ΠΑΝΕΠΙΣΤΗΜΙΟ
ΙΩΑΝΝΙΝΩΝ

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΔΙΟΙΚΗΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΛΓΟΡΙΘΜΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗ ΣΥΣΧΕΤΙΣΗ ΓΟΝΙΔΙΑΚΩΝ ΕΚΦΡΑΣΕΩΝ

Εμμανουέλα Πάχου

Επιβλέπων: Πέτρος Καρβέλης

Αναπληρωτής Καθηγητής

Άρτα, Ιούλιος, 2026

MACHINE LEARNING ALGORITHMS FOR GENE EXPRESSION CORRELATION

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Άρτα, 03/07/2026

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπων καθηγητής

   Πέτρος Καρβέλης

2. Μέλος επιτροπής

   Νικόλαος Γιαννακέας

3. Μέλος επιτροπής

   Αλέξανδρος Τζάλλας

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα μεταπτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Επίθετο, Όνομα
Πάχου Εμμανουέλα

Υπογραφή

ΕΥΧΑΡΙΣΤΙΕΣ

Με αυτή την πτυχιακή εργασία, θα ήθελα να ευχαριστήσω όλους εκείνους που συνέβαλαν, άμεσα και έμμεσα, στην πραγματοποίησή της. Πρώτα απ' όλα, ευχαριστώ θερμά τον επιβλέποντα καθηγητή Κ. Πέτρο Καρβέλη για την συνεργασία και τον χρόνο που διέθεσε στην καθοδήγηση αυτής της εργασίας. Ιδιαίτερες ευχαριστίες οφείλω στην οικογένειά μου για την συμπαράσταση και την υποστήριξη που μου προσέφεραν καθ' όλη τη διάρκεια των σπουδών μου. Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου, που στάθηκαν στο πλευρό μου και με ενθάρρυναν να μην τα παρατάω ποτέ.

ΠΕΡΙΛΗΨΗ

Στόχος της παρούσας πτυχιακής εργασίας είναι η σχεδίαση και η δημιουργία μιας αυτόματης υπολογιστικής ροής (data pipeline) με σκοπό την εύρεση και την ιεράρχηση γονιδίων που σχετίζονται με τη Διαταραχή Αυτιστικού Φάσματος. Λόγω της μεγάλης πολυπλοκότητας των βιολογικών δεδομένων, η χρήση σύγχρονων μεθόδων ανάλυσης δικτύων κρίνεται απαραίτητη, καθώς οι παραδοσιακές στατιστικές μέθοδοι παρουσιάζουν σημαντικούς περιορισμούς.

Όσον αφορά τη μεθοδολογία, αναπτύχθηκε ένας προγραμματιστικός μηχανισμός που βασίζεται στον αλγόριθμο μάθησης αναπαράστασης γράφων node2vec. Ο αλγόριθμος εφαρμόστηκε πάνω στο μεγάλο βιολογικό δίκτυο BioSNAP και κατάφερε να μετατρέψει την τοπολογική δομή του δικτύου σε μαθηματικά διανύσματα (embeddings). Στη συνέχεια, μέσω της μετρικής του συνημίτονου ομοιότητας (Cosine Similarity), υπολογίστηκε η εγγύτητα αυτών των διανυσμάτων με τον κόμβο-στόχο του Αυτισμού.

Τα αποτελέσματα έδειξαν ότι το μοντέλο κατάφερε να κατατάξει στις πρώτες θέσεις γονίδια που είναι διεθνώς αναγνωρισμένα από την κλινική βιβλιογραφία για τη σύνδεσή τους με τον Αυτισμό, βασιζόμενο αποκλειστικά και μόνο στη δομή του δικτύου. Παράλληλα, η συγκριτική ανάλυση με εναλλακτικούς αλγορίθμους (DeepWalk, LINE) και κλασικούς δείκτες (Jaccard, Adamic-Adar) απέδειξε την ξεκάθαρη υπολογιστική υπεροχή και σταθερότητα του node2vec, ειδικά σε συνθήκες όπου τα βιολογικά δεδομένα είναι ελλιπή ή αραιά.

Συμπερασματικά, η εργασία αναδεικνύει την αξία του node2vec και παρουσιάζει μελλοντικές προοπτικές επέκτασης, όπως η μετάβαση σε Νευρωνικά Δίκτυα Γράφων (GNNs) και η ενσωμάτωση επιπλέον κλινικών και μοριακών δεδομένων (HPO, GTEx), ενισχύοντας την έρευνα στον τομέα της Ιατρικής Ακριβείας.

Λέξεις-κλειδιά: Ανάλυση Δικτύων, node2vec, Ομοιότητα Συνημίτονου, Αυτισμός, Βιοπληροφορική

ABSTRACT

The aim of this thesis is to design and create an automatic data pipeline for the discovery and prioritization of genes associated with Autism Spectrum Disorder. Due to the high complexity of biological data, the use of modern network analysis methods is considered necessary, as traditional statistical methods have significant limitations.

Regarding the methodology, a programming mechanism based on the node2vec graph representation learning algorithm was developed. The algorithm was applied on the large biological network BioSNAP and managed to transform the topological structure of the network into mathematical vectors (embeddings). Then, through the Cosine Similarity metric, the proximity of these vectors to the Autism target node was calculated.

The results showed that the model managed to rank genes that are internationally recognized by the clinical literature for their association with Autism in the first places, based solely on the network structure. At the same time, the comparative analysis with alternative algorithms (DeepWalk, LINE) and classic indicators (Jaccard, Adamic-Adar) demonstrated the clear computational superiority and stability of node2vec, especially in conditions where biological data are incomplete or sparse.

In conclusion, the work highlights the value of node2vec and presents future expansion prospects, such as the transition to Graph Neural Networks (GNNs) and the integration of additional clinical and molecular data (HPO, GTEx), enhancing research in the field of Precision Medicine.

Keywords: Network Analysis, node2vec, Cosine Similarity, Autism, Bioinformatics

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ/ΕΙΚΟΝΩΝ

1.Εισαγωγή

.

1.1 Υπολογιστικές προκλήσεις στην ανάλυση δεδομένων Αυτιστικού Φάσματος

Η μελέτη της Διαταραχής του Αυτιστικού Φάσματος (ASD) αποτελεί ένα από τα πιο πολύπλοκα προβλήματα στη σύγχρονη υπολογιστική βιολογία, κυρίως λόγω της πολυπαραγοντικής και ετερογενούς φύσης της. Η γενετική αρχιτεκτονική του αυτισμού περιγράφεται ως ένα ευρύ «αλληλομορφικό φάσμα», το οποίο αποτελείται από σπάνιες παραλλαγές σε εκατοντάδες γονίδια, αλλά και από έναν κοινό πολυγονιδιακό κίνδυνο σε χιλιάδες γενετικές θέσεις (loci) (Iakoucheva et al., 2019). Αυτό σημαίνει ότι η διαταραχή δεν οφείλεται σε ένα μεμονωμένο γονίδιο, αλλά προκύπτει από τη συνδυαστική δράση πολλών παραγόντων που αλληλοεπιδρούν μεταξύ τους.

Από προγραμματιστικής πλευράς, η δυσκολία στηρίζεται στο γεγονός ότι τα εμπλεκόμενα γονίδια παρουσιάζουν ισχυρή αλληλεξάρτηση σε επίπεδο μεταγραφικών και πρωτεϊνικών δικτύων. Τα γονίδια αυτά λειτουργούν ως ρυθμιστές της νευροανάπτυξης και των συναπτικών πρωτεϊνών (synaptic proteins), επηρεάζοντας συνολικά τη νευρική δραστηριότητα. Η ανάγκη να οριστούν διαφορετικοί υπότυποι (subtypes) του Αυτισμού με βάση τα χαρακτηριστικά των γονιδίων αναδεικνύει την ανάγκη υιοθέτησης καινοτόμων προσεγγίσεων στην ανάλυση δικτύων. Η χρήση των γράφων θεωρείται κρίσιμη για την κατανόηση των προβλημάτων που κρύβονται πίσω από τη διαταραχή, καθώς επιτρέπει την ταυτόχρονη μελέτη πολλών αλληλεπιδράσεων. Αυτές οι σύνθετες σχέσεις μεταξύ των γονιδίων συχνά δεν εντοπίζονται εύκολα με τις απλές στατιστικές μεθόδους που εξετάζουν κάθε στοιχείο ξεχωριστά.

1.2 Υπολογιστική Μοντελοποίηση Βιολογικών Συστημάτων και ο ρόλος της Επιστήμης Δεδομένων

Η εργασία βασίζεται στην παραδοχή ότι η βιολογία της Διαταραχής του Αυτιστικού Φάσματος μπορεί να αποτυπωθεί ως ένας ετερογενής γράφος G=(V,E), όπου V είναι το σύνολο των κόμβων (γονίδια, φαινότυποι) και E το σύνολο των αλληλεπιδράσεων (Barabási, 2016). Οι Zitnik et al. (2018) τονίζουν ότι η χρήση τέτοιων δικτύων είναι ο πιο αποτελεσματικός τρόπος για την ανάλυση σύνθετων παθήσεων. Αντί να εξετάζεται κάθε γονίδιο ξεχωριστά, το μοντέλο τα μελετά όλα μαζί ως ένα ενιαίο σύστημα. Αυτό επιτρέπει στον υπολογιστή να εντοπίζει «κρυφές» σχέσεις και μοτίβα που δεν είναι εμφανή με τις απλές στατιστικές μεθόδους, προσφέροντας μια βαθύτερη κατανόηση της βιολογίας του Αυτισμού.

Κεντρικό ρόλο σε αυτή τη μοντελοποίηση παίζει η αναγνώριση των hubs, δηλαδή των κόμβων εκείνων που διαθέτουν έναν υπερβολικά μεγάλο αριθμό συνδέσεων σε σχέση με τους υπόλοιπους (Barabási, 2016). Στην περίπτωση της ASD, η ύπαρξη αυτών των κεντρικών κόμβων υποδηλώνει ότι τα βιολογικά δίκτυα δεν είναι τυχαία (random graphs) αλλά ακολουθούν μια συγκεκριμένη οργάνωση ελεύθερης κλίμακας (scale-free) που στηρίζονται από τον νόμο της ισχύος (Barabási, 2016). Αυτό σημαίνει ότι τα περισσότερα γονίδια έχουν ελάχιστες συνδέσεις με άλλα γονίδια ή βιολογικά χαρακτηριστικά, ενώ ελάχιστα γονίδια-κλειδιά (hubs) συνδέονται με ένα τεράστιο πλήθος άλλων κόμβων. Συνεπώς, η υπολογιστική ανάδειξη των hubs είναι κρίσιμη, καθώς η θέση τους στο γράφο τους καθιστά «σημεία ελέγχου» για τη ροή της πληροφορίας. Μια λανθασμένη επεξεργασία ή ο αποκλεισμός αυτών των κόμβων κατά την προεπεξεργασία των δεδομένων μπορεί να οδηγήσει σε μεροληπτικά (biased) αποτελέσματα και απώλεια της δομικής πληροφορίας του δικτύου (Zitnik et al., 2018).

Επιπλέον, τα συστήματα αυτά χαρακτηρίζονται από την ιδιότητα του «μικρού κόσμου» (small world property). Σύμφωνα με τον Barabási (2016), η ιδιότητα σημαίνει ότι η απόσταση μεταξύ οποιωνδήποτε δύο κόμβων στο δίκτυο παραμένει πολύ μικρή, παρά το μεγάλο συνολικό πλήθος των γονιδίων. Στην περίπτωση του Αυτισμού, αυτό το φαινόμενο υποδηλώνει ότι μια τοπική γενετική διαταραχή δεν παραμένει απομονωμένη, αλλά μπορεί να διαδοθεί ταχύτατα σε ολόκληρο το βιολογικό σύστημα μέσω πολύ σύντομων διαδρομών.

Αυτή η «ταχύτατη διάχυση» καθιστά την πρόβλεψη συνδέσμων (Link Prediction) μία σύνθετη υπολογιστική διαδικασία, καθώς απαιτεί την κατανόηση του πώς μια αλλαγή σε έναν κόμβο επηρεάζει τη συνολική συνδεσιμότητα (Zitnik et al., 2018). Η μοντελοποίηση αυτή επιτρέπει επίσης την ανάλυση της τοπολογικής ανθεκτικότητας (robustness) του συστήματος. Η ανθεκτικότητα αναφέρεται στην ικανότητα του βιολογικού δικτύου να παραμένει λειτουργικό ακόμη και όταν ορισμένοι κόμβοι (γονίδια) παρουσιάζουν δυσλειτουργία. Το φαινόμενο αυτό μπορεί να παρομοιαστεί με τη λειτουργία ενός συγκοινωνιακού δικτύου. Η απώλεια δευτερευόντων κόμβων δεν διαταράσσει τη συνολική ροή, καθώς το σύστημα διαθέτει εναλλακτικές διαδρομές για τη μεταφορά της πληροφορίας. Αντίθετα, η στόχευση των κεντρικών κόμβων (hubs) προκαλεί την κατάρρευση της συνοχής του δικτύου. Στη μελέτη της ASD, αυτό σημαίνει ότι ενώ το βιολογικό σύστημα αντέχει σε μεμονωμένες μικρές αλλαγές, είναι εξαιρετικά ευάλωτο σε μεταλλάξεις που «βλάπτουν» τα κεντρικά ρυθμιστικά γονίδια (Barabási, 2016, Iakoucheva et al., 2019).

Η διαδικασία αυτή επιτρέπει τη μετατροπή της δομής του γράφου σε αριθμητικά δεδομένα τα οποία ο υπολογιστής μπορεί να επεξεργαστεί εύκολα. Μέσω αλγορίθμων που «εξερευνούν» τις συνδέσεις του δικτύου (τυχαίοι περίπατοι), καταγράφεται η σημασία και η γειτονιά κάθε κόμβου (Grover & Leskovec, 2016). Με αυτό τον τρόπο, γονίδια με παρόμοια θέση στο δίκτυο ομαδοποιούνται μεταξύ τους. Αυτή η στατιστική επεξεργασία βοηθά στην αναγνώριση διακριτών υποτύπων της ASD, προσφέροντας μια συνολική εικόνα για το πώς οργανώνεται η διαταραχή σε βιολογικό επίπεδο (Iakoucheva et al., 2019).

Στο πλαίσιο της Επιστήμης Δεδομένων, η Διαταραχή Αυτιστικού Φάσματος (ASD) δεν προσεγγίζεται πλέον με στατικές μεθόδους, αλλά αντιμετωπίζεται ως ένα δυναμικό πρόβλημα Μηχανικής Μάθησης, κυρίως μέσω τεχνικών Μη Επιβλεπόμενης (Unsupervised) ή Ημιεπιβλεπόμενης (Semi-supervised) Μάθησης. Κεντρικό ρόλο αυτής της μεθοδολογίας αποτελεί η Μάθηση Αναπαραστάσεων (Representation Learning). Η τεχνική αυτή λύνει ένα βασικό πρόβλημα καθώς οι γράφοι, αν και περιέχουν πλούσια πληροφορία, είναι δύσκολα διαχειρίσιμοι από τους κλασικούς αλγορίθμους. Μέσω της Μάθησης Αναπαραστάσεων, η «τοπολογική εγγύτητα» των κόμβων -δηλαδή το πόσο κοντά συνδέονται δύο γονίδια στον γράφο- μετατρέπεται αυτόματα σε «ευκλείδεια απόσταση» μέσα σε έναν διανυσματικό χώρο (Goodfellow et al., 2016).

Στην πράξη, αυτό σημαίνει ότι ο αλγόριθμος αναθέτει σε κάθε γονίδιο μια σειρά από αριθμούς (διάνυσμα). Αν δύο γονίδια έχουν παρόμοιο βιολογικό ρόλο ή γειτονεύουν στο

δίκτυο, οι αριθμοί τους θα είναι κοντινοί. Η διαδικασία αυτή στοχεύει στη διατήρηση της γειτονιάς (neighborhood preservation), διασφαλίζοντας ότι η βιολογική πληροφορία της ASD δεν χάνεται κατά τη μετατροπή από γράφο σε αριθμούς (Grover & Leskovec, 2016). Η σημασία αυτής της μετατροπής είναι καθοριστική καθώς μόλις τα δεδομένα αποκτήσουν διανυσματική μορφή, μπορούν να εισαχθούν σε κλασικούς αλγόριθμους ταξινόμησης

Ωστόσο, η υπολογιστική αυτή υπεροχή αποκτά ουσιαστικό νόημα μόνο όταν εφαρμόζεται στο πεδίο της σύγχρονης Βιοπληροφορικής. Η ικανότητα των μοντέλων να επεξεργάζονται χιλιάδες αλληλεπιδράσεις ταυτόχρονα, επιτρέπει τη μετάβαση από τη στατική μελέτη μεμονωμένων βιολογικών μονάδων στην ανάλυση δυναμικών δικτύων. Στην επόμενη ενότητα, αναλύεται πώς οι αρχές της θεωρίας των γράφων εφαρμόζονται στην πράξη για την αποκωδικοποίηση της πολυπλοκότητας της ASD, μετατρέποντας τον τεράστιο όγκο δεδομένων σε αξιοποιήσιμη βιολογική γνώση.

Η σύγχρονη Βιοπληροφορική έχει αλλάξει ολοκληρωτικά, περνώντας από την απλή αρχειοθέτηση βιολογικών δεδομένων στην ενεργή και αυτοματοποιημένη εξόρυξη προτύπων (pattern mining). Η μετάβαση αυτή υπαγορεύεται από την ανάγκη διαχείρισης του τεράστιου όγκου δεδομένων που παράγουν οι τεχνολογίες υψηλής απόδοσης, όπου η πληροφορία δεν είναι πλέον γραμμική αλλά δικτυακή. Σύμφωνα με τους Zitnik et al. (2018) , η ανάλυση γράφων αποτελεί σήμερα το πλέον αποτελεσματικό πλαίσιο για τη μοντελοποίηση σύνθετων παθήσεων, καθώς επιτρέπει την ταυτόχρονη εξέταση χιλιάδων βιολογικών οντοτήτων και των μεταξύ τους αλληλεπιδράσεων.

Κεντρικό ρόλο σε αυτή την εξέλιξη παίζει η χρήση πολυτροπικών γράφων (multimodal graphs), οι οποίοι στην προκειμένη περίπτωση λαμβάνουν τη μορφή ενός διμερούς δικτύου (bipartite network). Σε αντίθεση με τους ομογενείς γράφους, οι πολυτροπικές δομές επιτρέπουν την ταυτόχρονη αναπαράσταση ετερογενών δεδομένων -όπως οι αλληλεπιδράσεις γονιδίων και ασθενειών- σε ένα ενιαίο υπολογιστικό σχήμα. Όπως επισημαίνουν οι Zitnik et al. (2018), αυτή η οργάνωση επιτρέπει στον υπολογιστή να «μοιράζεται» πληροφορίες ανάμεσα σε διαφορετικά είδη κόμβων (ασθένειες και γονίδια), αποκαλύπτοντας «κρυφές» σχέσεις που θα παρέμεναν αόρατες αν εξετάζονταν μεμονωμένα. Το σύνολο δεδομένων BioSNAP λειτουργεί ως ένας τέτοιος ψηφιακός χάρτης, συνδέοντας τη βιολογική πληροφορία με την υπολογιστική ανάλυση.

Προκειμένου ο υπολογιστής να είναι σε θέση να επεξεργαστεί αυτόν τον πολύπλοκο χάρτη, απαιτείται η διαδικασία της Μάθησης Αναπαραστάσεων (representation learning). Η έρευνα των Hamilton et al. (2017) προσέφερε τα απαραίτητα εργαλεία για τη μετατροπή αυτών των δικτύων σε αριθμητικές μορφές που οι αλγόριθμοι μπορούν να κατανοήσουν. Η μέθοδος αυτή βασίζεται στη λογική της «γειτονιάς» : για να οριστεί η λειτουργία ενός γονιδίου, ο υπολογιστής εξετάζει ποιες ασθένειες συνδέεται και ποια άλλα γονίδια παρουσιάζουν παρόμοια μοτίβα συνδεσιμότητας. Το συγκεκριμένο σύστημα είναι εξαιρετικά ευέλικτο, καθώς επιτρέπει την εξαγωγή συμπερασμάτων ακόμα και για νέα δεδομένα που δεν υπήρχαν κατά την αρχική εκπαίδευση του μοντέλου.

Επιπλέον, η χρήση των γράφων επιλύει σημαντικά ζητήματα υπολογιστικής ταχύτητας. Οι παραδοσιακές βάσεις δεδομένων (SQL) παρουσιάζουν περιορισμούς όταν καλούνται να εντοπίσουν σύνθετες σχέσεις ανάμεσα σε χιλιάδες στοιχεία, καθώς απαιτούν χρονοβόρες και δαπανηρές διαδικασίες συνένωσης πινάκων. Αντίθετα, οι δομές γράφων είναι σχεδιασμένες για την ταχεία πλοήγηση μέσα στο δίκτυο, επιτρέποντας την άμεση πρόσβαση στη «γειτονιά» κάθε κόμβου. (Zitnik et al, 2018).

Συμπερασματικά, η ανάλυση γράφων στη Βιοπληροφορική εισάγει μία νέα οπτική στην επιστήμη, καθώς η γνώση δεν αναζητείται πλέον σε μεμονωμένα στοιχεία, αλλά στον τρόπο που αυτά συνδέονται μεταξύ τους. Η στρατηγική αυτή επιτρέπει τη χρήση μοντέλων Link Prediction, τα οποία λειτουργούν ως ένας «έξυπνος μηχανισμός προτάσεων» για γενετικές συνδέσεις που δεν έχουν ακόμη επιβεβαιωθεί επίσημα. Μέσω αυτής της μεθοδολογίας, καθίσταται εφικτή η «χαρτογράφηση» του Αυτισμού όχι ως μια απομονωμένη πάθηση, αλλά ως μέρος ενός ευρύτερου γενετικού δικτύου. Με αυτό τον τρόπο, έρχονται στο φως σημαντικές πληροφορίες για τη δομή της διαταραχής, οι οποίες μέχρι σήμερα παρέμεναν κρυμμένες λόγω του τεράστιου όγκου των δεδομένων.

1.3 Σκοπός και Υπολογιστικοί Στόχοι: Ανάπτυξη μοντέλου Candidate Gene Ranking

Ο κεντρικός σκοπός της παρούσας εργασίας είναι η σχεδίαση και η υλοποίηση ενός έξυπνου υπολογιστικού μοντέλου, το οποίο θα έχει την ικανότητα να εντοπίζει νέες, άγνωστες έως

σήμερα βιολογικές συνδέσεις στο δίκτυο της Διαταραχής Αυτιστικού Φάσματος. Η μελέτη αυτή δεν περιορίζεται στην απλή καταγραφή των ήδη γνωστών δεδομένων, αλλά επιδιώκει να χρησιμοποιήσει την τεχνητή νοημοσύνη για να «προβλέψει» ποιες βιολογικές οντότητες, κυρίως γονίδια, σχετίζονται με συγκεκριμένες παθήσεις του φάσματος, ακόμη και αν αυτή η σχέση δεν έχει επιβεβαιωθεί ακόμα στο εργαστήριο. Αυτή η διαδικασία πρόβλεψης αποτελεί το επίκεντρο της έρευνας και στοχεύει στην αποκάλυψη της κρυφής αρχιτεκτονικής που συνδέει τις διαφορετικές οντότητες της διαταραχής.

Για την επίτευξη αυτού του σκοπού, η εργασία βασίζεται στην αξιοποίηση της βάσης δεδομένων BioSNAP, η οποία αποτελεί την κεντρική πηγή δεδομένων. Το συγκεκριμένο αρχείο επιτρέπει την κατασκευή ενός διμερούς δικτύου, όπου οι κόμβοι αντιπροσωπεύουν ασθένειες και γονίδια, ενώ οι ακμές μεταξύ τους αποτυπώνουν τις ήδη καταγεγραμμένες ιατρικές συσχετίσεις. Σύμφωνα με τους Zitnik et al. (2018), όπως προαναφέρθηκε, η μοντελοποίηση αυτή επιτρέπει στον υπολογιστή να εντοπίζει μοτίβα που συνδέουν διαφορετικές παθήσεις μέσω κοινών γενετικών παραγόντων.

Η υπολογιστική διαδικασία βασίζεται στην εφαρμογή του αλγορίθμου node2vec, ο οποίος επεξηγείται αναλυτικά στο κεφάλαιο 2.5. Τελικός στόχος είναι η εκπαίδευση του μοντέλου να προτείνει νέες συνδέσεις: αν ένα γονίδιο παρουσιάζει παρόμοια δικτυακή συμπεριφορά με γονίδια που ήδη σχετίζονται με τον Αυτισμό, το σύστημα το επισημαίνει ως ένα πιθανό νέο στόχο για έρευνα. Η διαδικασία ολοκληρώνεται με την οπτική αναπαράσταση των αποτελεσμάτων και τη χρήση μετρικών αξιολόγησης, μετατρέποντας τα στατικά δεδομένα του BioSNAP σε ένα δυναμικό εργαλείο πρόβλεψης που υποδεικνύει νέα μονοπάτια για τη μελέτη και την κατανόηση της γενετικής βάσης της διαταραχής.

1.4 Δομή της εργασίας

Η παρούσα πτυχιακή εργασία αποτελείται από έξι κεφάλαια, καθένα από τα οποία καλύπτει ένα συγκεκριμένο στάδιο της έρευνας, από τη θεωρητική θεμελίωση και τη μελέτη των δεδομένων μέχρι την υλοποίηση και την τελική αξιολόγηση του μοντέλου.

Στο Κεφάλαιο 2, πραγματοποιείται η βιβλιογραφική ανασκόπηση και η θεωρητική θεμελίωση της μελέτης. Αναλύονται οι έννοιες της Θεωρίας Γράφων, οι τεχνικές Μηχανικής

Μάθησης σε δίκτυα και η διαδικασία της Μάθησης Αναπαραστάσεων. Ιδιαίτερη έμφαση δίνεται στη μαθηματική ανάλυση του αλγορίθμου node2vec, ο οποίος αποτελεί τη βάση της υπολογιστικής προσέγγισης.

Στο Κεφάλαιο 3, παρουσιάζονται τα δεδομένα και η υπολογιστική υποδομή. Περιγράφεται αναλυτικά το σύνολο δεδομένων BioSNAP, η διαδικασία προεπεξεργασίας των αρχείων και η μοντελοποίηση του δικτύου σε περιβάλλον Python, με στόχο τη δημιουργία ενός λειτουργικού ψηφιακού γράφου.

Στο Κεφάλαιο 4, αναλύεται η μεθοδολογία και η υλοποίηση της Μηχανικής Μάθησης. Περιγράφεται η αρχιτεκτονική του συστήματος, η παραμετροποίηση του αλγορίθμου node2vec στα παραγόμενα διανύσματα για την πρόβλεψη νέων συνδέσεων.

Στο Κεφάλαιο 5, παρουσιάζονται τα πειραματικά αποτελέσματα. Μέσα από τις ποσοτικές αναλύσεις και τεχνικές οπτικοποίησης (όπως ο αλγόριθμος t-SNE), αξιολογείται η ποιότητα των αποτελεσμάτων και η ικανότητα του συστήματος να αναγνωρίζει τοπολογικά πρότυπα που σχετίζονται με τον Αυτισμό.

Η εργασία ολοκληρώνεται στο Κεφάλαιο 6, όπου πραγματοποιείται η τελική αξιολόγηση και συζήτηση των ευρημάτων. Συνοψίζοντας τα συμπεράσματα της μελέτης, επισημαίνονται οι περιορισμοί του μοντέλου και προτείνονται κατευθύνσεις για μελλοντική έρευνα στον τομέα της υπολογιστικής Βιοπληροφορικής.

2.Βιβλιογραφική Ανασκόπηση

2.1 Υπολογιστικές βάσεις δεδομένων και σχήματα δεδομένων

Η σύγχρονη βιοϊατρική έρευνα παράγει καθημερινά έναν τεράστιο όγκο δεδομένων, η διαχείριση των οποίων απαιτεί τη χρήση εξειδικευμένων υπολογιστικών υποδομών. Προκειμένου η πληροφορία που αφορά γονίδια και ασθένειες να θεωρηθεί αξιοποιήσιμη από αλγόριθμους Μηχανικής Μάθησης, πρέπει πρώτα να κωδικοποιηθεί σε πρότυπα σχήματα δεδομένων μέσω διεθνώς αναγνωρισμένων βάσεων. Στην παρούσα εργασία, ανάλυση βασίζεται σε δεδομένα που αντλούνται και τυποποιούνται από δύο κεντρικές βάσεις: το NCIBI (National Center for Biotechnology Information) και το σύστημα UMLS (Unified Medical Language System).

Το NCBI αποτελεί τον θεμελιώδη οργανισμό που εξειδικεύεται στην διαχείριση πληροφοριών που αφορούν γονίδια. Μέσω της υπηρεσίας Entrez Gene, αποδίδεται σε κάθε γονίδιο ένας μοναδικός ψηφιακός κωδικός, γνωστός ως Gene ID (Maglott et al. 2010). Η αυτής της κωδικοποίησης είναι κρίσιμη, καθώς η ονοματολογία των γονιδίων στη βιβλιογραφία συχνά παρουσιάζει συνωνυμίες ή ασάφειες που δυσκολεύουν την υπολογιστική επεξεργασία. Ο Gene ID λειτουργεί ως μία «ψηφιακή ταυτότητα» που επιτρέπει στον υπολογιστή να αναγνωρίζει με απόλυτη ακρίβεια τη βιολογική οντότητα, διευκολύνοντας τη διασύνδεση ετερογενών αρχείων δεδομένων, όπως το σύνολο BioSNAP.

Εκτός από την ταυτοποίηση των γονιδίων, είναι απαραίτητα να οργανωθούν σωστά και οι ιατρικοί όροι που αφορούν τις ασθένειες. Επειδή στην επιστημονική βιβλιογραφία μία πάθηση μπορεί να αναφέρεται με διαφορετικές ονομασίες, χρησιμοποιούνται ειδικά συστήματα (οντολογίες) που λειτουργούν ως «μεταφραστές», ενοποιώντας όλη αυτή την ορολογία κάτω από κοινούς κωδικούς. Το σύστημα UMLS, το οποίο συντηρείται από την Εθνική Βιβλιοθήκη Ιατρικής των ΗΠΑ (NLM), λειτουργεί ως ένας κεντρικός κόμβος (metathesaurus) που συνδέει πάνω από 200 διαφορετικά συστήματα ταξινόμησης, όπως το MeSH και το SNOMED CT (Bodenreider, 2004). Η κεντρική μονάδα του UMLS είναι ο κωδικός CUI (Concept Unique Identifier). Μέσω του CUI, κάθε πάθηση - συμπεριλαμβανομένης της Διαταραχής Αυτιστικού Φάσματος- αποκτά έναν μοναδικό αριθμό ταυτοποίησης. Όπως τεκμηριώνεται στη μελέτη του μοντέλου Decagon από τους

Zitnik et al. (2018), η χρήση των κωδικών CUI είναι απαραίτητη για τη δημιουργία αυτών των σύνθετων δικτύων. Χάρη σε αυτούς, ο υπολογιστής μπορεί να συνδέει τις εξωτερικές ενδείξεις μίας ασθένειας (τα συμπτώματα) με τα γονίδια, εξετάζοντάς τα όλα μαζί ως ένα ενιαίο σύστημα.

Η επιτυχής μετατροπή των πληροφοριών από τις προαναφερθείσες βάσεις σε έναν λειτουργικό γράφο προϋποθέτει την οργάνωσή τους σε ένα ενιαίο σχήμα δεδομένων (data schema). Το σχήμα αυτό λειτουργεί ως ένας δομημένος «χάρτης» που καθορίζει τον τρόπο σύνδεσης των γονιδίων με τις ασθένειες. , Με τη χρήση των μοναδικών κωδικών ταυτοποίησης από το NCBI και το UMLS, επιτυγχάνεται η ένωση ετερογενών πηγών πληροφορίας, εξαλείφοντας την πιθανότητα διπλότυπων εγγραφών ή σημασιολογικών συγχύσεων. Η συγκεκριμένη οργάνωση των δεδομένων αποτελεί το βασικότερο στάδιο της διαδικασίας, καθώς διαμορφώνει τη σταθερή ψηφιακή βάση πάνω στην οποία οικοδομείται ο γράφος και εφαρμόζονται στη συνέχεια οι αλγόριθμοι Μηχανικής Μάθησης.

2.2 Θεωρία Γράφων και Δικτύων: Μετρικές και Τοπολογία

Η ανάλυση του βιολογικού δικτύου ASD εστιάζει στην ποσοτική αξιολόγηση της δομής μέσω του Πίνακα Γειτνίασης (Adjacency Matrix) A. Στην περίπτωση διμερών (bipartite) δικτύων, όπου οι ακμές συνδέουν αποκλειστικά γονίδια με ασθένειες, η μαθηματική αναπαράσταση του συστήματος βασίζεται στον Πίνακα Πρόσπτωσης (Incidence Matrix) B. Σύμφωνα με τον μία χαρακτηριστική δομή μπλοκ (block structure) της μορφής:

$$A = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$$

Η διάταξη αυτή διασφαλίζει ότι οι αλληλεπιδράσεις περιορίζονται αποκλειστικά μεταξύ των διαφορετικών συνόλων κόμβων (γονιδίων και ασθενειών). Όπως επισημαίνουν οι Zitnik et al. (2019), η διατήρηση αυτής της εσωτερικής δομής κατά την οργάνωση των δεδομένων σε διακριτούς πίνακες (matrices) είναι κρίσιμη για την αποφυγή απώλειας πληροφορίας, ενώ η αραιή (sparse) μορφή του πίνακα επιτρέπει την αποδοτική επεξεργασία μεγάλων βιολογικών δεδομένων.

Η σπουδαιότητα κάθε κόμβου στο δίκτυο προσδιορίζεται μέσω των Μετρικών Κεντρικότητας (Centrality Indices). Σύμφωνα με τον Brandes (2001), οι δείκτες αυτοί δείχνουν πόσο σημαντικό είναι κάθε γονίδιο και κάθε ασθένεια μέσα στο δίκτυο, ανάλογα με το που βρίσκονται. Με αυτόν τον τρόπο, αναδεικνύονται εύκολα οι κόμβοι που έχουν τον πιο κρίσιμο ρόλο στη συνολική δομή. Πέρα από την απλή καταμέτρηση του πλήθους των συνδέσεων, ιδιαίτερη σημασία έχει η μετρική διαμεσολάβησης (Betweenness Centrality), η οποία μετρά πόσο συχνά ένας κόμβος λειτουργεί ως «γέφυρα» για τη ροή της πληροφορίας. Μαθηματικά, για έναν κόμβο , v ορίζεται ως:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Όπου σ_{st} είναι ο συνολικός αριθμός των συντομότερων μονοπατιών μεταξύ των κόμβων s και t, και $\sigma_{st}(v)$ ο αριθμός αυτών των μονοπατιών που διέρχονται από τον κόμβο v (Ma & Tang, 2021). Οι κόμβοι με υψηλή διαμεσολάβηση λειτουργούν ως κεντρικοί σύνδεσμοι μέσα στο δίκτυο. Όπως αναφέρεται από τους Zitnik et al. (2019), ο εντοπισμός αυτών των κρίσιμων γονιδίων (seed genes) είναι πολύ σημαντικός, καθώς η θέση τους βοηθά στο να παραμένει το βιολογικό σύστημα σταθερό και λειτουργικό, ακόμη και αν υπάρξουν γενετικές διαταραχές στο περιβάλλον του.

Τέλος, η τάση των κόμβων να δημιουργούν κλειστές ομάδες i με τους γείτονές τους υπολογίζεται μέσω του Τοπικού Συντελεστή Ομαδοποίησης (Local Clustering Coefficient). Σύμφωνα με τον Barabási (2016), για έναν κόμβο που έχει k_i συνδέσεις, ο συντελεστής C_i υπολογίζει την πυκνότητα των σχέσεων στο άμεσο περιβάλλον του και ορίζεται ως:

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

Όπου L_i αντιπροσωπεύει τον αριθμό των ακμών που υπάρχουν πραγματικά μεταξύ τον γειτόνων του συγκεκριμένου κόμβου.

Στα δίκτυα ελεύθερης κλίμακας, ο συντελεστής ομαδοποίησης εμφανίζει μία αντίστροφη εξάρτηση από τον βαθμό του κόμβου, η οποία ακολουθεί την κατανομή $C(k) \sim k^{-1}$. Η μαθηματική αυτή συμπεριφορά αποδεικνύει την ύπαρξη μια ιεραρχικής αρχιτεκτονικής, όπου οι κόμβοι με μικρό βαθμό οργανώνονται σε ιδιαίτερα πυκνές τοπικές υποομάδες.

Όπως τεκμηριώνεται από τους Barabási et al. (2011), στο πλαίσιο των σύνθετων ανθρώπινων ασθενειών, αυτές οι πυκνά συνδεδεμένες υποομάδες αναγνωρίζονται ως λειτουργικές ενότητες (functional modules) ή ενότητες νόσου (disease modules). Μέσα σε αυτές τις ομάδες, τα γονίδια δεν λειτουργούν απομονωμένα, αλλά συνεργάζονται στενά μεταξύ τους για να εκδηλωθούν τα χαρακτηριστικά του Αυτισμού. Αντίθετα, τα γονίδια-κλειδιά που έχουν πάρα πολλές συνδέσεις, δεν ανήκουν αποκλειστικά σε μία ομάδα· ο ρόλος τους είναι να λειτουργούν ως «γέφυρες» που ενώνουν τις διαφορετικές βιολογικές ομάδες μεταξύ τους. Η γνώση αυτής της δομής του δικτύου είναι απαραίτητη, καθώς πάνω σε αυτή την αρχιτεκτονική βασίζεται η εκπαίδευση των αλγορίθμων Μηχανικής Μάθησης που εξετάζονται παρακάτω.

2.3 Μηχανική Μάθηση σε Γράφους

Η παραδοσιακή Μηχανική Μάθηση έχει σχεδιαστεί κατά κύριο λόγο για να επεξεργάζεται δεδομένα του πραγματικού κόσμου που έχουν τη μορφή πινάκων (matrix), τανυστών (tensor) ή ακολουθιών (sequence). Η εφαρμογή αυτών των κλασικών τεχνικών στα βιολογικά δίκτυα παρουσιάζει σημαντικούς περιορισμούς, καθώς οι παραδοσιακοί αλγόριθμοι βασίζονται στη θεμελιώδη υπόθεση ότι τα δεδομένα είναι ανεξάρτητα και ισόνομα κατανεμημένα (independent and identically distributed -i.i.d) (Ma & Tang, 2021).. Στην περίπτωση όμως των δικτύων, οι κόμβοι είναι εγγενώς συνδεδεμένοι μεταξύ τους, γεγονός που καταρρίπτει την υπόθεση της ανεξαρτησίας. Για ξεπεραστεί αναπτύχθηκε το πεδίο της Μηχανικής Μάθησης σε Γράφους (Machine Learning on Graphs), το οποίο επιτρέπει την εκτέλεση υπολογιστικών εργασιών (computational tasks) απευθείας πάνω στη δομή του γράφου (Hamilton, 2020; Ma & Tang, 2021).

Σύμφωνα με τη βιβλιογραφία, οι βασικές υπολογιστικές εργασίες που καλείται να εκτελέσει η Μηχανική Μάθηση σε αυτό το πεδίο χωρίζονται σε δύο μεγάλες κατηγορίες: στις εργασίες που εστιάζουν στους κόμβους (node-focused tasks) και σε εκείνες που εστιάζουν στον γράφο (graph-focused tasks) (Ma & Tang, 2021). Όπως αναλύεται διεξοδικά από τον Hamilton (2020), οι διεργασίες αυτές εξειδικεύονται περαιτέρω σε συγκεκριμένα επίπεδα ανάλυσης:

- Ταξινόμηση κόμβων (Node Classification): Η διαδικασία κατά την οποία το μοντέλο καλείται να προβλέψει ή να συμπεράνει μία ιδιότητα ή μία ετικέτα (label) για έναν συγκεκριμένο κόμβο του δικτύου. Όπως επισημαίνεται, λόγω της ιδιαίτερης φύσης των γράφων, η Ταξινόμηση κόμβων προσεγγίζεται συχνά ως Ημί-Εποπτευόμενη Μάθηση (Semi-supervised Learning). Αυτό συμβαίνει διότι κατά την εκπαίδευση του μοντέλου είναι προσβάσιμος ολόκληρος ο γράφος, συμπεριλαμβανομένων των κόμβων χωρίς ετικέτα (test nodes), επιτρέποντας στον αλγόριθμο να αξιοποιήσει την πληροφορία της γειτονιάς τους για τη βελτίωση των προβλέψεων. Στη βιοπληροφορική, η εργασία αυτή βρίσκει εφαρμογή στον εντοπισμό και την κατάταξη γονιδίων που σχετίζονται με συγκεκριμένες ασθένειες (Ma & Tang, 2021).

- Πρόβλεψη Σχέσεων/Ακμών (Relation/Link Prediction): Σε αντίθεση με τις παραδοσιακές εφαρμογές, η πρόβλεψη σχέσεων στους γράφους παρουσιάζει μια μοναδική θεωρητική ιδιαιτερότητα, καθώς ξεφεύγει από τα αυστηρά όρια των κλασικών κατηγοριών της Μηχανικής Μάθησης. Η διεργασία αυτή συνδυάζει στοιχεία τόσο Εποπτευόμενης (Supervised) όσο και Μη Εποπτευόμενης (Unsupervised) μάθησης, ενώ για την επιτυχή υλοποίησή της απαιτούνται «επαγωγικές επαναλήψεις» (inductive biases) -δηλαδή ενσωματωμένοι κανόνες και υποθέσεις- που είναι αποκλειστικά σχεδιασμένοι για τη δομή και την αρχιτεκτονική των γράφων.

- Συσταδοποίηση και Ανίχνευση Κοινοτήτων (Clustering and Community Detection): Η εργασία αυτή αποτελεί μέθοδο Μη Εποπτευόμενης Μάθησης στους γράφους. Αντί για την πρόβλεψη συγκεκριμένων ιδιοτήτων, η προσέγγιση αυτή εστιάζει στην αυτόματη αναγνώριση κρυφών ομάδων (clusters), αξιοποιώντας αποκλειστικά και μόνο την υπάρχουσα τοπολογία του δικτύου. Η βασική παραδοχή της μεθόδου στηρίζεται στο ότι οι κόμβοι που παρουσιάζουν ισχυρή συνδεσιμότητα μεταξύ τους τείνουν να σχηματίζουν διακριτές και πυκνές κοινότητες. Στις βιολογικές εφαρμογές, η τεχνική αυτή αποδεικνύεται εξαιρετικά χρήσιμη, καθώς επιτρέπει τον εντοπισμό απομονωμένων λειτουργικών υποδικτύων ή συμπλεγμάτων μέσα σε ευρύτερα δίκτυα γενετικών αλληλεπιδράσεων.

- Ανάλυση σε επίπεδο γράφου: Η κατηγορία αυτή περιλαμβάνει ένα ευρύ φάσμα εφαρμογών όπου το ενδιαφέρον μετατοπίζεται από τα επιμέρους στοιχεία προς το δίκτυο (Ma & Tang, 2021). Παρότι το πεδίο περιλαμβάνει σύνθετες διεργασίες όπως η αντιστοίχιση γράφων (graph matching) και η υπολογιστική δημιουργία γράφων (Graph

generation), η πιο αντιπροσωπευτική εργασίας αφορά την Ταξινόμηση γράφων (Graph classification). Η τεχνική αυτή χρησιμοποιείται κυρίως για την αξιολόγηση και κατηγοριοποίηση ολόκληρων μοριακών ή χημικών δομών με βάση τις καθολικές τους ιδιότητες (Ma &Tang, 2021).

Η κατανόηση αυτών των υπολογιστικών εργασιών αναδεικνύει ότι η Μηχανική Μάθηση σε Γράφους δεν προσφέρει απλώς μια στατική απεικόνιση του δικτύου, αλλά αποτελεί ένα ενεργό εργαλείο ικανό να προβλέπει νέα βιολογικά δεδομένα (Hamilton, 2020). Στην παρούσα εργασία, η μεθοδολογία συνδυάζει αυτές τις δύο προσεγγίσεις, καθώς αξιοποιεί τη Μη-Εποπτευόμενη Συσταδοποίηση για την ομαδοποίηση των γονιδιακών διανυσμάτων και την Εποπτευόμενη μάθηση για την τελική πρόβλεψη των ακμών του Αυτισμού.

2.4 Τεχνικές Εκμάθησης Αναπαραστάσεων (Representation Learning) και Graph Embeddings

Η παραδοσιακή προσέγγιση για την εξαγωγή δομικών χαρακτηριστικών από ένα δίκτυο βασιζόταν σε χειροκίνητα σχεδιασμένα χαρακτηριστικά (hand-engineered features). Αντιπροσωπευτικά παραδείγματα τέτοιων μεθόδων αποτελούν ο βαθμός κόμβου (node degree) και η κεντρικότητα κόμβου (node centrality). Παρά τη χρησιμότητά τους, οι παραδοσιακές αυτές προσεγγίσεις εμφανίζουν σοβαρά μειονεκτήματα. Είναι άκαμπτες, καθώς δεν μπορούν να προσαρμοστούν αυτόματα μέσα από μία διαδικασία μάθησης, ενώ ο σχεδιασμός τους αποτελεί μία χρονοβόρα και δαπανηρή διαδικασία (Hamilton, 2020; Ma & Tang, 2021).

Για άρση αυτών των περιορισμών αναπτύχθηκε η Εκμάθηση Αναπαραστάσεων Γράφων. Στόχος της συγκεκριμένης προσέγγισης είναι να αντικαταστήσει αυτά τα χειροκίνητα χαρακτηριστικά με μεθόδους που μαθαίνουν αυτόματα να κωδικοποιούν τη δομική πληροφορία του γράφου (Hamilton, 2020; Ma & Tang, 2021).

Η βασική τεχνική για την επίτευξη αυτού του στόχου είναι η δημιουργία Ενσωματωμένων Γράφων (Graph Embeddings), μία διαδικασία που διαχωρίζει τα δεδομένα σε δύο διαφορετικά πεδία (domains) (Ma & Tang, 2021):

1. Το πεδίο του γράφου (graph domain): Το αρχικό δίκτυο, όπου οι κόμβοι συνδέονται μεταξύ τους με διακριτές τιμές.
2. Το πεδίο των ενσωματώσεων (embedding domain): Ο νέος χώρος, στον οποίο κάθε κόμβος μετατρέπεται πλέον σε ένα διάνυσμα.

Σύμφωνα με τους Ma & Tang (2021), ο κεντρικός στόχος είναι η σχεδίαση μιας διαδικασίας που μεταφέρει την πληροφορία από το δίκτυο στον διανυσματικό χώρο, διασφαλίζοντας ότι τα βασικά τοπολογικά χαρακτηριστικά του γράφου θα παραμείνουν αναλλοίωτα. Για να επιτευχθεί αυτό, η ανάπτυξη των αλγορίθμων βασίζεται σε δύο βασικά ερωτήματα: Ποια συγκεκριμένη πληροφορία πρέπει να διατηρηθεί από τον αρχικό γράφο και με ποιο υπολογιστικό μοντέλο θα γίνει η καταγραφή της.

Όσον αφορά το πρώτο ερώτημα, η έρευνα εστιάζει σε διαφορετικά στοιχεία που μπορούν να αντληθούν από ένα δίκτυο, όπως τους άμεσους γείτονες κάθε κόμβου (neighborhood information), τη λειτουργική του θέση (structural role), την κατάστασή του (node status), καθώς και τις ομάδες ή κοινότητες στις οποίες εντάσσεται (community information). Η καταγραφή αυτών των στοιχείων γίνεται μέσω της μεθοδολογίας των τυχαίων διαδρομών (random walks), η οποία αποτυπώνει ποιες ομάδες κόμβων συνυπάρχουν συχνά στις ίδιες διαδρομές του δικτύου. Χαρακτηριστικό παράδειγμα αυτής της προσέγγισης αποτελεί ο αλγόριθμος node2vec, ο οποίος εξετάζεται αναλυτικά στην επόμενη ενότητα.

Για να αποθηκευτούν αυτά τα δεδομένα, η πιο συνηθισμένη μέθοδος λειτουργεί σαν ένας πίνακας αναζήτησης (look-up table). Σε αυτόν τον πίνακα, κάθε κόμβος παίρνει μία δική του σειρά από αριθμούς, που αποτελεί το διάνυσμά του. Η βασική ιδέα για τη δημιουργία αυτού του πίνακα είναι ότι οι αριθμοί αυτοί πρέπει να μπορούν να αποδώσουν σωστά τη δομή του δικτύου. Πρακτικά, μία τέτοια αναπαράσταση θεωρείται πετυχημένη όταν οι αριθμοί της είναι αρκετοί για να «ξαναφτιάξουν» της πραγματικές συνδέσεις του γράφου. Η όλη διαδικασία βελτιώνεται συνεχώς διορθώνοντας τα λάθη που γίνονται κατά την ανακατασκευή (reconstruction error).

2.5 Ο αλγόριθμος node2vec: Μαθηματική Ανάλυση (Random Walks, Skip-gram Model)

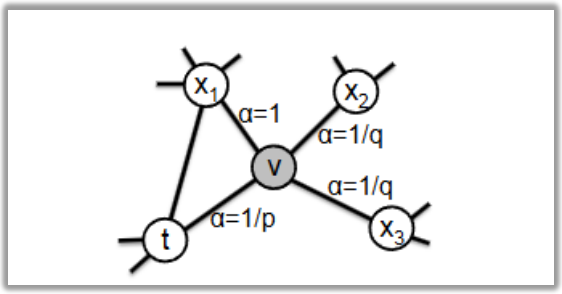
Η ανάγκη για την καταπολέμηση των περιορισμών των παραδοσιακών μεθόδων της στατιστικής και γραμμικής άλγεβρας για την διαχείριση των μεγάλων δικτύων οδήγησε στην ανάπτυξη του αλγορίθμου node2vec. Ιστορικά όμως, η εξέλιξη του αλγορίθμου ακολούθησε τρία στάδια.

Οι πρώτες προσεγγίσεις για την αναπαράσταση δικτύων βασίστηκαν σε κλασικές τεχνικές μείωσης διαστατικότητας, όπως η Ανάλυση Κυρίων Συνιστωσών (PCA) και οι Λαπλασιανοί Ιδιοχάρτες Laplacian Eigenmaps) (Belkin & Niyogi, 2003; Jolliffe, 2002). Αν και φάνηκαν αποδοτικές για την εξαγωγή σημαντικών γεωμετρικών στοιχείων, η υψηλή τους πολυπλοκότητα καθιστούσε αδύνατη την εφαρμογή τους σε μεγάλα δίκτυα. Συνεπώς, το 2013 παρουσιάστηκε για πρώτη φορά το μοντέλο Skip-Gram, το οποίο αποτελεί βασικό κομμάτι του αλγορίθμου Word2vec, και άλλαξε τα δεδομένα στην επεξεργασία φυσικής γλώσσας (NLP) (Mikolov et al., 2013). Το 2014, η φιλοσοφία του Word2vec μεταφέρθηκε από την ανάλυση κειμένων στη μελέτη δικτύων με την εισαγωγή του αλγορίθμου DeepWalk (Perozzi et al., 2014), ωστόσο, η απουσία ευελιξίας του οδήγησε τελικά στην δημιουργία του node2vec το 2016.

Ο αλγόριθμος node2vec αποτελεί μία από τις πιο δημοφιλείς και ευέλικτες μεθόδους για τη μετατροπή των κόμβων ενός δικτύου σε αριθμητικά διανύσματα (Grover & Leskovec, 2016). Ο κύριος στόχος του πλαισίου (framework) είναι να αντιστοιχίσει κάθε στοιχείο του γράφου σε έναν χώρο περιορισμένων διαστάσεων, με τέτοιο τρόπο ώστε να διατηρούνται οι σχέσεις και οι γειτονιές μεταξύ των κόμβων. Η βασική ιδιαιτερότητα και καινοτομία της συγκεκριμένης μεθόδου είναι ότι δεν χρησιμοποιεί έναν αυστηρό και περιορισμένο ορισμό για το τι θεωρείται «γειτονιά», αλλά δίνει τη δυνατότητα να αλλάζει τον τρόπο που εξερευνάται το δίκτυο, ανάλογα με το τι αναζητάται κάθε φορά.

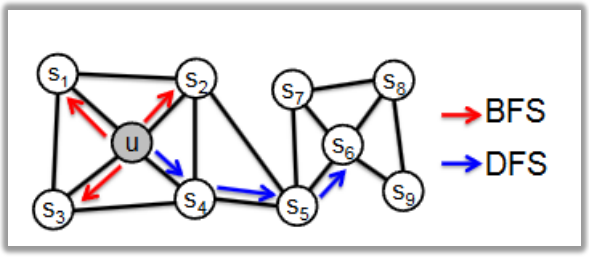
Για την ανακάλυψη της δομής γύρω από κάθε κόμβο, ο αλγόριθμος χρησιμοποιεί μεροληπτικές τυχαίες διαδρομές (biased random walks). Αυτό σημαίνει ότι κατά την περιήγηση στο δίκτυο από κόμβο σε κόμβο, η επιλογή του επόμενου βήματος δεν πραγματοποιείται τυχαία αλλά καθορίζεται από τις μαθηματικές παραμέτρους p και q.

Εικόνα 2.1: *Αναπαράσταση τυχαίων διαδρομών του node2vec από Grover & Leskovec*



Οι συγκεκριμένες παράμετροι επιτρέπουν στον αλγόριθμο να ελέγχει την κατεύθυνση των διαδρομών, εξισορροπώντας ανάμεσα στην καταγραφή της άμεσης τοπικής γειτονιάς (αναζήτηση κατά πλάτος -BFS) και στην εξερεύνηση βαθύτερων, καθολικών προτύπων μέσα στο δίκτυο (αναζήτηση κατά βάθος -DFS).

Εικόνα 2.2: *BFS και DFS από Grover & Leskovec*



Μετά την ολοκλήρωση των τυχαίων διαδρομών, οι σειρές των κόμβων που έχουν συγκεντρωθεί αντιμετωπίζονται σαν «προτάσεις» και οι κόμβοι σαν «λέξεις» μέσα σε ένα κείμενο. Στο σημείο αυτό, ο αλγόριθμος χρησιμοποιεί το μοντέλο Skip-gram, το οποίο βασίζεται στην υπόθεση ότι οι λέξεις που εμφανίζονται στα ίδια περιβάλλοντα τείνουν να έχουν και παρόμοια σημασία. Μεταφέροντας αυτή τη λογική στα δίκτυα, θεωρείται ότι οι κόμβοι που βρίσκονται συχνά μαζί στις ίδιες διαδρομές έχουν και παρόμοιες ιδιότητες.

Στην πράξη, το μοντέλο εξετάζει αυτές τις διαδρομές και προσπαθεί να μάθει τις διανυσματικές αναπαραστάσεις των κόμβων. Στόχος του είναι, όταν του δίνεται ένας συγκεκριμένος κόμβος, να μπορεί να προβλέψει με επιτυχία ποιοι άλλοι κόμβοι βρίσκονται γύρω του, μέσα σε ένα προκαθορισμένο όριο (context window). Για να γίνει αυτή η διαδικασία γρήγορα και σωστά από τον υπολογιστή, χρησιμοποιείται ο αλγόριθμος Στοχαστικής Καθόδου Κλίσης (SGD) μαζί με μία τεχνική που ονομάζεται αρνητική δειγματοληψία (negative sampling). Μέσα από αυτή τη διαδικασία, οι σχέσεις του γράφου μετατρέπονται τελικά σε συγκεκριμένες σειρές αριθμών, όπου οι κόμβοι με παρόμοια

συμπεριφορά καταλήγουν να έχουν κοντινές συντεταγμένες στον τελικό διανυσματικό χώρο.

Στη μελέτη για τον Αυτισμό, η εφαρμογή του συγκεκριμένου αλγορίθμου προσφέρει σημαντικά πλεονεκτήματα. Εξαιτίας της ικανότητάς του να ισορροπεί ανάμεσα στην τοπική και καθολική αναζήτηση, ο αλγόριθμος μπορεί να εντοπίσει με ακρίβεια τόσο τα γονίδια που συνεργάζονται άμεσα σε στενές βιολογικές ομάδες, όσο και τις ευρύτερες λειτουργικές ομοιότητες μεταξύ διαφορετικών βιολογικών μονοπατιών που συνδέονται με την διαταραχή.

2.6 Μετρική Ομοιότητας Διανυσμάτων: Ομοιότητα Συνημίτονου (Cosine Similarity)

Μετά την παραγωγή των διανυσματικών αναπαραστάσεων (node embeddings) από τον αλγόριθμο node2vec, η αξιολόγηση της τοπολογικής εγγύτητας μεταξύ των κόμβων στον πολυδιάστατο χώρο πραγματοποιείται μέσω μαθηματικών μετρικών ομοιότητας. Η πιο διαδεδομένη μεθοδολογία για τον σκοπό αυτό είναι η ομοιότητα συνημίτονου (Cosine Similarity).

Η διαδικασία βασίζεται στην αξιοποίηση των μονοπατιών του γράφου για την εξαγωγή της πληροφορίας. Αντί να εξετάζονται οι κόμβοι απομονωμένα, το σύστημα αναλύει τον τρόπο με τον οποίο οι ασθένειες και τα γονίδια συνδέονται ευρύτερα μέσα στο δίκτυο BioSNAP. Η εξόρυξη αυτή επιτυγχάνεται με τη χρήση του αλγορίθμου node2vec, ο οποίος μετατρέπει την τοπολογική θέση κάθε κόμβου σε ένα διάνυσμα αριθμών. Με τον τρόπο αυτό, οι πολύπλοκες σχέσεις του δικτύου μετατρέπονται σε μαθηματικά διανύσματα. Έτσι, οι οντότητες που έχουν παρόμοιο ρόλο ή θέση στο δίκτυο, αποκτούν και διανύσματα που βρίσκονται κοντά μεταξύ τους.

Για την ποσοτική αξιολόγηση των παραγόμενων διανυσμάτων και την εύρεση των στενότερων συσχετίσεων, χρησιμοποιείται η Ομοιότητα Συνημίτονου (Cosine Similarity) (Manning et al., 2008). Η μετρική αυτή υπολογίζει το συνημίτονο της γωνίας που σχηματίζουν τα διανύσματα δύο κόμβων στον πολυδιάστατο χώρο. Η μαθηματική σχέση για την ομοιότητα δύο διανυσμάτων $V(d_1)$ και $V(d_2)$ ορίζεται ως:

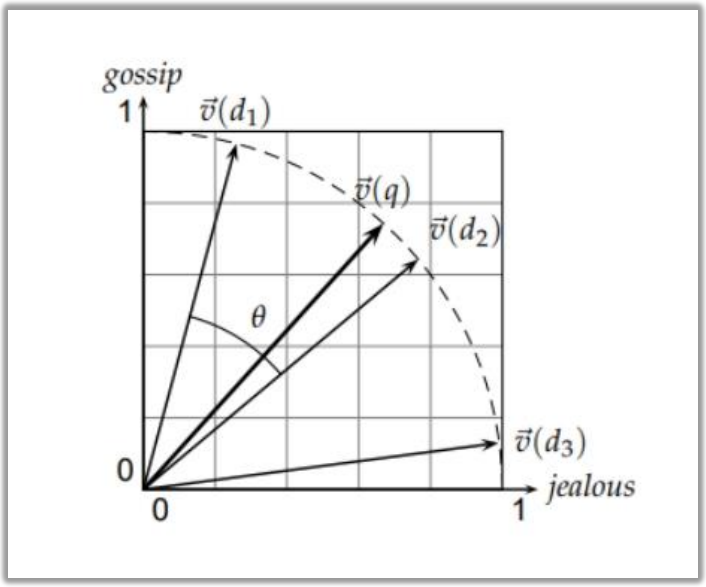
$$sim(d_1, d_2) = \frac{V(d_1) \cdot V(d_2)}{|V(d_1)||V(d_2)|}$$

Όπου:

- Ο αριθμητής αντιπροσωπεύει το εσωτερικό γινόμενο των διανυσμάτων $V(d_1)$ και $V(d_2)$ το οποίο για Μ συνιστώσες ορίζεται ως $\sum_{i=1}^{M} x_i y_i$.
- Ο παρονομαστής αποτελεί το γινόμενο των Ευκλείδειων μηκών των δύο διανυσμάτων, όπου το μήκος για κάθε διάνυσμα d με Μ συνιστώσες ορίζεται ως $\sqrt{\sum_{i=1}^{M} V_i^2(d)}$.

Η εισαγωγή του παρονομαστή επιτυγχάνει την κανονικοποίηση του μήκους των διανυσμάτων, μετατρέποντάς τα σε μοναδιαία διανύσματα. Με τον τρόπο αυτό, εξασφαλίζεται ότι η τελική μέτρηση δεν επηρεάζεται από το μέγεθος των διανυσμάτων – δηλαδή από το αν ένας κόμβος έχει υπερβολικά μεγάλο ή μικρό πλήθος συνολικών συνδέσεων – αλλά επικεντρώνεται αποκλειστικά στη γεωμετρική τους κατεύθυνση. Η γεωμετρική προσέγγιση αποτυπώνεται στην **Εικόνα 2.3**.

Εικόνα 2.3: *Γεωμετρική αναπαράσταση ομοιότητας συνημίτονου μεταξύ διανυσμάτων*



Οι τιμές της μετρικής κυμαίνονται αυστηρά μεταξύ 0 και 1 για θετικούς άξονες, όπου η τιμή $sim(d_1, d_2) = cos\,\theta$ εξαρτάται αποκλειστικά από τη γωνία θ. Τιμές που προσεγγίζουν την μονάδα υποδηλώνουν ότι η γωνία είναι πολύ μικρή, δηλαδή τα διανύσματα έχουν σχεδόν ίδια κατεύθυνση, γεγονός που ερμηνεύεται ως μέγιστη τοπολογική ομοιότητα. Στη μελέτη αυτή, με κεντρικό σημείο αναφοράς το διάνυσμα του Αυτισμού, υπολογίζεται η Ομοιότητα Συνημίτονου με κάθε άλλο κόμβο του δικτύου, επιτρέποντας την ιεραρχική κατάταξή τους από τον πιο σχετικό στον λιγότερο σχετικό.

2.7 Κριτική αποτίμηση υπαρχόντων υπολογιστικών μοντέλων

Η αξιολόγηση των διαθέσιμων μεθόδων αναδεικνύει μία ξεκάθαρη εξέλιξη, η οποία ξεκινά από τους απλούς στατιστικούς δείκτες και καταλήγει στα σύγχρονα αλγοριθμικά μοντέλα. Οι παραδοσιακές τεχνικές, που βασίζονται σε χειροκίνητα σχεδιασμένα μέτρα όπως ο βαθμός κόμβου και η κεντρικότητα κόμβου, παρουσιάζουν σοβαρά προβλήματα όπως η έλλειψη ευελιξίας. Αποτελούν στατιστικά στοιχεία που εξετάζουν απομονωμένα τους κόμβους και δεν επιτρέπουν τον διαμοιρασμό της πληροφορία μεταξύ τους (Hamilton, 2020). Επιπλέον η ανάγκη για χειροκίνητο σχεδιασμό αυτών των χαρακτηριστικών καθιστά τη διαδικασία μη αποδοτική για πραγματικά δίκτυα, τα οποία συχνά αποτελούνται από εκατομμύρια κόμβους και ακμές και απαιτούν υπολογιστικά κλιμακωτές προσεγγίσεις (Grover & Leskovec, 2016).

Η εισαγωγή αλγορίθμων που βασίζονται σε τυχαίους περιπάτους, όπως ο DeepWalk, αντιμετώπισε αυτή την υπολογιστική πρόκληση, χρησιμοποιώντας τον αλγόριθμο Hierarchical Softmax για την προσέγγιση των πιθανοτήτων (Hamilton, 2020). Παρά την επιτυχία του όμως, ο DeepWalk επικρίνεται για ακαμψία, καθώς η ομοιόμορφη κίνηση στο δίκτυο περιορίζει την ικανότητα του να προσαρμόζεται σε διαφορετικά δομικά πρότυπα .Επιπλέον, η χρήση του Hierarchical Softmax αποδεικνύεται λιγότερο αποδοτική υπολογιστικά σε σύγκριση με την τεχνική της αρνητικής δειγματοληψίας (Grover & Leskovec, 2016).

Από την άλλη πλευρά, ο αλγόριθμος node2vec ξεπερνά αυτούς τους περιορισμού, καθώς εισάγει μεγαλύτερη ευελιξία μέσω των μεροληπτικών διαδρομών και αξιοποιεί την αποδοτική τεχνική της αρνητικής δειγματοληψίας (Hamilton, 2020). Παρά τα

πλεονεκτήματα αυτά, η μελέτη των Hacker & Riek (2022), αναδεικνύει σοβαρές υπολογιστικές και δομικές αδυναμίες που σχετίζονται με τη σταθερότητα (stability) και την ποιότητα (quality) του αλγορίθμου. Λόγω των στοχαστικών παραγόντων του αλγορίθμου, παράγονται σημαντικά διαφορετικά διανύσματα ακόμα και με τις ίδιες ακριβώς υπερπαραμέτρους, ενώ ορισμένες επιλογές ρυθμίσεων οδηγούν σε εξαιρετικά χαμηλές αποδόσεις, οι οποίες δεν ξεπερνούν σε αποτελεσματικότητα μία εντελώς τυχαία πρόβλεψη.

Στο πλαίσιο της υπολογιστικής βιολογίας, οι παραπάνω διαφορές καθορίζουν την τελική επιλογή του μοντέλου. Τα βιολογικά δίκτυα περιέχουν συχνά μεγάλο βαθμό θορύβου και λανθασμένες καταγραφές συνδέσεων (Ma & Tang, 2021). Οι παραδοσιακές στατιστικές μέθοδοι αδυνατούν να ξεχωρίσουν τον θόρυβο, ενώ αλγόριθμοι όπως ο DeepWalk παρουσιάζουν θέματα ευελιξίας (Hamilton, 2020). Αντίθετα, ο node2vec προσφέρει μεγαλύτερη ανθεκτικότητα σε ελλιπή ή θορυβώδη δεδομένα (Grover & Leskovec, 2016). Ωστόσο, η διαπιστωμένη ευαισθησία και αστάθειά του απέναντι στις υπερπαραμέτρους επιβάλλουν ιδιαίτερη προσοχή. Για τον λόγο αυτό, κατά την αναζήτηση βιολογικών δεδομένων, κρίνεται αναγκαίο να αποφεύγεται η στήριξη σε μία μεμονωμένη εκτέλεση και προτείνεται η πραγματοποίηση πολλαπλών εκτελέσεων του αλγορίθμου, διασφαλίζοντας έτσι την εγκυρότητα των τελικών βιολογικών αποτελεσμάτων ((Hacker & Rieck, 2022).

3.Δεδομένα και Υπολογιστική Υποδομή

3.1 Πηγή και Χαρακτηριστικά του Συνόλου Δεδομένων (BioSNAP)

Η επιλογή ενός αξιόπιστου και κατάλληλα δομημένου συνόλου δεδομένων αποτελεί θεμελιώδη προϋπόθεση για την εξαγωγή έγκυρων συμπερασμάτων στην υπολογιστική βιολογία. Για τους σκοπούς της παρούσας πτυχιακής εργασίας χρησιμοποιήθηκαν δεδομένα από τη συλλογή BioSNAP (Biomedical Network Dataset Collection) του Πανεπιστημίου Stanford. Η συγκεκριμένη πλατφόρμα παρέχει εξειδικευμένα βιολογικά και ιατρικά δίκτυα, τα οποία αποτυπώνουν περίπλοκες κυτταρικές και γονιδιακές αλληλεπιδράσεις, καθιστώντας τα ιδανικά για την αξιολόγηση αλγορίθμων εκμάθησης Αναπαραστάσεων Γράφων.

Το συγκεκριμένο σύνολο δεδομένων που εξετάζεται προέρχεται από τη βάση DG-AssocMiner (Disease-Gene Associations) της συλλογής BioSNAP. Το αρχείο αυτό καταγράφει τις τεκμηριωμένες συσχετίσεις ανάμεσα σε διάφορες ασθένειες και στα αντίστοιχα γονίδια που συνδέονται με αυτές. Η πληροφορία είναι οργανωμένη σε τρεις βασικές στήλες: το μοναδικό αναγνωριστικό της ασθένειας (Disease ID), την επίσημη ονομασία της ασθένειας (Disease Name) και το αναγνωριστικό του γονιδίου (Gene ID). Μέσα από αυτό το ευρύτερο σύνολο δεδομένων, η παρούσα μελέτη απομονώνει και επικεντρώνεται στις βιολογικές οντότητες και τις συσχετίσεις που συνδέονται με τον Αυτισμό.

Σε ότι αφορά τα ποσοτικά και τοπολογικά χαρακτηριστικά του, το καθολικό δίκτυο της βάσης παρουσιάζει μια ιδιαίτερα πυκνή δομή. Τα αναλυτικά στοιχεία του δικτύου, οποία αποτελούν το υπολογιστικό υπόβαθρο της μελέτης, συνοψίζονται στον **Πίνακα 3.1.**

Πίνακας 3.1: Τοπολογικά και Ποσοτικά Χαρακτηριστικά του Καθολικού Δικτύου DG-AssocMiner (BioSNAP)

Χαρακτηριστικό Δικτύου	Τιμή/Μέτρηση
Συνολικό Πλήθος Κόμβων (Nodes)	7.813

Συνολικό Πλήθος Ακμών (Edges)	21.357
Κόμβοι Ασθενειών (Disease Nodes)	519
Κόμβοι Γονιδίων (Gene Nodes)	7.294
Ποσοστό Κόμβων σε Μέγιστη Συνεκτική Συνιστώσα (SCC)	100%
Διάμετρος Γράφου (Diameter)	8
Αποτελεσματική διάμετρος 90ού εκατοστημορίου (90-percentile Effective Diameter)	5.43

Όπως προκύπτει από τον Πίνακα 3.1, το 100% των κόμβων και των ακμών του δικτύου ανήκει στη Μέγιστη Συνεκτική Συνιστώσα (Strongly Connected Component -SCC), γεγονός που σημαίνει ότι δεν υφίστανται απομονωμένοι κόμβοι ή αποκομμένα υποδίκτυα. Η μέγιστη απόσταση μεταξύ δύο οποιονδήποτε κόμβων (διάμετρος του γράφου) ισούται με 8, ενώ η Αποτελεσματική διάμετρος 90ου εκατοστημορίου (90-percentile effective diameter) υπολογίζεται σε 5.43

Οι παραπάνω ποσοτικοί δείκτες και τα καθολικά στατιστικά στοιχεία αποτυπώνουν το υπολογιστικό μέγεθος του δικτύου, χωρίς όμως να φανερώνουν την εσωτερική αρχιτεκτονική του. Για να γίνει δυνατή η αξιοποίηση αυτών των δεδομένων και η μετέπειτα εφαρμογή αλγορίθμων εκμάθησης αναπαραστάσεων, είναι απαραίτητο να αναλυθεί ο τρόπος με τον οποίο οι δύο διαφορετικοί τύποι οντοτήτων συνδέονται και οργανώνονται δομικά, όπως εξετάζονται στην ενότητα που ακολουθεί.

3.2 Τοπολογική Δομή και Σχήμα Δεδομένων: Αναπαράσταση ως Bipartite Graph

Η κατανόηση της εσωτερικής οργάνωσης του συνόλου δεδομένων BioSNAP είναι καθοριστική για τη σωστή μοντελοποίησή του. Το συγκεκριμένο δίκτυο δομείται ως ένας διμερής γράφος (Barabási & Oltvai, 2004). Στη θεωρία δικτύων, ένας γράφος

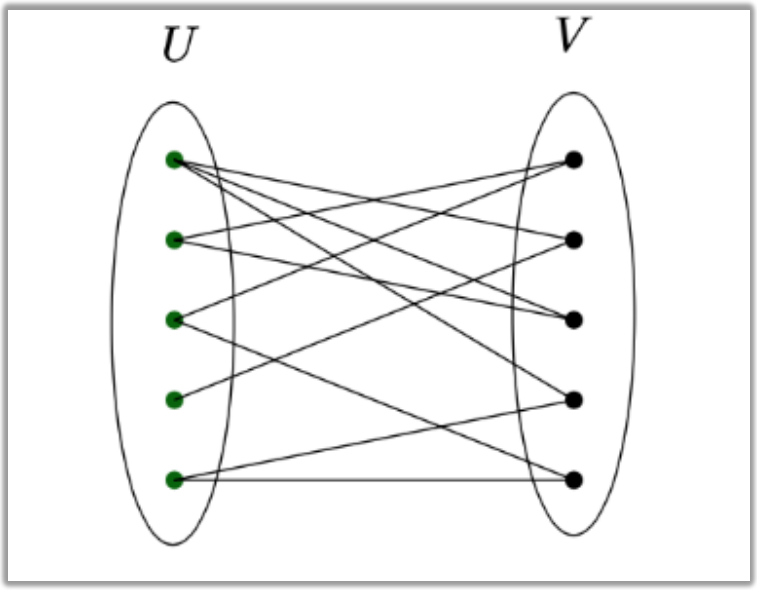
χαρακτηρίζεται ως διμερής όταν το σύνολο των κόμβων διαχωρίζεται σε δύο διακριτές και ανεξάρτητες ομάδες, με τέτοιο τρόπο ώστε κάθε σύνδεση να ενώνει αποκλειστικά έναν κόμβο της πρώτης ομάδας με έναν κόμβο της δεύτερης ομάδας (Barabási, 2016).

Στην παρούσα μελέτη, οι δύο αυτές ομάδες καθορίζονται άμεσα από τη βιολογική φύση και τη δομή του συνόλου δεδομένων DG-AssocMiner. Η πρώτη ομάδα περιλαμβάνει αποκλειστικά τις ασθένειες (με κεντρικό σημείο αναφοράς τον Αυτισμό), ενώ η δεύτερη ομάδα περιλαμβάνει τα γονίδια που σχετίζονται με αυτές. Το βασικό τοπολογικό χαρακτηριστικό αυτού του σχήματος είναι ότι δεν υπάρχουν απευθείας συνδέσεις μεταξύ κόμβων που ανήκουν στην ίδια ομάδα. Αυτό σημαίνει ότι το στο δίκτυο δεν υπάρχουν ακμές που να ενώνουν μία ασθένεια απευθείας με μία άλλη ασθένειας, ούτε γονίδια που να αλληλεπιδρούν απευθείας μεταξύ τους, καθιστώντας τις σχέσεις αυστηρά ετερογενείς.

Αυτή η διμερής δομή προσφέρει ένα σημαντικό πλεονέκτημα στην αποτύπωση των βιολογικών συσχετίσεων, καθώς επιτρέπει την καθαρή μελέτη των αλληλεπιδράσεων ανάμεσα σε δύο εντελώς διαφορετικά επίπεδα πληροφορίας – το φαινότυπο (ασθένειες) και το μοριακό (γονίδια). Η ύπαρξη μιας ακμής υποδηλώνει μία τεκμηριωμένη βιολογική συσχέτιση, όπου ένα συγκεκριμένο γονίδιο σχετίζεται άμεσα με μία ασθένεια. Κατά συνέπεια, αν δύο διαφορετικές ασθένειες συνδέονται με το ίδιο γονίδιο, η μεταξύ τους σχέση δεν φαίνεται με μία απευθείας γραμμή, αλλά προκύπτει έμμεσα, επειδή και οι δύο «ακουμπούν» στο ίδιο γονίδιο.

Η συγκεκριμένη δικτυακή διάταξη και ο τρόπος με τον οποίο οι συνδέσεις διαπερνούν τις δύο ομάδες κόμβων αποτυπώνονται στην **Εικόνα 3.1**. Η οπτικοποίηση αυτή καθιστά σαφές ότι, παρά την απουσία ομογενών συνδέσεων, το δίκτυο διατηρεί μία ισχυρή εσωτερική συνοχή. Η κατανόηση αυτής της διμερούς αρχιτεκτονικής, είναι απαραίτητη, καθώς καθορίζει τον τρόπο με τον οποίο οι αλγόριθμοι τυχαίων περιπάτων θα πλοηγηθούν αργότερα στο δίκτυο, μεταπηδώντας εναλλάξ από το επίπεδο των ασθενειών στο επίπεδο των γονιδίων, με σκοπό την εξαγωγή των τοπολογικών χαρακτηριστικών.

Εικόνα 3.1: *Σχηματική αναπαράσταση διμερούς γράφου*



3.3 Διαχείριση Υπολογιστικού Πλαισίου (Global Topology) και Προεπεξεργασία

Για την επιτυχή εκτέλεση των υπολογιστικών πειραμάτων, η απλή καταχώρηση του αρχικού αρχείου δεδομένων δεν επαρκεί· απαιτείται μία συστηματική διαδικασία προεπεξεργασίας. Στόχος της διαδικασίας αυτής είναι ο περιορισμός του ευρύτερου δικτύου γύρω από την διαταραχή, δημιουργώντας ένα στοχευμένο υπογράφημα (subgraph) που εστιάζει στον Αυτισμό. Στο υπολογιστικό πλαίσιο της ανάλυσης σύνθετων δικτύων, η επιλογή και η διαχείριση της καθολικής τοπολογίας (global topology) είναι καθοριστικής σημασίας, καθώς ορίζει τη δομή μέσω της οποίας διασκορπίζεται η πληροφορία γρήγορα σε ολόκληρο το σύστημα (Bastos-Filho et al., 2011).

Η ροή της προεπεξεργασίας για την εξαγωγή αυτού του πλαισίου βασίζεται στα δεδομένα του πίνακα DG-AssocMiner και υλοποιείται σε δύο διακριτά στάδια φιλτραρίσματος:

1. Πρώτο στάδιο (Απομόνωση γονιδίων): Αρχικά, ο καθολικός πίνακας φιλτράρεται με τη βάση τη στήλη των ασθενειών, κρατώντας μόνο τις εγγραφές που αφορούν την τιμή “Autistic Disorder”. Από το φιλτράρισμα αυτό εξάγεται η πλήρης λίστα με τα μοναδικά αναγνωριστικά των γονιδίων (Gene IDs) που σχετίζονται άμεσα με τον Αυτισμό.

2. Δεύτερο στάδιο (Διεύρυνση δικτύου): Στη συνέχεια, η λίστα αυτή χρησιμοποιείται ως κριτήριο για ένα φιλτράρισμα στο αρχικό, πλήρες σύνολο δεδομένων. Στο βήμα αυτό, αναζητούνται και εισάγονται στο DataFrame όλες οι υπόλοιπες ασθένειες της βάσης δεδομένων οι οποίες συνδέονται με τουλάχιστον ένα από τα γονίδια της λίστας του Αυτισμού.

Με την ολοκλήρωση αυτής της διαδικασίας, απορρίπτονται οι πληροφορίες και οι κόμβοι του αρχείου που δεν έχουν καμία σύνδεση με τη μελέτη, μειώνοντας τον υπολογιστικό φόρτο. Ταυτόχρονα, εξασφαλίζεται ότι το τελικό δίκτυο διατηρεί την απαραίτητη καθολική δομή, επιτρέποντας στους τυχαίους περιπάτους του node2vec να κινηθούν σε βάθος και να χαρτογραφήσουν σωστά τη θέση του Αυτισμού στο ευρύτερο βιολογικό περιβάλλον.

3.4 Μοντελοποίηση και Υλοποίηση του Δικτύου σε περιβάλλον Python (NetworkX)

Μετά την ολοκλήρωση της προεπεξεργασίας και τη δημιουργία του τελικού στοχευμένου πίνακα δεδομένων, το επόμενο στάδιο περιλαμβάνει τη φυσική κατασκευή και υλοποίηση του γράφου σε προγραμματιστικό περιβάλλον. Για τη διαδικασία αυτή χρησιμοποιείται η γλώσσα προγραμματισμού Python η οποία, σε συνδυασμό με τη βιβλιοθήκη NetworkX (Hagberg et al, 2008), αποτελεί ένα από τα σημαντικότερα εργαλεία στον τομέα εφαρμογών Βιοπληροφορικής

Η NetworkX είναι μία βιβλιοθήκη ανοιχτού κώδικα υψηλού επιπέδου, σχεδιασμένη για τη δημιουργία, τον χειρισμό και τη μελέτη της δομής σύνθετων δικτύων. Στο πλαίσιο της παρούσας εργασίας, η επιλογή της βασίστηκε στην ικανότητά της να διαχειρίζεται ετερογενείς δομές δεδομένων, όπως οι διμερείς γράφοι, και να αποδίδει δυναμικά ιδιότητες (attributes) στους κόμβους και τις ακμές του δικτύου. Εσωτερικά, η NetworkX βασίζεται σε δομές δεδομένων της Python (dictionaries), γεγονός που επιτρέπει την ταχύτατη αναζήτηση και προσθήκη στοιχείων με βέλτιστη υπολογιστική πολυπλοκότητα.

Η διαδικασία της υλοποίησης ξεκινά με την δημιουργία ενός εντελώς άδειου γράφου στη μνήμη του υπολογιστή, ο οποίος στην αρχή δεν περιέχει κανέναν κόμβο και καμία σύνδεση.

Στη συνέχεια, ο κώδικας διαβάζει μία προς μία τις γραμμές του φιλτραρισμένου πίνακα δεδομένων, απομονώνοντας την ασθένεια και το αντίστοιχο γονίδιο κάθε καταγραφής.

Κατά τη διάρκεια αυτής της διαδικασίας, ο αλγόριθμος ελέγχει αν ένας κόμβος υπάρχει ήδη στον γράφο. Αν μία ασθένεια ή ένα γονίδιο εμφανίζεται σε πολλές διαφορετικές γραμμές του πίνακα, ο υπολογιστής δεν ξαναδημιουργεί τον κόμβο, αλλά δημιουργεί μία ακμή για τον συνδέσει.

Το τελικό αποτέλεσμα είναι η δημιουργία ενός ολοκληρωμένου ψηφιακού δικτύου, το οποίο αποτυπώνει πιστά τις βιολογικές σχέσεις γύρω από τον Αυτισμό και είναι έτοιμο για την εφαρμογή αλγορίθμων ανάλυσης.

4.Υπολογιστική Υλοποίηση και Παραμετροποίηση του Αλγορίθμου Node2Vec

4.1 Αρχιτεκτονική του Συστήματος και Pipeline Επεξεργασίας

Η υπολογιστική προσέγγιση που αναπτύχθηκε στην παρούσα εργασία βασίζεται σε μία συγκεκριμένη Αρχιτεκτονική Συστήματος, συγκεκριμένα τριών επιπέδων, η οποία υλοποιεί ένα Data Pipeline σε περιβάλλον Python. Η επιλογή αυτής της αρχιτεκτονικής δομής έγινε για τον διαχωρισμό των εισαχθέντων δεδομένων από τους αλγόριθμους και το τελικό επίπεδο παρουσίασης των αποτελεσμάτων. Με αυτό τον τρόπο το σύστημα δεν αποτελεί απλά ένα σενάριο κώδικα (script), αλλά ένα δομημένο υπολογιστικό περιβάλλον, όπου κάθε επίπεδο εκτελεί έναν συγκεκριμένο ρόλο και τροφοδοτεί το επόμενο μέσα από την χρήση του Data Pipeline.

Η αρχιτεκτονική του συστήματος οργανώνεται στα ακόλουθα τρία κεντρικά επίπεδα, τα οποία εκτελούνται σειριακά μέσα από πέντε (5) διαδοχικά βήματα:

1. Επίπεδο Δεδομένων (Data layer): Το επίπεδο αυτό αποτελεί την αφετηρία του συστήματος και είναι υπεύθυνο για αποκλειστικά για την εισαγωγή, τον έλεγχο καθώς και την προετοιμασία των δεδομένων. Στο συγκεκριμένο επίπεδο εκτελείται το πρώτο βήμα, όπου είναι η Φόρτωση και ο Καθαρισμός των δεδομένων. Το σύστημα πραγματοποιεί έλεγχο για την διαθεσιμότητα του αρχείου **'DG-AssocMiner_miner-disease-gene.tsv'** μέσω της βιβλιοθήκης **os**. Εφόσον το αρχείο εντοπιστεί, η βιβλιοθήκη **pandas** αναλαμβάνει να το φορτώσει στη μνήμη σε μορφή πίνακα (DataFrame) με τις στήλες **'Disease_ID'**, **'Disease_Name'** και **'Gene_ID'**. Σε αυτό το σημείο πραγματοποιείται ο καθαρισμός δεδομένων (data sanitization) με την μέθοδο **'.str.replace('"', '')'**, η οποία αφαιρεί τα διπλά εισαγωγικά που προκαλούν θόρυβο στα ονόματα των ασθενειών, και την **'.str.strip()'**, η οποία απομακρύνει τυχόν περιττά κενά διαστήματα.

Τμήμα κώδικα 1: *Προεπεξεργασία δεδομένων BioSNAP*

```
file_path = 'DG-AssocMiner_miner-disease-gene.tsv'

if not os.path.exists(file_path):
    print(f"Σφάλμα: Το αρχείο {file_path} δεν βρέθηκε στον φάκελο.")
else:
    print("Φόρτωση δεδομένων...")
    df = pd.read_csv(file_path, sep='\t', comment='#', header=None,
                     names=['Disease_ID', 'Disease_Name', 'Gene_ID'])

    df['Disease_Name'] = df['Disease_Name'].astype(str).str.replace('"', '').str.strip()
```

2. Επίπεδο Επεξεργασίας (Processing Layer): Αποτελεί το σημαντικότερο κομμάτι του συστήματος, καθώς δέχεται τα πιο καθαρισμένα δεδομένα από το Data Layer και αναλαμβάνει όλο το δύσκολο αλγοριθμικό κομμάτι της μοντελοποίησης. Το επίπεδο χωρίζεται σε δύο βασικά υποσυστήματα, με πρώτο να είναι το Σύστημα Μοντελοποίησης του γράφου και δεύτερο η Εκμάθηση Αναπαραστάσεων με την εισαγωγή του αλγορίθμου node2vec. Στο κομμάτι της μοντελοποίησης, όπου αποτελεί και το δεύτερο βήμα της επεξεργασίας, ο κώδικα μεταφράζει τις γραμμές του πίνακα σε μία αντικειμενοστραφή δομή δικτύου. Αρχικά γίνεται η αρχικοποίηση ενός μη-κατευθυνόμενου γράφου με την χρήση της NetworkX, μέσα από την εντολή **‘nx.Graph()’**. Στη συνέχεια, με τη χρήση μίας επανάληψης (**for loop**) που βασίζεται στη συνάρτηση **‘df.iterrows()’** , το σύστημα τρέχει σειριακά τον πίνακα δεδομένων γραμμή προς γραμμή. Η συνάρτηση αυτή χωρίζει το DataFrame, επιτρέποντας στον κώδικα να απομονώσεις σε κάθε βήμα τα περιεχόμενα της εκάστοτε γραμμής (row), αγνοώντας των δείκτη της. Έτσι, κάθε ζεύγος ασθένειας (**row['Disease_Name']**) και γονιδίου (**row['Gene_ID']**) εξετάζεται μεμονωμένα και συνδέεται αυτόματα με μία ακμή (edge) μέσω της εντολής **‘G.add_edge()’**. Για λόγους ομοιομορφίας και ορθού διαχωρισμού των κόμβων, στα αναγνωριστικά των γονιδίων προστίθεται αυτόματα το πρόθεμα **"Gene_"**. Το αποτέλεσμα αυτής της επανάληψης είναι η κατασκευή ενός διμερούς γράφου, ο οποίος περιλαμβάνει το σύνολο όλων των βιολογικών συσχετίσεων.

Τμήμα κώδικα 2: *Δημιουργία και αρχικοποίηση του γράφου*

```
print("Δημιουργία γράφου...")
G = nx.Graph()
for _, row in df.iterrows():
    disease = row['Disease_Name']
    gene = "Gene_" + str(row['Gene_ID'])
    G.add_edge(disease, gene)

print(f"Ο γράφος δημιουργήθηκε με {G.number_of_nodes()} κόμβους.")
```

Στο τρίτο βήμα της επεξεργασίας αποτελεί η παραμετροποίηση και η εκπαίδευση του μοντέλου node2vec ο οποίος θα αναλυθεί περαιτέρω στο υποκεφάλαιο 4.2.

3. Επίπεδο Εφαρμογής και Παρουσίασης (Presentation Layer): Το επίπεδο αυτό αναλαμβάνει την εξαγωγή της βιολογικής γνώσης και τη γεωμετρική απεικόνιση του χώρου των embeddings. Αντίστοιχα με το προηγούμενο επίπεδο, και αυτό χωρίζεται σε δύο υποσυστήματα, το πρώτο αναζητά και εκτυπώνει την Ομοιότητα Συνημίτονου και το δεύτερο αναλαμβάνει την Οπτικοποίηση t-SNE η οποία καταλαμβάνει διαφορετικό υποκεφάλαιο της εργασίας.

Το πρώτο υποσύστημα, και το τέταρτο βήμα της επεξεργασίας, το σύστημα εκτελεί μία αυτοματοποιημένη αναζήτηση εντοπίζοντας τον κόμβο-στόχο που σχετίζεται με τον Αυτισμό (αναζητώντας τις λέξεις-κλειδιά "autism" ή "autistic"). Μόλις εντοπιστεί ο επίσημος όρος ('Autistic Disorder'), η συνάρτηση **‘.wv.most_similar()’** υπολογίζει την ομοιότητα συνημίτονου ανάμεσα στο διάνυσμα του Αυτισμού και όλων των άλλων κόμβων, επιστρέφοντας τη λίστα κατάταξης με τα 10 επικρατέστερα υποψήφια γονίδια

Τμήμα κώδικα 3: *Υπολογισμός Cosine Similarity*

```
search_term = None
for node in model.wv.index_to_key:
    if "autism" in node.lower() or "autistic" in node.lower():
        search_term = node
        break

if search_term:
    print(f"\n--- ΑΠΟΤΕΛΕΣΜΑΤΑ COSINE SIMILARITY ΓΙΑ: {search_term} ---")
    similar = model.wv.most_similar(search_term, topn=10)
    for name, score in similar:
        print(f"{name}: {score:.4f}")
else:
    print("\nΟ όρος 'Autism' δεν βρέθηκε.")
```

4.2 Παραμετροποίηση και Εκπαίδευση του Αλγορίθμου Node2Vec

Όπως προαναφέρθηκε, στο τρίτο βήμα της επεξεργασίας ο γράφος που δημιουργήθηκε εισάγεται στην κλάση node2vec, με σκοπό να μετατραπούν οι συνδέσεις του δικτύου σε μία σειρά από αριθμούς. Η διαδικασία αυτή βασίζεται στην εκτέλεση μεροληπτικών τυχαίων περιπάτων, οι οποίοι ελέγχονται από συγκεκριμένες υπερπαραμέτρους (hyperparameters) που καθορίζουν την εξερεύνηση του δικτύου BioSNAP.

Η παράμετρος **‘walk_length=20’** καθορίζει ότι κάθε τυχαίος περίπατος θα έχει μήκος 20 διαδοχικών βημάτων από κόμβο σε κόμβο, επιτρέποντας την ισορροπημένη εξερεύνηση τόσο της τοπικής γειτονιάς όσο και της ευρύτερης δομής του γράφου.

Ο αριθμός αυτών των διαδρομών ορίζεται από την παράμετρο **‘num_walks=100’**, η οποία επιβάλλει την εκκίνηση 100 ανεξάρτητων περιπάτων από κάθε κόμβο του δικτύου, διασφαλίζοντας τη συγκέντρωση ενός πλούσιου όγκου στατιστικών δεδομένων για τη γεωμετρία του γράφου.

Βέβαια, πέρα από τις παραμέτρους που αναφέρθηκαν παραπάνω, οι σημαντικότεροι παράμετροι για την υλοποίηση του αλγορίθμου είναι οι ***p*** και ***q***. Στην παρούσα υλοποίηση, οι παράμετροι επιστροφής p (return parameter) και μετακίνησης q (in-out parameter)

διατηρούν τις προεπιλεγμένες τους τιμές (p=1.0 και q=1.0). Η παράμετρος p ελέγχει την πιθανότητα επιστροφής του τυχαίου περιπάτου στον αμέσως προηγούμενο κόμβο. Αντίστοιχα, η παράμετρος q καθορίζει εάν η περιήγηση θα περιοριστεί τοπικά στην ίδια γειτονιά, προσεγγίζοντας τη συμπεριφορά της αναζήτησης σε πλάτος (BFS), ή εάν θα απομακρυνθεί προς βαθύτερα επίπεδα του δικτύου, προσομοιώνοντας την αναζήτηση σε βάθος (DFS) (Grover & Leskovek, 2016).

Τέλος, μέσω της μεθόδου **'f()'** οι αλυσίδες των κόμβων που παράχθηκαν εισάγονται στο μοντέλο **Skip-Gram** της βιβλιοθήκης **Gensim** (Rehurek & Sojka, 2010). Εκεί, με την συνθήκη **'window=10'**, το μοντέλο εξετάζει τη μέγιστη απόσταση 10 κόμβων εντός του ίδιου περιπάτου για να βελτιστοποιήσει τις διανυσματικές αναπαραστάσεις (Mikolov et al., 2013), ενώ η επιλογή **'min_count=1'** εγγυάται ότι κανένας κόμβος δεν θα αποκλειστεί από την εκπαίδευση, ακόμη και αν διαθέτει μία μόνο σύνδεση.

Το τελικό αποτέλεσμα θα είναι η παραγωγή ενός σταθερού πίνακα συνεχών διανυσμάτων διάστασης ***d=64***, ο οποίος αποθηκεύεται στο αντικείμενο **model**. Ενώ στην αρχική βιβλιογραφία των Grover & Leskovec (2016) ως τυπική τιμή αναφοράς χρησιμοποιούνται οι 128 διαστάσεις, στην παρούσα υλοποίηση επιλέχθηκαν οι 64 διαστάσεις. Η συγκεκριμένη παραμετροποίηση αποτελεί καθιερωμένη πρακτική σε βιοπληροφορικές αναλύσεις δικτύων παρόμοιας κλίμακας (όπως το BioSNAP), καθώς προσφέρει την ιδανική υπολογιστική ισορροπία, αποτρέποντας το φαινόμενο της υπερπροσαρμογής (overfitting) ,μειώνοντας τις απαιτήσεις μνήμης.

Τμήμα κώδικα 4: *Παραμετροποίηση και εκπαίδευση Node2vec*

```
print("\nΕκπαίδευση μοντέλου Node2Vec...")
node2vec = Node2Vec(G, dimensions=64, walk_length=20, num_walks=100, workers=4)
model = node2vec.fit(window=10, min_count=1)
```

5.Πειραματικά Αποτελέσματα

5.1 Τοπολογική Συμπεριφορά του Αλγορίθμου στο Δίκτυο Εισόδου

Η παρουσίαση των αποτελεσμάτων ξεκινά με την εξέταση του τρόπου με τον οποίο η ίδια η μορφή του γράφου BioSNAP επηρέασε τη δημιουργία των διανυσμάτων. Όπως αναλύθηκε σε προηγούμενο κεφάλαιο, το συγκεκριμένο δίκτυο είναι αρκετά αραιό αλλά περιέχει ορισμένους κεντρικούς κόμβους με πάρα πολλές συνδέσεις, όπως η ασθένεια του Αυτισμού. Κατά τη διάρκεια του πειράματος, η ιδιαίτερη αυτή δομή καθόρισε πλήρως τη διαδρομή που ακολούθησαν οι τυχαίοι περίπατοι δεύτερης τάξης του node2vec, καθώς οι πιθανότητες μετακίνησης του αλγορίθμου εξαρτώνται άμεσα από το πώς συνδέονται οι γειτονικοί κόμβοι μεταξύ τους (Grover & Leskovec, 2016). Εξαιτίας του μεγάλου αριθμού των συνδέσεων που έχουν οι κεντρικές ασθένειες, οι περίπατοι πέρασαν πολύ συχνά από αυτές, μαζεύοντας έτσι πληροφορίες για τις άμεσες και έμμεσες βιολογικές τους σχέσεις.

Στο σημείο αυτό, η επιλογή να διατηρηθούν οι προεπιλεγμένες τιμές στις παραμέτρους p και q αποδείχτηκε ιδιαίτερα σημαντική. Η ρύθμιση αυτή επέτρεψε στον αλγόριθμο να μην «κολλήσει» αποκλειστικά γύρω από μεγάλους και πυκνούς κόμβους, αλλά να συνεχίσει την περιήγησή του με φυσικό τρόπο μέσα στο δίκτυο. Με αυτόν τον τρόπο εξασφαλίστηκε ότι ακόμα και γονίδια με ελάχιστες συνδέσεις, τα οποία είναι λιγότερο μελετημένα, καταγράφηκαν σωστά μέσα στον διανυσματικό χώρο. Τελικά, η εκπαίδευση του μοντέλου Skip-Gram βασίστηκε καθαρά στη γεωμετρία του γράφου, μετατρέποντας τις συνδέσεις του δικτύου σε συγκεκριμένες αποστάσεις μέσα στον χώρο των 64 διαστάσεων. Αυτή η διαδικασία δημιούργησε τη βάση για να βρεθούν οι σημαντικές βιολογικές συσχετίσεις που παρουσιάζονται στη συνέχεια της εργασίας.

5.2 Εξαγωγή Συσχετίσεων και Κατάταξη Υποψήφιων Γονιδίων

Μετά την ολοκλήρωση της εκπαίδευσης του αλγορίθμου node2vec, οι τοπολογικές ιδιότητες του δικτύου BioSNAP έχουν μετατραπεί επιτυχώς σε διανύσματα σταθερού μεγέθους εντός του χώρου αναπαράστασης. Το επόμενο στάδιο της πειραματικής διαδικασίας αφορά την αξιοποίηση αυτών των διανυσμάτων για την ανακάλυψη νέων βιολογικών συσχετίσεων, με επίκεντρο τη νευροαναπτυξιακή διαταραχή του Αυτισμού. Για την ποσοτική αποτίμηση της συνάφειας μεταξύ της ασθένειας-στόχου και των γονιδίων του γράφου, χρησιμοποιήθηκε η μετρική της ομοιότητας συνημίτονου. Η συγκεκριμένη μετρική υπολογίζει το συνημίτονο της γωνίας που σχηματίζουν τα αντίστοιχα διανύσματα στον χώρο των 64 διαστάσεων, λαμβάνοντας τιμές στο διάστημα [0,1], όπου 1 υποδηλώνει απόλυτη διανυσματική ταύτιση και ευθυγράμμιση.

Μέσω της συνάρτησης '.most.similar()', η οποία εξηγήθηκε στο προηγούμενο κεφάλαιο, το σύστημα υπολόγισε το Cosine Similarity ανάμεσα στο διάνυσμα του όρου 'Autistic Disorder' και του συνόλου των διαθέσιμων γονιδίων του δικτύου. Στον ακόλουθο **Πίνακα 5.1** παρουσιάζονται τα δέκα επικρατέστερα γονίδια, τα οποία εμφάνισαν τον υψηλότερο βαθμό ομοιότητας με την ασθένεια-στόχο.

Πίνακας 5.1: Τα 10 Κορυφαία Υποψήφια Γονίδια με βάση το Cosine Similarity

Κατάταξη	Αναγνωριστικό Γονιδίου (Gene ID)	Επίσημο Σύμβολο Γονιδίου (Gene Symbol)	Σκορ Ομοιότητας Συνημίτονου
1	Gene_282553	AUTS8	0.9686
2	Gene_8456	FOXN1	0.9661
3	Gene_152789	JAKMIP1	0.9661
4	Gene_64326	COP1	0.9654
5	Gene_54538	ROBO4	0.9644

6	Gene_9228	DLGAP2	0.9639
7	Gene_9732	DOCK4	0.9637
8	Gene_643586(*)	PKMP1	0.9634
9	Gene_6999	TDO2	0.9626
10	Gene_1600	DAB1	0.9625

5.3 Βιολογική Τεκμηρίωση και Αξιολόγηση Αποτελεσμάτων

Η αξιοπιστία ενός υπολογιστικού μοντέλου στη βιοπληροφορική εξαρτάται από την ικανότητά του να παράγει αποτελέσματα με πραγματικό βιολογικό νόημα. Για τον λόγο αυτό, στο παρόν υποκεφάλαιο πραγματοποιείται ποιοτική αξιολόγηση των δέκα πρώτων γονιδίων της λίστας, συγκρίνοντάς τα με τη διεθνή ιατρική και γενετική βιβλιογραφία. Σημειώνεται ότι η ταυτοποίηση και οι επίσημες ονομασίες των γονιδίων αντλήθηκαν και επαληθεύτηκαν σύμφωνα με τα πρότυπα αναφοράς της βάσης δεδομένων NCBI. Η ανάλυση δείχνει ότι ο αλγόριθμος node2vec, βασιζόμενος αποκλειστικά και μόνο στον τρόπο που συνδέονται οι κόμβοι στο δίκτυο BioSNAP, κατάφερε να ανακαλύψει ξανά γονίδια που είναι αποδεδειγμένα κρίσιμα για τις διαταραχές του Αυτιστικού Φάσματος:

1. **Gene_282553 (AUTS8 - Autism Susceptibility Candidate 8):** Το γονίδιο αυτό κατέλαβε την πρώτη θέση στην κατάταξη με σκορ 0.9686. Η ανάδειξή του στην κορυφή αποτελεί ισχυρή επιβεβαίωση για το μοντέλο, καθώς το AUTS8 (αλλιώς AUTS2) αποτελεί έναν διεθνώς αναγνωρισμένο γενετικό τόπο (locus) που έχει ταυτοποιηθεί για την άμεση προδιάθεση και ευπάθεια εμφάνισης Αυτισμού. Οι συγκεκριμένες περιοχές των χρωμοσωμάτων και ο τρόπος που κληρονομούνται σε οικογένειες με Αυτισμό έχουν μελετηθεί αναλυτικά μέσα από μεγάλες γενετικές έρευνες (Auranen et al., 2002).
2. **Gene_8456 (FOXN1- forkhead box N1):** Το γονίδιο αυτό κατέλαβε τη δεύτερη θέση με σκορ 0.9661. Αν και το FOXN1 μελετάται παραδοσιακά για τη σημασία του στο ανοσοποιητικό σύστημα, σύγχρονες έρευνες το συνδέουν πλέον άμεσα με τον Αυτισμό.

Επιστημονικά δεδομένα αποδεικνύουν ότι το γονίδιο αυτό εκφράζεται κανονικά στον αναπτυσσόμενο εγκέφαλο, και η δυσλειτουργία του, σε συνδυασμό με άλλες γενετικές αλλαγές, μπορεί να επηρεάσει αρνητικά τη δημιουργία των νευρικών κυκλωμάτων (Moreno-Ramos et al., 2015).

3. **Gene_152789 (JAKMIP1- janus kinase and microtubule interacting protein 1):** Εντοπίστηκε στην τρίτη θέση με σκορ 0.9661. Στη σύγχρονη νευροβιολογία, το γονίδιο JAKMIP1 παίζει κρίσιμο ρόλο στην ανάπτυξη του εγκεφάλου, καθώς δρα ως ρυθμιστής που ελέγχει την παραγωγή πρωτεϊνών στα νευρικά κύτταρα. Η γενετική δυσλειτουργία του γονιδίου οδηγεί σε ανεξέλεγκτη υπερπαραγωγή πρωτεϊνών στις νευρικές συνάψεις, γεγονός που σχετίζεται άμεσα με την εμφάνιση στερεοτυπικών συμπεριφορών και ελλειμμάτων στην κοινωνική αλληλεπίδραση (Berg et al., 2015).

4. **Gene_64326 (COP1- COP1 E3 ubiquitin ligase):** Κατέλαβε την τέταρτη θέση με σκορ ομοιότητας 0.9654. Το γονίδιο αυτό κωδικοποιεί μία πρωτεΐνη που δρα ως E3 ubiquitin ligase και συμμετέχει ενεργά στο μονοπάτι της ουβικιτινίωσης (ubiquitination). Εκτενείς γονιδιωματικές μελέτες δείχνουν ότι η απώλεια της μοριακής ισορροπίας των πρωτεϊνών λόγω δυσλειτουργίας του COP1 επηρεάζει αρνητικά τη συναπτική πλαστικότητα, αποτελώντας κεντρικό βιολογικό μηχανισμό για την εκδήλωση του Αυτισμού (Glessner et al., 2009).

5. **Gene_54538 (ROBO4 - Roundabout Guidance Receptor 4):** Κατέλαβε την πέμπτη θέση με σκορ 0.9644. Το γονίδιο αυτό ανήκει σε μία οικογένεια υποδοχέων που είναι καθοριστικοί για την καθοδήγηση των αξόνων των νευρώνων (axon guidance). Βοηθάει τα νευρικά κύτταρα να αναπτυχθούν προς τις σωστές κατευθύνσεις ώστε να δημιουργηθεί ένα σωστό δίκτυο στον εγκέφαλο. Γενετικές αναλύσεις δείχνουν ότι συγκεκριμένες παραλλαγές του γονιδίου ROBO4 παρουσιάζουν σημαντική στατιστική συσχέτιση με τον Αυτισμό (Anitha et al., 2008).

6. **Gene_9228 (DLGAP2- DLG associated protein 2):** Βρίσκεται στην έκτη θέση της κατάταξης με σκορ 0.9639. Το γονίδιο αυτό κωδικοποιεί μία δομική πρωτεΐνη υποστήριξης (scaffold protein) στις ενώσεις των νευρώνων. Πειραματικά δεδομένα αποδεικνύουν ότι η έλλειψη λειτουργίας του DLGAP2 προκαλεί ελλείμματα στη συναπτική επικοινωνία σε συγκεκριμένες περιοχές του εγκεφάλου και σε οδηγεί σε προβλήματα συμπεριφοράς, επιβεβαιώνοντας ότι αποτελεί σημαντικό παράγοντα κινδύνου για το φάσμα του Αυτισμού (Jiang-Xie et al., 2014).

7. **Gene_9732 (DOCK4- dedicator of cytokinesis 4):** Κατέλαβε την έβδομη θέση με σκορ 0.9637. Το γονίδιο αυτό είναι απαραίτητο για τη σωστή ανάπτυξη των δενδριτών και τον σχηματισμό των νευρικών συνάψεων. Η διακοπή της λειτουργίας του DOCK4 έχει συνδεθεί κλινικά με τη χαρακτηριστική συμπτωματολογία του Αυτισμού, καθώς και με μαθησιακές δυσκολίες (Pagnamenta et al., 2010).

8. **Gene_643586 (PKMP1- pyruvate kinase M1/2 pseudogene 1):** Εντοπίστηκε στην όγδοη θέση με σκορ 0.9634. Πρόκειται για ένα ψευδογονίδιο. Από υπολογιστική άποψη, η ανάδυσή του είναι πολύ σημαντική, καθώς μεγάλες στατιστικές μελέτες συσχέτισης (GWAS) δείχνουν ότι αυτή η περιοχή του γονιδιώματος κρύβει σήματα κινδύνου για τον αυτισμό. Ο αλγόριθμος κατάφερε να εντοπίσει αυτή τη σχέση αναλύοντας αποκλειστικά τις συνδέσεις στον γράφο **(**Xia et al., 2013).

9. **Gene_6999 (TDO2- tryptophan 2,3-dioxygenase):** Κατέλαβε την ένατη θέση στην κατάταξη με σκορ 0.9626. Το γονίδιο αυτό κωδικοποιεί ένα ένζυμο που ελέγχει τον μεταβολισμό της τρυπτοφάνης, η οποία είναι ο πρόδρομος της σεροτονίνης. Η ανάδυσή του επιβεβαιώνεται από εκτενείς οικογενειακές στατιστικές αναλύσεις, όπως το τεστ TDT, οι οποίες δείχνουν ότι συγκεκριμένοι απλότυποι και παραλλαγές του TDO2 παρουσιάζουν σημαντική στατιστική συσχέτιση με τις διαταραχές του Αυτιστικού Φάσματος (Nabi et al., 2004).

10. **Gene_1600 (DAB1- DAB adaptor protein 1):** Εντοπίστηκε στην δέκατη θέση με σκορ 0.9625. Στη βιβλιογραφία, το γονίδιο αυτό είναι γνωστό για τον ρόλο που παίζει στη σωστή μετακίνηση και τοποθέτηση των κυττάρων κατά την ανάπτυξη του εγκεφάλου. Η γεωμετρική εγγύτητα (vector proximity) του DAB1 με τα υπόλοιπα γονίδια-στόχους στον διανυσματικό χώρο συμφωνεί με τα στατιστικά δεδομένα, τα οποία συνδέουν τυχόν σφάλματα σε αυτό το μονοπάτι με αυξημένο κίνδυνο εμφάνισης Αυτισμού (Fatemi, 2005).

Συμπερασματικά, το γεγονός ότι το μοντέλο Node2Vec κατάφερε να εντοπίσει και να ιεραρχήσει γονίδια με αποδεδειγμένη σύνδεση με τον αυτισμό, βασιζόμενο αποκλειστικά και μόνο στη δομή του δικτύου (network topology), επιβεβαιώνει την ορθότητα της υπολογιστικής ροής δεδομένων. Η ανάλυση αυτή αναδεικνύει την ισχύ των αλγορίθμων Μηχανικής Μάθησης σε Γράφους (Graph Representation Learning) στην ανακάλυψη κρυμμένων στατιστικών μοτίβων μέσα από σύνθετα δίκτυα δεδομένων.

5.4 Οπτικοποίηση Διανυσματικού Χώρου μέσω t-SNE

Μετά την εκπαίδευση του μοντέλου node2vec, κάθε κόμβος του δικτύου έχει μετατραπεί σε ένα διάνυσμα πολλών διαστάσεων (high-dimensional embedding). Επειδή η απευθείας οπτικοποίηση και ανάλυση δεδομένων σε τόσες πολλές διαστάσεις είναι αδύνατη από τον άνθρωπο, κρίνεται απαραίτητη η χρήση τεχνικών μείωσης διαστάσεων (dimensionality reduction). Για τον σκοπό αυτό, χρησιμοποιείται ο μη-γραμμικός αλγόριθμος **t-SNE** (t-Distributed Stochastic Neighbor Embedding).

Ο αλγόριθμος αναλαμβάνει να προβάλει τα πολυδιάστατα διανύσματα στον δισδιάστατο χώρο, φροντίζοντας παράλληλα να διατηρήσεις τις τοπικές ομοιότητες των δεδομένων. Με απλά λόγια, κόμβοι που βρίσκονταν κοντά στον πολυδιάστατο χώρο επειδή είχαν παρόμοιο τοπολογικό ρόλο στο δίκτυο, τοποθετούνται κοντά και στο τελικό δισδιάστατο γράφημα.

Για την υλοποίηση αυτού του σταδίου, το σύστημα μετατρέπει τα διανύσματα σε πίνακα NumPy (**np.array**) και ενεργοποιεί το αλγόριθμο t-SNE της **'scikit-learn'**. Η γραμμή **'perp_value = min(30, len(all_nodes) - 1)'** ορίζει στο περίπου πόσους κοντινούς γείτονες θα λάβει υπόψη ο αλγόριθμος για κάθε σημείο. Ο αλγόριθμος ορίζει ότι ο τελικός χώρος εξόδου θα αποτελείται από 2 διαστάσεις (X και Y), διατηρώντας τις σχετικές αποστάσεις των σημείων στον χώρο[1]. Μετά υπάρχει η παράμετρος **'random_state=42'** η οποία χρησιμοποιείται για να δείξει ότι κάθε φορά που τρέχει ο κώδικας οι τελείες θα κάθονται ακριβώς στις ίδιες θέσεις. Στη συνέχεια, χρησιμοποιείται η μέθοδος PCA για την αρχική τοποθέτηση των σημείων–εξασφαλίζοντας ότι η οπτικοποίηση θα είναι πιο σταθερή. Η εντολή **'tsne.fit_transform(vectors_np)'** εκτελεί τον μαθηματικό μετασχηματισμό. Ο αλγόριθμος υπολογίζει τις νέες συντεταγμένες των κόμβων βάσει των πιθανοτήτων γειτνίασής τους, διασφαλίζοντας ότι τα στοιχεία που παρουσιάζουν παρόμοια συμπεριφορά θα τοποθετηθούν κοντά το ένα στο άλλο, επιτρέποντας έτσι τον άμεσο εντοπισμό οπτικών συστάδων (Kobak & Berens, 2019; Pedregosa et al., 2011; Van der Maaten & Hinton, 2008).

[1] Στην προγραμματιστική υλοποίηση μέσω της βιβλιοθήκης scikit-learn η διατήρηση των δύο διαστάσεων για την οπτικοποίηση επιτυγχάνεται με τον καθορισμό της παραμέτρου n_components=2

Τμήμα κώδικα 5: *Κώδικας αλγορίθμου t-SNE*

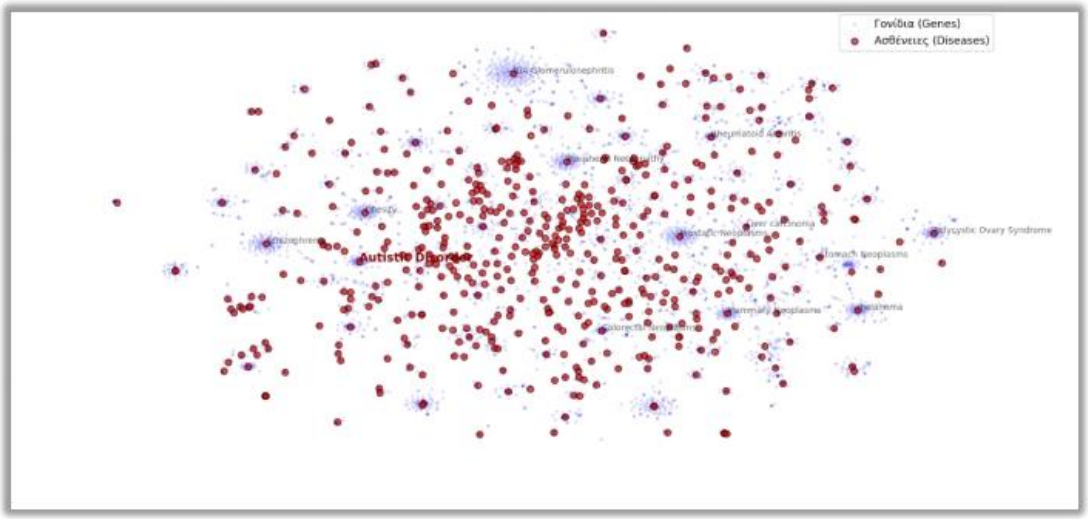
```
all_nodes = [node for node in model.wv.index_to_key if node in G.nodes()]
vectors = [model.wv[node] for node in all_nodes]
vectors_np = np.array(vectors)

perp_value = min(30, len(all_nodes) - 1) if len(all_nodes) > 1 else 1
tsne = TSNE(n_components=2, random_state=42, init='pca', learning_rate='auto', perplexity=perp_value)
Y = tsne.fit_transform(vectors_np)

all_diseases_set = set(df['Disease_Name'].unique())
disease_indices = [i for i, node in enumerate(all_nodes) if node in all_diseases_set]
gene_indices = [i for i, node in enumerate(all_nodes) if node not in all_diseases_set]

plt.figure(figsize=(14, 10))

# Σχεδίαση Γονιδίων (Μπλε - Genes)
plt.scatter(Y[gene_indices, 0], Y[gene_indices, 1],
            c='blue', edgecolors='none', alpha=0.15, s=8, label='Γονίδια (Genes)')
# Σχεδίαση Ασθενειών (Κόκκινο - Diseases)
plt.scatter(Y[disease_indices, 0], Y[disease_indices, 1],
            c='red', edgecolors='black', alpha=0.7, s=35, label='Ασθένειες (Diseases)')

# Προσθήκη ετικετών (Labels)
count = 0
for i, label in enumerate(all_nodes):
    if label == search_term:
        plt.annotate(label, xy=(Y[i, 0], Y[i, 1]), fontsize=11, weight='bold', color='darkred')
    elif label in all_diseases_set and count < 12:
        plt.annotate(label, xy=(Y[i, 0], Y[i, 1]), fontsize=8, alpha=0.6)
        count += 1

plt.title("t-SNE Visualization: Διαχωρισμός Ασθενειών (Κόκκινο) και Γονιδίων (Μπλε)", fontsize=14)
plt.legend(loc='upper right')
plt.axis('off')

print("\nΕμφάνιση γραφήματος...")
plt.show()
```

Το τελικό τμήμα του κώδικα αναλαμβάνει τη γραφική αναπαράσταση των αποτελεσμάτων μέσω της βιβλιοθήκης **matplotlib** , δημιουργώντας ένα δισδιάστατο διάγραμμα διασποράς (scatter plot). Τα γονίδια σχεδιάζονται με μικρό μέγεθος και χαμηλή διαφάνεια (**alpha=0.15**) για την αποφυγή του υπερκορεσμού των σημείων, ενώ οι ασθένειες εμφανίζονται ως μεγαλύτερες κόκκινες κουκκίδες για άμεσο οπτικό διαχωρισμό. Παράλληλα, χρησιμοποιείται η συνάρτηση **'plt.annotate'** για την ανάδειξη του όρου-στόχου του Αυτισμού με έντονα γράμματα για να ξεχωρίζει αμέσως, προκειμένου η εικόνα να μην γίνει δυσανάγνωστη. Τέλος, η εντολή **'plt.axis('off')'** αφαιρεί τους άξονες, καθώς η βιολογική πληροφορία κρύβεται αποκλειστικά στη σχετική γεωμετρική εγγύτητα και στις συστάδες των κόμβων (Hunter, 2007; Wilke, 2019; Moon et al., 2019).

Η γραφική αυτή απόδοση αποτυπώνει με σαφήνεια τα πειραματικά αποτελέσματα, καθώς οι κόμβοι αναπαρίστανται ως σημεία στο επίπεδο. Η οπτικοποίηση αποκαλύπτει τη διαμόρφωση ξεκάθαρων συστάδων, γεγονός που επιβεβαιώνει ότι ο node2vec κατάφερε να διαχωρίσει και να ομαδοποιήσει τα δεδομένα με τα κοινά τους χαρακτηριστικά.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η θέση των δέκα κορυφαίων γονιδίων που αναλύθηκαν στο υποκεφάλαιο 5.3. Στο γράφημα παρατηρείται ότι τα γονίδια αυτά δεν είναι διάσπαρτα στην τύχη, αλλά εντοπίζονται συγκεντρωμένα στην ίδια γεωμετρική περιοχή. Η οπτική αυτή εγγύτητα αποτελεί την τελική υπολογιστική επιβεβαίωση της ροής δεδομένων, καθώς αποδεικνύει γραφικά ότι η μαθηματική ομοιότητα των διανυσμάτων μεταφράζεται σε κοινή βιολογική συνάφεια. Συνεπώς, αναδεικνύεται η επιτυχία του μοντέλου στην αναγνώριση υποψηφίων γονιδίων για τις διαταραχές του Αυτιστικού Φάσματος.

Εικόνα 5.1: *Οπτικοποίηση των embeddings με τον αλγόριθμο t-SNE*



6.Αξιολόγηση και Συζήτηση

6.1 Υπολογιστική Αποτίμηση: Γενίκευση του μοντέλου σε άγνωστες συσχετίσεις

Η επιτυχία ενός μοντέλου μηχανικής μάθησης βασίζεται στην ικανότητά του να γενικεύει. Στην επιστήμη των δεδομένων, αυτό σημαίνει ότι το μοντέλο πρέπει να μπορεί να εφαρμόσει όσα «έμαθε» κατά την εκπαίδευσή του σε νέα δεδομένα, τα οποία δεν έχει ξαναδεί, και να βγάλει σωστά αποτελέσματα χωρίς να κάνει απλή αποστήθιση.

Στη συγκεκριμένα πειραματική διαδικασία, ο αλγόριθμος node2vec εκπαιδεύτηκε αποκλειστικά με βάση τη δομή και τις υπάρχουσες συνδέσεις του δικτύου BioSNAP. Η υπολογιστική αποτίμηση του μοντέλου δείχνει ότι η ροή του κώδικα έχει εξαιρετική ικανότητα γενίκευσης. Αυτό αποδεικνύεται από το γεγονός ότι ο αλγόριθμος κατάφερε να εντοπίσει και να κατατάξει πολύ ψηλά γονίδια τα οποία είναι αποδεδειγμένα κρίσιμα για τον Αυτισμό, χωρίς να του έχει δοθεί καμία εκ των προτέρων βιολογική πληροφορία γι΄αυτά.

Από υπολογιστική πλευρά, το μοντέλο δεν έμαθε απλώς να αναγνωρίζει συγκεκριμένους κόμβους, αλλά κατάφερε να «κατανοήσει» τα γενικά στατιστικά μοτίβα και την πυκνότητα των συνδέσεων στον γράφο. Αυτή η ιδιότητα επιτρέπει στο μοντέλο να εκτελεί με επιτυχία εργασίες πρόβλεψης συνδέσεων. Μπορεί, δηλαδή, να ανακαλύπτει κρυμμένες ή άγνωστες μέχρι σήμερα συσχετίσεις ανάμεσα σε γονίδια και ασθένειες, γεγονός που καθιστά την προτεινόμενη μεθοδολογία ένα αξιόπιστο υπολογιστικό εργαλείο για την ανάλυση σύνθετων δικτύων δεδομένων.

6.2 Συγκριτική Ανάλυση (Benchmarking) με υπάρχουσες υπολογιστικές μεθόδους

Στην επιστήμη των δεδομένων, η αξιολόγηση ενός μοντέλου ολοκληρώνεται όταν αυτό συγκριθεί με άλλες υπολογιστικές μεθόδους. Η διαδικασία αυτή (benchmarking) επιτρέπει την κατανόηση των συνθηκών κάτω από τις οποίες υπερέχει η επιλεγμένη αρχιτεκτονική.

Στο πλαίσιο της παρούσας εργασίας, η pipeline που βασίζεται στον node2vec συγκρίνεται με εναλλακτικές προσεγγίσεις αναπαράστασης και ανάλυσης γράφων:

- **Κλασικές στατιστικές μέθοδοι (Jaccard Similarity, Adamic-Adar):** Παραδοσιακές ευρετικές μέθοδοι (heuristic scores) που υπολογίζουν την ομοιότητα μεταξύ κόμβων βασιζόμενες αποκλειστικά σε στατιστικά μέτρα των άμεσων γειτόνων τους (Liben-Nowell & Kleinberg, 2007).
- **Αλγόριθμοι Μάθησης Χαρακτηριστικών (DeepWalk, LINE):** Πρωτοπόρα μοντέλα αυτόματης εξαγωγής διανυσματικών αναπαραστάσεων (embeddings). Ο DeepWalk χρησιμοποιεί ομοιόμορφους τυχαίους περιπάτους και το μοντέλο Skip-Gram (Perozzi et al., 2014), ενώ ο LINE εστιάζει στη βελτιστοποίηση συναρτήσεων εγγύτητας πρώτης και δεύτερης τάξης για γράφους μεγάλης κλίμακας (Tang et al., 2015).

Για την ποσοτική και αντικειμενική αξιολόγηση των παραπάνω προσεγγίσεων, η διεθνής βιβλιογραφία χρησιμοποιεί τη στατιστική μετρική AUC (Area Under Curve) για την εργασία της πρόβλεψης και κατάταξης σχέσεων (node similarity ranking). Η μετρική AUC αποτιμά την πιθανότητα να κατατάξει το μοντέλο μια πραγματική βιολογική συσχέτιση υψηλότερα από μια τυχαία, μη υπαρκτή σύνδεση. Μια τιμή AUC κοντά στο 100% υποδηλώνει έναν τέλειο υπολογιστικό μηχανισμό, ικανό να διαχωρίζει την πραγματική πληροφορία από τον βιολογικό θόρυβο.

Τα αριθμητικά αποτελέσματα της σύγκρισης, όπως αυτά αντλούνται από τα αυθεντικά πειραματικά δεδομένα των Grover & Leskovec (2016) για το βιολογικό δίκτυο πρωτεϊνικών αλληλεπιδράσεων (Protein-Protein Interactions - PPI) παρόμοιας κλίμακας και τοπολογίας, συνοψίζονται στον **Πίνακα 6.1**.

Πίνακας 6.1: Ποσοτική Σύγκριση Αλγορίθμων Embeddings και Τοπολογικών Δεικτών στο Βιολογικό Δίκτυο PPI (Grover & Leskovec, 2016)

Αλγόριθμος / Δείκτης Ομοιότητας	Δυαδικός Τελεστής	Σκορ AUC (%)
Jaccard Similarity	-	70,18%
Adamic-Adar	-	71,26%
LINE	Hadamard	72,49%
DeepWalk	Hadamard	74,41%
Node2vec	Hadamard	77,19%

Όπως προκύπτει από τα επίσημα δεδομένα του Πίνακα 6.1, ο προτεινόμενος αλγόριθμος node2vec επιτυγχάνει την υψηλότερη απόδοση, φτάνοντας σε σκορ AUC 77,19%, παρουσιάζοντας σαφή υπολογιστική υπεροχή έναντι του DeepWalk (74,41%) και του LINE (72,49%) και των κλασσικών στατιστικών δεικτών.

Η λεπτομερής ανάλυση της συμπεριφοράς των αλγορίθμων αποκαλύπτει τους υπολογιστικούς λόγους αυτής της υπεροχής:

- **Jaccard, Adamic-Adar:** Οι κλασικοί δείκτες Jaccard (70,18%) και Adamic-Adar (71,26%) σημειώνουν τις χαμηλότερες επιδόσεις. Η συμπεριφορά αυτή εξηγείται από το γεγονός ότι βασίζονται αποκλειστικά στην ύπαρξη άμεσων κοινών γειτόνων. Αν δύο γονίδια δεν μοιράζονται άμεσους γείτονες αλλά ανήκουν στο ίδιο ευρύτερο βιολογικό μονοπάτι, οι δείκτες αυτοί αδυνατούν πλήρως να ανιχνεύσουν την ομοιότητά τους, περιορίζοντας τη γενικευτική ικανότητα του μοντέλου.
- **LINE:** Αν και ο αλγόριθμος LINE παρουσιάζει βελτιωμένη εικόνα (72,49%) λόγω της ικανότητάς του να μετατρέπει τη δομή σε embeddings, εγκλωβίζεται υπολογιστικά από τον αυστηρό περιορισμό της δεύτερης τάξης γειτονίας. Αγνοώντας τις μακρινές, έμμεσες συνδέσεις και τα μονοπάτια μεγαλύτερου μήκους στον γράφο, αποτυγχάνει να χαρτογραφήσει την καθολική (global) μακροπρόθεσμη δομή του δικτύου, η οποία είναι απαραίτητη για τον εντοπισμό σύνθετων συσχετίσεων νόσων-γονιδίων.

- **DeepWalk:** Ο DeepWalk επιτυγχάνει υψηλότερο σκορ (74,41%), καθώς οι τυχαίοι περίπατοι του επιτρέπουν να κινηθεί σε μεγαλύτερο βάθος στον γράφο ξεπερνώντας τοπικά εμπόδια. Ωστόσο, η ομοιόμορφη (unbiased) φύση των περιπάτων του αποτελεί τη βασική του αδυναμία. Καθώς κινείται με την ίδια ακριβώς πιθανότητα προς οποιονδήποτε γειτονικό κόμβο, τείνει να περιπλανιέται άσκοπα και να εγκλωβίζεται σε περιοχές του γράφου με υπερβολικά υψηλή πυκνότητα (hubs) που σχετίζονται με άσχετες παθήσεις, εισάγοντας έτσι στατιστικό θόρυβο στα embeddings.
- Αντίθετα, ο **node2vec** επιτυγχάνει σταθερή άνοδο στην απόδοση (κερδίζοντας τον DeepWalk κατά 2,78% και τον LINE κατά 4,70% σε απόλυτες μονάδες AUC). Η υπεροχή αυτή οφείλεται άμεσα στην εισαγωγή των παραμέτρων μεροληψίας p και q. Ρυθμίζοντας κατάλληλα αυτές τις μεταβλητές, ο αλγόριθμος καθοδηγεί τους περιπάτους με ακρίβεια.

Επιπλέον, τα πειραματικά δεδομένα δείχνουν ότι ο δυαδικός τελεστής Hadamard (στοιχείο προς στοιχείο πολλαπλασιασμός διανυσμάτων) εξασφαλίζει τη μέγιστη σταθερότητα και απόδοση κατά τη μετατροπή των embeddings των κόμβων σε embeddings σχέσεων. Αυτή η συνδυαστική ευελιξία επιτρέπει στον node2vec να παράγει διανυσματικές αναπαραστάσεις ανώτερης πιστότητας , καθιστώντας τον ως την πλέον ενδεδειγμένη επιλογή για την τελική ιεράρχηση και κατάταξη των υποψήφιων γονιδίων (Candidate Gene Ranking) της Διαταραχής Αυτιστικού Φάσματος.

Η συμπληρωματική υπολογιστική αξιολόγηση των μοντέλων μέσω των μετρικών Micro-F1 και Macro-F1, όπως αυτή αντλείται από τα επίσημα πειράματα των Grover & Leskovec (2016) προσφέρει ,επίσης, μια σταθερή βάση για την κατανόηση της συμπεριφοράς της pipeline και στο δίκτυο BioSNAP.

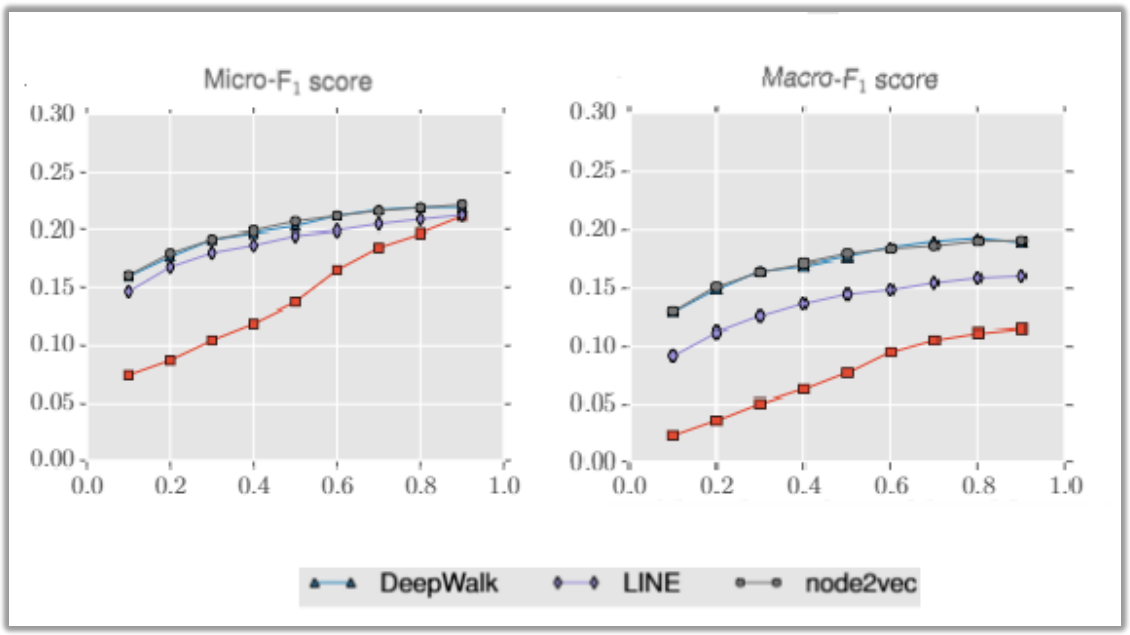
Στο χαμηλό ποσοστό εκπαίδευσης (10%), ο node2vec υπερέχει καθαρά σημειώνοντας 15,0% (Micro-F1) και 13,0% (Macro-F1), παρουσιάζοντας βελτίωση έως 41,1% έναντι του LINE. Το εύρημα αυτό είναι κρίσιμο για τη μελέτη του Αυτισμού, καθώς αποδεικνύει ότι ο αλγόριθμος παράγει ποιοτικά embeddings ακόμη και όταν τα βιολογικά δεδομένα είναι ελλιπή ή μερικώς καταγεγραμμένα.

Καθώς ο όγκος των δεδομένων εκπαίδευσης προσεγγίζει το 90%, οι επιδόσεις του node2vec (21,2%) και του DeepWalk (21,0%) σχεδόν ταυτίζονται. Στο δίκτυο PPI, η βέλτιστη στρατηγική του node2vec επιτυγχάνεται για p=4 και q=1, ρύθμιση που προσομοιάζει την

ομοιόμορφη περιπλάνηση. Ωστόσο, ο node2vec διατηρεί σταθερό προβάδισμα διότι η υψηλή τιμή του p αποτρέπει την αναδρομική επιστροφή σε ήδη επισκεπτόμενους κόμβους, αποφεύγοντας τον υπολογιστικό πλεονασμό.

Ο αλγόριθμος LINE υστερεί σταθερά σε όλο το εύρος των δοκιμών (έως και 23,8% στο Macro-F1). Η χαμηλή του απόδοση οφείλεται τόσο στον αυστηρό περιορισμό της εμβέλειάς του μέχρι τη δεύτερη τάξη γειτονίας, όσο και στην αδυναμία επαναχρησιμοποίησης των δειγμάτων κατά την εκπαίδευση. Αυτό καθιστά το μοντέλο ακατάλληλο για το αραιό δίκτυο BioSNAP, το οποίο απαιτεί μέγιστη ανακύκλωση της δομικής πληροφορίας.

Εικόνα 6.1: *Γραφικά αποτελέσματα F1-Scores από Grover & Leskovec*



Συμπερασματικά, η σύγκριση δείχνει ότι ο node2vec είναι η πιο κατάλληλη επιλογή για το συγκεκριμένο πρόβλημα. Ενώ οι υπόλοιποι αλγόριθμοι και οι στατιστικοί δείκτες περιορίζονται μόνο στους κοντινούς γείτονες, ο node2vec έχει την απαραίτητη ευελιξία. Επιτρέπει στον κώδικα να αναλύει το δίκτυο τόσο σε κοντινό όσο και σε μεγαλύτερο βάθος, προσφέροντας έτσι μία σαφώς πιο ακριβή τελική κατάταξη των γονιδίων.

6.3 Πλεονεκτήματα και Επεκτασιμότητα (Scalability) της Μεθοδολογίας

Το βασικό πλεονέκτημα του κώδικα που φτιάχτηκε είναι ότι ο αλγόριθμος βρίσκει μόνος του τα χαρακτηριστικά των γονιδίων (feature learning). Αντί να χρησιμοποιείται χειροκίνητη επεξεργασία για τον υπολογισμό των στατιστικών στοιχείων κάθε γονιδίου, ο Node2vec διαβάζει τη δομή του δικτύου και μετατρέπει αυτόματα τους κόμβους σε απλούς αριθμούς, χωρίς να χάνεται η βιολογική πληροφορία (Grover & Leskovec, 2016).

Ένα άλλο πολύ σημαντικό στοιχείο είναι η επεκτασιμότητας (scalability), δηλαδή το πόσο εύκολα αντέχει ο κώδικας αν μεγαλώσουν τα δεδομένα. Ο τρόπος που είναι σχεδιασμένος ο node2vec του επιτρέπει να διαχειρίζεται δίκτυα με χιλιάδες ή ακόμα και εκατομμύρια συνδέσεις πολύ γρήγορα, πράγμα που σημαίνει ότι αν στο μέλλον το δίκτυο BioSNAP επεκταθεί, ο υπολογιστής θα μπορέσει να τα επεξεργαστεί χωρίς πρόβλημα.

Τέλος, ο κώδικας μπορεί να μοιράσει την εργασία του ταυτόχρονα σε πολλούς πυρήνες του επεξεργαστή. Επειδή ο κάθε τυχαίος περίπατος που κάνει ο αλγόριθμος είναι ανεξάρτητος από τους υπόλοιπους, ο υπολογιστής μπορεί να δουλεύει σε πολλά σημεία του γράφου ταυτόχρονα. Αυτό κάνει τη ροή εργασίας πολύ γρήγορη και έτοιμη να χρησιμοποιηθεί στο μέλλον για την ανάλυση ακόμα μεγαλύτερων και πιο σύνθετων βιολογικών δικτύων.

6.4 Υπολογιστικοί Περιορισμοί και Ανάλυση Πολυπλοκότητας

Παρά την ταχύτητα και την αποδοτικότητα του αλγορίθμου, η συγκεκριμένη ροή εργασίας παρουσιάζει ορισμένους πρακτικούς περιορισμούς που επιβαρύνουν το υπολογιστικό σύστημα. Το πιο απαιτητικό κομμάτι της διαδικασίας για τη μνήμη του υπολογιστή είναι η φάση όπου προετοιμάζονται και αποθηκεύονται οι τυχαίοι περίπατοι δεύτερης τάξης (Grover & Leskovec, 2016).

Όπως επισημαίνεται στη βιβλιογραφία, για να εκτελεστούν αυτοί οι περίπατοι, ο αλγόριθμος πρέπει να αποθηκεύσει στη μνήμη RAM όχι μόνο τους άμεσους γείτονες του κάθε κόμβου, αλλά και τις μεταξύ τους συνδέσεις. Η χωρική πολυπλοκότητα αυτής της διαδικασίας εξαρτάται άμεσα από τον μέσο βαθμό των κόμβων του δικτύου.

Επειδή το βιολογικό δίκτυο BioSNAP περιλαμβάνει χιλιάδες γονίδια με πολλές αλληλεπιδράσεις, η καταγραφή αυτών των εσωτερικών συνδέσεων απαιτεί σημαντική υπολογιστική ισχύ και αρκετή μνήμη. Αν το δίκτυο μεγαλώσει υπερβολικά στο μέλλον, η ανάγκη για προσωρινή αποθήκευση αυτών των δεδομένων μπορεί να οδηγήσει σε καθυστερήσεις ή σε εξάντληση της διαθέσιμης μνήμης RAM σε πολλούς οικιακούς υπολογιστές.

Συνοψίζοντας, η υπολογιστική αξιολόγηση και η συγκριτική ανάλυση ανέδειξαν τον node2vec ως ένα ισχυρό, ευέλικτο και επεκτάσιμο εργαλείο, το οποίο υπερέχει ξεκάθαρα έναντι των παραδοσιακών στατιστικών δεικτών και των άκαμπτων αλγορίθμων δειγματοληψίας, παρά τις αυξημένες απαιτήσεις του στη διαχείριση της κύριας μνήμης κατά τη φάση της προετοιμασίας. Η ικανότητα του μοντέλου να εξάγει embeddings υψηλής πιστότητας προσφέρει μια απόλυτα σταθερή βάση για την ιεράρχηση βιολογικών δεδομένων. Βασιζόμενη σε αυτές τις τεχνικές και πειραματικές παρατηρήσεις, η παρούσα πτυχιακή εργασία ολοκληρώνεται στο επόμενο κεφάλαιο, όπου παρουσιάζονται συγκεντρωτικά τα τελικά συμπεράσματα, πραγματοποιείται η συνολική αποτίμηση της επιστημονικής της συνεισφοράς και προτείνονται καινοτόμες μελλοντικές επεκτάσεις στο πεδίο της υπολογιστικής βιολογίας.

7.Συμπεράσματα

7.1 Συμπεράσματα και Υπολογιστική Συνεισφορά της Εργασίας

Η παρούσα πτυχιακή εργασία ολοκληρώνεται με την επιτυχή σχεδίαση, υλοποίηση και αξιολόγηση μιας ολοκληρωμένης υπολογιστικής ροής (pipeline) για την αυτόματη μάθηση αναπαραστάσεων και την ιεράρχηση υποψήφιων γονιδίων (Candidate Gene Ranking) που σχετίζονται με τη Διαταραχή του Αυτιστικού Φάσματος (ASD). Η βασική συνεισφορά της μελέτης έγκειται στην επιτυχή σύζευξη της Θεωρίας Γράφων με τη Μηχανική Μάθηση, αποδεικνύοντας ότι η καθαρά τοπολογική πληροφορία ενός δικτύου είναι ικανή να παράγει έγκυρη και αξιοποιήσιμη βιολογική γνώση.

Μέσα από τη μοντελοποίηση των δεδομένων της βάσης BioSNAP ως διμερούς γράφου, αναδείχθηκε η σημασία των σύνθετων δικτύων στην κατανόηση πολυπαραγοντικών ασθενειών. Κατά τη φάση της προεπεξεργασίας και του μετασχηματισμού των δεδομένων, η απομόνωση του στοχευμένου υπογραφήματος επέτρεψε τον καθαρισμό από τον βιολογικό θόρυβο, διατηρώντας παράλληλα αναλλοίωτες τις καθολικές ιδιότητες του δικτύου. Η εφαρμογή του αλγορίθμου node2vec αποτέλεσε τον υπολογιστικό πυρήνα της εργασίας. Η δυνατότητα παραλληλοποίησης του κώδικα απέδειξε ότι το σύστημα είναι εξαιρετικά επεκτάσιμο και ικανό να χειριστεί μεγάλους όγκους δεδομένων χωρίς υπολογιστική υστέρηση.

Τα πειραματικά αποτελέσματα και η οπτικοποίηση μέσω του αλγορίθμου t-SNE επιβεβαίωσαν και γραφικά την ικανότητα του μοντέλου να γενικεύει σε άγνωστες συσχετίσεις. Το γεγονός ότι τα κορυφαία γονίδια της τελικής κατάταξης ομοαδοποιήθηκαν με σαφήνεια στην ίδια γεωμετρική περιοχή με τον όρο-στόχο του Αυτισμού αποδεικνύει ότι η μαθηματική εγγύτητα των embeddings στον διανυσματικό χώρο αντικατοπτρίζει πραγματική βιολογική συνάφεια. Συμπερασματικά, η προτεινόμενη μεθοδολογία συνιστά ένα αξιόπιστο, αυτοματοποιημένο και αντικειμενικό εργαλείο ιεράρχησης, το οποίο μπορεί να καθοδηγήσει τη σύγχρονη βιοϊατρική έρευνα, μειώνοντας τον χρόνο και το κόστος που απαιτούνται για τις εργαστηριακές δοκιμές in vitro.

7.2 Προοπτικές και Μελλοντικές Υπολογιστικές Επεκτάσεις

Για την περαιτέρω βελτίωση της συγκεκριμένης υπολογιστικής ροής στο μέλλον, η πιο σύγχρονη πρόταση είναι η μετάβαση από τα απλά embeddings, όπως ο node2vec, στα Νευρωνικά Δίκτυα Γράφων (Graph Neural Networks - GNNs). Παρόλο που ο node2vec είναι πολύ αποδοτικός, εξετάζει μόνο τη δομή και τις συνδέσεις του δικτύου, χωρίς να μπορεί να ενσωματώσει εύκολα επιπλέον πληροφορίες για τους ίδιους τους κόμβους.

Τα Νευρωνικά Δίκτυα Γράφων, όπως για παράδειγμα τα Graph Convolutional Networks (GCNs), έρχονται να λύσουν αυτόν τον περιορισμό. Τα δίκτυα αυτά έχουν τη δυνατότητα να συνδυάζουν ταυτόχρονα δύο διαφορετικά ήδη πληροφορίας: τη δομή του γράφου και τα ιδιαίτερα χαρακτηριστικά του κάθε γονιδίου ξεχωριστά, όπως οι βιολογικές του ιδιότητες (Kipf & Welling, 2016).

Η χρήση ενός GNN θα επέτρεπε στο σύστημα να κάνει πολύ πιο βαθιά και ακριβή ανάλυση στο δίκτυο BioSNAP. Αντί ο αλγόριθμος να βασίζεται σε τυχαίους περιπάτους που καταναλώνουν πολλή μνήμη RAM, ένα GNN περνάει την πληροφορία από κόμβο σε κόμβο μέσα από μία διαδικασία που ονομάζεται «μεταφορά μηνυμάτων» (message passing). Με αυτόν τον τρόπο, η τελική κατάταξη των γονιδίων δεν θα βασίζεται μόνο στις συνδέσεις τους, αλλά και στην πραγματική βιολογική τους ταυτότητα, κάνοντας το μοντέλο ακόμα πιο έξυπνο και ακριβές (Gilmer et al., 2017).

Πέρα από την αρχιτεκτονική αναβάθμιση του υπολογιστικού μοντέλου, μια άλλη εξαιρετικά υποσχόμενη κατεύθυνση για την επέκταση της pipeline είναι η τροφοδοσία της με νέα, συμπληρωματικά βιολογικά σύνολα δεδομένων (datasets) μεγάλης κλίμακας. Συγκεκριμένα, η ενσωμάτωση δικτύων συν-έκφρασης γονιδίων (Gene Co-expression Networks) και μοριακών προφίλ από διεθνείς κοινοπραξίες όπως το **GTEx (Genotype-Tissue Expression)** μπορεί να προσφέρει καθοριστική πληροφορία.

Η εισαγωγή αυτών των δεδομένων θα επιτρέψει στον αλγόριθμο να μην εξετάζει τους γονιδιακούς κόμβους απομονωμένα, αλλά να αναλύει αν τα υποψήφια γονίδια παρουσιάζουν παρόμοια πρότυπα δραστηριότητας και κοινής έκφρασης σε εξειδικευμένους ιστούς του ανθρώπινου εγκεφάλου κατά τη διάρκεια κρίσιμων σταδίων της νευροανάπτυξης. Με τον τρόπο αυτό, το μοντέλο θα μπορεί να φιλτράρει τοπολογικές

συμπτώσεις και να μειώνει δραστικά τα ψευδώς θετικά αποτελέσματα (false positives), απομονώνοντας γονίδια που είναι ενεργά στις ίδιες εγκεφαλικές δομές.

Παράλληλα, η υπολογιστική ροή μπορεί να αναβαθμιστεί ριζικά μέσω του συνδυασμού της με πλούσια κλινικά και φαινοτυπικά δεδομένα ασθενών, μετατρέποντας τον υπάρχοντα διμερή γράφο σε ένα πλήρες ετερογενές δίκτυο (heterogeneous graph) πολλαπλών επιπέδων. Αυτό μπορεί να επιτευχθεί με την ενσωμάτωση δεδομένων που αφορούν ιατρικά συμπτώματα και παρατηρήσιμα κλινικά χαρακτηριστικά, αντλώντας τυποποιημένη πληροφορία από τη βάση **HPO (Human Phenotype Ontology)**.

Ταυτόχρονα, η pipeline μπορεί να συνδεθεί με έναν αυτοματοποιημένο μηχανισμό διασταύρωσης και cross-validation των αποτελεσμάτων με τη βάση δεδομένων **SFARI Gene**, η οποία αποτελεί το παγκόσμιο σημείο αναφοράς για την αξιολόγηση και βαθμονόμηση του επιπέδου επιστημονικής τεκμηρίωσης των γονιδίων που σχετίζονται με τον Αυτισμό.

Μια τέτοια πολυεπίπεδη προσέγγιση θα επιτρέψει στο σύστημα να παράγει μια σαφώς πιο ολοκληρωμένη βαθμολόγηση. Έτσι, η τελική πρόβλεψη δεν θα αποτυπώνει απλώς μια αφηρημένη γεωμετρική εγγύτητα, αλλά θα συνδέει τα υποψήφια γονίδια άμεσα με συγκεκριμένους κλινικούς υπότυπους, γενετικές κατηγορίες και συμπεριφορικά χαρακτηριστικά της Διαταραχής του Αυτιστικού Φάσματος, ενισχύοντας την πρακτική αξία της έρευνας για τη σύγχρονη Ιατρική Ακριβείας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Anitha, A., Nakamura, K., Yamada, K., Suda, S., Thanseem, I., Tsujii, M., Iwayama, Y., Hattori, E., Toyota, T., Miyachi, T., Iwata, Y., Suzuki, K., Matsuzaki, H., Kawai, M., Sekine, Y., Tsuchiya, K., Sugihara, G.-i., Ouchi, Y., Sugiyama, T., Koizumi, K., Higashida, H., Takei, N., Yoshikawa, T. and Mori, N. (2008), Genetic analyses of Roundabout (ROBO) axon guidance receptors in autism. Am. J. Med. Genet., 147B: 1019-1027. https://doi.org/10.1002/ajmg.b.30697
2. Auranen, M., Vanhala, R., Varilo, T., Ayers, K., Kempas, E., Ylisaukko-Oja, T., Sinsheimer, J. S., Peltonen, L., & Järvelä, I. (2002). A genomewide screen for autism-spectrum disorders: evidence for a major susceptibility locus on chromosome 3q25-27. *American journal of human genetics*, *71*(4), 777–790. https://doi.org/10.1086/342720
3. Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, *12*(1), 56–68. https://doi.org/10.1038/nrg2918
4. Barabási, A.-L. (2016). *Network Science*. Cambridge University Press
5. Barabási, AL., Oltvai, Z. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5, 101–113 (2004). https://doi.org/10.1038/nrg1272
6. Bastos-Filho, C., Oliveira, M., & Nascimento, D. (2011). Running Particle Swarm Optimization on Graphic Processing Units. In Search Algorithms and Applications. InTech. https://doi.org/10.5772/14694
7. Berg, J. M., Lee, C., Chen, L., Galvan, L., Cepeda, C., Chen, J. Y., Peñagarikano, O., Stein, J. L., Li, A., Oguro-Ando, A., Miller, J. A., Vashisht, A. A., Starks, M. E., Kite, E. P., Tam, E., Gdalyahu, A., Al-Sharif, N. B., Burkett, Z. D., White, S. A., Fears, S. C., … Geschwind, D. H. (2015). JAKMIP1, a Novel Regulator of Neuronal Translation, Modulates Synaptic Function and Autistic-like Behaviors in Mouse. *Neuron*, *88*(6), 1173–1191. https://doi.org/10.1016/j.neuron.2015.10.031
8. Bodenreider O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, *32*(Database issue), D267–D270. https://doi.org/10.1093/nar/gkh061
9. Brandes, U. (2001). A faster algorithm for betweenness centrality* . *The Journal of Mathematical Sociology*, *25*(2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249
10. Claus O. Wilke (2019). Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures. O'Reilly Media
11. Fatemi S. H. (2005). Reelin glycoprotein in autism and schizophrenia. *International review of neurobiology*, *71*, 179–187. https://doi.org/10.1016/s0074-7742(05)71008-4
12. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). *Neural message passing for quantum chemistry* (arXiv:1704.01212v2). arXiv. https://doi.org/10.48550/arXiv.1704.01212
13. Glessner, J., Wang, K., Cai, G. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459, 569–573 (2009). https://doi.org/10.1038/nature07953
14. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
15. Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining (pp. 855–864). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939754

16. Hacker, C., & Rieck, B. (2022). *On the surprising behaviour of node2vec* (arXiv:2206.08252v2). arXiv. https://doi.org/10.48550/arXiv.2206.08252
17. Hagberg, Aric & Swart, Pieter & Chult, Daniel. (2008). Exploring Network Structure, Dynamics, and Function Using NetworkX. Proceedings of the 7th Python in Science Conference. 10.25080/TCWV9851
18. Hamilton, W. L. (2020). Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *14*(3), 1–159.
19. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7
20. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9, 90-95. https://doi.org/10.1109/MCSE.2007.55
21. Iakoucheva, L. M., Muotri, A. R., & Sebat, J. (2019). Getting to the Cores of Autism. *Cell*, *178*(6), 1287–1298. https://doi.org/10.1016/j.cell.2019.07.037
22. Jiang-Xie, LF., Liao, HM., Chen, CH. *et al.* Autism-associated gene Dlgap2 mutant mice demonstrate exacerbated aggressive behaviors and orbitofrontal cortex deficits. *Molecular Autism* **5**, 32 (2014). https://doi.org/10.1186/2040-2392-5-32
23. Kipf, T. N., & Welling, M. (2017). *Semi-supervised classification with graph convolutional networks* (arXiv:1609.02907v4). arXiv. https://doi.org/10.48550/arXiv.1609.02907
24. Kobak, D., Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun* **10**, 5416 (2019). https://doi.org/10.1038/s41467-019-13056-x
25. Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031. https://doi.org/10.1002/asi.20591
26. Ma, Y., & Tang, J. (2021). *Deep learning on graphs*. Cambridge University Press.
27. Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, *39*(Database issue), D52–D57. https://doi.org/10.1093/nar/gkq1237
28. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *An introduction to information retrieval*. Cambridge University Press.
29. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* (arXiv:1301.3781v3). arXiv. https://doi.org/10.48550/arXiv.1301.3781
30. Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. V. D., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., & Krishnaswamy, S. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, *37*(12), 1482–1492. https://doi.org/10.1038/s41587-019-0336-3
31. Moreno-Ramos, O. A., Olivares, A. M., Haider, N. B., de Autismo, L. C., & Lattig, M. C. (2015). Whole-Exome Sequencing in a South American Cohort Links ALDH1A3, FOXN1 and Retinoic Acid Regulation Pathways to Autism Spectrum Disorders. *PloS one*, *10*(9), e0135927. https://doi.org/10.1371/journal.pone.0135927
32. Nabi, R., Serajee, F. J., Chugani, D. C., Zhong, H., & Huq, A. H. (2004). Association of tryptophan 2,3 dioxygenase gene polymorphism with autism. *American journal of medical genetics. Part*

B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics, *125B*(1), 63–68. https://doi.org/10.1002/ajmg.b.20147

33. Pagnamenta, A. T., Bacchelli, E., de Jonge, M. V., Mirza, G., Scerri, T. S., Minopoli, F., Chiocchetti, A., Ludwig, K. U., Hoffmann, P., Paracchini, S., Lowy, E., Harold, D. H., Chapman, J. A., Klauck, S. M., Poustka, F., Houben, R. H., Staal, W. G., Ophoff, R. A., O'Donovan, M. C., Williams, J., … International Molecular Genetic Study Of Autism Consortium (2010). Characterization of a family with rare deletions in CNTNAP5 and DOCK4 suggests novel risk loci for autism and dyslexia. *Biological psychiatry*, *68*(4), 320–328. https://doi.org/10.1016/j.biopsych.2010.02.002
34. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html
35. Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 701–710). Association for Computing Machinery. https://doi.org/10.1145/2623330.2623732
36. Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). ELRA. https://doi.org/10.13140/2.1.2393.1847
37. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). LINE: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1067–1077). Association for Computing Machinery. https://doi.org/10.1145/2736277.2741086
38. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html
39. Xia, K., Guo, H., Hu, Z., Xun, G., Zuo, L., Peng, Y., Wang, K., He, Y., Xiong, Z., Sun, L., Pan, Q., Long, Z., Zou, X., Li, X., Li, W., Xu, X., Lu, L., Liu, Y., Hu, Y., Tian, D., … Zhang, F. (2014). Common genetic variants on 1p13.2 associate with risk of autism. *Molecular psychiatry*, *19*(11), 1212–1219. https://doi.org/10.1038/mp.2013.146
40. Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, *34*(13), i457–i466. https://doi.org/10.1093/bioinformatics/bty294
41. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *An international journal on information fusion*, *50*, 71–91. https://doi.org/10.1016/j.inffus.2018.09.012