Simplicity within biological complexity

Bioinformatics Advances, 2025, **00**, vbae164
https://doi.org/10.1093/bioadv/vbae164
Advance Access Publication Date: 6 February 2025
Perspective

Systems biology

Simplicity within biological complexity

Nataša Pržulj [1,2,3,4,*] and Noël Malod-Dognin [2]

[1]Computational Biology Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, 00000, United Arab Emirates
[2]Barcelona Supercomputing Center, Barcelona 08034, Spain
[3]Department of Computer Science, University College London, London WC1E6BT, United Kingdom
[4]ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain

*Corresponding author. Computational Biology Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, 00000, United Arab Emirates. E-mail: natasa.przulj@mbzuai.ac.ae

Associate Editor: Sofia Forslund

Abstract

Motivation: Heterogeneous, interconnected, systems-level, molecular (multi-omic) data have become increasingly available and key in precision medicine. We need to utilize them to better stratify patients into risk groups, discover new biomarkers and targets, repurpose known and discover new drugs to personalize medical treatment. Existing methodologies are limited and a paradigm shift is needed to achieve quantitative and qualitative breakthroughs.

Results: In this perspective paper, we survey the literature and argue for the development of a comprehensive, general framework for embedding of multi-scale molecular network data that would enable their explainable exploitation in precision medicine in linear time. Network embedding methods (also called graph representation learning) map nodes to points in low-dimensional space, so that proximity in the learned space reflects the network's topology–function relationships. They have recently achieved unprecedented performance on hard problems of utilizing few omic data in various biomedical applications. However, research thus far has been limited to special variants of the problems and data, with the performance depending on the underlying topology–function network biology hypotheses, the biomedical applications, and evaluation metrics. The availability of multi-omic data, modern graph embedding paradigms and compute power call for a creation and training of efficient, explainable and controllable models, having no potentially dangerous, unexpected behaviour, that make a qualitative breakthrough. We propose to develop a general, comprehensive embedding framework for multi-omic network data, from models to efficient and scalable software implementation, and to apply it to biomedical informatics, focusing on precision medicine and personalized drug discovery. It will lead to a paradigm shift in the computational and biomedical understanding of data and diseases that will open up ways to solve some of the major bottlenecks in precision medicine and other domains.

1 Motivation

This century is the century of biology (Glover 2012, Venter and Cohen 2014), characterized by biotechnological advances that continue to amass various molecular and clinical data [Lappalainen *et al.* 2013, Papatheodorou *et al.* 2020, Kearnes *et al.* 2021, Oughtred *et al.* 2021, Uffelmann *et al.* 2021, Moreno *et al.* 2022, Pugh *et al.* 2022, Aleksander *et al.* 2023, Baron *et al.* 2024, Clough *et al.* 2024, The UK BIOBANK (https://www.ukbiobank.ac.uk/), PISTACHIO Database (https://www.nextmovesoftware.com/pistachio.html), REAXIS Database (https://www.elsevier.com/products/reaxys), The Cancer Genome Atlas Program (TCGA) (https://www.cancer.gov/ccg/research/genome-sequencing/tcga). Algorithmic and computational developments follow this data revolution striving to model and compute the amassed data to uncover new biomedical knowledge (Pržulj 2019). Versatile computational methods, also of ever-growing complexity, are being proposed to capture these ever-growing data and are incurring enormous computational and energy costs. Despite these substantial scientific labour and energy costs, the fundamental biomedical breakthroughs still elude us. Hence, we question whether omics data science is currently lost in its complexity and argue that it is the *simplicity* that is hidden within the complexity of the data that we should endeavour to uncover. Recall that it is *simplicity* (and not complexity) that has as an epistemological value long been taken as self-evident in science, a well-known philosophical principle illustrating it being Occam's razor (Chaitin 2004, Lichacz 2021).

However, the currently mainstream, black-box artificial intelligence/machine learning (AI/ML) models are not only excessively complex and incurring substantial energy costs but also have unexpected behaviours, leading not only to some significant dangers but also to surprising scientific discoveries, which we cannot control or account for (Beerens and Higham 2015, Bastounis *et al.* 2023). Hence, we should not rely on such models to draw conclusions from human health-related data, but seek to uncover simple and fundamental principles underlying the molecular organization of life that determine the higher order behaviour (Dhar and Giuliani 2010, Eftimie 2022, Glazier 2022) and that we can easily and cheaply exploit. Currently, there are only hints of such organizing principles and further developments are needed towards pioneering paradigm shifts to improve our understanding of the fundamental principles of biology, ultimately leading to the theory of molecular organization of life. We survey the state-of-the-art and propose modelling paradigms to facilitate the understanding of the simple underlying principles of multi-omics data organization leading to their controlled and sustainable exploitation and much-needed scientific breakthroughs.

Received: May 15, 2024; **Revised:** October 1, 2024; **Editorial Decision:** October 12, 2024; **Accepted:** October 23, 2024

2 Current state-of-the-art

2.1 Multi-omics data

Modern biotechnologies have uncovered a complex system of heterogeneous interacting molecular entities, including genes, proteins, and metabolites; they interact within the cells, bodies, and with the external environment to maintain life and function. For example, next-generation sequencing techniques [Next-Generation Sequencing Articles from across *Nature* Portfolio (https://www.nature.com/subjects/next-generation-sequencing)] have unlocked genome sequencing at mass scale and made it affordable. An integrated encyclopaedia of DNA elements in the human genome (ENCODE) (Hitz *et al.* 2023, The ENCODE Project (https://www.encodeproject.org/), bioRxiv ENCODE Channel (https://connect.biorxiv.org/relate/content/177) has systematically mapped DNA regions of transcription, transcription-factor association, chromatin structure, and histone modification, assigning biochemical functions for 80% of the genome outside of the well-studied protein-coding regions (genes) (Hitz *et al.* 2023). High-throughput experimental methods, such as Yeast two-hybrid assays (Young 1998, Ito *et al.* 2000, Uetz *et al.* 2000, Velásquez-Zapata *et al.* 2021) and affinity purification with mass spectrometry (Gavin *et al.* 2006, Krogan *et al.* 2006, Low *et al.* 2021), are widely used for identifying physical interactions (bindings) between proteins, protein–protein interactions (PPIs). Notably, these methods have been used to create proteome-scale interaction maps for humans and other species (Luck *et al.* 2020, Skinnider *et al.* 2021, Michaelis *et al.* 2023, Tang *et al.* 2023). Synthetic genetic array analysis and CRISPR functional genomics tools (Adames *et al.* 2019, Przybyla and Gilbert 2022) automate the combinatorial construction of defined mutants and enable the quantitative analysis of genetic interactions (GIs) that exist between pairs of genes which when mutated together produce a phenotype that is not expected from the phenotypes produced by mutations of each of the two genes individually (Baryshnikova *et al.* 2010, Costanzo *et al.* 2010, 2016, Zhao and Boeke 2020). Other experimental technologies, such as microarrays (Quackenbush 2001, Dahlquist *et al.* 2002, Zlobec *et al.* 2014), RNA-sequencing technologies (Hirschhorn and Daly 2005, Duerr *et al.* 2006, Marioni *et al.* 2008, Mortazavi *et al.* 2008, Wang *et al.* 2009, Macosko *et al.* 2015, Li and Wang 2021), various single-cell omics (Klein *et al.* 2015, Macosko *et al.* 2015, Stegle *et al.* 2015, Linnarsson and Teichmann 2016, Li and Wang 2021, Zhang and Qiao 2024), etc., have enabled construction of other omics layers in a cell, e.g. the genome, epigenome (Hammal *et al.* 2022, Hitz *et al.* 2023, The ENCODE Project, bioRxiv ENCODE Channel), transcriptome (Quackenbush 2001, Dahlquist *et al.* 2002, Hirschhorn and Daly 2005, Duerr *et al.* 2006, Marioni *et al.* 2008, Mortazavi *et al.* 2008, Wang *et al.* 2009, Macosko *et al.* 2015), proteome (Ito *et al.* 2000, Uetz *et al.* 2000, Giot *et al.* 2003, Li *et al.* 2004, Stelzl *et al.* 2005, Gavin *et al.* 2006, Krogan *et al.* 2006, Simonis *et al.* 2009, Arabidopsis Interactome Mapping Consortium 2011, Luck *et al.* 2020, Skinnider *et al.* 2021, Michaelis *et al.* 2023, Tang *et al.* 2023), metabolome (The Kyoto Encyclopedia of Genes and Genomes (KEGG), exposome (data on nutrition, toxic molecules, and radiation) (Kato *et al.* 2011, Kumar *et al.* 2012, Lambin *et al.* 2012, Niedzwiecki *et al.* 2019), microbiome [The National Institutes of Health (NIH) Microbiome Project (https://www.hmpdacc.org/)], allergome [The ALLERGOME Database (https://www.allergome.org/)], foodome [The FOODOME Project (https://www.barabasilab.com/science/project/foodome)], data on all drugs, their chemical similarities and bindings to protein targets (drug–target interactions, DTIs) [The DRUGBANK Database (https://go.drugbank.com/), The CHEMBL Database (https://www.ebi.ac.uk/chembl/), The PUBCHEM Database (https://pubchem.ncbi.nlm.nih.gov/], chemical reactions (Kearnes *et al.* 2021, PISTACHIO Database (https://www.nextmovesoftware.com/pistachio.html, REAXIS Database (https://www.elsevier.com/products/reaxys), pesticides [The CHEMBL Database (https://www.ebi.ac.uk/chembl/, National Pesticide Information Center, Databases for Chemical Information (http://npic.orst.edu/ingred/cheminfo.html, Pesticide, Bio-Pesticide and Veterinary Substances Properties Databases. http://sitem.herts.ac.uk/aeru/ppdb/], herbicides [The CHEMBL Database (https://www.ebi.ac.uk/chembl/], protein sequence (Wang *et al.* 2018b, UniProt Consortium 2022), structure (Burley *et al.* 2023) and engineering data (Wang *et al.* 2018a), functional annotations of genes (Gene Ontology, GO) (Aleksander *et al.* 2023) and diseases (Disease Ontology) (Baron *et al.* 2024), genome-wide association studies (Uffelmann *et al.* 2021), mutational data from The Cancer Genome Atlas (TCGA) that also contains epigenomic, transcriptomic, proteomic, and clinical data for 32 cancers (a landmark dataset for multi-omics methods development) [The Cancer Genome Atlas Program (TCGA) (https://www.cancer.gov/ccg/research/genome-sequencing/tcga], Pan-Cancer Atlas and AACR Project GENIE (a largest public, clinico-genomic cancer dataset) (Pugh *et al.* 2022), including genomic structural variation data (Lappalainen *et al.* 2013), versatile phenomic data from electronic health records [The UK BIOBANK (https://www.ukbiobank.ac.uk/), etc. These data are multi-scale, multi-omics layers describing cells, and health status (illustrated in Fig. 1), each of which is a result of a type of biotechnological measurement that produced it, each with its technological limitations and biases, and therefore measuring different aspects of the functioning of our cells, tissues, and bodies (Pržulj 2019). Substantial scientific efforts have been focusing on how to analyse these data to learn more about biology and curing diseases.

2.2 Biological network analytics

Networks (graphs) are universal descriptors of systems of interacting elements in biomedicine. Networks contain nodes, representing biological entities (e.g. genes, diseases, drugs), and edges, representing pairwise relationships between the entities (e.g. physical, functional, chemical) (Pržulj 2019). Widely used molecular networks include those capturing PPIs (Oughtred *et al.* 2021), GIs (Oughtred *et al.* 2021), co-expression of genes (COEX) in tissues (Papatheodorou *et al.* 2020, Moreno *et al.* 2022, Clough *et al.* 2024), or in single cells (Papatheodorou *et al.* 2020, Moreno *et al.* 2022, Clough *et al.* 2024), in health and disease. Analysing the wiring (also called topology, or structure) of networks by graph theoretic approaches to extract biological information contained in the wiring, mostly from a single type of a molecular network data, has been a topic of many studies (Pržulj 2019). More recently, significant efforts on developing data fusion algorithms for collectively mining these interconnected multi-omic data have been made (Gligorijević *et al.* 2016a, Malod-Dognin *et al.* 2019, Pržulj 2019, Zambrana *et al.* 2021, Xenos *et al.* 2023). However, there is an increasing realization that ever-growing multi-omic data, followed by models

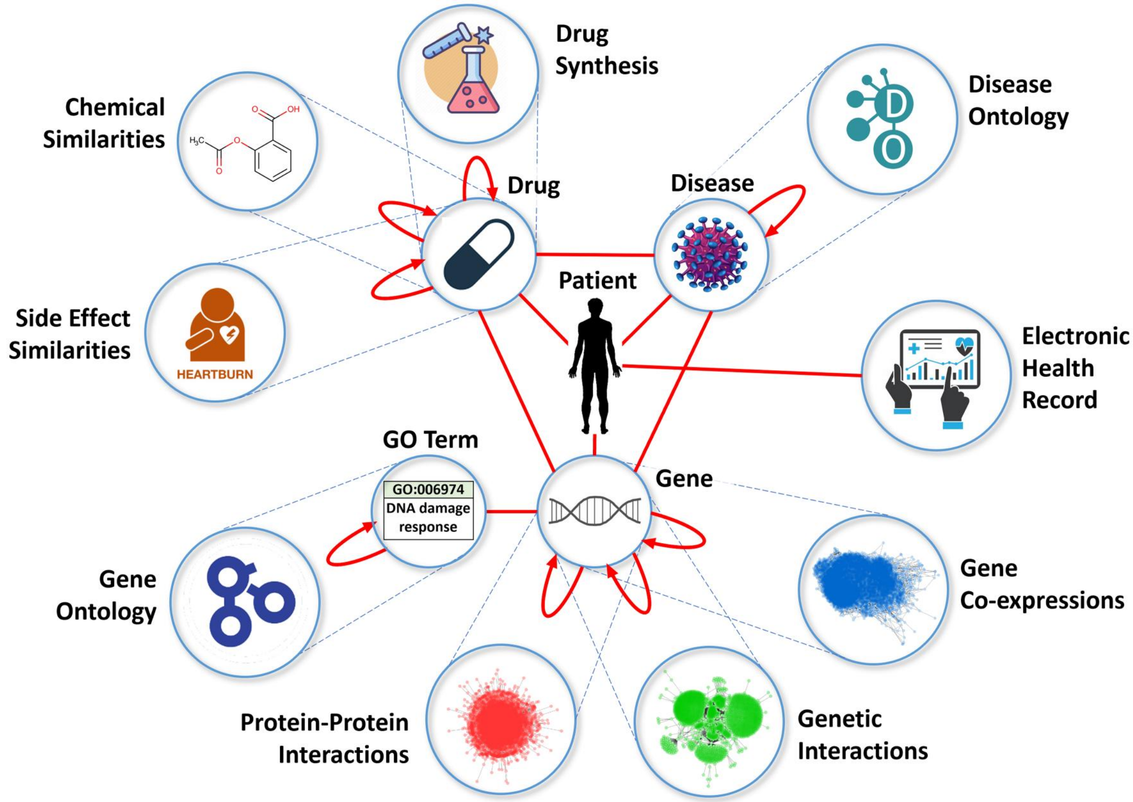


Figure 1. Illustration of the connectedness of biomedical network data.

and algorithms of ever-growing complexity are hitting the limits due to computational intractability (e.g. NP-hardness) and over-parametrization (e.g. see Section 2), having large carbon footprints. This is calling for alternative conceptual, modelling, and computational approaches, with those based on embedding of the data points into lower-dimensional spaces, that would allow for their easier computational analyses, being expected to have ground-breaking impacts in transforming the study of complex biomedical data (Nelson *et al.* 2019, Li *et al.* 2022).

In particular, graph theoretic methods applied to these data have fuelled biomedical discoveries, from uncovering biomarkers and relationships between diseases (Menche *et al.* 2015, Le and Dang 2016, Guo *et al.* 2019, Sumathipala *et al.* 2019) to repurposing of drugs (Cheng *et al.* 2019a, 2019b, Morselli Gysi *et al.* 2021), with algorithmic innovations including graphlets (higher order graph substructures) (Pržulj *et al.* 2004, Pržulj 2007, Milenković and Pržulj 2008, Windels *et al.* 2019, 2022), applications of random walks (Yang *et al.* 2019, Chen *et al.* 2020, Wong *et al.* 2020), kernels (Geng *et al.* 2020), and network propagation (Veselkov *et al.* 2019), having been used for capturing structural information from networks. Extracting these predetermined features from a network and feeding the feature vectors into AI/ML models is a common approach, despite handcrafting optimally predictive features across diverse types of omics networks and biomedical applications being challenging (Zhang *et al.* 2022). Hence, network embeddings (Nelson *et al.* 2019), also called graph representation learning (Li *et al.* 2022), have emerged as leading, cornerstone approaches, viewed as a bridge between network topology and classical ML, since ML is mostly able to process items in vector space. However, their development is challenging, because biomedical networks are very heterogeneous (Fig. 1), noisy, incomplete, and complex, have no node ordering or reference points, and usually contain biomedical text and other domain knowledge, which makes embedding tasks more complicated than in other application domains. For example, classic deep-learning (DL) neural network (NN) methods cannot handle such diverse structural properties and complex interactions, because they are designed for fixed-size grids (such as matrices of pixels in images and tabular datasets), or optimized for text and sequences. Hence, there is an increasing need for a deeper combination of network topology and ML (Su *et al.* 2020) into new modelling paradigms for biomedical systems that would parallel the successes of DL, which when applied to images and sequences have revolutionized image analysis and natural language processing (NLP).

2.3 Network embedding in computational biology

Network embedding methods represent the network nodes (or larger graph substructures) as points in low-dimensional space, leveraging the topology of the network, so that the points representing nearby or similarly wired nodes are put close in the space. Then, the vector representations of these points in space (corresponding to network nodes), a.k.a. embeddings, or embedding vectors, are used as input in traditional ML methods for downstream analyses, to link the network structure to molecular phenotypes, biological functions, or disease states. There is a wide range of network embedding methods in computational biomedicine, too many to describe here, that could roughly be classified by:

2.3.1 The type of algorithms for generating network embeddings

This is a large research field, which, despite significant efforts, is still lacking due to the computational intractability (e.g. NP-hardness) of the underlying problems on graphs. There are many diverse algorithms that can be summarized into seven major categories: (i) graph theoretic based (Pržulj 2019) (e.g. node degree, betweenness, graphlet counts, clustering); (ii) random walks, or network diffusion based (Yang *et al.* 2019, Chen *et al.* 2020, Wong *et al.* 2020) [e.g. Diffusion State Distance (Cao *et al.* 2013), GraphWave (Donnat *et al.* 2018)]; (iii) topological data analysis (TDA) based [e.g. persistent homology (Edelsbrunner and Harer 2008, 2022, Bubenik 2015, Hofer *et al.* 2017, 2019, 2020, Carrière *et al.* 2020) and Mapper (Singh *et al.* 2007, Nicolau *et al.* 2011, De Cecco *et al.* 2015, Madhobi *et al.* 2019, Wang *et al.* 2019b, Bodnar *et al.* 2021)]; (iv) manifold learning, or nonlinear dimensionality reduction [e.g. Isomap (Tenenbaum *et al.* 2000), Laplacian eigenmap (Belkin and Niyogi 2001) and t-SNE (Van der Maaten and Hinton 2008)]; (v) shallow NN embeddings [e.g. unsupervised Skip-gram (Mikolov *et al.* 2013b), semisupervised node2vec (Grover and Leskovec 2016)]; (vi) supervised graph neural networks (GNNs) (e.g. see Gilmer *et al.* 2017, Han *et al.* 2019, Ingraham *et al.* 2019, Xie *et al.* 2019); and (vii) generative NN models (e.g. variational graph autoencoder; Kipf and Welling 2016). Another rough categorization, often used in ML and Knowledge Discovery & Data Mining (KDD) communities (Goyal and Ferrara 2018, Cui *et al.* 2019, Su *et al.* 2020) without an emphasis on a particular application domain (which can be a drawback, as all these methods are heuristic due to computational intractability of the underlying problems that they approximately optimize, so their performance is strongly dependent on particular data), is based on commonly used models into: (i) factorization based; (ii) random walk based, and (iii) DL based.

2.3.2 The applications of embedding algorithms

There are many diverse applications, including predicting: protein structures, chemical compounds, PPIs, drug–drug interactions (synergistic or antagonistic), DTIs, new disease-related genes (e.g. new biomarkers, drug targets), drugs to repurpose for different diseases, disease co-morbidities, etc. (Su *et al.* 2020, Li *et al.* 2022). Hence, biomedical applications are more diverse and complex than those usually considered by application-agnostic ML&KDD communities, that mostly focus on node classification, link prediction, and community detection in general (Goyal and Ferrara 2018, Cui *et al.* 2019). Even when an application of a mainstream ML method in biology seems simple, the complexity of the resulting model is more than mind-boggling: e.g. transformers (a DL model designed to process sequential input data, primarily used in NLP and computer vision) were recently used to predict atomic-level protein structure from protein sequence by training a family of large language models (LLM) that utilize DL, with 15 billion parameters, on 138 million known protein sequences (Lin *et al.* 2023). In the same vein, LLMs have been used to embed genes and cells (Yang *et al.* 2022a, Cui *et al.* 2024, Chen and Zou 2024). For instance, genePT (Chen and Zou 2024) embeds genes based on their text summaries from NCBI's Gene database and embeds single cells from single-cell RNA-seq expression data by using Open AI's GPT 3.5, an LLM model having 175B parameters. Also, DeID-GPT (Liu *et al.* 2023b) de-identifies medical text (electronic health records) by using OpenAI's GPT4, a LLM model having 1.5 T parameters.

Not only these billion/trillion parameter models are computationally costly to train and use, incurring substantial carbon footprint (Faiz *et al.* 2024) but also several concerns arose about their usage for real-world and biomedical applications. First, recent studies showed that the performances of these complex methods are often on par with those of simpler, dedicated approaches (Abdelhamid *et al.* 2023, Boiarsky *et al.* 2023, Kedzierska *et al.* 2023, Liu *et al.* 2023a, Muscoloni and Cannistraci 2023). For instance, Boiarsky *et al.* (2023) showed that for annotating single cells from single-cell RNA-seq data, a simple L1 logistic regression can perform favourably against LLM-based approaches, such as scBERT(Yang *et al.* 2022b) and scGPT (Cui *et al.* 2024). In the same vein, Cannistraci-Hebb network automata, using only local network information and without training, can outperform AlphaFold2 intelligence in protein interaction prediction, while also having a substantial advantage in computational time (Abdelhamid *et al.* 2023). Another ongoing controversy of LLM-based approaches is their tendency to produce 'hallucinations', outputs that look coherent, but are factually incorrect, or nonsensical (Huang *et al.* 2023). These hallucinations raise concerns over the reliability of LLMs in real-world scenarios, and thus, limit their practical deployment (Huang *et al.* 2023).

In addition to the above-mentioned limitations, LLMs also exhibit unexpected and surprising behaviour. For instance, Meta's 15-billion-parameter model (Lin *et al.* 2023) [a competitor to RoseTTAFold (Baek *et al.* 2021) and Google's AlphaFold (Jumper *et al.* 2021)] is trained to predict the next amino acid in sequence, but surprisingly it can also predict the corresponding 3D protein structure. Similar unexpected behaviours leading to surprising discoveries happened in the field of NLP with utilization of NN-based embedding methods. For instance, recent successes in that field (document classification, named entity recognition, information retrieval) are attributed to the novel way of embedding words of text in a low-dimensional space, so that semantically similar words (e.g. synonyms, capital cities) are encoded by vector representations with low cosine similarity, e.g. by using Word2vec (Mikolov *et al.* 2013a) with the well-known Skip-Gram architecture, a 1-layer NN with a softmax function, seeking to maximize the dot product, $\vec{w} \cdot \vec{c}$, of vectors $\vec{w}$ and $\vec{c}$, corresponding to word pairs, (w, c), in low d-dimensional embedding space that co-occur in the lexical corpus. As this architecture is computationally intractable for a large lexical corpus (e.g. the size of the widely used Google News corpus is billions of words), it has been replaced by the Skip-Gram with negative sampling (SGNS) (Mikolov *et al.* 2013b). Word vectors learned by the Skip-Gram model can meaningfully be combined by using just simple vector addition (Mikolov *et al.* 2013b), enabling defining the vector representation of a phrase (sentence) as the average of the vector representations of its constituent words; this carries to paragraphs and entire texts (Le and Mikolov 2014b). Surprisingly, the embedding space of words exhibits a linear algebraic structure: words represented as vectors in a low-dimensional space (word embeddings) can be used to extract new knowledge (analogies) directly by linear operations on the word vectors ($\vec{v}$) in the space, the famous example being that the vector calculation of $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$ is closer to $\overrightarrow{queen}$ than any other word vector (Fig. 2). However, since the first observations of such linear analogy properties of word embeddings (Mikolov *et al.*

2013c), there have been controversies about the applicability of linear operations to solve real questions beyond simple examples (Rogers *et al.* 2017). A key highlighted issue in these earlier approaches is the so-called 'offset problem': for instance, solving the above analogy by using the linear operation $\overrightarrow{queen} \approx \overrightarrow{king} - \overrightarrow{man} + \overrightarrow{women}$ supposes that the embedding vectors of these four words form a parallelogram (Fig. 2) and that any four words that are in the same relationship should also form a parallelogram, ignoring the fact that more important words (that appear more frequently in the text) tend to have longer embedding vectors than infrequent words. To overcome this limitation, ML was applied to uncover a suitable linear operation solving a particular analogy question beyond the simple parallelogram (Lim *et al.* 2021).

2.4 Towards explainable, predictable, and controllable AI models

Other AI technologies also often have unexpected behaviours learned from the vast amounts of data that they analyse, so the AI community has begun to call for steps back from potentially dangerous, ever-larger, unpredictable, black-box models with emergent capabilities [Association for the Advancement of Artificial Intelligence. Working together on our future with AI. https://aaai.org/working-together-on-our-future-with-ai/, Future of Life Institute. Pause giant AI experiments: an open letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/]. Clearly, new paradigms are needed to overcome these obstacles, to go beyond the controversies, and unexpected behaviours leading to surprising discoveries and to provide an understanding of the principles/laws underlying the organization of the data, which will make all observations and models explainable, predictable, and controllable, hence enabling answering of all questions commonly asked from the data as simple applications of these laws. Elucidating the underlying laws and constructing explainable, controllable and environmentally sustainable models is the ambition to strive for.

Currently, there exist biological applications for which the best-performing embedding approaches yield nonlinear embedding spaces (e.g. Alanis-Lobato *et al.* 2015, Durán *et al.* 2021, Zhao *et al.* 2023). However, inspired by the successes

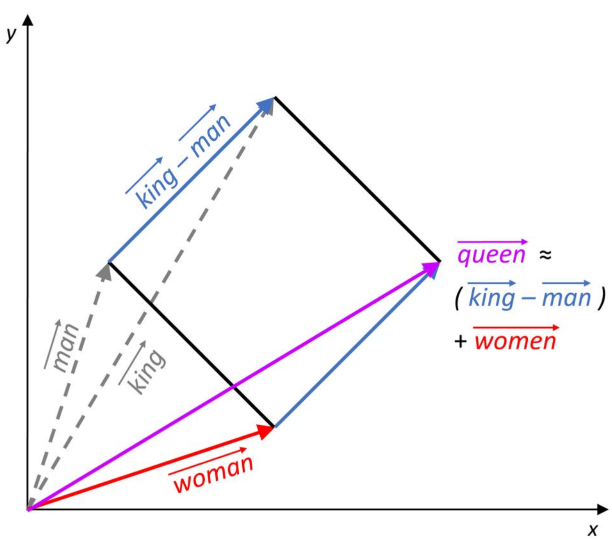


Figure 2. Illustration of the linear analogy 'queen is to women what king is to man'.

of word embedding methods in NLP, which revolutionized the field by embedding words into a low-dimensional space that exhibits linear algebraic structure (also having been applied to protein sequences, outlined in Section 2) and enables solving word similarities and analogies, efficient document classification, retrieval, etc., we hypothesize that it is also possible to embed all multi-omic data into a system of linearly connected, low-dimensional subspaces, each exhibiting linear algebraic structure. This could be viewed as taking projections of the entire high-dimensional space in which the biomedical entities reside, currently of unknown geometry and properties, into well-controlled and well-defined linear subspaces, which are linearly linked and traversed. Such a system will enable computationally efficient answering of all commonly asked biomedical and precision medicine questions by applying linear operations on the embedding vectors of the biomedical entities from this system.

Such a system could be constructed by generalizing the observed linear algebraic structure of the embedding space of words in a lexical corpus from the field of NLP to biomedical multi-omics network embedding research, applied to address real problems in precision medicine. As a proof of concept, we began with finding low-dimensional embeddings of various individual omics network layers into linear subspaces (Xenos *et al.* 2021, Doria-Belenguer *et al.* 2023, 2024). Interestingly, we demonstrated that embedding the proteins of a PPI network by decomposing matrix representations of the network with Non-negative Matrix Tri-Factorization (NMTF, detailed below) results in embedding spaces that are functionally organized. In particular, first, we demonstrated that in the resulting embedding spaces, proteins with similar biological annotations (functions) are embedded close in the embedding space (Xenos *et al.* 2021, Doria-Belenguer *et al.* 2023, 2024). Second, we demonstrated that the embedding space itself is functionally organized, with axes of the space capturing new, data-driven, higher-order biological functions, which we uncovered by joint embedding of genes and annotations into the same space (Doria-Belenguer *et al.* 2024). Importantly, we initiated the exploitation of the linear organization of these embedding spaces to answer biological questions. For instance, we showed that by applying simple vector operations on the embedding vectors of proteins we can uncover new members of protein complexes and predict new cancer-related genes (Xenos *et al.* 2021). As protein complexes in PPI networks are homophilic, meaning that adjacent proteins tend to perform the same biological function, direct application of NMTF to factorize the Positive Pointwise Mutual Information (PPMI, defined below) matrix of the PPI network based on random walks yielded this linearly exploitable subspace (Xenos *et al.* 2021). However, for heterophilic data, such as cancer driver genes (and equivalently proteins, as gene products), which have been known to exhibit similarity in the wiring patterns in the PPI network even when being nonadjacent in the network (i.e. they are heterophilic), we had to introduce a new version of the PPMI matrix based on randomly walking between nodes with similar graphlet-based wiring patterns to obtain the linear behaviour exploitable for predicting new cancer-related genes: we successfully predicted new cancer-related genes by performing linear operations on the embedding vectors of cancer driver genes in this new, graphlet-enabled embedding space (Xenos *et al.* 2021).

In addition to embedding one or few omic network types, e.g. PPI networks described above, we should also be embedding integrated/fused multi-omics network layers, also into linear subspaces, and we should find linear transformations that link these linear subspaces, enabling answering of all commonly asked biomedical questions by linear operations on the low-dimensional embedding vectors of biomedical entities. That is, we should untangle, 'linearize' and organize all multi-omics data into a low-dimensional, linear subspace system, enabling controllable and explainable answering of all common biomedical questions by simple linear operations on low-dimensional embedding vectors (hence, compute-space saving) of biomedical entities in linear time (hence, compute-time saving), enabling not only revolutionizing of biomedical data science but also saving computational space and power, i.e. reducing the carbon footprint. Encouraged by our recent preliminary feasibility studies that proposed a new, topology-constrained, matrix factorization (MF)-based embedding of the human PPI network into the low-dimensional embedding space exhibiting linear algebraic structure, which yielded new biomedical knowledge and identified new cancer-related genes and functions only by applying linear operations on the protein embedding vectors (Xenos *et al.* 2021, Doria-Belenguer *et al.* 2023, 2024), we expect that furthering such approaches will achieve quantitative and qualitative leaps both in modelling and algorithmic development and in precision medicine and other applications. We expect them to elucidate simple, linear organization of multi-omics data systems and enable effective utilization and exploitation of the multi-omic data in precision medicine.

To achieve that, we need new methodological advances towards a new representation and mining of the rich, multi-scale structure of multi-omic molecular organization within an integrated system, which will uncover a new, emergent nature of biological organization where linear organization holds. Importantly, the advances should be applied to explore some of the most challenging problems in different realms (e.g. in precision medicine) where the proposed methods can utilize the wealth of available multi-omic data to help address one of the greatest problems in all areas of science, the nature of complex organization. To realize the ambition, the new methodologies should be based on the four scientific pillars described below.

2.5 Pillar I. Characterizing biological network structure

Many computational problems on large networks are computationally intractable (e.g. subgraph isomorphism problem (Cook 1971), which underlies large network comparisons and makes them compute intensive). Also, because nature is variable and the biological data are noisy, traditional graph theoretic algorithms are of little use for network biology, and more flexible, intentionally approximate (heuristic) approaches are necessary. The area of developing such methods in biological and other domains is vast, too big to comprehensively survey here, some recent reviews including (Koutrouli *et al.* 2020, Liu *et al.* 2020, Guo *et al.* 2022). Hence, we summarize it and argue for extending only the state-of-the-art approaches relevant to demonstrating that the hypothesis of this manuscript holds.

Easily computable macroscopic statistical global properties of large networks have extensively been examined. The most widely used global network properties are the degree distribution, clustering coefficient, clustering spectra, network diameter, and various forms of network centralities (Newman 2003). The degree of a node is the number of edges touching the node and the degree distribution is the distribution of degrees of all nodes in the network. Many large real-world networks have non-Poisson degree distributions with a power-law tail, termed scale-free (Barabási and Albert 1999). However, networks with exactly the same degree distributions can have vastly different structures affecting their function (Pržulj *et al.* 2004, Li *et al.* 2005). The same holds for other global network properties (Pržulj *et al.* 2004). Furthermore, global network properties of largely incomplete molecular networks do not tell us much about the true structure of these networks; instead, they describe the network structure produced by the sampling biotechnologies used to obtain these networks (Han *et al.* 2005, Stumpf *et al.* 2005). Thus, global statistics on such incomplete data may be substantially biased, or even misleading with respect to the (currently unknown) full network. Conversely, certain neighbourhoods of these networks are well-studied (usually the regions of a network relevant to human disease). Since we have detailed knowledge of certain local areas of biological networks, but data outside these well-studied areas is currently incomplete, local statistics are likely to be more valid and meaningful. Furthermore, biological experiments for detecting molecular and other biochemical networks are of local nature. Thus, many have focused on developing tools based on local network structure. The state-of-the-art methods are based on graphlets, introduced in 2004 (Pržulj *et al.* 2004) and defined as small, connected, nonisomorphic, induced subgraphs of large networks (Pržulj *et al.* 2004). Graphlets and the symmetry groups in them, called automorphism orbits (introduced in 2007; Pržulj 2007), have extensively been used to develop many new tools for analysing structural properties of networks, some based on graphlet and orbit frequency distributions in a network (Pržulj 2007, Milenković and Pržulj 2008), others including graphlet Laplacians, eigencentralities, spectral clustering (Windels *et al.* 2019, 2022), etc. Graphlet statistics have also been used as kernels and feature vectors in various ML methods (Shervashidze *et al.* 2009, Vacic *et al.* 2010), and in the message-passing framework of GNNs, yielding better results in downstream analysis tasks (Bouritsas *et al.* 2023). They have been utilized and cited in around 21 000 research papers in various domains and in around 300 patents according to Google Scholar. Also, they were generalized to mine the multi-scale network organization: to hypergraphlets (Gaudelet *et al.* 2018, Lugo-Martinez *et al.* 2021) in hypergraphs, and simplets (Malod-Dognin and Pržulj 2019) in abstract simplicial complexes. Due to their large applicability and the above-mentioned increasing need for a deeper combination of network topology and ML (Su *et al.* 2020) for modelling biomedical systems, we propose that extending and utilizing graphlet-based methodologies would achieve breakthroughs.

2.6 Pillar II. Omic network data fusion

Omics network data fusion is another vast research area, too large to comprehensively survey here, recent surveys including (Gligorijević and Pržulj 2015, Ritchie *et al.* 2015, Gligorijević *et al.* 2016b, Picard *et al.* 2021, Reel *et al.* 2021, Agamah *et al.* 2022, Lee and Kim 2022, Vahabi and Michailidis 2022, Flores *et al.* 2023). Hence, after a brief general overview, we focus only on the subarea relevant to the future research perspectives presented in this manuscript.

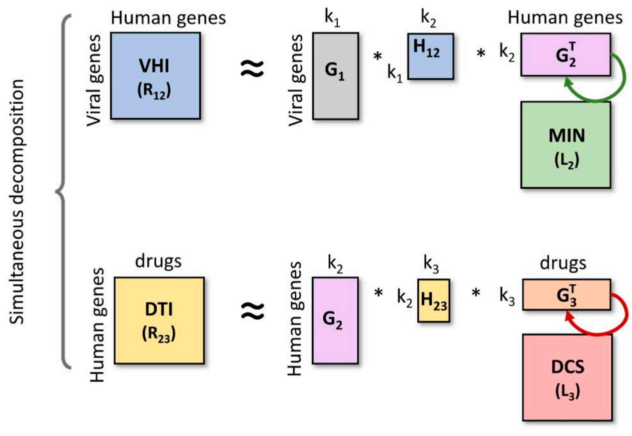
A wide variety of techniques that leverage the information contained in the relationships between omics data types have been driven by different biomedical applications (e.g. patient subtyping, biomarker discovery, *de novo* drug discovery, drug repurposing, and personalizing of treatment) and various challenges specific to multi-omics data analyses (e.g. omic data heterogeneity and high-dimensionality). The multitude of approaches motivates several rough categorizations into broader methodological groups, e.g. into early, middle, and late integrative methods, alternatively called concatenation-based, transformation-based, or model-based integration, respectively (Gligorijević and Pržulj 2015, Ritchie *et al.* 2015, Gligorijević *et al.* 2016b, Picard *et al.* 2021), because: early integration usually involves an initial concatenation of the features across the measured omics, followed by application of methods for analysing the resulting high-dimensional dataset; middle integration applies a transformation to represent a complex combination of the datasets before applying downstream analyses; a late integration method analyses each dataset separately and introduces a model or algorithm that combines the outputs of each individual analysis. Another characterization focuses on specific techniques, e.g. kernel learning, MF, graph/network representations, and deep learning. The categorizations are not mutually exclusive and a given method may be difficult to classify. Furthermore, the use of ML to analyse multi-omic data face key, unique, data-driven challenges, roughly summarized as: (i) data heterogeneity, sparsity, and outliers; (ii) class imbalance and over-fitting; (iii) much more features than the data (curse of dimensionality); (iv) computational and storage cost; (v) choosing an algorithm that works best for a given biomedical problem is hard, since ML methods are approximate (heuristic) due to computational intractability (e.g. NP-hardness) of the underlying problems (e.g. nonlinear optimization), so from the theory of computation we know that each method is guaranteed to fail on some particular examples, calling for a design of robust, application-specific methods.

MFs have largely demonstrated their usefulness in improving understanding of biological mechanisms from omic data, e.g.: Singular Value Decomposition (SVD) (Alter *et al.* 2000), Principal Component Analysis (PCA) (Wall *et al.* 2003) with the sparse and probabilistic variants (Meng *et al.* 2016), Independent Component Analysis (ICA) (Sompairac *et al.* 2019) and Nonnegative Matrix Factorizations (NMF) (Yang and Michailidis 2016, Boccarelli *et al.* 2018, Rappoport and Shamir 2018, Stein-O'Brien *et al.* 2018, Esposito *et al.* 2019, 2021). Each of these techniques is based on different constraints that characterize the final properties of the matrix factors, leading to different optimization problems and numerical algorithms being used. NMF has become a standard tool in the analysis of high-dimensional data, a comprehensive and up-to-date detail of the most important theoretical aspects, including geometric interpretation, nonnegative rank, complexity, and uniqueness, having been provided (Gillis 2020). Importantly, a very broad array of models belongs to the 'factorized model' family (Haeffele and Vidal 2015).

A particular variant for omics network fusion, Non-negative Matrix Tri-Factorization (NMTF) (Devarajan 2008), has been gaining in popularity. It was originally proposed for dimensionality reduction and co-clustering due to its connection with k-means clustering (Fei *et al.* 2008). It decomposes an $n \times m$ data matrix, A (e.g. the adjacency matrix of a biological network), representing the relations between n and m elements, into a product of three nonnegative, low-dimensional matrices, $G_{n\times k}$, $S_{k\times k_1}$ and $P_{m\times k_1}$, as $A \approx GSP^T$, by solving the optimization problem: $min(||A - GSP^T||_F^2 : G, S, P \geq 0)$. The low-dimensional matrix factors G and P are used to assign n data points into $k \ll n$ clusters and m data points into $k_1 \ll m$ clusters, respectively (the clustering property of NMTF) (Ding *et al.* 2006), and S is a $k \times k_1$ compressed representation of A linking clusters of G and P, and also allowing for different numbers of clusters in G and in P (k and k_1, respectively). G and P imply co-clustering. The reconstructed matrix GSP^T is more complete then the initial data matrix A, featuring new entries, unobserved in the data, which emerge from the latent structure captured by the low-dimensional matrix factors, hence inferring new predictions (the matrix completion property of NMTF; Gligorijević *et al.* 2016a).

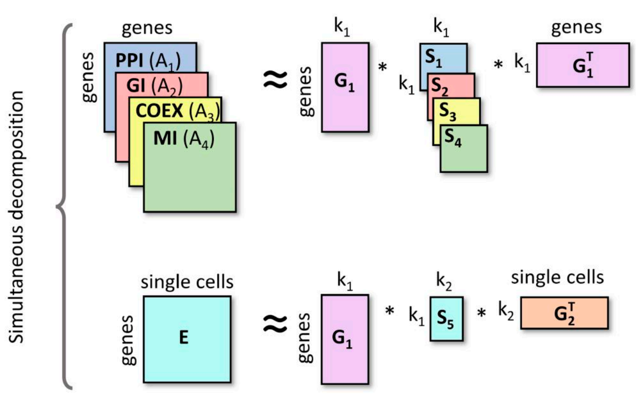
Importantly, NMTF allows for fusion of heterogeneous network data, by sharing matrix factors during decompositions. For instance, as illustrated in Fig. 3, to re-purpose known drugs for Covid-19, we recently factorized viral-host PPIs (VHI) simultaneously with DTIs as $VHI \approx G_1 H_{12} G_2^T$ and $DTI \approx G_2 H_{23} G_3^T$, where G_1 matrix clusters n viral proteins into k_1 clusters, G_2 clusters m human proteins into k_2 clusters, G_3 clusters q drugs into k_3 clusters, G_2 matrix is shared across the decomposition to ensure data fusion, and the reconstructed entities of the $G_2 H_{23} G_3^T$ matrix resulting from data fusion identify new DTIs used to re-purpose drugs (Zambrana *et al.* 2021). The only hyperparameters to tune were low-dimensional k_1, k_2 and k_3, with a total of about 2.6M parameters (the number of entries in the low-dimensional matrix factors G_1, G_2, G_3, H_{12} and H_{23}). As another example, we recently integrated single-cell gene expression data from patient-derived induced pluripotent stem cells together with four molecular interaction networks by using the NMTF model presented in Fig. 4 to uncover novel Parkinson's disease-related genes and pathways (Mihajlović *et al.* 2024). Again, and unlike the LLMs that require billions/trillions of parameters to tune (e.g. Cui *et al.* 2024, Lin *et al.* 2023, Liu *et al.* 2023b), our proposed NMTF model only requires two hyperparameters to tune, the low dimensionality parameters k_1 and k_2, and about 1.17M parameters (the number of entries in the low-dimensional matrix factors G_1, G_2, S_1, S_2, S_3, S_4 and S_5). These two examples illustrate that by careful modelling of the relationships between the data points, requiring deep understanding of the biomedical domain, with very few hyperparameters to tune, NMTF can achieve results superior to those of highly parametrized ML methods (Malod-Dognin *et al.* 2019, Isokääntä *et al.* 2020, Xenos *et al.* 2021, Zambrana *et al.* 2021, Doria-Belenguer *et al.* 2023, Mihajlović *et al.* 2024, Xenos *et al.* 2023).

There are several other reasons for which NMTF has increasingly become a methodology of choice for multi-omics data fusion. First, nonnegativity that applies to both bases and weights allows for meaningful interpretations. Second, modularity in the sense of the methodology easily enabling adding or removing omics data layers (illustrated by A_1 to A_4 and E in Fig. 4) is consistent with the natural human part-based learning processes (Esposito 2021). Third, NMTF-based methods overcome the shortcomings of the



$$J_1 = \min_{G_1,G_2,G_3,H_{12},H_{23}\geq 0} \begin{pmatrix} ||R_{12} - G_1 H_{12} G_2^T||_F^2 \\ +||R_{23} - G_2 H_{23} G_3^T||_F^2 \\ +tr(G_2^T L_2 G_2) + tr(G_3^T L_3 G_3) \end{pmatrix}$$

Figure 3. Illustration of the NMTF model from Zambrana *et al.* (2021). The viral host interactions, VHIs (represented by matrix R_{12}), are simultaneously decomposed with the DTIs, DTIs (represented by matrix R_{23}). The matrix factor G_2 is shared across decompositions to allow learning from all input matrices. The first graph regularization penalty (illustrated by the green arrow) is added so that the human genes that interact in the molecular interaction network, MIN (represented by its Laplacian matrix, L_2), are assigned similar low-dimensional embedding vectors in G_2. Similarly, the second graph regularization penalty (illustrated by the red arrow) is added so that the drugs that have similar chemical structures in the drug chemical similarity network, DCS (represented by its Laplacian matrix, L_3), are assigned similar low-dimensional embedding vectors in G_3. The hyperparameters k_1, k_2 and k_3 indicate the reduced dimensions of the embedding spaces of the viral proteins, human proteins, and drugs, respectively. The lower dimensional matrix factors, G_1, G_2, G_3, H_{12} and H_{23} are obtained by solving the corresponding minimization problem, J_1.



$$J_2 = \min_{G_1,G_2,S_1,S_2,S_3,S_4,S_5\geq 0} \begin{pmatrix} \sum_{i=1}^4 ||A_i - G_1 S_i G_1^T||_F^2 \\ +||E - G_1 S_5 G_2^T||_F^2 \end{pmatrix}$$

Figure 4. Illustration of the NMTF model from Mihajlović *et al.* (2024). The single-cell gene expression matrix, E(which can be thought of as capturing the phenotype), is simultaneously decomposed with four molecular interaction networks, PPI, GI, COEX, and MI, represented by their adjacency matrices, A_1, A_2, A_3, and A_4, respectively (which can be thought of as capturing the genotype, as they describe all possible interactions). The matrix factor G_1 is shared across decompositions to allow learning from all input matrices. The hyperparameters k_1 and k_2 indicate the reduced dimensions of the embedding spaces of human genes and single cells, respectively. The lower dimensional matrix factors, G_1, G_2, S_1, S_2, S_3, S_4 and S_5 are obtained by solving the corresponding minimization problem, J_2.

competing ML data fusion methods (Bayesian network-based and Kernel-based integration; Gligorijević and Pržulj 2004): they model inter-type data relations (e.g. patients-to-drugs) and intra-type data relations (e.g. PPIs) simultaneously, require no data transformation (hence, incur no information loss), and incorporate both intra- and inter-type networks at their full systems-level size, into a single integrative model from which new medical knowledge is inferred (Gligorijević and Pržulj 2015, Gligorijević *et al.* 2016b). Importantly, NMTF results in an integrated network in which: (i) some 'less important' (or topologically inconsistent) links across the networks do not appear (e.g. links appearing only in one network); (ii) links with structural support in most of the networks appear in the integrated network with enhanced association scores; (iii) new links emerge as a consequence of the matrix completion property as predictions that need to be further examined. That is, these methods give the highest weight to the information supported by evidence from multiple data, hence diminishing the effects of sparsity and noise in the data (Malod-Dognin *et al.* 2023). Although NMTF is equivalent to the 2-factor, NMF problem (Ding *et al.* 2006), $min(||D - GF^T||_F^2 : G, F \geq 0)$, which is NP-hard (Vavasis 2010), so we cannot expect to solve NMTF to optimality, availability of fast optimization methods (Čopar *et al.* 2019) coupled with exploitation of sparsity of the omic data and utilization of parallel matrix manipulations on high-performance computing (HPC) facilities make these methods not only applicable but also feasible for omics data fusion.

2.7 Pillar III. Omics network embedding methods (NLP-inspired)

The area of network embedding methods is also too large to comprehensively survey here, with recent reviews including (Goyal and Ferrara 2018, Cui *et al.* 2019, Su *et al.* 2020, Xu 2021, Li *et al.* 2022). They can roughly be divided into those based on traditional MFs, NNs, and random walks. The traditional MF-based methods come in many variants, e.g.: SVD (Alter *et al.* 2000), Isomap (Tenenbaum *et al.* 2000), Locally Linear Embedding (Roweis and Saul 2000). Those based on NNs include: multilayer perceptrons (Tang *et al.* 2015a), autoencoders (Kipf and Welling 2016), generative adversarial networks (Wang *et al.* 2019a), GNNs (Zhou *et al.* 2020) including graph convolutional networks (Kipf and Welling 2017), and geometric deep learning (Atz *et al.* 2021). A limitation of these methods is that they mainly focus on factorizing the first-order data matrix (adjacency matrix), despite limitations of representing networks with adjacency matrices and using as input into NNs, e.g. that many different adjacency matrices can represent the same graph connectivity, but do not produce the same result when input into a NN, i.e. they are not permutation invariant (Ying *et al.* 2018) (some works recently attempting to address it; Meltzer *et al.* 2019, Balan *et al.* 2022, Huang *et al.* 2022). Other limitations include: lack of robustness to noise and overfitting due to scarcity of labelled data, interpretability, and real networks usually being 'multiplex' (Yang *et al.* 2022), i.e. having incomplete multi-view representations from different relations (some works recently attempting to address it; Reel *et al.* 2021, Lee and Kim 2022, Wang *et al.* 2022, Flores *et al.* 2023).

Random-walk-based network embedding methods are inspired by the field of NLP, where words are represented as vectors in a low, d-dimensional space (word embeddings)

(Mikolov *et al.* 2013a) and new knowledge is extracted directly by linear operations on the vectors in the space. It has long been known that two words in a similar context have similar meanings (distributional hypothesis) (Harris 1954). Following this observation, Word2vec model was introduced (Mikolov *et al.* 2013a), which generates continuous representations of words as vectors in d-dimensional space, so that words that appear frequently in the same context are placed close in the space (detailed above). However, this architecture is computationally intractable for a large lexical corpus (e.g. the size of the widely used Google News corpus is billions of words) and therefore has been replaced by the SGNS (Mikolov *et al.* 2013b) and more robust versions of it were proposed (Armandpour *et al.* 2019). Importantly, this model enables semantic, context-based, word comparisons (similarity task) in the embedding space solely by computing the cosine similarity of their vector representations (Mikolov *et al.* 2013a). For example, Paris and Berlin, being capital cities, have similar vector representations and consequently, high cosine similarity. As noted above, word representations learned by the Skip-Gram model can also meaningfully be combined using vector addition (Mikolov *et al.* 2013b), enabling the defining of the vector representation of a phrase (sentence) as the average of the vector representations of its constituent words (Le and Mikolov 2014a); more recent works (e.g. Lin *et al.* 2017) train DL models to learn phrase representations that respect the word order in the sentence. These allow for analysing texts beyond the level of words, revealing semantic similarities between sentences, paragraphs, or entire documents. However, serious limitations exist about the applicability of linear operations to solve real questions beyond simple examples (Rogers *et al.* 2017), a key issue being the above-mentioned 'offset problem', among others. Analogously in networks, hub-nodes (of very high degree) occur on more random walks and hence, have longer embedding vectors than peripheral nodes. These limitations need to be addressed.

Following the success of the Skip-Gram model, various attempts have been made to generalize it and apply it to networks, e.g. DeepWalk (Perozzi *et al.* 2014), LINE (Tang *et al.* 2015c), PTE (Tang *et al.* 2015b), and Node2vec (Grover and Leskovec 2016). These methods rely on random walks to generate sequences of nodes, the equivalent of the lexical corpus, on which the Skip-Gram architecture can be applied. Interestingly, it was shown that the SGNS is implicitly factorizing a word-context matrix (Levy and Goldberg 2014), called PPMI matrix, defined as follows: for two words, w and c, $PPMI(w,c) = \max\left(0, \log \frac{(w,c)|C|}{|w||c|}\right)$, where |C| is the size of the corpus, (w,c) is the number of times the two words co-occur in the corpus, and |w| and |c| are the numbers of times the words w and c occur in the corpus. The cells of this matrix quantify how frequently two words of the lexical corpus co-occur in a sliding window compared with what would be expected if the occurrences of the words were independent. Subsequently, it was shown that the Skip-Gram-based network embeddings (in DeepWalk, LINE, PTE, and Node2vec) can be unified into the MF framework with closed forms (Qiu *et al.* 2018): they are implicitly factorizing a random-walk-based mutual information matrix, M (a diffusion on the original network), which is equivalent to the above described PPMI matrix, as its cells quantify how frequently two nodes of the network co-occur in a random walk compared with what would be expected if the occurrences of the nodes were independent. Formally, each entry of M, M_{ij}, is the logarithm of the average probability that node i randomly walks to node j in a fixed number of steps. A closed formula to approximate M was presented and used to generate the embedding space by applying SVD on matrix M (NetMF method and its approximation algorithm for computing network embedding) (Qiu *et al.* 2018). This provided the theoretical connections between Skip-gram-based network embedding algorithms and the theory of graph Laplacian (Qiu *et al.* 2018).

These methods embed network nodes in a low-dimensional space, so that nodes that are 'similar' are close in space, where 'similarity' means either belonging to the same network neighbourhood or 'community', or having similar network topology, e.g. being hub, or bridge nodes (structural/topological similarity). Then vector representations of nodes are used as input into ML models to predict protein functions, drug-disease associations, drug–drug interactions, and PPIs (see Su *et al.* 2020 for details). Hence, in both cases of words and networks, the Skip-Gram-based embeddings are approximating the exact factorization of the mutual information matrix. In NLP they use the NN-based embeddings rather than explicitly factorizing the PPMI matrix due to the size (factorizing a matrix of 1B× 1B words being computational intractable, as the time complexity of embedding data in k dimensions with the partial SVD is $O(k \times n^2)$; Kishore Kumar and Schneider 2017). In contrast, the size of the human PPI network and thus its PPMI matrix is currently $\approx 19,000 \times 19,000$, making its decompositions, either with SVD, or NMTF feasible (time complexity is $O(n^3)$). This motivates us to investigate whether the equivalent properties of word embeddings (molecular analogues to word-analogies, Fig. 2) also hold for molecular networks if we decompose their PPMI, or adjacency matrix representation, with an NMTF-based framework (Gligorijević *et al.* 2016a, Malod-Dognin *et al.* 2019, Zambrana *et al.* 2021, Xenos *et al.* 2023). In omics networks, the best equivalent to word analogy are functional modules, e.g. protein complexes and pathways in PPI networks. However, biological function is not only shared between the proteins that physically interact (Sharan *et al.* 2007) or participate in the same functional module (Chen *et al.* 2014) but also between proteins that have similar local wiring patterns regardless of their adjacency in the interaction network (Milenković and Pržulj 2008), best quantified by graphlet-based measures (Pržulj *et al.* 2004, Pržulj 2007, Milenković and Pržulj 2008, Windels *et al.* 2019, 2022), e.g. graphlet degree vectors (Milenković and Pržulj 2008). Thus, an important challenge is to generate embeddings that locate close in space nodes with similar wiring patterns.

There are several other reasons for which we argue that NMTF-based framework is the methodology of choice to embed each of the omics network layers individually (e.g. PPI, GI, COEX) into an embedding space whose properties we can control, that exhibits linear algebraic structure, which enables answering of various bio-medical questions by linear operations on the embedding vectors of nodes, and which importantly enables fusion of all omics networks into a linearly traversable system of embedding subspaces (detailed below). The seven constraints to enforce control for the system to be free of potentially dangerous, unexpected behaviours are: 1) nonnegativity of the coordinates of the embedding vectors (for easier interpretation), 2) bases (axes) of the spaces being orthogonal (to minimize entity dependencies in the embedding space), 3) sparsity (to capture the strongest signal), 4) enable data fusion, 5) enable 'walking' (by linear transformations) between the embedding spaces as needed to answer biomedical questions, i.e. directly investigating different types of entities in the same, or in different linear subspaces (modelling different types of data), 6) explainability, capturing how the entities relate (i.e. being eXplainable AI, XAI), and 7) few parameters to tune (to simplify the model). NMTF-based methods (Ding *et al.* 2006) are the only ones that can easily satisfy all seven constraints. They were successfully used in biology to analyse individual omics data types in isolation from each other, and also enable heterogeneous data fusion, including network topology as constraints, further motivating their utilization within new network embedding algorithms in biomedical applications (Gligorijević *et al.* 2016a, Vitali *et al.* 2018, Malod-Dognin *et al.* 2019). Importantly, explainability (constraint 6 above) allows control of the system's behaviour to coincide with what is expected (by examining the linear operation that gave the answer), while reduction of unexpected behaviour comes mostly from orthogonality (constraint 2 above) and the property of NMTF that it boosts the signal that comes from multiple data types, rather than outliers, resulting in more robust predictions (constraint 4 above).

2.8 Pillar IV. Finding optimal dimensionality of the embedding spaces

Another main challenge in graph embeddings is finding an optimal reduced dimension that is small enough to be efficient, but large enough to keep all of the necessary properties of the whole network. The choice of the dimension considerably influences the model's performance (Lai *et al.* 2016), with the embeddings with very low dimensionality typically not being expressive enough to capture the richness of the data (Patel and Bhattacharyya 2017), while embeddings with a very large dimensionality suffering from over-fitting. Also, large dimensionalities tend to increase model complexity, slow down training speed, and add inferential latency, all of which are constraints that can potentially limit model applicability and deployment (Wu *et al.* 2016). As this is an unresolved scientific question, the embedding space dimensionality is considered a hyperparameter of the model. In most existing methods, the dimensionality is selected by a grid search, by estimating the performance in downstream tasks, such as node classification, or link prediction. The choice of the optimal dimension is mainly evaluated for nonbiological networks, with few studies evaluating it on biological networks (Cannistraci *et al.* 2013, Yue *et al.* 2020). The approaches that estimate the intrinsic dimensionality of the data usually yield super low dimensionality (at most 20), while the data-driven methods yield up to 200/250 dimensions (Yin and Shen 2018). Recently proposed ultra-low dimensions (20 or fewer) are shown to preserve network structure, but perform poorly in downstream classification tasks (Chanpuriya *et al.* 2020). The popular approaches using hyperbolic space to embed networks are not applicable for omics networks. The model-based hyperbolic space embedding approaches (Serrano *et al.* 2008, Boguñá *et al.* 2020) assume that the number of squares or pentagons is higher than the number of triangles, which is not the case for our data (Seshadhri *et al.* 2020, Almagro *et al.* 2022). While it is not known if model-free hyperbolic space embedding methods (Muscoloni *et al.* 2017, Alanis-Lobato *et al.* 2018, Härtner *et al.* 2018, Poleksic 2023, Zhao *et al.* 2023) suffer from a similar issue, they tend to provide ultra-low dimensional embeddings that are not suited for downstream biological analyses. The vast majority of analyses that use Node2vec (Grover and Leskovec 2016) and DeepWalk (Perozzi *et al.* 2014) use the default dimensionalities of these methods, 128 or 256. Our initial, data-driven feasibility studies on PPI networks demonstrate that the lower the dimensionality, the less distinguishable the embedding vectors of genes, and as we increase the number of dimensions, the average distance between the embedding vectors increases and hence the space is more disentangled (Xenos *et al.* 2021, Doria-Belenguer *et al.* 2023), but after 250/300 dimensions the position of the embedding vectors in the embedding space does not change, suggesting this to be an upper limit for PPI networks (Xenos *et al.* 2021, Doria-Belenguer *et al.* 2023). Recently, the 2NN method (Facco *et al.* 2017) for estimating the dimensionality of datapoints in high-dimensional space has been successfully extended for estimating the dimensionality of network embedding spaces (Grindrod *et al.* 2023). While this approach has yet to be tested on real biological networks, it may be the basis of novel measures for estimating the optimal dimensionality of the system of embedding subspaces used to integrate the multi-omics datasets in precision medicine.

3 Perspectives and challenges

We propose that multi-disciplinary efforts are needed to address the above challenges and develop an advanced, explainable, and controllable network embedding methodology, free of potentially dangerous, unexpected behaviour. It will marry biomedical informatics with network science, linear algebra, and data fusion, unlocking foremost emerging interdisciplinary fields, precision medicine, and personalized drug discovery, enabling the extraction of new medical knowledge from all data collectively. To achieve this, innovation is needed in the following main areas.

3.1 Innovation in methods for embedding multi-omic networks into a liner subspace system

We should aim at bridging the gap between multi-omic network data and their network embedding-based biomedical interpretability by generalizing and uniting NMTF-based network embedding and network-science methods to propose new algorithmic and biological paradigms and solve real problems in precision medicine and other domains. We should generalize sophisticated network embedding methods to encompass and model the multi-scale structure of molecular organization within a linearly traversable system of embedding spaces exhibiting linear algebraic structure. This pretrained system should fuse all publicly available multi-omics data types and serve as a basis to uncover new biomedical paradigms and knowledge. It will enable us to find solutions to these generalizations that are as good as possible (locally optimal), which can be achieved by using the best practices from nonlinear optimization and the best HPC infrastructure, which given NP-hardness of these problems, is the best we can do. Encouraged by our recent feasibility studies yielding significant results for cancer (Xenos *et al.* 2021, Doria-Belenguer *et al.* 2023, 2024), we expect that such approaches will achieve quantitative and qualitative leaps in computational development and applications.

3.2 Innovation in data science and biological paradigms

The new methods will enable much-needed paradigm shifts in data science and biology. In data science, by simultaneously capturing the adjacency and higher-order structural network information along with the semantic information (node labels) within the new state-of-the-art network embedding algorithms (see Section 3.1), it will be possible to describe a general methodology for uncovering homophilic representations of multi-labelled networks, leading to functionally organized, linearly separable, embedding spaces suitable for applying linear operations. This would pave the road to generating a new generation of embedding methods that could drastically improve the accuracy of downstream analysis tasks for heterophilic data, which is a bottleneck of the current embedding methods. This, along with the proposed orthogonality of the embedding spaces (to minimize dependencies) (Xenos *et al.* 2021, Doria-Belenguer *et al.* 2023, 2024, Gureghian *et al.* 2023, Malod-Dognin *et al.* 2023), would further enable a shift in the exploration of the embedding spaces from objects' embedding vectors to the orthonormal axes of the space (Doria-Belenguer *et al.* 2024). In the biological domain, it would enable the much called for displacement of the dominant, sequence alignment-based construction of GO (Yu *et al.* 2019) and the species phylogenetic organization (Foley *et al.* 2023), resolving the controversial relationships and clarifying diversification (e.g. in the evolution of SARS-CoV-2 virus that caused the recent pandemic) by the proposed multi-omics data-driven approaches, guaranteed to provide better disentanglement of the functional and phylogenetic information.

3.3 Applications in precision medicine, drug discovery, and other domains

In this perspective, we focus on precision medicine applications where multi-omics data are embedded and mined to uncover patient disease subgroups, and for these subgroups to uncover disease genes and molecular mechanisms, drug repurposing, and drug discovery. Precision medicine also aims at discovering medical knowledge for a given patient or sample. Since the general framework that we propose also embeds each patient or sample in the common space together with the other biological entities, the framework can easily be adapted to uncover these additional types of precision medicine knowledge.

The proposed advances can be applied and validated on real multi-omic data. They will enable innovation in utilizing new AI frameworks, encompassing multi-omics 'big data', to enable continued growth in digital innovation applied to precision medicine. They will enable better subtyping of patients, identifying new biomarkers, drug-targets, and drugs to repurpose, expending beyond oncology, particularly in neurologic diseases affecting the ageing population: e.g. Alzheimer's disease, Parkinson's disease (Mihajlović *et al.* 2024), and glioblastoma cancer. The proposed directions would enable AI-assisted personalized drug discovery (Popova *et al.* 2018, Segler *et al.* 2018, Méndez-Lucio *et al.* 2020, Yasonik 2020, Grechishnikova 2021) (drugs for particular patients/patient subgroups) simultaneously with the other precision medicine tasks, at the same time and cost, from analysis of the same multi-omics data, within the same linear system (see Sections 3.1 and 3.2), unique thus far, further enabling better insights in medicine and drug discovery. Transformative potential of the proposed directions can also be envisioned in agriculture (Nicolopoulou-Stamati *et al.* 2016, Blois 2022, Mei *et al.* 2022, Pires *et al.* 2022, Oršolić *et al.* 2021) and protein engineering (Wang *et al.* 2018a, Greenhalgh *et al.* 2021, Freschlin *et al.* 2022).

Author contributions

Nataša Pržulj (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]) and Noël Malod-Dognin (Writing—original draft [equal], Writing—review & editing [equal])

Conflict of interest

None declared.

Funding

This work is supported by the European Research Council (ERC) Consolidator [770827], the Spanish State Research Agency and the Ministry of Science and Innovation MCIN [PID2022-141920NB-I00/AEI/10.13039/501100011033/FEDER], UE, and the Department of Research and Universities of the Generalitat de Catalunya code 2021 [SGR 01536].

References

Association for the Advancement of Artificial Intelligence. Working together on our future with AI. https://aaai.org/working-together-on-our-future-with-ai/ (5 April 2023, date last accessed).

The ALLERGOME Database. https://www.allergome.org/ (27 November 2023, date last accessed).

The Cancer Genome Atlas program (TCGA). https://www. cancer.gov/ccg/research/genome-sequencing/tcga (27 November 2023, date last accessed).

The CHEMBL Database. https://www.ebi.ac.uk/chembl/ (27 November 2023, date last accessed).

The ENCODE project. a. https://www.encodeproject.org/ (27 November 2023, date last accessed).

bioRxiv ENCODE channel. b. https://connect.biorxiv.org/relate/content/177 (27 November 2023, date last accessed).

The DRUGBANK Database. https://go.drugbank.com/ (27 November 2023, date last accessed).

Future of Life Institute. Pause giant AI experiments: an open letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (22 March 2023, date last accessed).

The FOODOME Project. https://www.barabasilab.com/science/proj ect/foodome (27 November 2023, date last accessed).

The Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database. https://www.genome.jp/kegg/pathway.html (27 November 2023, date last accessed).

Next-Generation Sequencing Articles from across Nature Portfolio. https://www.nature.com/subjects/next-generation-sequencing (27 November 2023, date last accessed).

The National Institutes of Health (NIH) Microbiome Project. https://www.hmpdacc.org/ (27 November 2023, date last accessed).

National Pesticide Information Center, Databases for Chemical Information. http://npic.orst.edu/ingred/cheminfo.html (27 November 2023, date last accessed).

PISTACHIO Database. https://www.nextmovesoftware.com/pistachio.html (27 November 2023, date last accessed).

Pesticide, Bio-Pesticide and Veterinary Substances Properties Databases. http://sitem.herts.ac.uk/aeru/ppdb/ (27 November 2023, date last accessed).

The PUBCHEM Database. https://pubchem.ncbi.nlm.nih.gov/ (27 November 2023, date last accessed).

REAXIS Database. https://www.elsevier.com/products/reaxys (27 November 2023, date last accessed).

The UK BIOBANK. https://www.ukbiobank.ac.uk/ (27 November 2023, date last accessed).

Abdelhamid I, Muscoloni A, Rotscher DM, *et al.* Network shape intelligence outperforms AlphaFold2 intelligence in vanilla protein interaction prediction. bioRxiv, 2023, 2023–308, preprint: not peer reviewed.

Adames NR, Gallegos JE, Peccoud J. Yeast genetic interaction screens in the age of CRISPR/Cas. *Curr Genet* 2019;**65**:307–27.

Agamah FE, Bayjanov JR, Niehues A, *et al.* Computational approaches for network-based integrative multi-omics analysis. *Front Mol Biosci* 2022;**9**:967205.

Alanis-Lobato G, Cannistraci CV, Eriksson A, *et al.* Highlighting nonlinear patterns in population genetics datasets. *Sci Rep* 2015; **5**:8140.

Alanis-Lobato G, Mier P, Andrade-Navarro M. The latent geometry of the human protein interaction network. *Bioinformatics* 2018; **34**:2826–34.

Aleksander SA, Balhoff J, Carbon S, *et al.*; Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. *Genetics* 2023; **224**:iyad031.

Almagro P, Boguñá M, Serrano MÁ. Detecting the ultra low dimensionality of real networks. *Nat Commun* 2022;**13**:6096.

Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* 2000;**97**:10101–6.

Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an Arabidopsis interactome map. *Science* 2011; **333**:601–7.

Armandpour M, Ding P, Huang J, *et al.* Robust negative sampling for network embedding. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. **33**. 2019, 3191–8.

Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. *Nat Mach Intell* 2021;**3**:1023–32.

Baek M, DiMaio F, Anishchenko I, *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6. https://doi.org/10.1126/science.abj8754

Balan R, Haghani N, Singh M. Permutation invariant representations with applications to graph deep learning. arXiv, https://doi.org/2203.07546, 2022, preprint: not peer reviewed.

Barabási A-L, Albert R. Emergence of scaling in random networks. *Science* 1999;**286**:509–12.

Baron JA, Johnson CS-B, Schor MA, *et al.* The DO-KB Knowledgebase: a 20-year journey developing the disease open science ecosystem. *Nucleic Acids Res* 2024;**52**:D1305–14.

Baryshnikova A, Costanzo M, Kim Y, *et al.* Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat Methods* 2010;**7**:1017–24.

Bastounis A, Gorban AN, Hansen AC, *et al.* The boundaries of verifiable accuracy, robustness, and generalisation in deep learning. In: Iliadis L, Papaleonidas A, Angelov P, Jayne C (eds), *Artificial Neural Networks and Machine Learning – ICANN 2023*. Cham: Springer Nature Switzerland, 2023, 530–41.

Beerens L, Higham DJ. Vulnerability analysis of transformer-based optical character recognition to adversarial attacks. arXiv, https://doi.org/1506.07540, 2015, preprint: not peer reviewed.

Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Adv Neural Inf Process Syst* 2001;**14**.

Blois M. Following several fallow decades, herbicide companies are searching for new modes of action. *Chem Eng News* 2022; **100**:22–8.

Boccarelli A, Esposito F, Coluccia M, *et al.* Improving knowledge on the activation of bone marrow fibroblasts in MGUS and MM disease through the automatic extraction of genes via a nonnegative matrix factorization approach on gene expression profiles. *J Transl Med* 2018;**16**:217–6.

Bodnar C, Cangea C, Liò P. Deep graph mapper: seeing graphs through the neural lens. *Front Big Data* 2021;**4**:680535.

Boguñá M, Krioukov D, Almagro P, *et al.* Small worlds and clustering in spatial networks. *Phys Rev Res* 2020;**2**:023040.

Boiarsky R, Singh N, Buendia A, *et al.* A deep dive into single-cell RNA sequencing foundation models. bioRxiv, https://doi.org/2023.10.19.563100, 2023, preprint: not peer reviewed. https://doi.org/10.1101/2023.10.19.563100, https://www.biorxiv.org/content/early/2023/10/23/2023.10.19.563100

Bouritsas G, Frasca F, Zafeiriou S, *et al.* Improving graph neural network expressivity via subgraph isomorphism counting. *IEEE Trans Pattern Anal Mach Intell* 2023;**45**:657–68.

Bubenik P. Statistical topological data analysis using persistence landscapes. *J Mach Learn Res* 2015;**16**:77–102.

Burley SK, Bhikadiya C, Bi C, *et al.* RCSB protein data bank (RCSB.org): delivery of experimentally-determined pdb structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res* 2023; **51**:D488–508.

Cannistraci CV, Alanis-Lobato G, Ravasi T. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics* 2013;**29**:i199–209.

Cao M, Zhang H, Park J, *et al.* Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS One* 2013;**8**:e76339.

Carrière M, Chazal F, Ike Y, *et al.* PERSLAY: a neural network layer for persistence diagrams and new graph topological signatures. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, 2786–96.

Chaitin G. On the intelligibility of the universe and the notions of simplicity, complexity, and irreducibility. In: Hogrebe W (ed.), *Grenzen und grenzüberschreitungen: XIX. Deutscher kongress für philosophie, bonn, 23.-27. September 2002Vorträge und kolloquien*. Berlin (Boston): Akademie Verlag, 2004, 517–34. https://doi.org/10.1515/9783050081939.517, https://doi.org/10.1515/9783050081939.517

Chanpuriya S, Musco C, Sotiropoulos K, *et al.* Node embeddings and exact low-rank representations of complex networks. *Adv Neural Inf Process Syst* 2020;**33**:13185–98.

Chen B, Fan W, Liu J, *et al.* Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. *Brief Bioinform* 2014;**15**:177–94.

Chen YT, Zou J. GenePT: a simple but hard-to-beat foundation model for genes and cells built from chatgpt. bioRxiv, https://doi.org/10.16.562533, 2024, preprint: not peer reviewed. https://doi.org/10.1101/2023.10.16.562533, https://www.biorxiv.org/content/early/2023/10/19/2023.10.16.562533

Chen Z-H, You Z-H, Guo Z-H, *et al.* Prediction of drug–target interactions from multi-molecular network based on deep walk embedding model. *Front Bioeng Biotechnol* 2020;**8**:338.

Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. *Nat Commun* 2019a;**10**:1197.

Cheng F, Lu W, Liu C, *et al.* A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat Commun* 2019b;**10**:3476.

Clough E, Barrett T, Wilhite SE, *et al.* NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Res* 2024;**52**:D138–44.

Cook SA. The complexity of theorem proving procedures. In: *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing*. 1971, 151–8.

Čopar A, Zupan B, Zitnik M. Fast optimization of non-negative matrix tri-factorization. *PLoS One* 2019;**14**:e0217994.

Costanzo M, Baryshnikova A, Bellay J, *et al.* The genetic landscape of a cell. *Science* 2010;**327**:425–31.

Costanzo M, VanderSluis B, Koch EN, *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* 2016;**353**:aaf1420.

Cui H, Wang C, Maan H, *et al.* scGPT: towards building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1470–80.

Cui P, Wang X, Pei J, *et al.* A survey on network embedding. *IEEE Trans Knowl Data Eng* 2019;**31**:833–52.
Dahlquist KD, Salomonis N, Vranizan K, *et al.* GENMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* 2002;**31**:19–20.
De Cecco L, Nicolau M, Giannoccaro M, *et al.* Head and neck cancer subtypes with biological and clinical relevance: meta-analysis of gene-expression data. *Oncotarget* 2015;**6**:9627–42.
Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol* 2008; **4**:e1000029.
Dhar PK, Giuliani A. Laws of biology: why so few? *Syst Synth Biol* 2010;**4**:7–13.
Ding C, Li T, Peng W, *et al.* Orthogonal nonnegative matrix tri-factorizations for clustering. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2006, 126–35.
Donnat C, Zitnik M, Hallac D, *et al.* Learning structural node embeddings via diffusion wavelets. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, 1320–9.
Doria-Belenguer S, Xenos A, Ceddia G, *et al.* A functional analysis of omic network embedding spaces reveals key altered functions in cancer. *Bioinformatics* 2023;**39**:btad281.
Doria-Belenguer S, Xenos A, Ceddia G, *et al.* The axes of biology: a novel axes-based network embedding paradigm to decipher the functional mechanisms of the cell. *Bioinform Adv* 2024;**4**:vbae075. 2024, preprint: not peer reviewed. https://doi.org/10.1101/2023.07.31.551263
Duerr RH, Taylor KD, Brant SR, *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 2006;**314**:1461–3.
Durán C, Ciucci S, Palladini A, *et al.* Nonlinear machine learning pattern recognition and bacteria-metabolite multilayer network analysis of perturbed gastric microbiome. *Nat Commun* 2021;**12**:1926.
Edelsbrunner H, Harer J. Persistent homology – a survey. *Contemp Math* 2008;**453**:257–82.
Edelsbrunner H, Harer JL. *Computational Topology: An Introduction*. American Mathematical Society, 2022.
Eftimie R. Grand challenges in mathematical biology: integrating multi-scale modeling and data. *Front Appl Math Stat* 2022;**8**:1010622.
Esposito F. A review on initialization methods for nonnegative matrix factorization. Towards omics data experiments. *Mathematics* 2021; **9**:1006.
Esposito F, Buono ND, Selicato L. Nonnegative matrix factorization models for knowledge extraction from biomedical and other real world data. *Proc Appl Math Mech* 2021;**20**:e202000032.
Esposito F, Gillis N, Del Buono N. Orthogonal joint sparse NMF for microarray data analysis. *J Math Biol* 2019;**79**:223–47.
Facco E, d'Errico M, Rodriguez A, *et al.* Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci Rep* 2017;**7**:12140.
Faiz A, Kaneda S, Wang R, *et al.* LLMCarbon: modeling the end-to-end carbon footprint of large language models. In: *The Twelfth International Conference on Learning Representations (ICLR)*. Appleton, Wisconsin, USA: International Conference on Learning Representations (ICLR), 2024, 1387.
Fei W, Tao L, Changshui Z. Semi-supervised clustering via matrix factorization. In: *Proceedings of the SIAM International Conference on Data* Mining. 2008.
Flores JE, Claborne DM, Weller ZD, *et al.* Missing data in multi-omics integration: recent advances through artificial intelligence. *Front Artif Intell* 2023;**6**:1098308.
Foley NM, Mason VC, Harris AJ, *et al.* Zoonomia Consortium‡. A genomic timescale for placental mammal evolution. *Science* 2023; **380**:eabl8189.
Freschlin CR, Fahlberg SA, Romero PA. Machine learning to navigate fitness landscapes for protein engineering. *Curr Opin Biotechnol* 2022;**75**:102713.
Gaudelet T, Malod-Dognin N, Pržulj N. Higher-order molecular organization as a source of biological function. *Bioinformatics* 2018; **34**:i944–53.
Gavin A-C, Aloy P, Grandi P, *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006;**440**:631–6.
Geng C, Jung Y, Renaud N, *et al.* ISCCORE: a novel graph kernel-based function for scoring protein–protein docking models. *Bioinformatics* 2020;**36**:112–21.
Gillis N. *Nonnegative Matrix Factorization*. Society for Industrial and Applied Mathematics (SIAM), 2020.
Gilmer J, Schoenholz SS, Riley PF, *et al.* Neural message passing for quantum chemistry. In: *International Conference on Machine Learning*. PMLR, 2017, 1263–72.
Giot L, Bader JS, Brouwer C, *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* 2003;**302**:1727–36.
Glazier DS. Variable metabolic scaling breaks the law: from 'Newtonian' to 'Darwinian' approaches. *Proc R Soc B* 2022; **289**:20221605.
Gligorijević V, Malod-Dognin N, Pržulj N. Patient-specific data fusion for cancer stratification and personalised treatment. In: *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, 2016a, 321–32.
Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. *Proteomics* 2016b;**16**:741–58.
Gligorijević V, Pržulj N. Computational methods for integration of biological data. In: Bodiroga-Vukobrat DRN, Pavelic K, Sander GG (eds), *Personalised Medicine: A New Medical and Social Challenge*, Chapter 8. Springer Verlag, 2004.
Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. *J R Soc Interface* 2015;**12**:20150571.
Glover A. The 21st century: the age of biology. In: *OECD Forum on Global Biotechnology*. 2012. https://www.oecd.org/sti/emerging-tech/A%20Glover.pdf
Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: a survey. *Knowl Based Syst* 2018;**151**:78–94.
Grechishnikova D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Sci Rep* 2021;**11**:321.
Greenhalgh J, Saraogee A, Romero PA. *Data-Driven Protein Engineering*, Chapter 6. John Wiley & Sons, Ltd, 2021, 133–51.
Grindrod P, Higham DJ, de Kergorlay H-L. Estimating network dimension when the spectrum struggles. *R Soc Open Sci* 2024;**11**:230898.
Grover A, Leskovec J. NODE2VEC: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2016, 855–64.
Guo M, Yu Y, Wen T, *et al.* Analysis of disease comorbidity patterns in a large-scale China population. *BMC Med Genomics* 2019; **12**:177–10.
Guo MG, Sosa DN, Altman RB. Challenges and opportunities in network-based solutions for biological questions. *Brief Bioinform* 2022;**23**:bbab437.
Gureghian V, Herbst H, Kozar I, *et al.* A multi-omics integrative approach unravels novel genes and pathways associated with senescence escape after targeted therapy in NRAS mutant melanoma. *Cancer Gene Ther* 2023;**30**:1–16.
Haeffele BD, Vidal R. Global optimality in tensor factorization, deep learning, and beyond. arXiv, https://doi.org/1506.07540, 2015, preprint: not peer reviewed.
Hammal F, De Langen P, Bergon A, *et al.* ReMap 2022: a database of human, mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res* 2022;**50**:D316–25.
Han J-DJ, Dupuy D, Bertin N, *et al.* Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* 2005;**23**:839–44.
Han P, Yang P, Zhao P, *et al.* GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, 705–13.

Harris ZS. Distributional structure. *Word* 1954;**10**:146–62.
Härtner F, Andrade-Navarro MA, Alanis-Lobato G. Geometric characterisation of disease modules. *Appl Netw Sci* 2018;**3**:10–7.
Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;**6**:95–108.
Hitz BC, Lee J-W, Jolanki O, *et al.* The ENCODE uniform analysis pipelines. *Res Sq* 2023;
Hofer C, Graf F, Rieck B, *et al.* Graph filtration learning. In: *International Conference on Machine Learning*. PMLR, 2020, 4314–23.
Hofer C, Kwitt R, Niethammer M, *et al.* Deep learning with topological signatures. *Adv Neural Inf Process Syst* 2017;**30**.
Hofer CD, Kwitt R, Niethammer M. Learning representations of persistence barcodes. *J Mach Learn Res* 2019;**20**:1–45.
Huang L, Yu W, Ma W, *et al.* A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv, https://doi.org/2311.05232, 2023, preprint: not peer reviewed.
Huang Z, Wang Y, Li C, *et al.* Going deeper into permutation-sensitive graph neural networks. In: *International Conference on Machine Learning*. PMLR, 2022, 9377–409.
Ingraham J, Garg V, Barzilay R, *et al.* Generative models for graph-based protein design. *Adv Neural Inf Process Syst* 2019;**32**.
Isokääntä S, Kari E, Buchholz A, *et al.* Comparison of dimension reduction techniques in the analysis of mass spectrometry data. *Atmos Meas Tech* 2020;**13**:2995–3022.
Ito T, Tashiro K, Muta S, *et al.* Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci USA* 2000;**97**:1143–7.
Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2, https://doi.org/10.1038/s41586-021-03819-2
Kato H, Takahashi S, Saito K. Omics and integrated omics for the promotion of food and nutrition science. *J Tradit Complement Med* 2011;**1**:25–30.
Kearnes SM, Maser MR, Wleklinski M, *et al.* The open reaction database. *J Am Chem Soc* 2021;**143**:18820–6.
Kedzierska KZ, Crawford L, Amini AP, *et al.* Assessing the limits of zero-shot foundation models in single-cell biology. bioRxiv, https://doi.org/2023.10.16.561085, 2023, preprint: not peer reviewed. https://doi.org/10.1101/2023.10.16.561085, https://www.biorxiv.org/content/early/2023/11/05/2023.10.16.561085
Kipf TN, Welling M. Variational graph auto-encoders. arXiv, https://doi.org/1611.07308, 2016, preprint: not peer reviewed.
Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: *Proceedings of the 5th International Conference on Learning Representations*. 2017.
Kishore Kumar N, Schneider J. Literature survey on low rank approximation of matrices. *Lin Multilin Algebra* 2017;**65**:2212–44.
Klein AM, Mazutis L, Akartuna I, *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015; **161**:1187–201.
Koutrouli M, Karatzas E, Paez-Espino D, *et al.* A guide to conquer the biological network era using graph theory. *Front Bioeng Biotechnol* 2020;**8**:34.
Krogan NJ, Cagney G, Yu H, *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006; **440**:637–43.
Kumar V, Gu Y, Basu S, *et al.* Radiomics: the process and the challenges. *Magn Reson Imaging* 2012;**30**:1234–48.
Lai S, Liu K, He S, *et al.* How to generate a good word embedding. *IEEE Intell Syst* 2016;**31**:5–14.
Lambin P, Rios-Velazquez E, Leijenaar R, *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 2012;**48**:441–6.
Lappalainen I, Lopez J, Skipper L, *et al.* DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res* 2013; **41**:D936–41.
Le D-H, Dang V-T. Ontology-based disease similarity network for disease gene prediction. *Vietnam J Comput Sci* 2016;**3**:197–205.
Le Q, Mikolov T. Distributed representations of sentences and documents. In: *International Conference on Machine Learning*. PMLR, 2014a, 1188–96.
Le Q, Mikolov T. Distributed representations of sentences and documents. In: Xing EP, Jebara T (eds), *Proceedings of the 31st International Conference on Machine Learning*, Vol. **32**. Bejing, China: PMLR, 2014b, 1188–96.
Lee D, Kim S. Knowledge-guided artificial intelligence technologies for decoding complex multiomics interactions in cells. *Clin Exp Pediatr* 2022;**65**:239–49.
Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization. *Adv Neural Inf Process Syst* 2014;**27**.
Li L, Alderson D, Doyle JC, *et al.* Towards a theory of scale-free graphs: definition, properties, and implications. *Internet Math* 2005; **2**:431–523.
Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. *Nat Biomed Eng* 2022;**6**:1353–69.
Li S, Armstrong CM, Bertin N, *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* 2004;**303**:540–3.
Li X, Wang C-Y. From bulk, single-cell to spatial RNA sequencing. *Int J Oral Sci* 2021;**13**:36.
Lichacz P. Epistemic simplicity – a virtue or a vice? *Metaphilosophy* 2021;**52**:200–19.
Lim S, Prade H, Richard G. Classifying and completing word analogies by machine learning. *Int J Approx Reason* 2021;**132**:1–25.
Lin Z, Akin H, Rao R, *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**: 1123–30. https://doi.org/10.1126/science.ade2574
Lin Z, Feng M, Santos CNd, *et al.* A structured self-attentive sentence embedding. In: *Proceedings of the 5th International Conference on Learning Representations*. 2017.
Linnarsson S, Teichmann SA. Single-cell genomics: coming of age. *Genome Biol* 2016;**17**:97.
Liu C, Ma Y, Zhao J, *et al.* Computational network biology: data, models, and applications. *Phys Rep* 2020;**846**:1–66.
Liu T, Li K, Wang Y, *et al.* Evaluating the utilities of large language models in single-cell data analysis. bioRxiv, https://doi.org/2023.09.08.555192, 2023a, preprint: not peer reviewed. https://doi.org/10.1101/2023.09.08.555192, https://www.biorxiv.org/content/early/2023/11/10/2023.09.08.555192
Liu Z, Yu X, Zhang L, *et al.* DeID-GPT: zero-shot medical text de-identification by gpt-4. arXiv, https://doi.org/2303.11032, 2023b, preprint: not peer reviewed.
Low TY, Syafruddin SE, Mohtar MA, *et al.* Recent progress in mass spectrometry-based strategies for elucidating protein–protein interactions. *Cell Mol Life Sci* 2021;**78**:5325–39.
Luck K, Kim D-K, Lambourne L, *et al.* A reference map of the human binary protein interactome. *Nature* 2020;**580**:402–8.
Lugo-Martinez J, Zeiberg D, Gaudelet T, *et al.* Classification in biological networks with hypergraphlet kernels. *Bioinformatics* 2021; **37**:1000–7.
Macosko EZ, Basu A, Satija R, *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**:1202–14.
Madhobi KF, Kamruzzaman M, Kalyanaraman A, *et al.* A visual analytics framework for analysis of patient trajectories. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2019, 15–24.
Malod-Dognin N, Ceddia G, Gvozdenov M, *et al.* A phenotype driven integrative framework uncovers molecular mechanisms of a rare hereditary thrombophilia. *PLoS One* 2023;**18**:e0284084.
Malod-Dognin N, Petschnigg J, Windels SF, *et al.* Towards a data-integrated cell. *Nat Commun* 2019;**10**:805.
Malod-Dognin N, Pržulj N. Functional geometry of protein interactomes. *Bioinformatics* 2019;**35**:3727–34.
Marioni JC, Mason CE, Mane SM, *et al.* RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**:1509–17.

Mei L-C, Chen H-M, Dong A-Y, *et al.* Pesticide informatics platform (PIP): an international platform for pesticide discovery, residue, and risk evaluation. *J Agric Food Chem* 2022;**70**:6617–23.
Meltzer P, Mallea MDG, Bentley PJ. Pinet: a permutation invariant graph neural network for graph classification. arXiv, https://doi.org/1905.03046, 2019, preprint: not peer reviewed.
Menche J, Sharma A, Kitsak M, *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* 2015; **347**:1257601.
Méndez-Lucio O, Baillif B, Clevert D-A, *et al.* De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat Commun* 2020;**11**:10.
Meng C, Zeleznik OA, Thallinger GG, *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016;**17**:628–41.
Michaelis AC, Brunner A-D, Zwiebel M, *et al.* The social and structural architecture of the yeast protein interactome. *Nature* 2023; **624**:192–200.
Mihajlovic K, Ceddia G, Malod-Dognin N, *et al.* Multi-omics integration of scRNA-seq time series data predicts new intervention points for Parkinson's disease. bioRxiv, https://doi.org/2023.12.12.570554, 2023, preprint: not peer reviewed. https://doi.org/10.1101/2023.12.12.570554
Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. arXiv, https://doi.org/1301.3781, 2013a, preprint: not peer reviewed.
Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 2013b;**26**.
Mikolov T, Yih W-T, Zweig G. Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2013c, 746–51.
Milenković T, Pržulj N. Uncovering biological network function via graphlet degree signatures. *Cancer Inform* 2008;**6**:CIN.S680–S680.
Moreno P, Fexova S, George N, *et al.* Expression atlas update: gene and protein expression in multiple species. *Nucleic Acids Res* 2022; **50**:D129–40.
Morselli Gysi D, do Valle Í, Zitnik M, *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc Natl Acad Sci USA* 2021;**118**:e2025581118.
Mortazavi A, Williams BA, McCue K, *et al.* Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 2008;**5**:621–8.
Muscoloni A, Cannistraci CV. "Stealing fire or stacking knowledge" by machine intelligence to model link prediction in complex networks. *Iscience* 2023;**26**:105697.
Muscoloni A, Thomas JM, Ciucci S, *et al.* Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nat Commun* 2017;**8**:1615.
Nelson W, Zitnik M, Wang B, *et al.* To embed or not: network embedding as a paradigm in computational biology. *Front Genet* 2019; **10**:381.
Newman ME. The structure and function of complex networks. *SIAM Rev* 2003;**45**:167–256.
Nicolau M, Levine AJ, Carlsson G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc Natl Acad Sci USA* 2011;**108**:7265–70.
Nicolopoulou-Stamati P, Maipas S, Kotampasi C, *et al.* Chemical pesticides and human health: the urgent need for a new concept in agriculture. *Front Public Health* 2016;**4**:148.
Niedzwiecki MM, Walker DI, Vermeulen R, *et al.* The exposome: molecules to populations. *Annu Rev Pharmacol Toxicol* 2019; **59**:107–27.
Oršolić D, Pehar V, Šmuc T, *et al.* Comprehensive machine learning based study of the chemical space of herbicides. *Sci Rep* 2021; **11**:11479.
Oughtred R, Rust J, Chang C, *et al.* The BIOGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* 2021;**30**:187–200.
Papatheodorou I, Moreno P, Manning J, *et al.* Expression atlas update: from tissues to single cells. *Nucleic Acids Res* 2020;**48**:D77–83.
Patel K, Bhattacharyya P. Towards lower bounds on number of dimensions for word embeddings. In: *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2017, 31–6.
Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2014, 701–10.
Picard M, Scott-Boyer M-P, Bodein A, *et al.* Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* 2021;**19**:3735–46.
Pires DE, Stubbs KA, Mylne JS, *et al.* CROPCSM: designing safe and potent herbicides with graph-based signatures. *Brief Bioinform* 2022;**23**:bbac042.
Poleksic A. Hyperbolic matrix factorization improves prediction of drug-target associations. *Sci Rep* 2023;**13**:959.
Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. *Sci Adv* 2018;**4**:eaap7885.
Pržulj N. Biological network comparison using graphlet degree distribution. *Bioinformatics* 2007;**23**:e177–83.
Pržulj N. *Analyzing Network Data in Biology and Medicine: An Interdisciplinary Textbook for Biological, Medical and Computational Scientists*. Cambridge University Press, 2019.
Pržulj N, Corneil DG, Jurisica I. Modeling interactome: scale-free or geometric? *Bioinformatics* 2004;**20**:3508–15.
Przybyla L, Gilbert LA. A new era in functional genomics screens. *Nat Rev Genet* 2022;**23**:89–103.
Pugh TJ, Bell JL, Bruce JP, *et al.*; AACR Project GENIE Consortium, Genomics and Analysis Working Group. AACR project GENIE: 100,000 cases and beyond. *Cancer Discov* 2022;**12**:2044–57.
Qiu J, Dong Y, Ma H, *et al.* Network embedding as matrix factorization: unifying DeepWalk, LINE, PTE, and Node2vec. In: *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. 2018, 459–67.
Quackenbush J. Computational analysis of microarray data. *Nat Rev Genet* 2001;**2**:418–27.
Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* 2018;**46**:10546–62.
Reel PS, Reel S, Pearson E, *et al.* Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol Adv* 2021;**49**:107739.
Ritchie MD, Holzinger ER, Li R, *et al.* Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet* 2015; **16**:85–97.
Rogers A, Drozd A, Li B. The (too many) problems of analogical reasoning with word vectors. In: *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*. 2017, 135–48.
Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000;**290**:2323–6.
Segler MH, Kogej T, Tyrchan C, *et al.* Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent Sci* 2018;**4**:120–31.
Serrano MÁ, Krioukov D, Boguná M. Self-similarity of complex networks and hidden metric spaces. *Phys Rev Lett* 2008;**100**:078701.
Seshadhri C, Sharma A, Stolman A, *et al.* The impossibility of low-rank representations for triangle-rich complex networks. *Proc Natl Acad Sci USA* 2020;**117**:5631–7.
Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol* 2007;**3**:88.
Shervashidze N, Vishwanathan S, Petri T, *et al.* Efficient graphlet kernels for large graph comparison. In: *Artificial Intelligence and Statistics*. PMLR, 2009, 488–95.
Simonis N, Rual J-F, Carvunis A-R, *et al.* Empirically controlled mapping of the *Caenorhabditis elegans* protein–protein interactome network. *Nat Methods* 2009;**6**:47–54.
Singh G, Mémoli F, Carlsson GE, *et al.* Topological methods for the analysis of high dimensional data sets and 3D object recognition. In: Eurographics Symposium on Point-Based Graphics. 2007, 91–100.
Skinnider MA, Scott NE, Prudova A, *et al.* An atlas of protein-protein interactions across mouse tissues. *Cell* 2021;**184**:4073–89.e17.

Sompairac N, Nazarov PV, Czerwinska U, *et al.* Independent component analysis for unraveling the complexity of cancer omics datasets. *Int J Mol Sci* 2019;**20**:4414.

Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015;**16**:133–45.

Stein-O'Brien GL, Arora R, Culhane AC, *et al.* Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet* 2018;**34**:790–805.

Stelzl U, Worm U, Lalowski M, *et al.* A human protein–protein interaction network: a resource for annotating the proteome. *Cell* 2005;**122**:957–68.

Stumpf MP, Wiuf C, May RM. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci USA* 2005;**102**:4221–4.

Su C, Tong J, Zhu Y, *et al.* Network embedding in biomedical data science. *Brief Bioinform* 2020;**21**:182–97.

Sumathipala M, Maiorino E, Weiss ST, *et al.* Network diffusion approach to predict lncRNA disease associations using multi-type biological networks: LION. *Front Physiol* 2019;**10**:888.

Tang H-W, Spirohn K, Hu Y, *et al.* Next-generation large-scale binary protein interaction network for *Drosophila melanogaster*. *Nat Commun* 2023;**14**:2162–216.

Tang J, Deng C, Huang G-B. Extreme learning machine for multilayer perceptron. *IEEE Trans Neural Netw Learn Syst* 2015a;**27**:809–21.

Tang J, Qu M, Mei Q. PTE: predictive text embedding through large-scale heterogeneous text networks. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2015b, 1165–74.

Tang J, Qu M, Wang M, *et al.* LINE: large-scale information network embedding. In: *Proceedings of the 24th International Conference on World Wide Web*. 2015c, 1067–77.

Tenenbaum JB, de Silva V, Langford JC. Global geometric framework for nonlinear dimensionality reduction. *Science* 2000;**290**:2319–23.

Uetz P, Giot L, Cagney G, *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000;**403**:623–7.

Uffelmann E, Huang QQ, Munung NS, *et al.* Genome-wide association studies. *Nat Rev Methods Primers* 2021;**1**:59.

UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 2022;**51**:D523–31. https://doi.org/10.1093/nar/gkac1052, https://doi.org/10.1093/nar/gkac1052

Vacic V, Iakoucheva LM, Lonardi S, *et al.* Graphlet kernels for prediction of functional residues in protein structures. *J Comput Biol* 2010;**17**:55–72.

Vahabi N, Michailidis G. Unsupervised multi-omics data integration methods: a comprehensive review. *Front Genet* 2022;**13**:854752.

Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.

Vavasis SA. On the complexity of nonnegative matrix factorization. *SIAM J Optim* 2010;**20**:1364–77.

Velásquez-Zapata V, Elmore JM, Banerjee S, *et al.* Next-generation yeast-two-hybrid analysis with Y2H-SCORES identifies novel interactors of the MLA immune receptor. *PLoS Comput Biol* 2021;**17**:e1008890.

Venter CJ, Cohen D. The century of biology. *N Perspect Q* 2014;**31**:28–37.

Veselkov K, Gonzalez G, Aljifri S, *et al.* HYPERFOODS: machine intelligent mapping of cancer-beating molecules in foods. *Sci Rep* 2019;**9**:9237.

Vitali F, Marini S, Pala D, *et al.* Patient similarity by joint matrix trifactorization to identify subgroups in acute myeloid leukemia. *JAMIA Open* 2018;**1**:75–86.

Wall ME, Rechtsteiner A, Rocha LM. Singular value decomposition and principal component analysis. In: *A Practical Approach to Microarray Data Analysis*. Springer, 2003, 91–109.

Wang C, Chang P, Ary M, *et al.* PEBANK: a comprehensive database for protein engineering and design. *Biophys J* 2018a;**114**:411a.

Wang C, Xu C, Yao X, *et al.* Evolutionary generative adversarial networks. *IEEE Trans Evol Computat* 2019a;**23**:921–34.

Wang CY, Chang PM, Ary ML, *et al.* PROTBANK: a repository for protein design and engineering data. *Protein Sci* 2018b;**27**:1113–24.

Wang Q, Fang Y, Ravula A, *et al.* Deep partial multiplex network embedding. In: *Companion Proceedings of the Web Conference 2022*. 2022, 1053–62.

Wang R, Li S, Cheng L, *et al.* Predicting associations among drugs, targets and diseases by tensor decomposition for drug repositioning. *BMC Bioinformatics* 2019b;**20**:628–19.

Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.

Windels SF, Malod-Dognin N, Pržulj N. Graphlet Laplacians for topology-function and topology-disease relationships. *Bioinformatics* 2019;**35**:5226–34.

Windels SF, Malod-Dognin N, Pržulj N. Graphlet eigencentralities capture novel central roles of genes in pathways. *PLoS One* 2022;**17**:e0261676.

Wong L, You Z-H, Guo Z-H, *et al.* MIPDH: a novel computational model for predicting microRNA–mRNA interactions by DeepWalk on a heterogeneous network. *ACS Omega* 2020;**5**:17022–32.

Wu Y, Schuster M, Chen Z, *et al.* Google's neural machine translation system: bridging the gap between human and machine translation. arXiv, https://doi.org/1609.08144, 2016, preprint: not peer reviewed.

Xenos A, Malod-Dognin N, Milinković S, *et al.* Linear functional organization of the omic embedding space. *Bioinformatics* 2021;**37**:3839–47.

Xenos A, Malod-Dognin N, Zambrana C, *et al.* Integrated data analysis uncovers new COVID-19 related genes and potential drug repurposing candidates. *Int J Mol Sci* 2023;**24**:1431.

Xie Y, Peng J, Zhou Y. Integrating protein-protein interaction information into drug response prediction by graph neural encoding. Research Square, https://doi.org/rs.2.18936, 2019, preprint: not peer reviewed https://doi.org/10.21203/rs.2.18936/v1, https://doi.org/10.21203/rs.2.18936/v1

Xu M. Understanding graph embedding methods and their applications. *SIAM Rev* 2021;**63**:825–53.

Yang C, Xiao Y, Zhang Y, *et al.* Heterogeneous network representation learning: a unified framework with survey and benchmark. *IEEE Trans Knowl Data Eng* 2022a;**34**:4854–73.

Yang F, Wang W, Wang F, *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* 2022b;**4**:852–66.

Yang K, Wang R, Liu G, *et al.* HERGEPRED: heterogeneous network embedding representation for disease gene prediction. *IEEE J Biomed Health Inform* 2019;**23**:1805–15.

Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 2016;**32**:1–8.

Yasonik J. Multiobjective de novo drug design with recurrent neural networks and nondominated sorting. *J Cheminform* 2020;**12**:14.

Yin Z, Shen Y. On the dimensionality of word embedding. *Adv Neural Inf Process Syst* 2018;**31**.

Ying Z, You J, Morris C, *et al.* Hierarchical graph representation learning with differentiable pooling. *Adv Neural Inf Process Syst* 2018;**31**.

Young KH. Yeast two-hybrid: so many interactions,(in) so little time. *Biol Reprod* 1998;**58**:302–11.

Yu MK, Ma J, Ono K, *et al.* DDOT: a Swiss army knife for investigating data-driven biological ontologies. *Cell Syst* 2019;**8**:267–73.e3.

Yue X, Wang Z, Huang J, *et al.* Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* 2020;**36**:1241–51.

Zambrana C, Xenos A, Böttcher R, *et al.* Network neighbors of viral targets and differentially expressed genes in COVID-19 are drug target candidates. *Sci Rep* 2021;**11**:18985.

Zhang D, Qiao L. Microfluidics coupled mass spectrometry for single cell multi-omics. *Small Methods* 2024;**8**:e2301179.

Zhang Z, Cui P, Zhu W. Deep learning on graphs: a survey. *IEEE Trans Knowl Data Eng* 2022;**34**:249–70.

Zhao Y, Boeke JD. CRISPR–Cas12a system in fission yeast for multiplex genomic editing and CRISPR interference. *Nucleic Acids Res* 2020;**48**:5788–98.

Zhao Y, Zhang S, Xu J, *et al.* Spatial reconstruction of oligo and single cells by de novo coalescent embedding of transcriptomic networks. *Adv Sci* 2023;**10**:2206307.

Zhou J, Cui G, Hu S, *et al.* Graph neural networks: a review of methods and applications. *AI Open* 2020;**1**:57–81.

Zlobec I, Suter G, Perren A, *et al.* A next-generation tissue microarray (ngTMA) protocol for biomarker studies. *J Vis Exp* 2014;**91**:e51893.

Bioinformatics Advances, 2025, 00, 1–16
https://doi.org/10.1093/bioadv/vbae164
Perspective