Γεφυρώνοντας μεγάλα ομικά, γενετικά και ιατρικά δεδομένα για την ευρεία εφαρμογή της Ιατρικής Ακριβείας στην Ελλάδα

ΠΑΡΑΔΟΤΕΟ Π6.1

«Μοντέλα Μηχανικής Μάθησης για Ταξινόμηση Ασθενών και Πρόβλεψη Κινδύνου»

Φορέας	Πανεπιστήμιο Πατρών
Τύπος Παραδοτέου	Έκθεση
Τίτλος Παραδοτέου	Μοντέλα Μηχανικής Μάθησης για Ταξινόμηση Ασθενών και Πρόβλεψη Κινδύνου
Ενότητα Εργασίας	ΕΕ6: Ανάπτυξη Μοντέλων Μηχανικής Μάθησης για Εξόρυξη Γνώσης και Κατηγοριοποίησης Ασθενών

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Funded by the European Union
NextGenerationEU

Σύνοψη Ενότητας Εργασίας 6, Δράσης 6.1 Πανεπιστήμιο Πατρών

Παραδοτέο 6.1

Τίτλος Παραδοτέου: Μοντέλα Μηχανικής Μάθησης για Ταξινόμηση Ασθενών και Πρόβλεψη Κινδύνου

Στη δράση 6.1 αναπτύχθηκαν εργαλεία βασιζόμενα στην προσέγγιση της «Ιατρικής Δικτύων» (Network Medicine). Η κεντρική φιλοσοφία απορρίπτει τον βιολογικό αναγωγισμό, αντιμετωπίζοντας τις φαινοτυπικές εκδηλώσεις σύνθετων ασθενειών ως αναδυόμενες ιδιότητες διαταραγμένων βιολογικών δικτύων και όχι ως μεμονωμένα γονιδιακά συμβάντα. Η μεθοδολογία αναπτύσσεται σε τέσσερις βασικούς άξονες:

1) Ταξινόμηση Γονιδιακών Παραλλαγών (Nodes): Αναπτύχθηκε μια υπολογιστική διαδικασία για την ταξινόμηση Παραλλαγών Αβέβαιης Σημασίας (VUS) στα γονίδια BRCA1/2. Το μοντέλο αποφεύγει την κυκλικότητα (circularity) μη χρησιμοποιώντας κλινικά κριτήρια (ACMG-AMP) ως δεδομένα εισόδου, αλλά βασίζεται αποκλειστικά σε φυσικοχημικές ιδιότητες και την εξελικτική διατήρηση της πρωτεΐνης. Χρησιμοποιώντας έναν συνδυασμό Random Forests και Deep Learning (Wavelet-Attention ResNet-18), το μοντέλο επιτυγχάνει ακρίβεια 95%, εξασφαλίζοντας έγκυρα δεδομένα εισόδου για τα δίκτυα.

2) Ανάλυση Τοπολογίας Δικτύου (Edges): Εισάγεται η χρήση Κρυφών Μαρκοβιανών Μοντέλων (HMMs) για την ανίχνευση δομικών μοτίβων (motifs) σε θορυβώδη βιολογικά δίκτυα. Η μέθοδος επιτρέπει την «ασαφή ταύτιση» (soft matching), αναγνωρίζοντας λειτουργικά μοτίβα ακόμη και όταν υπάρχουν ελλιπή δεδομένα ή αβεβαιότητα στις αλληλεπιδράσεις πρωτεϊνών.

3) Έλεγχος Δικτύων (Network Logic): Για την επίλυση του προβλήματος του «επιλεκτικού ελέγχου» (στόχευση ασθενών κυττάρων χωρίς βλάβη στα υγιή), εφαρμόζεται Ακέραιος Γραμμικός Προγραμματισμός (ILP) σε δίκτυα Boolean. Η μέθοδος μετατρέπει τους λογικούς κανόνες σε γραμμικές ανισότητες, επιτρέποντας τον εντοπισμό ελάχιστων παρεμβάσεων για τη θεραπευτική τροποποίηση της συμπεριφοράς του δικτύου.

4) Συστημική Βιολογία & Μεταγραφωματική: Εφαρμόζονται τεχνικές Μηχανικής Μάθησης σε δεδομένα από το PsychENCODE. Χρησιμοποιούνται «Masked Denoising Autoencoders» που ενσωματώνουν βιολογική γνώση (αλληλεπιδράσεις μεταγραφικών παραγόντων) στην αρχιτεκτονική του νευρωνικού δικτύου, αποφεύγοντας τη λογική του «black box». Η ανάλυση ανέδειξε γονίδια κινδύνου όπως τα KDM1A και E2F1.

Συνολικά, το παραδοτέο αυτό παρέχει μια ολοκληρωμένη υπολογιστική εργαλειοθήκη που διατρέχει όλο το φάσμα της βιολογικής πολυπλοκότητας —από το αμινοξύ έως το δίκτυο— εξασφαλίζοντας έγκυρα δεδομένα εισόδου και ισχυρές μεθόδους ανάλυσης για τα στάδια της κλινικής μετάφρασης που ακολουθούν στην Δράση 6.2.

Deliverable 6.1: Machine Learning Models for Patient Classification and Risk Prediction

Emergent Properties and Network Medicine in GoMedPrecision

Anchored in the GoMedPrecision program, this research employs a "network medicine" approach to translate high-dimensional data into clinical diagnoses. The central philosophy underpinning this body of work is the systematic rejection of biological reductionism. The etiology of complex diseases—ranging from hereditary cancers and metabolic syndromes to neuropsychiatric disorders—cannot be fully elucidated by studying genes or metabolites in isolation. Instead, these phenotypes are viewed as emergent properties of perturbed biological networks. Consequently, we have developed a modular yet integrated computational pipeline that traverses the full spectrum of biological complexity: from single amino acid substitutions to the wiring of gene regulatory circuits and, ultimately, the stratification of human patients based on multiplex heterogeneous networks.

The portfolio is structured around methodological pillars: At the most fundamental level, the program addresses the "node" of the network. By integrating physicochemical properties and evolutionary conservation into machine learning architectures, we have developed tools to classify Variants of Uncertain Significance (VUS) in BRCA1/2 with high confidence. This ensures that the inputs for network models are accurate and clinically validated. Moving from nodes to edges, the program introduces novel mathematical formalisms to understand network topology and dynamics. The use of Hidden Markov Models (HMMs) to treat graph topologies as probabilistic sequences allows for the detection of "fuzzy" motifs in noisy biological data. Simultaneously, the application of Integer Linear Programming (ILP) to Boolean networks solves the "selective control" problem, providing a mathematical basis for targeting diseased cells while sparing healthy tissue.

These theoretical tools are subsequently applied to the complex transcriptomic landscape of neuropsychiatric disorders. We employ Masked Denoising Autoencoders and Explainable AI to prioritize risk genes. Crucially, these models are not "black boxes"; they are constrained by biological priors, ensuring that the learned features represent genuine regulatory mechanisms rather than statistical artifacts.

Genomic Variant Interpretation

Structure-Informed Machine Learning and Deep Learning for High-Confidence BRCA1/2 Missense Variant Classification

Context and Clinical Significance

The clinical management of hereditary breast and ovarian cancer is frequently hampered by the detection of Variants of Uncertain Significance (VUS) in the BRCA1 and BRCA2 genes. While truncating mutations are readily classified as pathogenic due

to their obvious impact on protein length and function, missense variants—where a single amino acid is substituted for another—often result in ambiguous clinical reports. This leaves patients and clinicians in a state of uncertainty regarding risk management strategies such as prophylactic surgery or enhanced surveillance. Existing computational predictors often suffer from circularity, utilizing clinical classifications as training features which leads to target leakage, and ascertainment bias, where models trained on global populations fail to recognize local founder mutations specific to certain ethnic or geographic groups. Our study addresses these critical gaps by developing a robust, structure-informed computational pipeline [6,7].

Dataset Description

To train robust machine learning models for classifying BRCA1 and BRCA2 variants, a comprehensive feature set was constructed using a wide variety of biologically relevant descriptors. A total of 977 missense variant sequences from BRCA1 and BRCA2 were retrieved from the ClinVar database [11]. ClinVar provides a massive, community-driven archive of human genetic variants with interpretations aggregated from submitters worldwide, offering broad coverage of established pathogenic and benign variants across diverse populations. Complementing this, CanVaS serves as a high-quality, population-specific reference database, documenting the germline genetic variation spectrum of Greek cancer patients with extensive phenotypic and segregation data [1]. This local cohort is particularly valuable for capturing founder mutations and population-specific variants that may be underrepresented globally. The dataset was constructed by integrating these two distinct sources to maximize both population specificity and global coverage. First, we incorporated the complete set of unique BRCA1/2 missense variants identified in the Greek population from the CanVaS cohort. To augment this dataset and ensure robust training, we then merged these with non-redundant missense variants retrieved from the global ClinVar database (v20250106). This merge process prioritized the high-confidence local annotations from CanVaS while using the broader ClinVar dataset to fill gaps, resulting in a consolidated final dataset of 977 unique variants.

Methodological Innovation

This work distinguishes itself by developing a strictly structure-informed, ab initio classification pipeline. We deliberately avoid using ACMG-AMP clinical criteria as input features to prevent the circularity inherent in models that learn human curation rules rather than biological truths. Instead, we engineer a rich feature set derived solely from the fundamental biophysics of the protein. The input for the models is a 51-residue peptide window centered on the mutation. Each residue is quantified using high-dimensional descriptors from the AAindex database, capturing properties such as hydrophobicity, steric volume, and charge. Furthermore, Position Specific Scoring Matrices (PSSMs) are generated via PSI-BLAST to capture the evolutionary constraints acting on each residue, distinguishing between neutral drift and pathogenic disruption by analyzing vertebrate proteomes [6].

The methodology is visually summarized in the pipeline diagram below (Fig. 1), which illustrates how physicochemical and evolutionary features are processed into matrices and vectors for different model architectures.

As shown in Figure 1, we employ a sophisticated dual-track modeling approach. "Classical" machine learning models, specifically Random Forests, are trained on vector-based sequence descriptors (Fig.1 III), while a Wavelet-Attention ResNet-18 deep learning model is trained on matrix-based physicochemical representations (Fig.1

I). The wavelet transform allows the neural network to capture both local residue interactions (high frequency) and domain-level structural stability (low frequency) by decomposing the physicochemical signal. This hybrid strategy leverages the interpretability of random forests with the feature extraction power of deep convolutional networks.

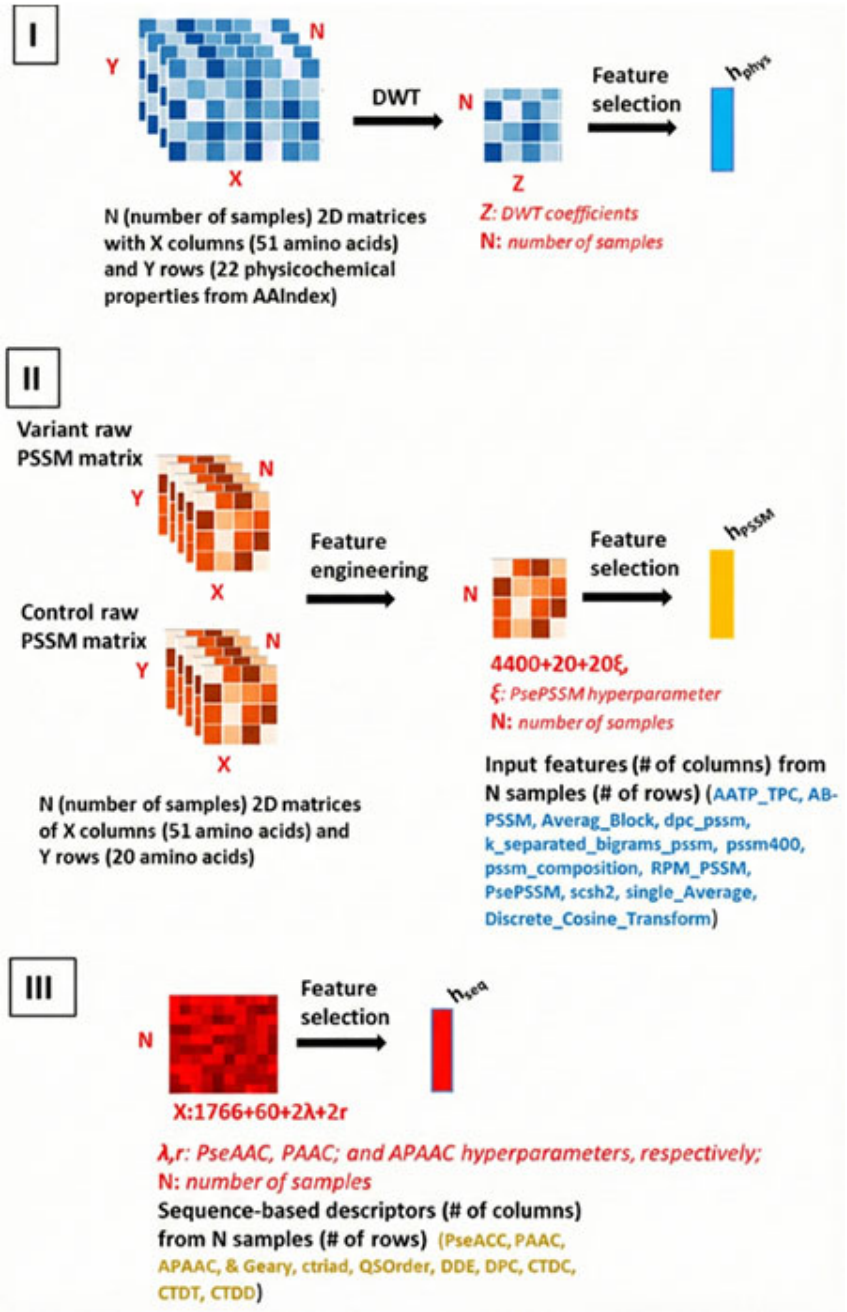

Figure 1: Diagram illustrating types and dimensions of input features used in the models. Physicochemical and evolutionary conservation features are represented as matrices (I, II), while sequence-based descriptors are encoded as vectors (III).

To handle the inherent class imbalance in the dataset—where benign variants often outnumber pathogenic ones—we utilized the Synthetic Minority Over-sampling Technique (SMOTE) within a nested pipeline approach. Crucially, this oversampling was performed exclusively on the training subset of each cross-validation fold to prevent data leakage, ensuring that the validation fold remained composed entirely of original, non-synthetic biological data. Feature selection was rigorously applied using four distinct methods: variance thresholding to remove low-variability features, mutual information analysis to select the top 300 most relevant features, correlation-based filtering to remove multicollinearity, and Principal Component Analysis (PCA) to capture 80% of the variance.

Results and Impact

We prioritized clinical safety by defining a "high-confidence" threshold, accepting only predictions where the model probability exceeded 80%. A voting ensemble constructed via AutoGluon, aggregating the best Random Forest and Deep Learning models, achieved a balanced accuracy of 95% with a pathogenic recall of 0.91. This indicates that the ensemble model is highly reliable in f lagging deleterious mutations, a critical requirement for clinical diagnostics where false negatives can be fatal. The model demonstrated an exceptional ability to filter benign variants with high precision and recall, which is essential for reducing the burden of unnecessary clinical surveillance.

Theoretical Network Mechanics

Network Motif Detection Using Hidden Markov Models

Context and Challenge

Biological networks are built from small, recurring subgraphs called motifs (e.g., feed-forward loops, bi-fans) that perform specific information-processing functions, such as filtering transient signals or generating pulse-like responses. Detecting these motifs usually involves "hard" graph isomorphism counting, which is computationally expensive and intolerant to noise. In biological data, where edges are often uncertain or missing due to experimental limitations like low confidence protein-protein interactions, strict isomorphism can miss functional motifs that are structurally imperfect. This report introduces a novel probabilistic method for detecting network motifs using Hidden Markov Models (HMMs), treating graph topology as a sequence data problem [2].

Methodological Innovation

We propose treating graph topology as a symbolic sequence. We employ a sliding window approach to linearize the adjacency matrix of the graph into symbolic strings representing the pattern of edges (activation, inhibition, no connection). An HMM is then trained to recognize specific motif patterns against a background model. The core advantage of this approach is "soft matching." Instead of a binary yes/no decision, the HMM outputs a log-likelihood score indicating how well a subgraph fits a motif profile. This allows the method to detect motifs even in the presence of noise or missing edges.

The process of generating these symbolic sequences is critical to the methodology and is depicted below. Figure 2 demonstrates how the adjacency matrix is traversed to create the input for the HMM. The HMM framework naturally incorporates edge weights, allowing for the detection of "intensity" motifs where the strength of interaction is as

important as the connectivity itself. The method includes a redundancy removal step to identify and discard isomorphic and automorphic variants before HMM scoring, which significantly reduces computational overhead. Significance is assessed by comparing motif frequencies and intensity scores against randomized null networks that preserve degree distribution and weight composition. We also detail different criteria for subgraph overlap, ranging from allowing shared nodes and edges to requiring completely disjoint subgraphs, offering flexibility depending on the specific biological question.

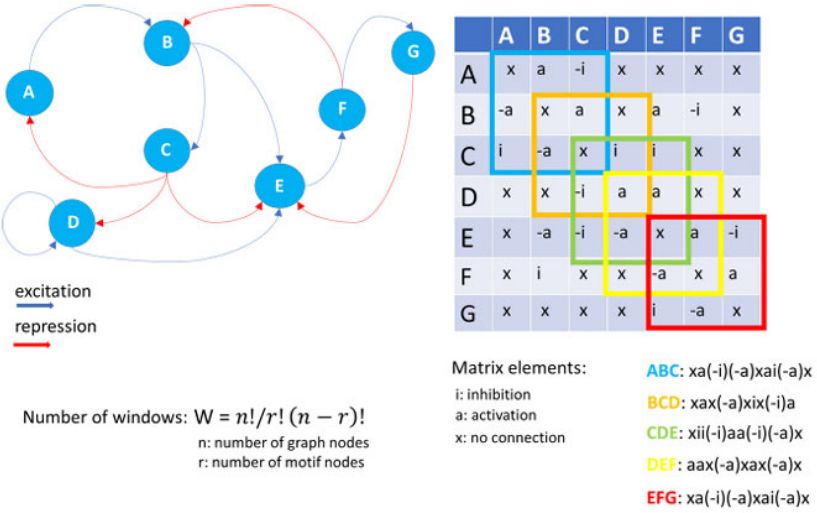

Figure 2: Applying a sliding window to generate candidate network motifs. Each window yields a symbolic subgraph string.

Key Findings

The HMM pipeline was validated on a 253-node directed random graph and successfully recovered known 4-node motifs such as the "tri-fan" and "bi-fan". A significant finding was the method's ability to distinguish between motifs that are overrepresented by count versus those that are overrepresented by intensity. For instance, the tri-fan motif (M1) was the most abundant by count with a Z-score of 16.68, yet its mean intensity was below expectation. Conversely, the bi-fan-plus-tail motif (M3) showed modest over-weighting. This nuance, distinguishing between structural abundance and functional intensity, is something that traditional binary counting methods would miss. We argue that this probabilistic framework is particularly well-suited for analyzing protein-protein interaction networks and social graphs where data uncertainty is high. This tool provides the GoMedPrecision program with a resilient method for structural analysis, allowing for the identification of functional building blocks in noisy biological networks. Areal-world case in biology is the human protein–protein interactome: Rolland et al. constructed a large-scale binary interactome but stressed that datasets remain incomplete and interactions have varying confidence across assays, so treating edges as strictly present/absent can miss biologically relevant patterns [5]. When searching for regulatory motifs (e.g., feed-forward loops, bi-fans), weak or missing edges cause exact binary counting to under-detect functional instances; using weighted or likelihood-based scoring (e.g., intensity/coherence or HMM-derived likelihood ratios) incorporates edge confidence and partial matches, enabling recovery of biologically meaningful motifs that binary counts would miss.

By Soft / weighted / likelihood-based detection, we mean replacing hard binary isomorphism tests with continuous scoring that:

1. Soft-matches — tolerates partial or noisy matches (missing/extra edges, label noise) and assigns scores reflecting degree-of-fit rather than accept/reject decisions;
2. Incorporates edge weights — uses edge confidences or strengths when computing motif significance (e.g., via intensity/coherence measures or weighted likelihoods); and
3. Uses probabilistic models — ranks or tests candidates by likelihoods or likelihood ratios produced by generative models (e.g., HMMs), enabling principled thresholds, p-value estimation, and model selection (AIC/BIC).

Integer Linear Programming for Contrasting State Interventions in Boolean Networks

Context and Challenge

A central challenge in pharmacology is selectivity: how to target a diseased cell (e.g., cancer) without harming a healthy cell, given that both share similar signaling networks. Traditional drug discovery targets individual molecules, but biological networks are robust and often rewire themselves to bypass blockade. This work proposes targeting the logic of the network itself. We address the "selective modulation" problem: finding a minimal set of nodes to perturb (knock out or overexpress) that forces a target node to a desired state in a disease network while leaving it unaffected in a healthy network.

Methodological Innovation

We present a mathematical framework based on Integer Linear Programming (ILP). We model gene regulatory networks as Boolean networks, where gene states are binary (0 or 1). The core innovation is the translation of Boolean logic rules into a system of linear inequalities. For example, an AND gate $Z = X \wedge Y$ is converted into constraints like $Z \leq X$, $Z \leq Y$, and $Z \geq X+Y-1$. This transformation allows us to leverage powerful ILP solvers to handle the combinatorial complexity of network interventions [4].

Key Findings

The method was validated on synthetic networks and a real-world Colitis-Associated Colon Cancer (CACC) network [9]. It successfully identified non-obvious intervention strategies, such as the combined modulation of SOCS, GSK3B, PTEN, P53, and IL10 to inhibit the proliferative node PROL in the cancer network without disrupting the healthy counterpart. In another case study involving metabolic KEGG networks, the method successfully identified node sets that could repress pyruvate production in one network variant while preserving it in another. The benchmarks demonstrated that the ILP-based approach offers superior scalability compared to brute-force EM enumeration, making it feasible to analyze larger, more complex biological networks. The EM verification step, while computationally intensive, scales polynomially with network size, avoiding the super-exponential explosion of full enumeration. This framework provides a rigorous mathematical foundation for precision medicine within the GoMedPrecision program, enabling the design of "smart" drug combinations that exploit the specific topological rewiring of disease states to achieve therapeutic selectivity.

Systems Biology & Transcriptomics

Exploration of Disease-Associated Gene Modules Using Graph Theory, Co Expression Networks, and Dimensionality Reduction

Context and Challenge

Complex disorders are often characterized by subtle perturbations across thousands of genes rather than single monogenic defects. Traditional differential expression analysis often fails to capture the coordinated nature of these perturbations. This study addresses the complexity of transcriptomics using the massive PsychENCODE dataset [10] as a model, aiming to identify coherent gene modules that drive disease phenotypes. We argue that relying on a single computational method can lead to biased results; therefore, we employ a triangulation strategy using three orthogonal approaches [8].

Methodological Approach

The study utilized RNA-seq data from 261 healthy controls and 153 patients. A critical component of the work was the rigorous preprocessing pipeline, which employed Surrogate Variable Analysis (SVA) to correct for latent batch effects. We specifically identified and corrected for artifacts arising from heterogeneous rRNA-depletion protocols (Ribo-Zero Gold vs. HMR), a technical confounder that could otherwise masquerade as biological signal, demonstrating the importance of handling hidden technical variation in large-scale datasets. The three analytical pipelines were designed to capture different aspects of network biology. First, a graph-theoretical approach used the igraph package to construct gene co-expression networks from differentially expressed genes (DEGs). By applying Prim's algorithm to generate Minimum Spanning Trees (MSTs), we reduced network density to reveal core topological structures and computed centrality measures (degree, betweenness, eigenvector) to identify "hub" genes. Second, Weighted Gene Co-expression Network Analysis (WGCNA) was used to define modules of highly correlated genes and relate them to the disease phenotype. We ensured the network approximated a scale-free topology by optimizing the soft-thresholding power ($R^2 > 0.9$). Third, Principal Component Analysis (PCA) provided a dimensionality reduction perspective, isolating genes that contributed most significantly to the variance separating patients from controls.

Key Findings and Biological Convergence

Each method revealed distinct but complementary pathological insights. The graph-based approach highlighted hub genes involved in synaptic signaling and mRNA export, identifying AKT3 as a critical regulatory node. AKT3 was identified as a central node in the pathogenic network but of secondary importance in controls, suggesting it as a potential therapeutic target with reduced off-target risks. WGCNA identified modules enriched for immune response and epigenetic regulation. The relationship between these modules was visualized to identify key regulatory hubs (Fig. 3).

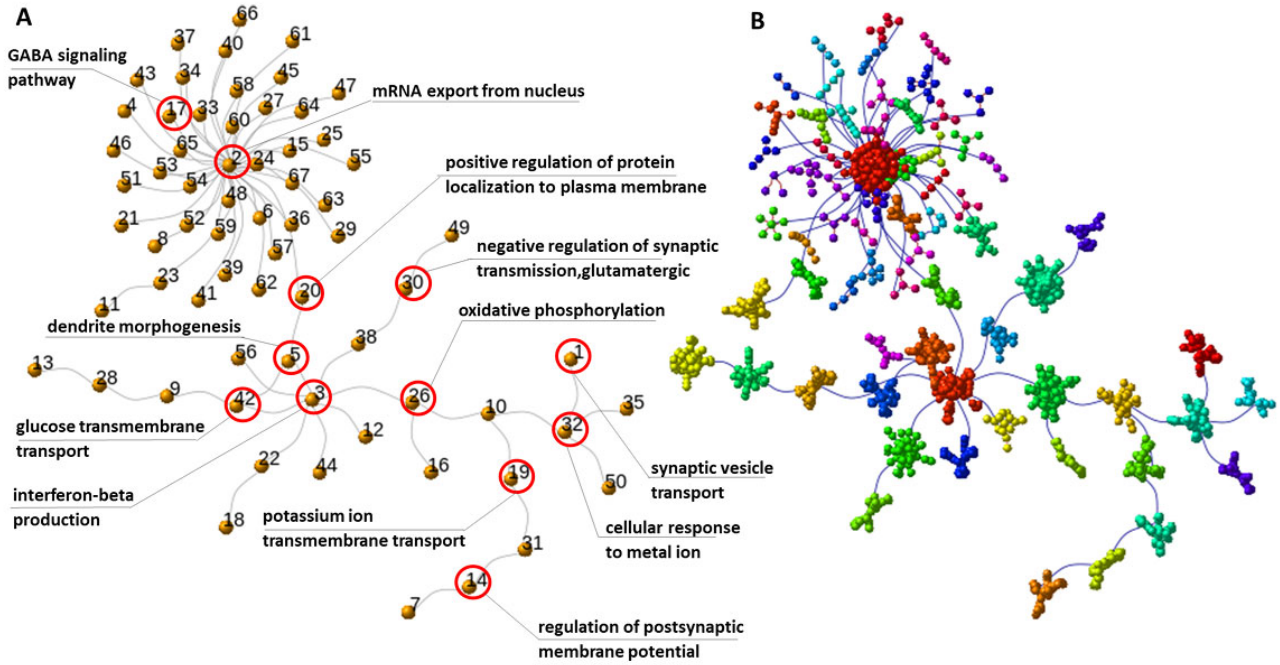

Figure 3: Network analysis in schizophrenia: We scaled the network for better visualization (B) and assigned a different community number to each module for enrichment analysis (A).

We propose a multi-dimensional ranking system for therapeutic targets that combines network topology, co-expression membership, and druggability. This study emphasizes the importance of using multiple analytical lenses to uncover robust biological signatures in complex disorders.

Consensus-Based Identification of Risk Genes Using Masked Denoising Autoencoders and Explainable Machine Learning

Context and Innovation

While the previous work focused on unsupervised module detection, this study introduces a supervised, deep learning approach to gene prioritization. We critique standard deep neural networks for their lack of interpretability and "black box" nature. To overcome this, we developed a "Masked Denoising Autoencoder" (DAE) architecture where the connections between layers are not fully connected but are instead masked by a binary matrix representing known Transcription Factor (TF) to Target Gene (TG) interactions. This biological constraint forces the neural network to learn features that are biologically plausible, effectively embedding a gene regulatory network within the neural architecture [3].

Methodological Framework

The pipeline was applied to ten balanced transcriptomic datasets from PsychENCODE [10] to ensure robustness against demographic confounders such as assay, sex, ethnicity, and age. We implemented a multi-stage training process. First, Restricted Boltzmann Machines (RBMs) were pre-trained with the TF-TG mask to initialize weights, capturing known regulatory links. These were then stacked into a Denoising

Autoencoder, which was trained to reconstruct gene expression profiles from corrupted inputs, forcing the model to learn robust feature representations. Finally, the encoder weights were transferred to a feedforward classifier. To ensure interpretability, we employed SHAP (SHapley Additive exPlanations) values to quantify the contribution of each gene to the model's predictions. This was complemented by an XGBoost classifier, which served as a high-performance baseline.

Figure 4 illustrates how the input layer is connected to the hidden layer via a biologically defined mask, followed by sparse layers that are pruned to retain only the most significant weights.

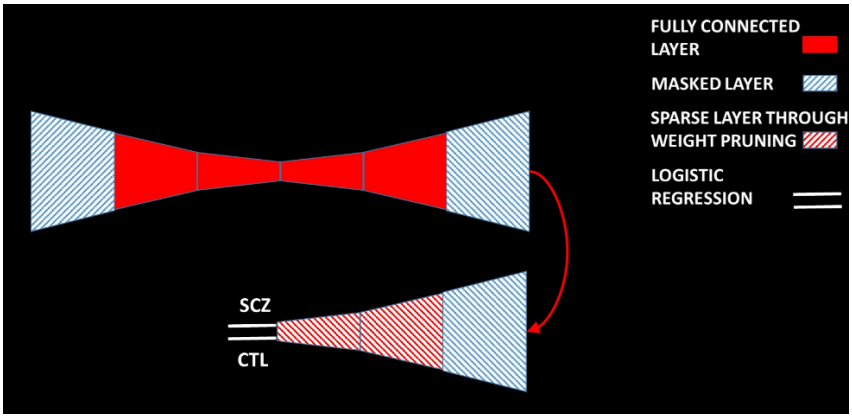

Figure 4: Sparsity after masking and pruning. In the masked layer the gradients can propagate only through specific weights defined by the applied mask. In the second and third layers sparsity is enforced by pruning the smallest weights.

Results and Integration

The sparse, biologically masked DAEs achieved classification accuracies comparable to fully connected networks (approximately 70–80%) but with far greater interpretability. The analysis consistently highlighted chromatin-remodeling factors, particularly KDM1A and E2F1, as top-ranked risk genes across multiple data splits and model architectures. Validation using protein protein interaction networks and regulatory motif analysis confirmed that these prioritized genes form coherent functional clusters. Specifically, E2F1 was found to participate in TF-miRNA motifs, linking transcriptional control to post-transcriptional regulation.

References

[1] Kalfakakou D., Fostira F., Papathanasiou A., Apostolou P., Dellatola V., Gavra I. E.,Vlachos I. S., Scouras Z. G., Drosopoulou E., Yannoukakos D., Konstantopoulou I. CanVaS:Documenting the genetic variation spectrum of Greek cancer patients. Human Mutation. 2021 Sep;42(9):1081–1093. doi:10.1002/humu.24249. PMID: 34174131.

[2] Bampos C., Megalooikonomou V. Network motif detection using Hidden Markov Models. Scientific Reports. 2025;15:41954. doi:10.1038/s41598-025-25936-y. PMID: 41290858.

[3] Bampos C., Megalooikonomou V. Consensus-Based Identification of Schizophrenia Risk Genes Using Masked Denoising Autoencoders and Explainable Machine Learning. Paper/Presentation (session S.4.1), Proc. IEEE 25th International Conference on Bioinformatics and BioEngineering (BIBE), Athens, Greece, 6–8 Nov 2025. See BIBE 2025 program (Athens, Nov 2025).

[4] Bampos C., Megalooikonomou V. (in press). Integer Linear Programming for Contrasting State Interventions in Boolean Networks. PeerJ. Accepted for publication (in press).

[5] T. Rolland, M. Ta¸san, and B. et. al. Charloteaux. A proteome-scale map of the human interactome network. Cell, 159(5):1212–1226, 2014.

[6] Kalfakakou D., Bampos C., Papathanasiou A., Fostira F., Apostolou P., Konstantopoulou I., Pavlopoulos G. A., Megalooikonomou V., Yannoukakos D. A neural network architecture approach for variant prioritization and annotation using Canvas, a population-specific cancer patient database, as a training set. Presented at: European Human Genetics Conference (ESHG) 2025. Milan, Italy, May 24–27, 2025. Presentation EP17.012.

[7] Bampos C., Megalooikonomou V. (under review). Structure-Informed Machine Learning and Deep Learning for High-Confidence BRCA1/2 Missense Variant Classification. Bioinforrmatics. Under Review.

[8] Bampos C., Megalooikonomou V. (under review). Exploration of Schizophrenia-Associated Gene Modules Using Graph Theory, Co-Expression Networks, and Dimensionality Reduction. Plos One. Under Review.

[9] Lu J, Zeng H, Liang Z, Chen L, Zhang L, Zhang H, Liu H, Jiang H, Shen B, Huang M, Geng M, Spiegel S, Luo C. 2015a. Network modeling reveals the mechanism underlying colitis-associated colon cancer and identifies novel combinatorial anticancer targets. Scientific Reports 5(1):14739 DOI 10.1038/srep14739.

[10] Akbarian, S., Liu, C., Knowles, J. et al. The PsychENCODE project. Nat Neurosci 18, 1707–1712 (2015). https://doi.org/10.1038/nn.4156

[11] M.J Landrum, J.M Lee, G.R Riley, et al. Clinvar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res, 42:D980–D985, 2014.