, the most common approach in metabolomics employs pathway enrichment analysis to compare the 2 samples for identifying affected pathways. However, this strategy grounds on existing pathway knowledge, limiting its potential. To fully leverage metabolomics, a paradigm shift towards comprehensive metabolite analysis, beyond known pathways, is imperative, promising deeper insights into cellular metabolism and advancing preventive healthcare and pharmaceutical research.

In this paper, we leverage molecular structural encodings, machine learning (ML) models, and explainability techniques to capitalise on a broader spectrum of data acquired with mass spectrometry, extending to all metabolites with known structures. This study builds upon our recent work, where we introduced a novel approach that uses ML models trained on molecular fingerprints of metabolites to predict whether metabolite levels significantly differ between the sample under study and a control [2]. This method explores the relationships between chemical structures and

Christel Sirocchi and Federica Biancucci contributed equally to this work.

✉ Christel Sirocchi
c.sirocchi2@campus.uniurb.it

✉ Federica Biancucci
federica.biancucci@uniurb.it

[1] Department of Pure and Applied Sciences, University of Urbino, Piazza della Repubblica 13, 61029 Urbino, Italy

[2] Department of Biomolecular Sciences, University of Urbino, Via Saffi 2, 61029 Urbino, Italy

Published online: 20 September 2024

metabolic responses to specific experimental conditions in a data-driven manner, thereby opening new avenues for understanding metabolic pathways and identifying biomarkers. The proposed approach was evaluated using untargeted mass spectrometry data, comparing a cellular model of Ataxia Telangiectasia (AT)–a rare neurodegenerative disorder caused by mutations in the Ataxia Telangiectasia Mutated (ATM) gene–with a control group to identify down-regulated metabolites relevant to the disease [3].

In this study, we extend our previous research by focusing on metabolite structural representation. We further explore the efficacy of molecular fingerprinting as a structural encoding method, and compare it with graph representation, adopting the same case study for validation.

The study begins by exploring the effectiveness of the Morgan fingerprint, a well-established molecular fingerprinting technique widely utilised in drug discovery, for encoding the structural features of metabolites [4]. The suitability of this fingerprinting method for providing unique structural encodings for the detected metabolites is carefully evaluated. Resolution limitations, leading to duplicate fingerprints for distinct metabolites, are addressed by proposing extensions to the Morgan fingerprint. Contextually, the interpretability of this fingerprinting method, intended as the correspondence between fingerprint bits and chemical substructures, is also investigated, revealing that the use of the Morgan hashing algorithm diminishes interpretability by mapping multiple substructures to the same bit. To address this, an enhanced fingerprint inspired by Morgan but maintaining a 1-to-1 correspondence between bits and substructures is proposed. In addition to the structured encoding provided by the Morgan fingerprint and its variations, the study explores the representation of metabolites as graphs, which offer a more natural representation of their structure.

ML classifiers trained on the chosen structural encodings to predict down-regulated metabolites achieve satisfactory performance, providing evidence that the different structural encodings of a metabolite holds predictive value over its response to a particular condition. Feature importance analysis computed for the best-performing models identifies metabolites known to participate in affected pathways, thereby validating existing knowledge, as well as groups of metabolites not previously associated with AT, opening up novel opportunities for further investigation.

In summary, this article harnesses a variety of structural encodings and ML models for the analysis of large metabolomic datasets. By comparing the efficacy of structured and unstructured representations, the study aims to gain insights into the relative strengths and limitations of different encoding approaches in capturing the structural diversity of metabolites.

2 Background

2.1 Metabolomics in healthcare and drug discovery

Metabolomics, as the quantitative study of small molecule substrates and products of cellular metabolism, occupies a unique position in the -omics landscape due to its proximity to the phenotype [1]. The metabolome, representing the final product of genomic, transcriptomic, and proteomic processes, provides a direct readout of the physiological state of an organism [5]. Metabolomic profiling of diseased and healthy tissues can help uncover the disease mechanisms of action and identify metabolic signatures, aiding the identification of potential drug targets [6]. Additionally, metabolomics can help assess the effects of candidate treatments, evaluating the response at the metabolic level [7]. Therefore, metabolomics serves as an indispensable tool in preventive healthcare as well as pharmaceutical research and development, with the potential to enable timely disease diagnosis, early detection, effective drug testing, disease monitoring, and personalised treatment strategies [8].

The potential of metabolomics lies in characterising and quantifying the metabolites present in a particular biological system using a combination of analytical tools [8]. Central to this endeavour are the advancements in mass spectrometry technologies, including in-line chromatographic separation modes, ionisation techniques, mass analysers, and detection methods. Untargeted High-Resolution Mass Spectrometry (HR-MS) has emerged as a powerful tool, capable of simultaneously detecting a vast array of metabolites, thereby facilitating the identification of metabolic alterations and the discovery of novel metabolites [9]. However, the analysis and biological interpretation of the resulting large and complex datasets remains challenging.

In metabolomics studies, a typical approach involves comparing samples from a normal state to those from a perturbed state, often induced by genetic knockout or the administration of specific treatments [9]. Pathway enrichment analysis is the prevalent method for comparing such samples, as it identifies pathways that exhibit a higher degree of overlap with significantly under or over-expressed metabolites than would be expected by chance. This analysis aids in pinpointing the affected pathways and unravelling the underlying metabolic mechanisms. However, this approach poses various challenges, as it heavily relies on the existing knowledge of biological pathways, which is not comprehensive or fully annotated. Furthermore, it should be applied to metabolomic data with caution, as the method was primarily developed for transcriptomic data and guidelines for best practices in metabolic pathway enrichment analysis are still lacking [10]. Lastly, enrichment results were found to be quite sensitive to the pathway definitions used by different metabolomic databases [11]. Consequently, despite significant technological advancements, the full potential of metabolomics remains untapped, as metabolic data analysis often leans on the limited knowledge of known pathways. To achieve a comprehensive understanding of cellular metabolism, there is a pressing need to extend beyond the boundaries of known pathways and consider all detected metabolites in the analyses.

2.2 The role of machine learning in metabolomics

As a consequence of this overall picture, as in many other biomedical fields [12], metabolomics has turned to ML to navigate the complexity of its domain, although its use is still in its early stages and limited to certain applications [13]. In this context, the use of ML is fairly established at the level of data preprocessing, in tasks such as baseline correction, noise filtering, peak detection and alignment, data normalisation and scaling, retention time prediction, and handling of missing data [14, 15].

Most recent explorations have deployed advanced ML techniques to the identification of metabolites from their mass spectra, an area that is actively studied in metabolomics and quantum chemistry but has traditionally received limited attention from the ML community [16]. For instance, neural networks that integrate domain-specific knowledge have been employed to annotate mass spectrometry peaks with chemical structures [17]. Additionally, efficient graph neural network architectures mapping molecular graphs to probability distributions over chemical formulas have been developed for spectrum prediction [16]. Furthermore, ML models trained on molecular descriptors have been utilised to predict collision cross-section measurements, which provide supplementary structural information to mass spectrometry [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 [19]. Despite the contributions of ML to metabolomic data analysis, it is crucial to acknowledge that the biological insights derived from these studies are limited.

2.3 Metabolite structure encoding

When metabolites lack annotations of their roles in enzymatic reactions and metabolic pathways, chemical similarity emerges as a valuable tool for unravelling potential relationships with other metabolites. This approach leverages the known tendency for chemically similar compounds to be found in close proximity within metabolic pathways [20]. Chemical structures can be analysed to identify enriched chemical features within a specific experimental condition, thereby providing insights into the affected cellular processes. The structure of metabolites can be represented in two primary ways: as molecular graphs or using fingerprints. Molecular graphs depict the atoms and bonds within a molecule as nodes and edges, respectively, while fingerprints are vectors that capture the presence or absence of specific structural features within a molecule. The Morgan fingerprint is widely utilised for its ability to offer a compact, informative, and easily processed format, making it well-suited for analysis by ML models. In generating a Morgan fingerprint, an initial identifier is assigned based on atomic properties such as atom type, atomic number, charge, and hybridisation state. Subsequently, for each specified radius up to a predefined value, neighbourhood information around each atom is iteratively updated. A hash function is then applied to generate a numerical value, which is used to set bits in a fingerprint vector of a fixed size [21]. However, this representation has several drawbacks. The fixed vector length can limit its ability to capture nuanced differences between molecules of varying sizes and complexities. Additionally, the fingerprint vector tends to be sparse, with many bits remaining unset for most molecules. Furthermore, during the hashing process, information about specific atomic arrangements can be lost, and different structures may be mapped onto the same bit.

Recognising the aforementioned challenges in metabolomics, the scarcity of literature leveraging ML in this domain, and the crucial role of metabolite structures in their functionalities, we recently proposed training ML classifiers on standard molecular fingerprinting techniques to gain insights into affected cellular processes [2]. Here, we extend our preliminary investigations by comparing various structural encodings in terms of resolution and interpretability, encompassing both fingerprinting and graph representations, and introducing novel encodings that address the encountered limitations. We also investigate additional ML models and feature importance methods to identify relevant chemical substructures enriched under specific experimental conditions. To our knowledge, no study has provided such an in-depth investigation of metabolite structural encodings for ML applications, with a focus on both model accuracy and interpretability. We believe that this study of metabolite structural representation is crucial for advancing the use of ML in metabolomics.

3 Data and methods

This section presents the dataset used in the investigation and the methods applied. Section 3.2 covers sample preparation, data collection, and data preprocessing steps to format data for ML analysis (metabolite chemical structures and the corresponding metabolic response). Section 3.2 discusses structural encoding techniques explored, including established methods like Morgan and graph representations, and a novel encoding called enhanced Morgan. Section 3.3 details the experimental setup for model training, ML models used, and methods for interpreting feature importance. Figure 1 illustrates these methodological steps.

3.1 Data collection and preprocessing

The study utilised fibroblasts AT GM00648 as the cellular model for AT and AG09429 as a control. Metabolite analysis was conducted in triplicate using the UHPLC Vanquish system with an Accucore 150 amide HILIC column. LC was coupled to an Orbitrap Exploris 240 mass spectrometer equipped with an H-ESI source, operating in positive and negative modes, scanning the 80–800 m/z range. Metabolites were putatively annotated using Compound Discoverer 3.2 (Thermo Fisher Scientific) by matching detected MS2 spectra with the referential spectra of compounds in mzCloud, mzVault, and ChemSpider fragmentation libraries, resulting in the annotation of 4643 chemical structures. To enhance the precision of metabolite annotation, the mass of each detected compound was compared to the mass of the matched compound recorded in the ChemSpider database. Metabolites exhibiting a delta mass exceeding 5 ppm were excluded from the dataset. Duplicate molecules were filtered, retaining the one with the highest peaks, yielding a set of 2453 distinct metabolites. Among these, only 156 metabolites were successfully assigned a KEGG ID, enabling pathway enrichment analysis. The ratios between the measured quantities in the diseased and healthy conditions were calculated for each metabolite also using Compound Discoverer 3.2 and hypothesis testing was performed using a 2-tailed student's t-test, with corresponding p-values adjusted by the Benjamini-Hochberg algorithm. The target for the classification task is binary and indicates whether the metabolite is significantly down-regulated, i.e. its adjusted p-value is below 0.05 and the ratio of diseased to healthy is less than 1. Here, the focus on down-regulation stems from the disease's known tendency to inhibit cellular activities [3].

As we are considering only metabolites with a real-valued ratio between measured quantities in diseased and control samples, we are analysing metabolites that could be detected and quantified both in the AT cellular model and the control, excluding metabolites that could not be detected in one condition.

3.2 Structural encoding

Three different structural encoding classes – both from literature and newly devised – have been adopted for molecular representation. The goal is to define and test the most comprehensive, interpretable and unique representation for each molecular structure, facilitating the learning of structural patterns linked to affected metabolites during training, and enhancing the interpretability of predictions.

Morgan with variations Chemical structures of annotated metabolites were encoded using the Morgan molecular fingerprinting method with a radius of 2 and 1024 bits, accounting for molecule chirality. Two versions of this fingerprinting method were computed using the RDKit cheminformatics Python library [22]: *Binary Morgan*, which detects the presence or absence of substructures within the molecules and represents them as binary values, and *Count Morgan*, which records the occurrence count of each substructure. An additional structural encoding, termed *Binary Morgan with added properties*, was devised to enhance the Binary Morgan fingerprint by incorporating 3 features: Molecular Weight (MW), partition coefficient (logP), and retention time (RT).

Enhanced Morgan A modified version of the Binary and Count Morgan fingerprints, termed *Enhanced Binary Morgan* and *Enhanced Count Morgan* respectively, was developed to address the reduced interpretability associated with hashed fingerprints. The fingerprint was designed based on the following observations based on previous studies on Morgan encoding [23]:

- substructures with radius zero are present in the majority of molecules and are not effective in differentiating chemical structures;
- substructures up to radius 2 are sufficiently discriminative, and radii of 3 and above do not offer a significant increase in performance;
- most substructures are very rare, present in only a handful of metabolites, and therefore do not contribute to pattern identification.

The fingerprint generation process proceeds as follows: for each metabolite and for each atom, the substructures centred at that atom with radii 1 and 2 are selected. Substructures are used to generate one-hot feature vectors indicating the presence or absence of each substructure in the metabolite for the binary version, and the count of occurrences of each substructure in the metabolite for the count version. Features that are identical across all metabolites are collapsed into a single feature to reduce redundancy. Features that have non-zero values in at least 10 metabolites (indicating the presence of the substructure in at least 10 metabolites) are retained. The resulting fingerprint comprises 1080 features, of the same order of magnitude as the hashed Morgan fingerprint. The code to generate the proposed fingerprinting method is available on Github: https://github.com/ChristelSirocchi/unique-morgan/.

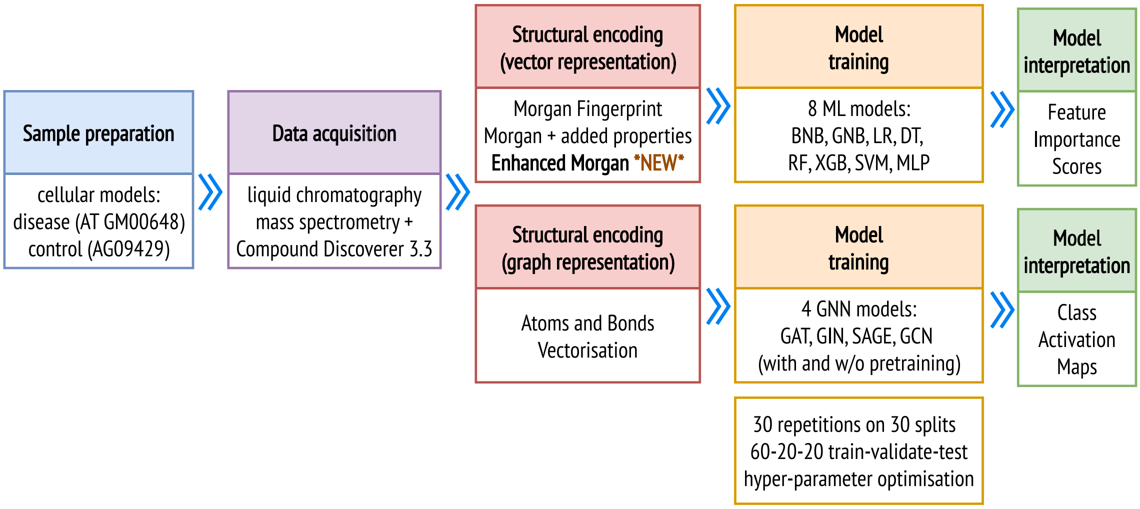

Fig. 1 Diagram summarising the research workflow

Molecular graph Chemical structures were encoded as graphs using the "from_smiles" function from PyTorch Geometric [24]. Each atom in the molecule corresponds to a node in the graph, with associated node features capturing attributes like atomic number, chirality, degree, formal charge, number of hydrogens, number of radical electrons, hybridisation, aromaticity, and ring membership. Meanwhile, bonds between atoms are translated into edges in the graph, with edge features describing properties such as bond type, stereochemistry, and conjugation.

3.3 ML models and feature importance

Morgan fingerprints and variations served as input to eight ML models: Bernoulli Naive Bayes (BNB), Gaussian Naive Bayes (GNB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), XGBoost (XGB), Support Vector Classifier (SVC), and MultiLayer Perceptron (MLP). Grid-search with cross-validation was applied for hyperparameter optimisation maximising $F1$ using an ample range of parameters commonly used in literature. Models were trained with sample weights inversely proportional to the class frequency, to enhance recall. The feature importance was determined by various feature importance computation methods depending on the ML algorithm used. Logistic Regression (LR) leveraged model coefficients to determine feature importance, while all tree-based algorithms (DT, RF, XGB) relied on impurity-based feature importances. SVC, BNB and GNB employed permutation methods, averaging importance over 5 permutations, while MLP adopted a gradient-based approach, where feature importance is calculated based on the gradients of the MLP model's output with respect to the input data.

Molecular graphs served as the input for Graph Neural Networks (GNNs), with each GNN model comprising 3 graph convolutional layers of varying types (GCN [25], SAGE [26], GAT [27], GIN [28]), each followed by a rectified linear unit (ReLU) activation function. The initial layer performed graph convolution with 16 output channels, while subsequent layers increased the output channels to 32 and 64, respectively, capturing progressively complex features. Following the final convolutional layer, global mean pooling was employed to aggregate node-level representations into a unified graph-level representation. The pooled representation was then fed into a linear layer with 64 input features and 2 output features for the binary classification task. The optimal configuration for the number of layers and layer size was determined using grid-search combined with 5-fold cross-validation. During training, all models were trained on batches of size 32 for 100 epochs.

Inspired by GNN pretraining solutions [29–31], we also pretrained a specific GNN architecture with the goal of leveraging the knowledge acquired during this initial phase to enhance the final classification capabilities of our model. The pretraining task was formulated as a self-supervised task, aiming to reconstruct the input metabolite structure, using a dataset of 217,761 metabolites from the Human Metabolome Database [32]. The GNN architecture used in this phase consists of an encoder that extracts significant features from the input to generate a latent representation, coupled with a decoder that, from such latent representation, reconstructs the original input. Employing mean-squared error as the loss function for this self-supervised pretraining task, we configured the model using grid-search combined with 5-fold cross-validation. The selected architecture considers a total of six GCN layers, where 3 of these are dedicated to the encoder and an additional 3 to the decoder. The layer sizes for the encoder are {16, 32, 64}, while for the decoder they are {32, 16, C_{out}}, where C_{out} corresponds to the number of node features. Moreover, specific residual connections [33] are incorporated between the encoder and decoder layers to facilitate training convergence. Lastly, for fine-tuning, the model utilises only the 3 encoder layers and incorporates an additional dense layer which acts as the classification head of the whole architecture.

Class Activation Maps (CAMs) were computed for each GNN model to identify atoms within each molecule that most contributed to the classification output. CAMs were computed by multiplying the feature maps from the last convolutional layer with the weights of the final linear layer. CAMs for all samples were concatenated and then normalised to ensure values are within the [0,1] range and averaged across all iterations.

Each model was trained 30 times on 30 different split of the dataset. In the case of count and added features, data was normalised to the interval [0,1]. The dataset was divided into a training set (80%) and a testing set (20%), with stratification to ensure an equal proportion of samples from each class. Model performance analysis encompasses a wide range of metrics particularly relevant to biomedical applications. Accuracy (A) provides a straightforward measure of overall correctness but can be overly optimistic with imbalanced classes, common in biomedical datasets where diseased samples are often fewer than healthy ones. Balanced Accuracy (BA) addresses this by averaging accuracies across classes. Receiver Operating Characteristic Area Under the Curve (ROC) evaluates a model's ability to distinguish between classes and is often the default metric in biomedical studies. Matthew's Correlation Coefficient (MCC), ranging from -1 to 1, is increasingly favoured in clinical settings for its ability to handle class imbalance and provide a conservative evaluation of model performance. F1-score (F1), the harmonic mean of precision and recall, balances the trade-off between correctly identifying positive samples and minimising false positives. Precision (P) focuses on minimising false positives, while recall (R) emphasises detecting all positive instances, crucial in biomedical tasks such as identifying disregulated metabolites. For F1, P, and R, the metric was computed and averaged across the 2 classes. To compare model performances, a paired t-test was employed. In each results table, a bold value indicates that the performance index of the corresponding model is statistically significantly better than the other models, as determined by the t-test with a significance level of 0.05.

4 Results and discussion

4.1 Evaluation of structural encoding

The binary Morgan fingerprint was evaluated in terms of its resolution power, i.e. the ability to provide molecules with a unique encoding. Out of 2453 metabolites, 2265 were assigned unique fingerprints, leaving 188 metabolites that shared identical fingerprints with at least one other metabolite (as seen in Table 1). Further investigation into these groups of molecules with identical fingerprints unveiled a consistent pattern: they consisted of molecules that were identical except for the length of hydrocarbon chains. This is unsurprising given that the selected fingerprint method resolves substructures of diameter 4 and cannot account for longer repetitive structures, which are rare in libraries screened for drug discovery. Figure 2a showcases a group of metabolites with varied hydrocarbon chain lengths but identical structural encodings. To enhance the resolution of the structural encoding, 2 strategies are proposed: (a) to adopt a count fingerprint rather than a binary encoding, capturing not just the presence of substructures but also their quantity, and (b) to incorporate the local information detected by the binary fingerprint with global properties, either measured during data acquisition or computed from the structure. Considering that the length of the hydrocarbon chain influences the molecule's weight, polarity, and interaction with chromatography phases, the molecular weight, partition coefficient (logP), and retention time (RT) are selected as additional features. Both approaches provide unique encoding for metabolites in the dataset, effectively discriminating molecules with varying hydrocarbon chain lengths. The unique representations provided by the proposed strategies are exemplified in Fig. 2b and c for the group of metabolites in Fig. 2a.

Morgan fingerprinting was also evaluated in terms of interpretability, intended as the correspondence between fingerprint bits and molecular substructures. Figure 3a illustrates the distribution of the number of distinct substructures associated with each bit in the binary Morgan fingerprint. On average, each bit corresponds to about 15 different substructures within the dataset, thus far from a 1-to-1 correspondence that would make the fingerprint fully interpretable. Figure 3b presents the distribution of the fraction of metabolites where the fingerprint bit sets the most frequent substructure. Typically, the most frequent substructure is observed in approximately 25% of the metabolites linked to the respective bit. However, the skewed distribution indicates that a minority of fingerprint bits sets one predominant substructure found in most or almost all metabolites. Figure 3c exemplifies these results for bit 370 where the most frequent substructure (a carbonyl group) occurs in about half of the molecules, and the next 2 most frequent substructures share with it some struc-

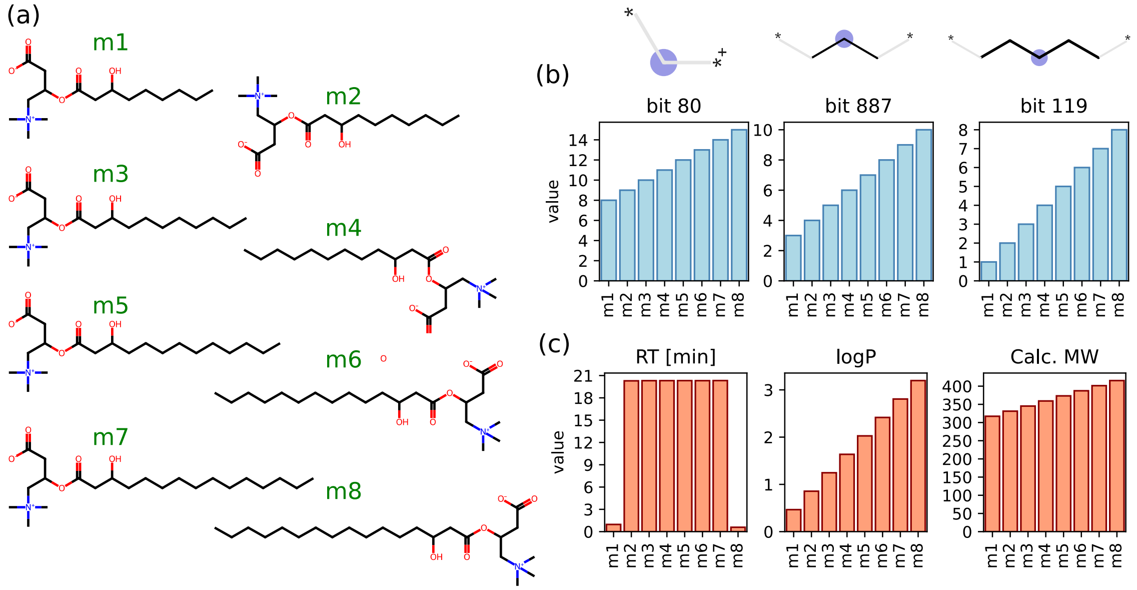

Fig. 2 **a** Group of metabolites with identical Morgan fingerprint, **b** number of repetitions of given substructures (corresponding to the bits 80, 887, and 119 of the fingerprint) and **c** global physico-chemical properties measured or computed for each molecule. Count fingerprints and added global properties can both provide a unique structural encoding for groups of molecules with varying hydrocarbon chain lengths

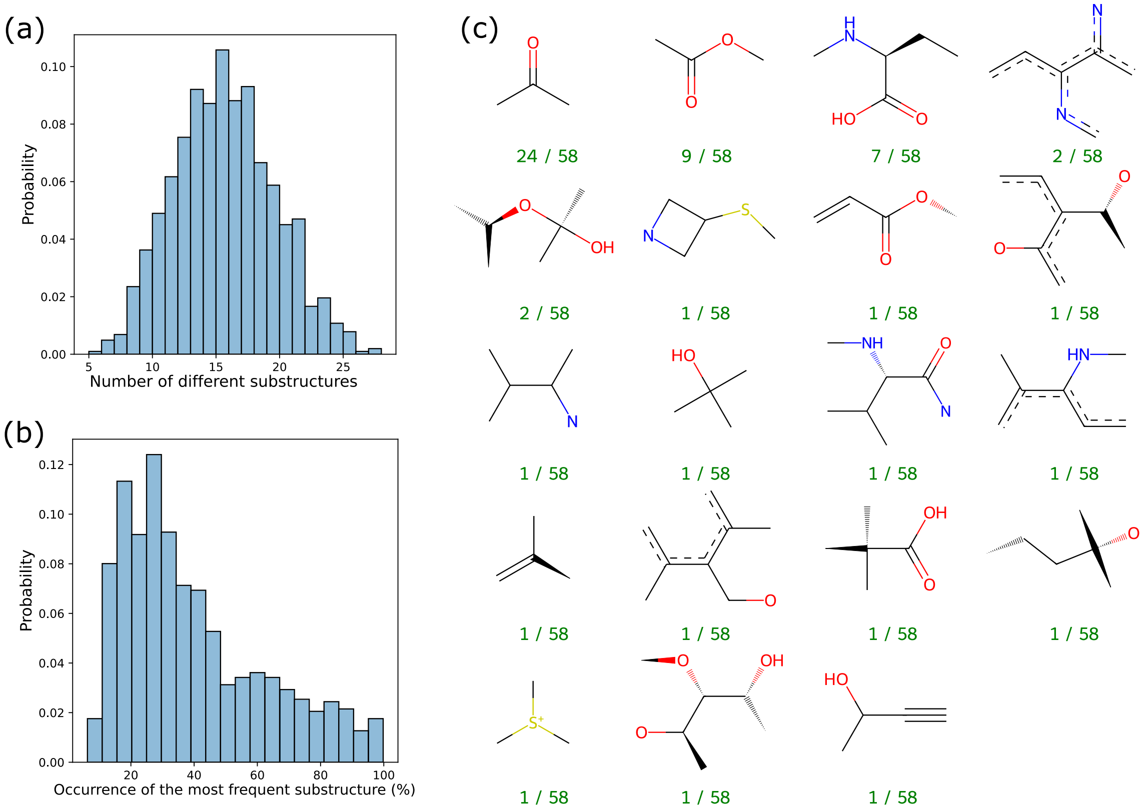

Fig. 3 **a** Distribution of the number of distinct substructures associated with each bit in the binary Morgan fingerprint. **b** Distribution of the percentage of metabolites where the bit corresponds to the most frequent substructure. **c** Substructures corresponding to bit 370, along with their respective frequencies within the dataset

Table 1 The table displays the number of unique encodings for metabolites detected by mass spectrometry within the considered AT case study, categorised by target class (down-regulated metabolites, which we aim to identify, and all other metabolites) as well as for all metabolites

Target	Metabolites	SMILE		Formula		KEGG ID		Binary Morgan	
Down-regulated	369	369	(100%)	297	(80%)	74	(20%)	362	(92%)
Other metabolites	2084	2084	(100%)	1916	(92%)	82	(4%)	1903	(98%)
Total	2453	2453	(100%)	2213	(90%)	156	(6%)	2265	(92%)

The considered standard encodings include SMILE, chemical formula, KEGG ID, and Binary Morgan

tural similarities. In contrast, infrequent substructures set by the fingerprint bit display completely different properties.

4.2 Performance of trained models

The performance metrics for ML models trained on various structural encodings are summarised in Tables 2 and 3. Among models trained on Binary Morgan fingerprints, RF and SVC stand out as the top performers, exhibiting significantly superior metrics compared to other models. For Count Morgan fingerprints, RF emerges as the best performing model, showcasing superior performances across most metrics. While the Count Morgan representation shows an improvement over the binary fingerprint, the differences are not statistically significant, suggesting that although it provides unique encoding, it does not substantially enhance the overall representation. Models trained on Morgan fingerprints augmented with additional properties demonstrate promising results, with RF and XGB emerging as the most effective. These models consistently achieve the best performance metrics across all structural encodings, indicating that incorporating supplementary chemical properties alongside the structural representation offered by Morgan fingerprints yields the most significant benefits. Among models trained on molecular graphs, GNNs utilising SAGE convolutional layers exhibit the best performance by achieving higher MCC and F1 scores. However, their performance remains below that of models trained on Morgan and its variations, suggesting that further refinement of molecular representations (in terms of node and edge features) and convolutional layers is needed. Moreover, pretraining GNNs in a self-supervised manner on a large dataset of metabolites did not yield significant performance improvements.

Given the superior performance of the Morgan representation, further work was devoted to the study of more interpretable variants. The performance of models trained on the Enhanced Binary and Count Morgan fingerprints with respect to their standard counterparts varies across models. While some models experience decreases in performance (BNB, GNB, LR, SVC) and others show improvements (DT, XGB, MLP). For RF, the top-performing model, the observed differences are not statistically significant. This suggests that the proposed representations can capture a wide range of structural features like Morgan fingerprints, with the advantage of interpretability, and can be leveraged for the unbiased analysis of metabolomics datasets.

4.3 Interpretation of trained models

To gain insights into the chemical structures contributing to the classification, feature importance analysis was first conducted on the models trained with the binary Morgan fingerprint. The analysis of the top 10 features for each model indicates that although models tend to have similar performance within a 5% difference in performance metric, they leverage different bits to inform classification. Figure 4a shows that the top 10 fingerprint bits often differ among models. However, this could be due to the fact that some fingerprints correspond to similar structures and are used by different models to leverage the same information.

A set of 14 fingerprint bits found in the top 10 of at least 2 models are illustrated in Fig. 4b and mapped onto Nicotinamide Adenine Dinucleotide (NAD) in Fig. 4c, an essential antioxidant cofactor whose role in AT as a result of impaired response to reactive oxygen species has been thoroughly characterised [3]. Upon closer examination of affected metabolites containing these bits (192, 814, 932, 786, 640 and 935), several nucleotides and nucleotide-containing compounds were identified. These findings align with previous observations of perturbations in the levels of pyridine nucleotide levels in AT cells [34]. Additionally, affected metabolites containing structures represented by other important bits (119, 794 and 591) encompass a diverse array of lipids, with a prominent presence of phospholipids. The metabolism of these lipids has been reported to be disrupted in the context of the disease [35].

The analysis of class activation maps on Graph Neural Networks (GNN) shows that different GNN architectures identify the same regions of the molecule as important for classification, as shown in Fig. 4d on NAD. Moreover, the relevant areas identified by GNN correspond to the same configurations identified by important fingerprint bits from other models, suggesting consistent patterns of importance across different structural encodings. Finally, the top-10 features of the RF model trained on the enhanced binary Morgan fingerprint, shown in Fig. 4e, identified similar structural patterns.

Table 2 The table presents the performance of ML models trained using state-of-the-art molecular structure encoding techniques

	Binary Morgan						
Metric	A	BA	ROC	MCC	F1	R	P
Bernoulli Naive Bayes	0.794	0.693	0.693	0.333	0.659	0.693	0.644
Gaussian Naive Bayes	0.778	0.690	0.690	0.321	0.647	0.690	0.637
Logistic Regression	0.761	0.697	0.697	0.318	0.641	0.697	0.628
Decision Tree	0.754	0.653	0.653	0.258	0.616	0.653	0.609
Random Forest	**0.824**	0.682	0.682	0.350	**0.673**	0.682	**0.668**
XGBoost	0.794	0.677	0.677	0.311	0.650	0.677	0.637
Support Vector	0.788	**0.707**	**0.707**	0.348	0.662	**0.707**	0.646
MultiLayer Perceptron	0.813	0.591	0.591	0.207	0.599	0.591	0.618
	Count Morgan						
Metric	A	BA	ROC	MCC	F1	R	P
Bernoulli Naive Bayes	0.794	0.693	0.693	0.333	0.659	0.693	0.644
Gaussian Naive Bayes	0.794	0.639	0.639	0.275	0.624	0.639	0.643
Logistic Regression	0.779	0.699	0.699	0.330	0.653	0.699	0.637
Decision Tree	0.755	0.640	0.640	0.238	0.608	0.640	0.601
Random Forest	**0.825**	0.686	0.686	**0.354**	**0.675**	0.686	**0.670**
XGBoost	0.804	0.689	0.689	0.337	0.663	0.689	0.650
Support Vector	0.815	0.681	0.681	0.336	0.666	0.681	0.656
MultiLayer Perceptron	0.809	0.600	0.600	0.216	0.606	0.600	0.617

Table 2 continued

	Binary Morgan						
Metric	A	BA	ROC	MCC	F1	R	P
Bernoulli Naive Bayes	0.794	0.693	0.693	0.333	0.659	0.693	0.644
Gaussian Naive Bayes	0.767	0.699	0.699	0.326	0.645	0.699	0.634
Logistic Regression	0.759	0.702	0.702	0.324	0.642	0.702	0.630
Decision Tree	0.775	0.723	0.723	0.362	0.661	0.723	0.647
Random Forest	**0.831**	0.700	0.700	0.380	0.688	0.700	**0.681**
XGBoost	0.809	**0.733**	**0.733**	**0.400**	**0.689**	**0.733**	0.672
Support Vector	0.786	0.709	0.709	0.348	0.662	0.709	0.645
MultiLayer Perceptron	0.815	0.595	0.595	0.215	0.603	0.595	0.623

	Graph						
Metric	A	BA	ROC	MCC	F1	R	P
GCN (pretrained)	0.722	0.680	0.680	0.283	0.610	0.680	0.612
GCN	0.718	0.675	0.675	0.272	0.605	0.675	0.606
GAT	0.691	0.688	0.688	0.282	0.595	0.688	0.606
GIN	0.721	0.669	0.669	0.268	0.605	0.669	0.607
SAGE	0.742	0.689	0.689	**0.301**	**0.625**	0.689	0.621

Reference metrics have been calculated for each model. A bold value in a column indicates that the metric of the corresponding model is better than all the other models for a given structural encoding, with statistical significance determined via a paired t-test at a significance level of 0,05. An underlined value indicates that the metric is significantly better across all models and structural encodings

Table 3 The table presents the performance of ML models trained on our newly defined enhanced version of the Morgan fingerprint

	Enhanced Binary Morgan							
Metric	A	BA	ROC	MCC	F1	R	P	
Bernoulli Naive Bayes	0.764	0.649	0.649	0.253	0.618	0.649	0.608	↓
Gaussian Naive Bayes	0.812	0.615	0.615	0.240	0.618	0.615	0.627	↓
Logistic Regression	0.750	0.678	0.678	0.286	0.625	0.678	0.615	↓
Decision Tree	0.775	0.668	0.668	0.289	0.635	0.668	0.624	↑
Random Forest	**0.827**	0.674	0.674	0.340	**0.669**	0.674	**0.667**	≈
XGBoost	0.793	0.690	0.690	0.329	0.657	0.690	0.642	↑
Support Vector	0.779	0.683	0.683	0.310	0.643	0.683	0.632	↓
MultiLayer Perceptron	0.796	0.607	0.607	0.215	0.607	0.607	0.609	↑
	Enhanced Count Morgan							
Metric	A	BA	ROC	MCC	F1	R	P	
Bernoulli Naive Bayes	0.764	0.649	0.649	0.253	0.618	0.649	0.608	↓
Gaussian Naive Bayes	0.820	0.578	0.578	0.189	0.584	0.578	0.617	↓
Logistic Regression	0.766	0.691	0.691	0.312	0.641	0.691	0.628	↓
Decision Tree	0.763	0.664	0.664	0.274	0.624	0.664	0.615	↑
Random Forest	0.819	0.686	0.686	0.347	**0.671**	0.686	**0.663**	≈
XGBoost	0.797	0.695	0.695	0.337	0.661	0.695	0.646	↑
Support Vector	0.793	0.680	0.680	0.315	0.651	0.680	0.638	↓
MultiLayer Perceptron	0.803	0.618	0.618	0.239	0.618	0.618	0.622	↑

Reference metrics have been calculated for each model. A bold value in a column indicates that the metric of the corresponding learning scheme is better than all the other learning schemes, on the given dataset, with statistical significance determined via a paired t-test at a significance level of 0,05. The arrow in the last column denotes whether the model's performance is significantly higher or lower compared to the same model trained on the standard Morgan fingerprint

The 1st, 3rd and 6th most important substructures identified by the model point to amino acids, dipeptides, and their derivatives. Further analysis of metabolites containing these substructures unveiled over 100 down-regulated amino acids, dipeptides, and their derivatives. While the role of the tripeptide glutathione in oxidative stress within the context of AT has been well-characterised [36], the functions of other peptides remain unknown, presenting a promising avenue for future investigations.

4.4 Opportunities for further improvement

Several avenues to further improve model performance can be explored. Alternative hashed fingerprints frequently employed in drug discovery, such as Daylight, Atop Pair, and Topological Torsion, can provide higher encoding resolution. Additionally, fine-tuning parameters such as fingerprint size and radius can positively impact model performance. Expanding the metabolic dataset by integrating measurements from different chromatographic columns, such as HILIC and C18, would effectively double the dataset size and enhance its diversity. The number of features can be reduced by eliminating fingerprint bits that are predominantly unset (mostly zeros) and/or those that exhibit a correlation coefficient with the target vector below a specified threshold, thereby reducing the dimensionality of the dataset while preserving the interpretability of features. A feature correlation study can be conducted to eliminate highly correlated features and reduce collinearity. Lastly, it should be noted that classification tasks are intrinsically more challenging for complex conditions with multiple affected pathways and a diverse range of affected metabolites like in AT. Consequently, achieving excellent classification performance in such a complex disease is challenging. Conversely, classification tasks may be relatively easier in experimental conditions where only a few metabolic pathways are affected.

5 Conclusions

This research explores different molecular structure representations and ML models for analysing large metabolomic datasets. Through the exploration of structured and unstructured representations such as molecular fingerprints and graphs, the study provides insights into the relative strengths and limitations of different encoding approaches in capturing the structural diversity of metabolites. By leveraging these representations, the study effectively predicts down-

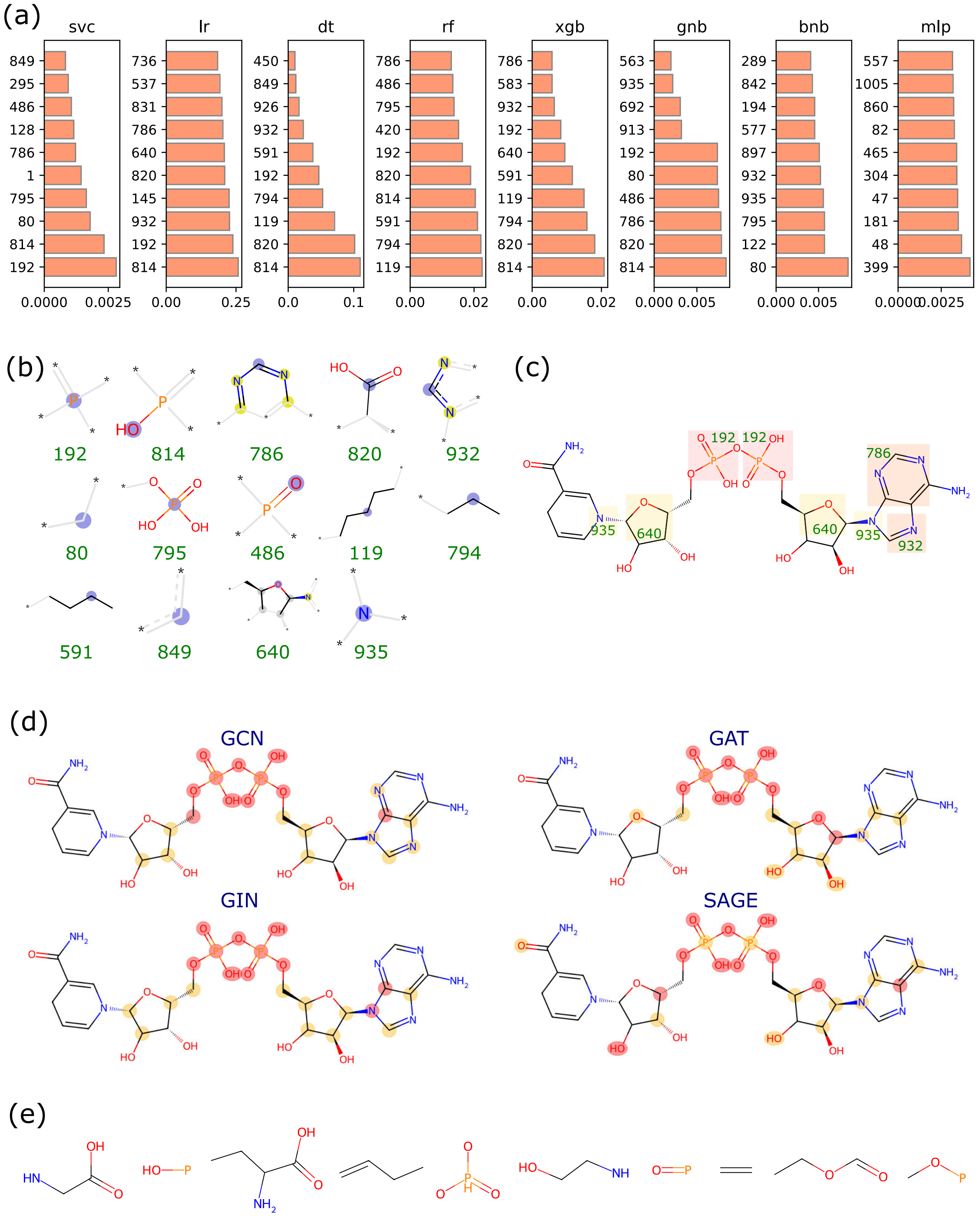

Fig. 4 **a** Top-10 features identified by 8 ML models trained on binary Morgan fingerprints. **b** Fingerprint bits corresponding to features present in the top-10 of at least 2 models and **c** mapped onto Nicotinamide Adenine Dinucleotide (NAD+). **d** Class activation maps computed for 4 GNN architectures (where red and yellow atoms have CAM values in the first and second quartile, respectively). **e** Substructures corresponding to the top-10 feature in Random Forest (RF) trained on enhanced binary Morgan fingerprints

regulated metabolites in the disease under study. Feature importance analysis provides insights into the cellular processes affected by the disease, validating existing knowledge and uncovering novel associations. This study serves as a foundation for future research exploring more sophisticated structural encodings, diverse ML models, and advanced explainability techniques. These explorations are essential for the ongoing development of metabolomics as a powerful tool for enhancing the understanding of cellular metabolism and its implications for human health.

Acknowledgements 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.

Funding Open access funding provided by Università degli Studi di Urbino Carlo Bo within the CRUI-CARE Agreement.

References

1. Holmes, E., Wilson, I.D., Nicholson, J.K.: Metabolic phenotyping in health and disease. Cell **134**(5), 714–717 (2008)
2. Sirocchi, C., Biancucci, F., Donati, M., Bogliolo, A., Magnani, M., Menotta, M., Montagna, S.: Exploring machine learning for untargeted metabolomics using molecular fingerprints. Comput. Methods Progr. Biomed. **250**, 108163 (2024)
3. Ricci, A., Biancucci, F., Morganti, G., Magnani, M., Menotta, M.: New human atm variants are able to regain atm functions in ataxia telangiectasia disease. Cell. Mol. Life Sci. **79**(12), 601 (2022)
4. Morgan, H.L.: The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J. Chem. Doc. **5**(2), 107–113 (1965)
5. Peng, B., Li, H., Peng, X.-X.: Functional metabolomics: from biomarker discovery to metabolome reprogramming. Protein Cell **6**(9), 628–637 (2015)
6. Harrigan, G.G., Goodacre, R.: Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis. Springer, Heidelberg, Germany (2003)
7. Griffiths, W.J.: Metabolomics. Metabonomics and Metabolite Profiling. Royal Society of Chemistry, London, UK (2007)
8. Puchades-Carrasco, L., Pineda-Lucena, A.: Metabolomics in pharmaceutical research and development. Curr. Opin. Biotechnol. **35**, 73–77 (2015)
9. Drexler, D.M., Reily, M.D., Shipkova, P.A.: Advances in mass spectrometry applied to pharmaceutical metabolomics. Anal. Bioanal. Chem. **399**, 2645–2653 (2011)
10. Wieder, C., Frainay, C., Poupin, N., Rodríguez-Mier, P., Vinson, F., Cooke, J., Lai, R.P., Bundy, J.G., Jourdan, F., Ebbels, T.: Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. PLoS Comput. Biol. **17**(9), 1009105 (2021)
11. Karp, P.D., Midford, P.E., Caspi, R., Khodursky, A.: Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. BMC Genom. **22**, 1–11 (2021)
12. Sirocchi, C., Bogliolo, A., Montagna, S.: Medical-informed machine learning: integrating prior knowledge into medical decision systems. BMC Med. Inform. Decis. Mak. **24**(Suppl 4), 186 (2024). https://doi.org/10.1186/s12911-024-02582-4
13. Galal, A., Talal, M., Moustafa, A.: Applications of machine learning in metabolomics: disease modeling and classification. Front. Genet. **13**, 1017340 (2022)
14. Liebal, U.W., Phan, A.N., Sudhakar, M., Raman, K., Blank, L.M.: Machine learning applications for mass spectrometry-based metabolomics. Metabolites **10**(6), 243 (2020)
15. Kontou, E.E., Walter, A., Alka, O., Pfeuffer, J., Sachsenberg, T., Mohite, O.S., Nuhamunada, M., Kohlbacher, O., Weber, T.: Umetaflow: an untargeted metabolomics workflow for high-throughput data processing and analysis. J. Cheminform. **15**(1), 52 (2023)
16. Murphy, M., Jegelka, S., Fraenkel, E., Kind, T., Healey, D., Butler, T.: Efficiently predicting high resolution mass spectra with graph neural networks. In: International Conference on Machine Learning, pp. 25549–25562 (2023). PMLR
17. Goldman, S., Wohlwend, J., Stražar, M., Haroush, G., Xavier, R.J., Coley, C.W.: Annotating metabolite mass spectra with domain-inspired chemical formula transformers. Nat. Mach. Intell. **5**(9), 965–979 (2023)
18. Asef, C.K., Rainey, M.A., Garcia, B.M., Gouveia, G.J., Shaver, A.O., Leach, F.E., III., Morse, A.M., Edison, A.S., McIntyre, L.M., Fernández, F.M.: Unknown metabolite identification using machine learning collision cross-section prediction and tandem mass spectrometry. Anal. Chem. **95**(2), 1047–1056 (2023)
19. Anwardeen, N.R., Diboun, I., Mokrab, Y., Althani, A.A., Elrayess, M.A.: Statistical methods and resources for biomarker discovery using metabolomics. BMC Bioinform. **24**(1), 1–18 (2023)
20. Barupal, D.K., Haldiya, P.K., Wohlgemuth, G., Kind, T., Kothari, S.L., Pinkerton, K.E., Fiehn, O.: Metamapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. BMC Bioinform. **13**(1), 1–15 (2012)
21. Glen, R.C., Bender, A., Arnby, C.H., Carlsson, L., Boyer, S., Smith, J.: Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. IDrugs **9**(3), 199 (2006)
22. Landrum, G.: Rdkit documentation. Release **1**(1–79), 4 (2013)
23. Sirocchi, C., Biancucci, F., Donati, M., D'Amore, N., Benedetti, R., Bogliolo, A., Ferretti, S., Magnani, M., Menotta, M., Suffian, M., Montagna, S.: Machine learning-enabled prediction of metabolite response in genetic disorders. In: Calimeri, F., Dragoni, M., Stella, F. (eds.) 2nd AIxIA Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2023). CEUR Workshop Proceedings, vol. 3578, pp. 1–9. Sun SITE Central Europe, RWTH Aachen University, Germany (2023). Workshop co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023)
24. Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric. arXiv preprint arXiv:1903.02428 (2019)
25. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

26. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst.**30** (2017)
27. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
28. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)
29. Sun, R., Dai, H., Yu, A.W.: Does gnn pretraining help molecular representation? Adv. Neural Inf. Process. Syst. **35**, 12096–12109 (2022)
30. Xia, J., Zhu, Y., Du, Y., Li, S.Z.: Pre-training graph neural networks for molecular representations: retrospect and prospect. In: ICML 2022 2nd AI for Science Workshop (2022)
31. Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., Leskovec, J.: Strategies for pre-training graph neural networks. arXiv preprint arXiv:1905.12265 (2019)
32. Wishart, D.S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B.L., et al.: Hmdb 5.0: the human metabolome database for 2022. Nucleic Acids Res. 50(D1), 622–631 (2022)
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
34. Stern, N., Hochman, A., Zemach, N., Weizman, N., Hammel, I., Shiloh, Y., Rotman, G., Barzilai, A.: Accumulation of DNA damage and reduced levels of nicotine adenine dinucleotide in the brains of atm-deficient mice. J. Biol. Chem. **277**(1), 602–608 (2002)
35. Yorek, M.A., Dunlap, J.A., Manzo-Fontes, A., Bianchi, R., Berry, G.T., Eichberg, J.: Abnormal myo-inositol and phospholipid metabolism in cultured fibroblasts from patients with ataxia telangiectasia. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids 1437(3), 287–300 (1999)
36. Degan, P., d'Ischia, M., Pallardó, F.V., Zatterale, A., Brusco, A., Calzone, R., Cavalieri, S., Kavaklı, K., Lloret, A., Manini, P., et al.: Glutathione levels in blood from ataxia telangiectasia patients suggest in vivo adaptive mechanisms to oxidative stress. Clin. Biochem. **40**(9–10), 666–670 (2007)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.