Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Multi-Omics biological embeddings for ML-models

Lennart B. Otte[a], Christer Hogstrand[b], Adil Mardinoglu[c,d] and Miao Guo[a*]

[a] King's College London, Department of Engineering, London, UK
[b] King's College London, Department of Analytical, Environmental and Forensic Sciences, London, UK
[c] Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral & Craniofacial Sciences, King's College London, London, SE1 9RT, United Kingdom
[d] Science for Life Laboratory, KTH - Royal Institute of Technology, Stockholm, Sweden

* Corresponding Author: miao.guo@kcl.ac.uk

ABSTRACT

Machine learning algorithms have led to the development of numerous vector embeddings for biological entities such as metabolites, proteins, genes, and enzymes. However, these embeddings often lack contextual information due to their specialized focus on individual omics. Disease progression and biosynthesis pathways are increasingly understood through complex, multi-layered networks that integrate diverse omics data and intricate signaling and reaction sequences. Capturing these relationships in a meaningful way requires embeddings that account for both functional and multi-modal dependencies. We propose an embedding approach that unifies these different biological modalities by treating them as directions in a shared space rather than as isolated data types. Similar to how word embeddings in natural language processing reveal meaningful relationships (e.g., Tokyo – Japan + UK = London, indicating a directional representation of capitals), we can model genes and proteins in a way that captures their inherent connections. A gene implies information about the protein it encodes, and vice versa, forming a structured and interpretable representation of biological pathways. Our model, inspired by NLP techniques, breaks down pathway sequences into contextual pairs spanning different omics types. By aligning pathway steps in proximity, the embeddings reflect biologically relevant relationships, enhancing their interpretability and utility. Because these embeddings are generated from pathway sequences, they can be applied to optimize reaction pathways, aiding retrosynthesis in microbiomes, drug development, and even human health interventions.

Keywords: Chemical fingerprints, Biological Pathways, multi-omics, Drug Discovery, Biosynthesis

INTRODUCTION

Machine Learning models can be fitted to predict a diverse set of properties or interactions between proteins, genes, small molecules etc. In every case a digital vector representation is necessary to represent physical/experimental data in a machine learning model. Specialised fingerprints exist for the different modalities (genes, proteins, small molecules) that aim to capture their own intricacies (amino acid sequence for a protein, nucleotide sequence for a gene or SMILES for a small molecule). However, in a bioprocess we use multiple omics together to promote the synthesis of targets through a biological pathway. Consequentially, predicting multi-omics interactions in a bioreactor requires a mapping between multiple spaces – the embedding space of proteins, of genes and small molecules. To increase the interpretability, efficiency and accuracy of predicting relations between multi-omics data we propose a new embedding space that encapsulates multi-omics layers together. Traversing a direction encoding omics-type allows us to find interacting compounds.

Bioprocesses

Bioprocesses underpin a vast array of engineering or medical challenges such as synthesis trough engineered *E. Coli* strains [12], degradation and valorisation of waste material and disease progression [11]. The rise of high throughput analysis of genomics, proteomics etc. has vastly increased our ability to probe complex

systems. The next step is to optimise processes further using the new insights. An early approach is Flux Balance analysis where multi-omics pathways are modelled to increase the flux along preferred reactions that produce the desired products [7]. A bottleneck to FBA is the fact that detailed pathway knowledge is required. Identifying all pathways first to then eliminated unwanted fluxes involves inference across multi-omics. The proposed work facilitates the processing of multi-omics data by placing them in the same embedding space, where proximity in space codes for interaction strengths.

Fingerprints in biological Networks

Biological systems whether we mean an organism-scale (e.g. microbiome), an organ or tissue scale, exhibit complex interaction networks [3, 4]. Different modalities like genes, proteins and small molecules interact e.g. to synthesise new proteins or start an inflammation reaction. To accurately model reaction cascades, signalling and metabolic networks we need to respect every modality (omics layer) as it can make or break the system's function [4, 6]. Drug design as well as bioreactors are increasingly relying on more complex networks for synthesis of proteins or the treatment of disease [11, 12]. In many more challenging cases long cascades and networks need to be optimised to improve a condition or achieve a higher yield. The analysis and generation of pathways using Machine Learning requires a digital vector-based representation of the compound. The approaches to obtain a descriptive vector for molecules can be grouped into different categories: 1. Chemical structure based [2] 2. Functional Group based [9] 3. Property based 4. Learned fingerprints [8]. Each of the fingerprints (MACCSKeys, Protein fingeprints etc.) have an applicability domain to a specific biological entity or an application in e.g. specifically drug like molecules. Due to the connected nature of multi-omics in biological networks, we propose a fingerprint that encapsulates multi-omics layers and can thus model an entire biological system within the same space. As the availability of established bioprocesses and pathway networks is still limited we aim to facilitate the discovery and analysis of biological systems.

METHODS

Creating a descriptive embedding model based on pathway data requires a large pool of pathway diagrams. The pathway data are downloaded from the Pathway database (PDB) MetaCyc [1] and then sampled for the use in the embedding model. The model itself is based on the Word2Vec [5] architecture which has been used to find descriptive embeddings for natural language. Fig.1 exemplifies the complex interaction network found in biological systems.

Traditional structure-based fingerprints such as Morgan's (ECFP) fingerprint [2] use an iterative graph algorithm to derive a local neighborhood embedding which is aggregated and across all neighborhoods to define the final vector embedding. Its hyperparameters can control the influence of distant atoms to the local neighborhood. A fingerprint for functional groups is MACCS Keys [9]. They use a predefined set of substructures which are matched against the given molecule. If a substructure is matched the corresponding bit in the fingerprint vector is assigned a "1". Learned fingerprints [8] can be found e.g. by using an LLM which trains over SMILES string. The internal representation which the model has learned after converging can be extracted and used in Cheminformatics tasks. Our hypothesis is that from the reaction networks themselves (Fig. 1) comprising proteins, genes and small molecules we can extract semantic properties and develop a common embedding space for all involved components.

Data Preparation

Downloading every pathway diagram contained in MetaCyc [1] yields a CSV file containing substrates, products, enzymes and genes. From this data we can infer which genes are translated into which proteins, which of these proteins are enzymes for a reaction and what substrates and products that reaction consists of. Using these relations, we can find possible subsequent interactions for a starting point i.e. interactions of the same pathway and related to a product of the previous reaction. After compiling the list of every interaction and its possible successors we sample a sequence of interactions in the pathway space. We start by picking a random reaction and then create a chain until no succeeding interaction can be found or the *sequence_length* limit is reached [Fig.1].

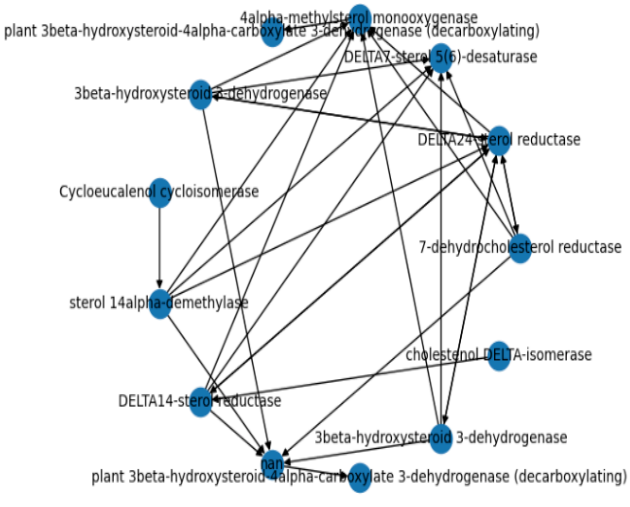

Figure 1. Processed reaction interaction network of cholesterol biosynthesis

The sampled trajectory through the space is further divided into context windows using the *context_window_size* parameter. From the sampled sequence of interactions in the pathway space we pick pairs of related nodes within the context window (regardless of which omics layer they represent). Our final training data are comprised of a list of related nodes and is supplement with a list of random unrelated nodes.

Embedding Model

A shallow autoencoder network is used to compress the initial one-hot encoded nodes into a lower dimensional space. We optimise the network by using an objective that aims to align vectors of related nodes and misalign unrelated embedding vectors. By using both positive and negative examples we maximise the usage of our available training data and add additional information for the model to learn. The model structure is illustrated in Fig.2. Different linear layers are used for the target

Figure Embedding Model

```
1:  class BioEmbeddingModel
2:      function init (vocab_size, embed_dim)
3:          target_embedding ← Linear(vocab_size, embed_dim)
4:          context_embedding ← Linear(vocab_size, embed_dim)
5:          activation function ← tanh
6:      function Forward (input, context)
7:          target_embeddings ← target_embedding(input)
8:          context_embeddings ← context_embedding(context)
9:          target_embeddings ← tanh(target_embeddings)
10:         context_embeddings ← tanh(context_embeddings)
11:         return Dots(target_embeddings, context_embeddings)
12:     function GetWordEmbedding (word)
13:         embedding ← target_embedding(word)
14:         return tanh(embedding)
```

Figure 2. Embedding Model

At the start of the training every node is assigned a one-hot encoded vector. At every iteration we create batch of positive and negative context nodes. The target word is given to the target network which compresses the vector of length *num_nodes* to the *embedding dimension* size (128). The context words are fed to the context layer which compresses the one hot encoding in the same way. Finally, we calculate the dot product between the target and context. The loss is calculated between the dot products of the context and target node embedding and the *learning_target*. The target is 1 if the node pair was a positive pair i.e. the nodes were in the same context in a pathway and 0 if the node were negative i.e. unrelated/ not-interacting. Our loss function is the mean squared error between the prediction and target. The objective aims to find embeddings that allow for a direct inference of its context in pathways.

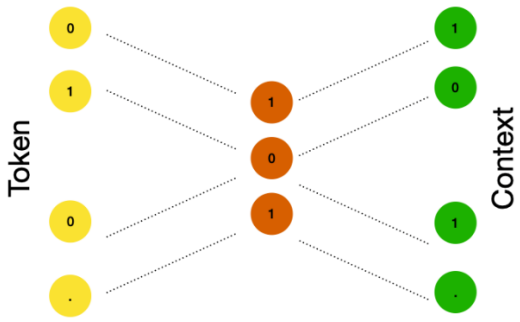

Figure 3. Embedding Model Objective Schematic

RESULTS

Training the model according to the objective listed in yields new 128-dimensional vectors for every node of a MetaCyc [1] pathway. To visualise the newly found space we first run a t-SNE dimensionality reduction algorithm. The dimensionality reduced vectors are plotted in a two-dimensional space (Fig. 4a, b). In the space we find distinct islands which according to our objective we expect to represent interacting/ related pathway nodes across all omics-layers. The validation of our hypothesis is carried out by running a clustering algorithm which then enables us to analyse the properties of our clusters/ islands Fig. 4a. The coloured section in the plot show related nodes (highly aligned embedding vectors). In contrast to Fig. 4a we visualise omics-layers in Fig.4b. by coloring genes, proteins and small molecules.

Embedding Space

We find that different pathway modalities are well distributed in the space despite genes being more abundant in the left half (Fig.4b). Therefore, in most clusters (Fig. 4a) we can identify closely related genes, proteins, enzymes and small molecules easily by searching in the neighborhood. To find the meaning of the clusters themselves we run 1. look at direct links between the most similar embeddings and 2. run a pathway enrichment analysis.

Links between aligned nodes

As an example, we take UDP-alpha-D-galactose and its related node in the embedding space which encodes phosphoglucomutase-1 (PGM1). Their proximity in the space is valid because MetaCyc [1] indeed states UDP-alpha-D-galactose as an intermediate in carbohydrate metabolism, and its relation to phosphoglucomutase-1 (PGM1) lies in its role in glucose and galactose metabolism. Their relation is defined by the following characteristics:

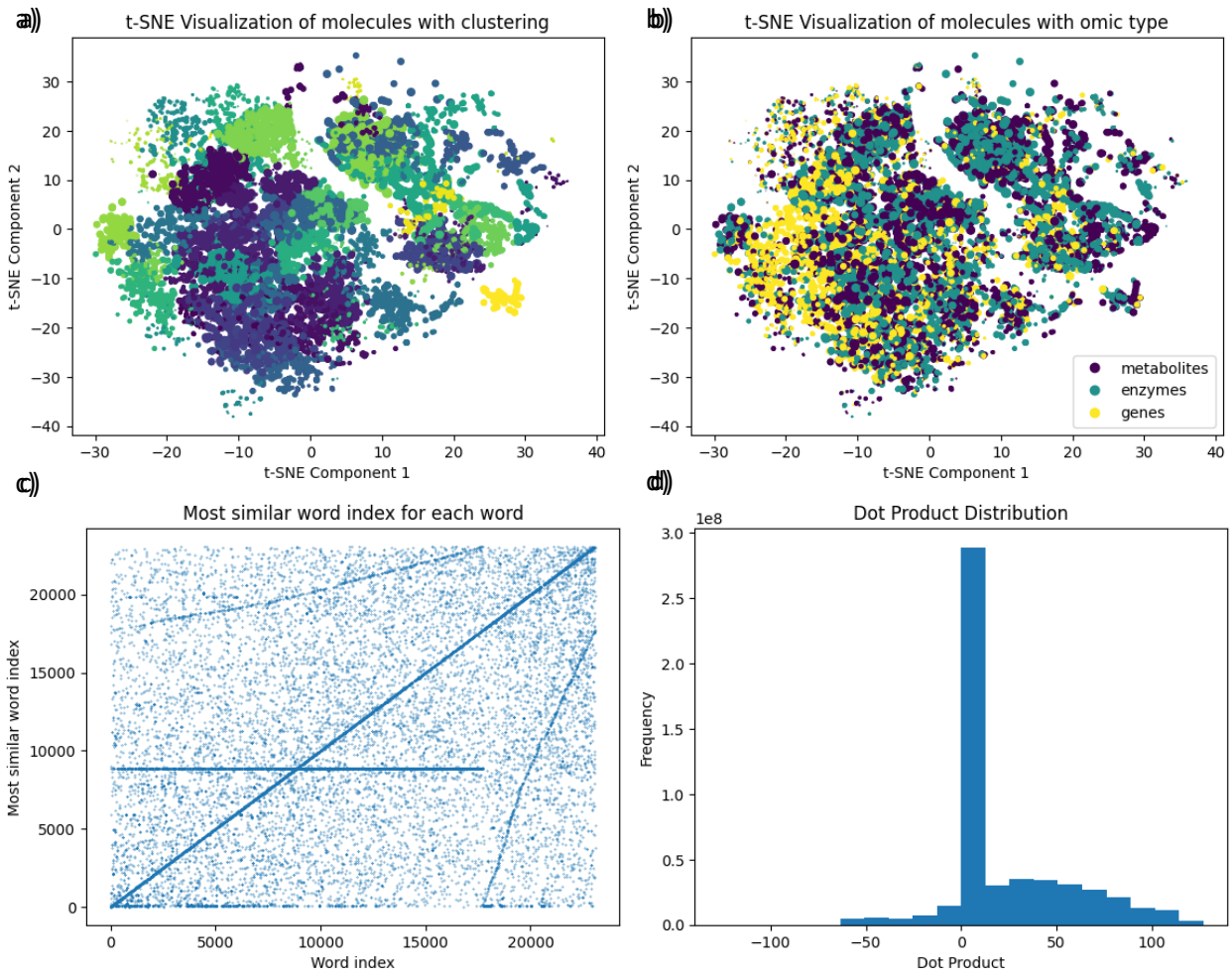

Figure 4: Resulting embedding space after running the embedding model

1. **PGM1 function**: Phosphoglucomutase-1 catalyzes the reversible conversion between glucose-1-phosphate (G1P) and glucose-6-phosphate (G6P). This step is crucial for glycogen breakdown, glycolysis, and the generation of nucleotide sugars such as UDP-glucose.
2. **UDP-galactose synthesis**: UDP-alpha-D-galactose is derived from UDP-glucose through the action of UDP-galactose-4-epimerase. UDP-glucose is synthesized from G1P by the enzyme UDP-glucose pyrophosphorylase, and PGM1 provides the necessary G1P by converting G6P.
3. **Interconnection**: PGM1 indirectly supports the biosynthesis of UDP-galactose because it ensures the availability of G1P, a precursor for the production of UDP-glucose, which is required for UDP-galactose formation. Therefore, PGM1 plays a foundational role in the metabolic pathways that produce UDP-galactose.

Similar results are found for other pairs of embeddings and thus verify that the model correctly relates different omics to aligned embeddings based on a functional connection.

In Fig. 4c we visualise the most related pathway node for every node. We achieved achieve a good coverage of the space because we limited the effect of highly common intermediates such as water, oxygen or carbon dioxide by under-sampling them. The model has therefore given an equal weighting to all related nodes. The diagonal has been set to zero. What looks like a diagonal in the picture originates from the fact that the nodes are implicitly ordered because we assign the indices as we process pathways. Hence, a similar index results in a higher likelihood of appearing first in a similar pathway. The overall distribution of dot products between all embeddings (Fig. 4d) follows the shape of a mixture of two distribution. One distribution belonging to related nodes with a smaller but wider peak and the other distribution of unrelated nodes with a sharp narrow peak.

Pathway Enrichment Analysis

Based on the pathway enrichment analysis for the clusters in Fig. 4 we found that clusters share a set of common characteristics. Cluster 0 in the example contains

biosynthesis pathways focusing on Pyrimidine Nucleotide Biosynthesis, Pyrimidine Deoxyribonucleotide Biosynthesis and related processes. Based on the functional connections revealed by the enriched pathways within clusters we have shown that closely aligned nodes interact and share similar behaviours/ functions. The interactions can be grouped into the following categories:

1. **Shared Enzymes and Genes**: dut (Deoxyuridine 5'-triphosphate nucleotidohydrolase): Present in multiple pyrimidine-related pathways (e.g. pyrimidine biosynthesis and degradation)
2. **Key Reactions:** Conversion of nucleotides or nucleotide precursors like UMP, dUMP, and dTTP. Hydrolysis of diphosphates, ensuring the regulation of nucleotide intermediates.
3. **Cofactors and Byproducts:** Common usage of ADP, diphosphate, and AMP. Generation or utilization of cofactors like NADP+ and CO2 in related biosynthetic processes.
4. **Organisms and Applications:** The pathways include contributions from diverse organisms) highlighting conserved metabolic roles in nucleotide synthesis and homeostasis.
5. **Interconnection with Metabolic Networks:** These processes are interconnected with biosynthetic pathways like baumannoferrin biosynthesis and teichoic acid biosynthesis, emphasizing the role of nucleotides as substrates for broader cellular functions (e.g. synthesis of wall components or coenzymes).

DISCUSSION

We have shown that clustering/ proximity in the embedding implies an interaction between the nodes (See Embedding Space Analysis). Unlike specific protein sequence, SMILES or gene embeddings we can directly link instances of genes, proteins and small molecules to learn about their relationship by computing the alignment of their embedding vectors/ proximity in the embedding space (Fig. 4b).

The amount of meaningful interactions an embedding vector captures depends on the data availability. Annotated pathways are limited, especially when we move away from model organisms or well understood disease metabolic networks. If we only witness a certain function of a protein in one domain and see no other examples of its behaviour in other contexts then it is unlikely that its embedding can generalise well to other domains. Additionally, Networks from MetaCyc [1] are biased in the sense that they have undergone a step of curation and are not direct results of experiments but rather a product of experimental studies and its subsequent analysis. Through this process we also lose some information. The networks found in MetCyc [1] are binary and no interaction strength or confidence of interactions can be inferred. We lose information on lower confidence interactions that still add nuanced detail to the characterisation of the nodes. Additionally, some interactions that were perhaps not relevant in pathway can be omitted making it more difficult to assess unknown or new pathways.

High throughput experimental results

By providing additional experimental data from high throughput differential expression, we can reduce bias and increase the amount and diversity of training data. Currently, the focus lies on annotated and curated pathways from PDBs. The interactions contained in these databases are reliable reactions that support the higher order function of a pathway. However, these pathways networks have limited availability especially outside model organisms. Through additional experimental methods we can establish more linkages between metabolites, genes and proteins [10]. The interpretation of differential expression data is different however because correlations between two analytes do not necessitate a linkage by chemical reactions as it is the case in PDBs. In the context of this work the different interpretation is not an issue because we only consider the context of two analytes in a training objective. If two analytes are differentially expressed it means that a specific context change has had an effect on the analytes and there must be a pathway that co-affects them.

Unknown Nodes

To find fingerprints for nodes not found in MetaCyc [1] we can provide the model with the node context. MetaCyc [1] is a large database of metabolic networks, and it contains many proteins, genes and small molecules. However, given the vast size of the chemical/ biological universe there are still a significant number of nodes which are not present. In the case that related nodes within the training set are known we can still find an embedding for them by feeding in all the related nodes to the encoder network by summing their one-hot encodings. A more generalizable approach to finding embeddings of unknown compounds would be to define a mapping from atom/ sequence space to the embedding space. Especially a bijective mapping can enable the embedding to be extended to generative applications. If we assume a generator model creates molecules in the embedding space with given properties, then these are unlikely to have been part of the training data. Thus, we require a mapping back from the embedding space to SMILES or sequences. To implement the additional mapping, we can create 3 new networks for each omics layer that can learn to map into the embedding space.

FUTURE WORK

To address the issues of the limited network availability and bias as well as the loss of data due to binary edges we can supplement curated biological pathways with general high throughput experiments of protein-protein interactions, gene-protein interactions and metabolite interactions. Using data directly from experiments removes the need of curation from a Pathway Databases which reduces bias. Additionally, we can reduce our reliance on PDBs to contain data for the domain we are interested in. Certain disease networks might be very well represented in a database but venturing into a lesser-known disease requires data not yet found in Pathway databases. Apart from chemical reactions in metabolic networks gene regulation with transcription factors and cell signalling can be included to widen the applicability domain of the model as well as producing more nuanced embeddings.

Structural and functional data

An additional structural, functional or compositional component in the embedding vector adds information to the training data that the embedding model can learn. The sequence (nucleotides, amino acids) provides information on the presence of functional groups, can show how related two genes are or give information on the 3D structure. To an extent these datapoint are implicitly present in the pathway network. However, Pathway Databases are never complete. They represent a subset of highly curated reaction networks, and we lose information by only teaching the model to reproduce these relations instead of also identifying new relationships purely on functional properties.

Application to Pathway Discovery

The proposed fingerprints offer direct relationships between multi-omics data. Using the fingerprints to identify missing edges/ interactions in metabolic networks is a fitting application to test their predictive strength. Going even further we can use the fingerprints as the basis for generative models to create new pathway networks. As the interactions between nodes lie at the core of this fingerprint, we expect generative models to benefit from this single space embedding that capture multiple omics layers at once.

ACKNOWLEDGEMENTS

This work was kindly supported by the NMES faculty scholarship from King's College London.

REFERENCES

1. Ron Caspi, Tomer Altman, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic acids research, 42(D1):D459–D471, 2014
2. Manish Kumar. A beginner's guide for understanding extended-connectivity fingerprints(ecfps), Mar 2021.
3. Xinmeng Liao, Mehmet Ozcan, et al. Open moa: revealing the mechanism of action (moa) based on network topology and hierarchy. Bioinformatics, 39(11):btad666, 2023.
4. Javier Lopez-Ibáñez, Florencio Pazos, and Monica Chagoyen. Predicting biological pathways of chemical compounds with a profile-inspired approach. BMC bioinformatics, 22(1):320, 2021.
5. Tomas Mikolov, Ilya Sutskever, Kai Chen, et al. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26, 2013.
6. Kumar Saurabh Singh, Justin JJ van der Hooft, et al. Integrative omics approaches for biosynthetic pathway discovery in plants. Natural Product Reports, 39(9):1876–1896, 2022.
7. Raman, K. and Chandra, N., 2009. Flux balance analysis of biological systems: applications and challenges. Briefings in bioinformatics, 10(4), pp.435-449.
8. Honda S, Shi S, Ueda HR. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. arXiv preprint arXiv:1911.04738. 2019.
9. Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. Journal of Chemical Information and Computer Sciences, 42(6), 1273-1280.
10. Subramanian I, Verma S, et al.. Multi-omics data integration, interpretation, and its application. Bioinformatics and biology insights. 2020 Jan;14:1177932219899051.
11. Muller, E., Shiryan, I. and Borenstein, E., 2024. Multi-omic integration of microbiome data for identifying disease-associated modules. Nature Communications, 15(1), p.2621.
12. Hussain, M.H., Mohsin, et al., 2022. Multiscale engineering of microbial cell factories: A step forward towards sustainable natural products industry. Synthetic and systems biotechnology, 7(1), pp.586-601.