ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

2η Έκθεση Προόδου

Νοέμβριος 2021 – Ιανουάριος 2023

Εκπονούμενη Διατριβή:

Υπολογιστική ανάλυση 1D και 3D δεδομένων για την επαναστόχευση φαρμάκων στον σακχαρώδη διαβήτη

Ονοματεπώνυμο Υποψήφιου Διδάκτορα:
Σωτήριος Ουζούνης

Αριθμός Μητρώου:
2003

Επιβλέπων Καθηγητής:
Καλατζής Ιωάννης, Καθηγητής

Τριμελής Συμβουλευτική Επιτροπή:
Καλατζής Ιωάννης, Καθηγητής
Κατσίλα Θεοδώρα, Εντεταλμένη Ερευνήτρια, Εθνικό Ίδρυμα Ερευνών
Ζουμπουλάκης Παναγιώτης, Αναπληρωτής Καθηγητής

1. Εισαγωγή

Στόχος της διατριβής αυτής είναι η ανάπτυξη μεθοδολογίας ενοποίησης και ανάλυσης ετερογενών δεδομένων μεγάλου όγκου στον σακχαρώδη διαβήτη η οποία εκτιμάται πως θα συμβάλλει στην επαναστόχευση φαρμάκων με βέλτιστη ακρίβεια και αξιοπιστία. Για αυτό, εφαρμόζεται για πρώτη φορά η συνδυαστική ανάλυση βιοιατρικών δεδομένων από δημόσια αποθετήρια δεδομένων ομικών τεχνολογιών, φαρμακογονιδιωματικής πληροφορίας, δεδομένων ευρεσιτεχνιών βιοδραστικών μορίων, κλινικών δοκιμών και δεδομένων δομικής βιολογίας.

Κατά την εκπόνηση της διατριβής, σε συνέχεια των πεπραγμένων, όπως αυτά περιγράφονται στην 1η έκθεση προόδου, τα βήματα που ακολουθηθήκαν είναι τα εξής:

- Εμπλουτισμός των δεδομένων, που είχαν συλλεχθεί και ενοποιηθεί σε μια βάση δεδομένων γράφου, με επικαιροποιημένα δεδομένα από νέα αποθετήρια.
- Ορισμός σημασιολογικών συσχετίσεων μεταξύ των δεδομένων που έχουν συλλεχθεί και αναλυθεί, ώστε η βάση δεδομένων γράφου να μετατραπεί σε γνωσιακό γράφο.
- Ανάπτυξη πλαισίου εργασίας μηχανικής μάθησης για την πρόβλεψη της αναστολής ή μη επιλεγμένων ισομορφών του κυτοχρώματος P450, με σκοπό την πρόβλεψη τοξικότητας των επαναστοχευμένων χημικών μορίων.
- Εμπλουτισμούς του γνωσιακού γράφου με προβλέψεις για τη δράση των χημικών μορίων στις επιλεγμένες ισομορφές του κυτοχρώματος P450.
- Εφαρμογή τεχνικών μηχανικής μάθησης σε δεδομένα γράφων, με σκοπό την πρόβλεψη αλληλεπίδρασης φαρμάκου-πρωτεΐνης.

2. Υλικά και Μέθοδοι

2.1 Εμπλουτισμός δεδομένων

Στο πλαίσιο περαιτέρω εμπλουτισμού της υπάρχουσας πληροφορίας, που είχε συλλεχθεί για τα χημικά μόρια, έγινε συλλογή φαρμακογονιδιωματικών δεδομένων από το αποθετήριο ENSEMBL (1). Η εξόρυξη των δεδομένων έγινε συνδυαστικά μέσω επερωτήσεων στο API του αποθετηρίου και τεχνικών συγκομιδής δεδομένων (data scraping). Τα δεδομένα, που συλλέχθηκαν αφορούν φαρμακογονιδιωματικές παραλλαγές με έμφαση στις παρανοηματικές μεταλλάξεις. Συμπληρωματικά σε αυτά τα δεδομένα λήφθηκαν φαρμακογονιδιωματικές συστάσεις από την πλατφόρμα CPIC (2). Οι συστάσεις αυτές αφορούν αλληλεπιδράσεις φαρμάκου-πρωτεΐνης, ενώ από την πλατφόρμα RefSeq (3) ανακτήθηκαν αλληλουχίες φαρμακογονιδίων και κατόπιν επεξεργασίας, διατηρήθηκε μόνο το τμήμα της αλληλουχίας, το οποίο περιλαμβάνει την περιοχή της μετάλλαξης. Επιπρόσθετα, συλλέχθηκαν δεδομένα από τις πλατφόρμες ClinVar (4) και dbSNP (5) οι οποίες φέρουν πληροφορίες για την κλινική σημασία της μετάλλαξης και τη συχνότητα εμφάνισης αυτής σε διαφορετικές ομάδες του πληθυσμού.

Όσον αφορά τις πληροφορίες που αφορούν τα χημικά μόρια, συλλέχθηκαν επιπλέον δεδομένα από το αποθετήριο ChEMBL (6). Πιο συγκεκριμένα, συλλέχθηκαν πειραματικά δεδομένα μέσω API τα οποία αφορούν, είτε την φαρμακευτική απόκριση των χημικών μορίων σε κυτταρικές σειρές, είτε πειραματικές τιμές πρόσδεσης σε συγκεκριμένες πρωτεΐνες-στόχους. Αντίστοιχα, συλλέχθηκαν επικαιροποιημένα δεδομένα από την πλατφόρμα mirTarBase (7). Παράλληλα, συλλέχθηκαν συσχετίσεις γονιδίων-καρκινικών τύπων μέσω της TCGA. Τέλος, από την πλατφόρμα OpenTargets (8) λήφθηκαν δεδομένα πρωτεΐνης στόχου-ασθένειας στον άνθρωπο. Η OpenTargets ενοποιεί και αναλύει μεγάλο όγκο δεδομένων από ετερογενείς πηγές, με σκοπό να παρέχει τις εν λόγω συσχετίσεις, υποστηρίζοντας την αναγνώριση και προτεραιοποίηση στόχων. Οι πηγές δεδομένων, που αξιοποιεί η OpenTargets, προκειμένου να παράξει συσχετίσεις ποικίλουν και συνδυάζονται

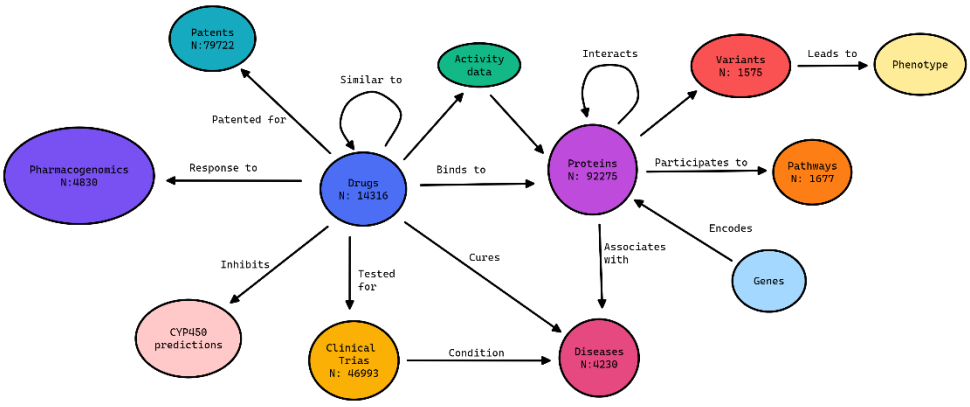
για την τελική πρόβλεψη. Τα δεδομένα αυτά μπορούν να κατηγοριοποιηθούν στις παρακάτω κύριες ομάδες, με βάση τον τύπο πληροφορίας που προσφέρουν:

1. Γενετικές συσχετίσεις
2. Σωματικές μεταλλάξεις
3. Φαρμακευτικές πληροφορίες
4. Βιολογία συστημάτων και βιολογικά μονοπάτια
5. Έκφραση RNA
6. Εξόρυξη κειμένου
7. Ζωικά μοντέλα

Το βάρος, που χρησιμοποιείται για την εκάστοτε πηγή, είναι διαφορετικό και έχει οριστεί από τους δημιουργούς της πλατφόρμας, ώστε να εξαχθεί το τελικό σκορ συσχέτισης στόχου-ασθένειας. Από το σύνολο των δεδομένων της OpenTargets αποφασίστηκε να συλλεχθούν, μόνο, οι συσχετίσεις στόχου-ασθένειας από την ομάδα φαρμακευτικών πληροφοριών. Ως πηγή φαρμακευτικών πληροφοριών η OpenTargets χρησιμοποιεί την ChEMBL. Τα στοιχεία, που προσφέρει η ChEMBL για τις συσχετίσεις πρωτεϊνικού στόχου-ασθένειας, προκύπτουν από τη σχέση μεταξύ των εγκεκριμένων φαρμάκων και των αντίστοιχων ασθενειών, τις οποίες στοχεύουν, όπως έχουν καταγραφεί σε κλινικές δοκιμές.

2.2 Ορισμός σημασιολογικών συσχετίσεων για το σχηματισμό γνωσιακού γράφου

Το σύνολο των δεδομένων τα οποία έχουν συλλεχθεί και ενοποιηθεί σε μια βάση δεδομένων γράφου μέσω της πλατφόρμας Neo4j, όπως περιγράφεται στην 1^{η} έκθεση προόδου, μετασχηματίστηκε, ώστε να εισαχθούν σημασιολογικές συσχετίσεις μεταξύ των ετερογενών οντοτήτων του γράφου. Ενδεικτικό παράδειγμα του μετασχηματισμού αυτού αποτελεί η Εικόνα 1, η οποία απεικονίζει μέρος του συνολικού γνωσιακού γράφου.

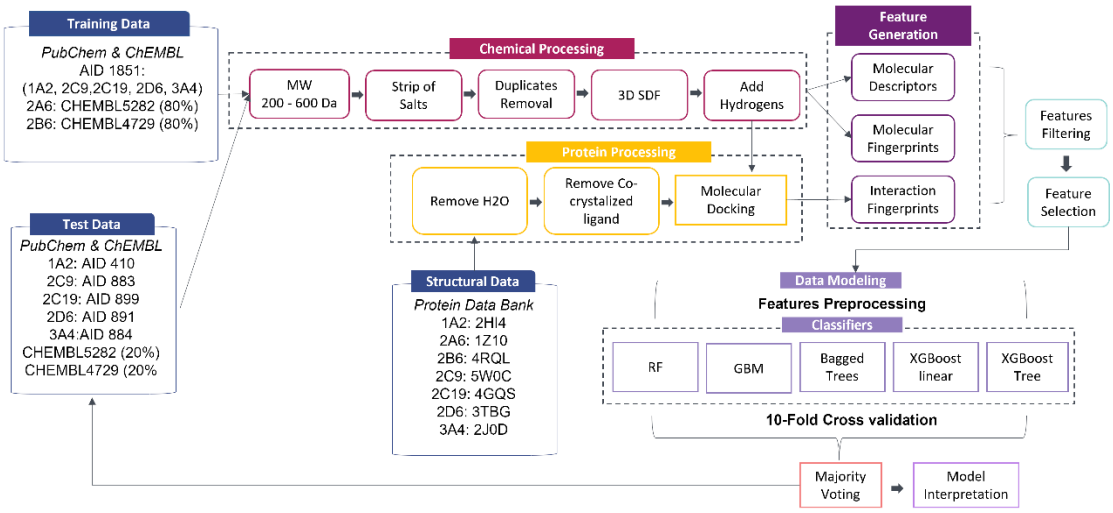


Εικόνα 1. Παράδειγμα γνωσιακού γράφου

2.3 Ανάπτυξη πλαισίου εργασίας μηχανικής μάθησης για την πρόβλεψη της αναστολής ή μη επιλεγμένων ισομορφών του κυτοχρώματος P450

Για το σχεδιασμό των μοντέλων μηχανικής μάθησης ως χαρακτηριστικά (μεταβλητές εισόδου) αξιοποιήθηκαν τα μοριακά αποτυπώματα (molecular fingerprints) των χημικών μορίων. Πιο συγκεκριμένα, από κάθε χημική ένωση υπολογίστηκαν ποικίλα αριθμητικά χαρακτηριστικά (μοριακοί περιγραφείς, descriptors). Οι περιγραφείς αυτοί μπορεί να είναι μονοδιάστατοι (0D ή 1D) ή να είναι δισδιάστατοι (2D) και προέρχονται από την τοπολογία της δομής του μορίου. Επιπλέον, υπάρχουν τρισδιάστατοι (3D) περιγραφείς, οι οποίοι εξάγονται από τη γεωμετρία της χημικής ένωσης (3D συντεταγμένες), αλλά και τετραδιάστατοι (4D), που υπολογίζονται με βάση τις πιθανές διαμορφώσεις, που έχει ένα χημικό μόριο στον χώρο. Συμπληρωματικά στα χαρακτηριστικά αυτά υπολογίστηκαν και μοριακά αποτυπώματα αλληλεπίδρασης χημικού μορίου-στόχου. Τα αποτυπώματα αλληλεπίδρασης κωδικοποιούν τις αλληλεπιδράσεις μεταξύ των ατόμων/χημικών ομάδων του χημικού μορίου και της πρωτεΐνης-στόχου, με βάση την απόσταση που έχουν κατά την μοντελοποίησή τους στον τρισδιάστατο χώρο. Κατά αυτόν τον τρόπο κωδικοποιείται η δομική και χωρική πληροφορία, όχι μόνο του χημικού μορίου, αλλά και της περιοχής πρόσδεσης στον πρωτεϊνικό στόχο. Η πληροφορία που φέρει το νέο αυτό μοριακό αποτύπωμα δεν μπορεί να προκύψει από τις τρισδιάστατες δομές του χημικού μορίου ή της πρωτεΐνης ξεχωριστά, αλλά παρά μόνο από την προσομοίωση της πρόσδεσης αυτών. Συνεπώς, το μοριακό αποτύπωμα αλληλεπίδρασης προέκυψε από τα πειράματα μοριακής πρόσδεσης μεταξύ των χημικών μορίων και των επιλεγμένων ισομορφών του κυτοχρώματος P450 (1A2, 2A6, 2B6, 2C9, 2C19, 2D6 και 3A4).

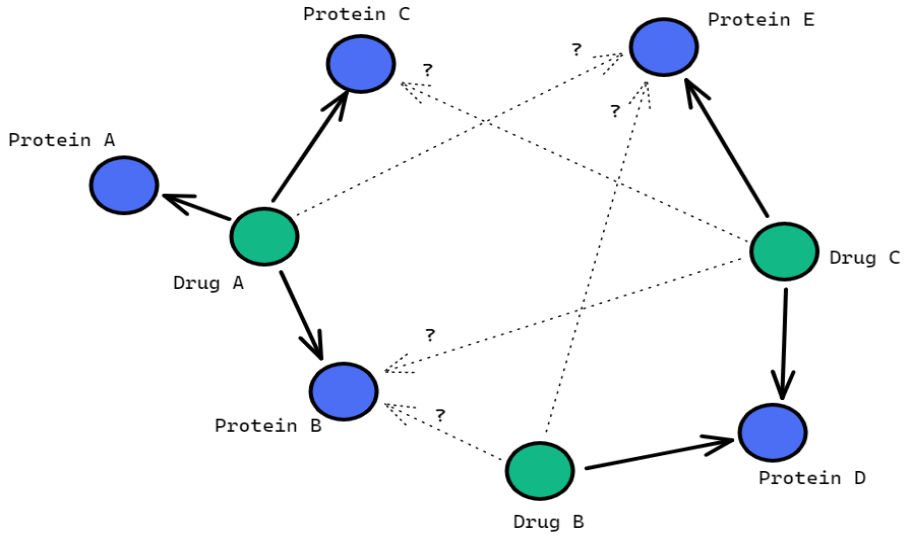
Γνωρίζοντας πως τα CYP450 ισοένζυμα παρουσιάζουν σημαντικές δομικές διαφορές στην περιοχή πρόσδεσης, αξιολογήθηκε ανά περίπτωση ποιος τύπος αποτυπώματος ή/και μοριακός περιγραφέας μπορεί να μοντελοποιήσει καλύτερα το βιολογικό αποτέλεσμα. Για το σκοπό αυτό και προκειμένου να ορίζεται πώς ο τύπος χαρακτηριστικών επηρεάζει την τελική πρόβλεψη, αλλά και για να ισχυροποιηθεί το αποτέλεσμα της πρόβλεψης, υιοθετήθηκε ένα σύστημα ταξινόμησης με βάση την πλειοψηφία. Πιο αναλυτικά, αντί σε κάθε περίπτωση να αξιολογείται η ικανότητα πρόβλεψης του συστήματος από τον καλύτερο ταξινομητή, η τελική ταξινόμηση έγινε με βάση την πλειοψηφία της κλάσης που επέλεξαν οι πέντε ταξινομητές υπό μελέτη. Η τελική πιθανότητα, βάση της οποίας επιλέχθηκε η κλάση ταξινόμησης προέκυψε από τον μέσο όρο των πιθανοτήτων των πέντε ταξινομητών. Η μεθοδολογία του πλαισίου εργασίας που αναπτύχθηκε περιγράφεται στην Εικόνα 2.



Εικόνα 2. Σχηματική αναπαράσταση πλαισίου εργασίας μηχανικής μάθησης για την ταξινόμηση αναστολέων (ή μη) των επιλεγμένων ισομορφών του κυτοχρώματος P450

2.4 Τεχνικές μηχανικής μάθησης σε δεδομένα γράφων

Από τον ενιαίο, ομογενοποιημένο γνωσιακό γράφο, εστιάζοντας σε υποδίκτυα αυτού, μπορούν να εξαχθούν, είτε μέσω στατιστικών τεχνικών, είτε μέσω μεθόδων μηχανικής μάθησης, συσχετίσεις οι οποίες δεν υφίστανται, αλλά είναι πιθανό να ισχύουν. Ως ςκ τούτου, μπορεί να εφαρμοστεί η τεχνική πρόβλεψης σύνδεσης (link prediction), βάση της οποίας εξετάζονται οι τοπικές διασυνδέσεις μεταξύ των κόμβων και με βάση τα χαρακτηριστικά που φέρουν μπορούν να προβλεφθούν συνδέσεις, οι οποίες δεν είναι γνωστές, αλλά εν δυνάμει υφίστανται. Ακολουθώντας την τεχνική αυτή, αρχικά, διερευνήθηκε το υποδίκτυο το οποίο αποτελείται, μόνο, από τις συνδέσεις μεταξύ χημικών μορίων και πρωτεϊνών-στόχων. Στόχος ήταν - με βάση τις ήδη γνωστές σχέσεις - να προβλεφθούν νέες σχέσεις μεταξύ των δυο αυτών οντοτήτων του γράφου. Ενδεικτικά, η ιδέα αυτή αποτυπώνεται στην Εικόνα 3.



Εικόνα 3. Σχηματική αναπαράσταση πρόβλεψης σύνδεσης

Για την υλοποίηση της παραπάνω διαδικασίας ακολουθήθηκαν τα εξής βήματα:

- Απομόνωση και εξαγωγή του υποδικτύου ενδιαφέροντος (φάρμακο-πρωτεΐνη) από το σύνολο των δεδομένων του γνωσιακού γράφου.
- Χαρακτηρισμός των σχέσεων του υποδικτύου, με βάση την κλάση στην οποία ανήκουν. Σε πρώτη φάση μοντελοποιήθηκε ως πρόβλημα δυαδικής ταξινόμησης και συνεπώς, όσα ζεύγη φαρμάκου-πρωτεΐνης είναι γνωστά χαρακτηρίστηκαν ως θετική κλάση. Ως αρνητική κλάση ορίστηκαν τα ζεύγη, που με βάση τις πειραματικές τους τιμές (IC50, EC50, Ki), δεν υφίσταται διασύνδεση μεταξύ φαρμάκου-πρωτεΐνης.
- Υλοποιήθηκε η εξαγωγή χαρακτηριστικών, με βάση τοπικές στατιστικές μετρήσεις των αποστάσεων μεταξύ των ζευγών φαρμάκου-πρωτεΐνης, με τις τεχνικές Node2Vec και FastRP.
- Οι μετρήσεις που εξήχθησαν χαρακτηρίζουν τον κάθε κόμβο και όχι το ζεύγος. Συνεπώς, επόμενο βήμα ήταν ο συνδυασμός των ζευγών μέσω πολλαπλασιασμού των διανυσμάτων των χαρακτηριστικών του κάθε κόμβου.
- Έχοντας τα ζεύγη φαρμάκου-πρωτεΐνης και τα χαρακτηριστικά αυτών, σχεδιάστηκαν και εκπαίδευτηκαν ταξινομητές, οι οποίοι διακρίνουν τα ζεύγη τα οποία συνδέονται από εκείνα, που δε συνδέονται.

3. Αποτελέσματα

3.1 Απόδοση του συστήματος ταξινόμησης αναστολέων επιλεγμένων ισομορφών του κυτοχρώματος P450

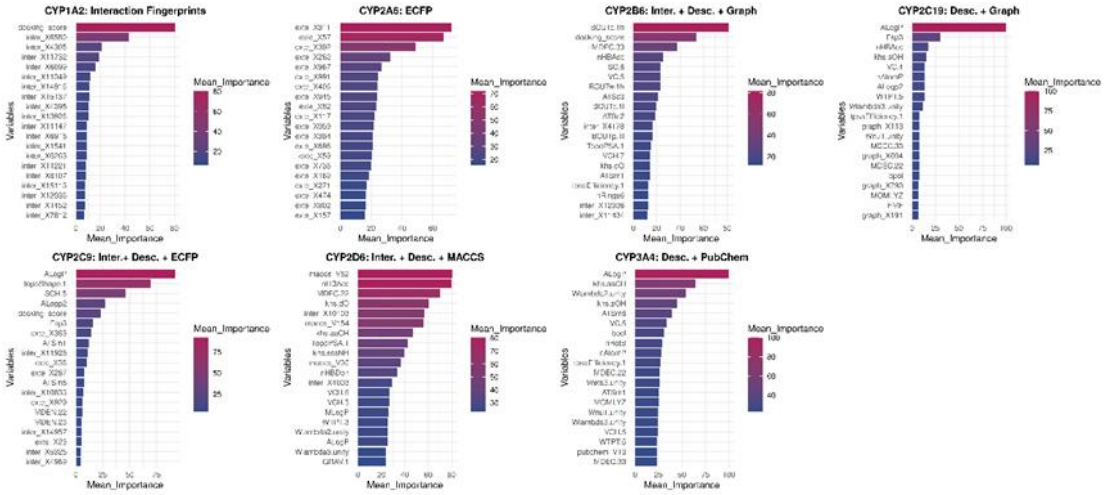
Στον Πίνακα 1 δίδονται τα αποτελέσματα της αξιολόγησης των προβλέψεων κατά την αξιολόγηση με την μέθοδο «10-fold cross validation» και στο εξωτερικό σετ ελέγχου για κάθε ισομορφή, με βάση τον βέλτιστο τύπο χαρακτηριστικών. Συγκρίνοντας την απόδοση για κάθε τύπο χαρακτηριστικού ανά ισομορφή του κυτοχρώματος P450 μεταξύ του σετ εκπαίδευσης και του σετ εξωτερικής δοκιμής, προσδιορίστηκαν τα βέλτιστα μοντέλα για κάθε ισομορφή. Ως βέλτιστη απόδοση θεωρήθηκε ότι η ακρίβεια, τόσο στο σετ εκπαίδευσης, όσο και στο σετ ελέγχου θα πρέπει να είναι υψηλή, αλλά χωρίς μεγάλη διαφορά μεταξύ των δυο σετ. Σημειώνεται, εδώ, πως μια σημαντική διαφορά (> 20%) θα μπορούσε να υποδεικνύει υπερπροσαρμογή (overfitting). Επιπλέον, οι τύποι χαρακτηριστικών που παρείχαν τις καλύτερες επιδόσεις στο σετ ελέγχου, συγκριτικά με το σετ εκπαίδευσης, θεωρήθηκαν ως μη βέλτιστοι, επειδή η απόδοση θα μπορούσε να υποδηλώνει υπο-προσαρμογή του μοντέλου.

Ισομορφές κυτοχρώματος P450	Τύπος Χαρακτηριστικών	10-fold cross validation					Εξωτερικό σετ ελέγχου				
		ACC (%)	SE (%)	SP (%)	MCC	AUC	ACC (%)	SE (%)	SP (%)	MCC	AUC
CYP1A2	PLEC	95.82	98.42	91.53	0.92	0.99	92.12	91.55	83.33	0.82	0.97
CYP2A6	ECFP	99.03	100	90.48	0.95	1	97.37	100	71.43	0.83	0.95
CYP2B6	PLEC + Desc. + Graph	90.23	100	66	0.76	0.99	88.09	87.50	63.64	0.68	0.89
CYP2C19	Desc. + Graph	93.92	94.30	91.92	0.88	0.99	76.72	57.14	60.87	0.43	0.80
CYP2C9	PLEC + Desc. + ECFP	96.42	99.89	87.90	0.91	0.99	81.27	38.37	33.67	0.25	0.73
CYP2D6	PLEC + Desc. + MACCS	92.23	100	45.49	0.65	0.99	89.24	64.51	43.48	0.47	0.88
CYP3A4	Desc. + PubChem	88.34	97.22	59.46	0.70	0.99	79.03	83.54	34.82	0.46	0.85

Πίνακας 1. Βέλτιστη απόδοση ανά ισομορφή για το σετ εκπαίδευσης και εξωτερικού ελέγχου

Ο Πίνακας 1 υποδηλώνει την ικανότητα γενίκευσης του επιλεγμένου τύπου χαρακτηριστικών για κάθε ισομορφή. Στον πίνακα δίδονται για κάθε ισομορφή, ο τύπος χαρακτηριστικών που οδηγεί στη βέλτιστη απόδοση και αξιολογείται μέσω της ακρίβειας, της ευαισθησίας, της ειδικότητας, της συσχέτισης Matthews και της περιοχής υπό την καμπύλη (AUC), τόσο σε δεδομένα εκπαίδευσης, όσο και σε δεδομένα ελέγχου. Η ισομορφή CYP1A2, που χρησιμοποιεί, μόνο, τα αποτυπώματα αλληλεπίδρασης, επιτυγχάνει ακρίβεια 95,82% στη 10-fold cross validation και 92,12% στο εξωτερικό σετ, με συντελεστή συσχέτισης Matthews ίσο προς 0,92 και 0,82, αντίστοιχα, για τα δύο σύνολα δεδομένων. Απόδοση, που υποδεικνύει την ικανότητα γενίκευσης, που παρέχει αυτός ο τύπος χαρακτηριστικών για το CYP1A2. Παρόμοιες επιδόσεις επιτυγχάνονται από τα CYP2A6 και CYP2B6 με την ακρίβεια κατά την 10-fold cross validation να είναι 100% και για τις δύο ισομορφές, ενώ η ακρίβεια στο εξωτερικό σετ ελέγχου ήταν 97,37% για το CYP2A6 και 88,09% για το CYP2B6. Η συσχέτιση Matthews ήταν αρκετά καλή στην περίπτωση του CYP2A6, αποδίδοντας 0,95 στο σετ εκπαίδευσης και 0,83 στο σετ ελέγχου, ενώ η συσχέτιση Matthews για το CYP2B6 ήταν 0,76 στην εκπαίδευση και 0,68 στο σετ ελέγχου. Αυτή η απόδοση επιτεύχθηκε, μόνο, με τα αποτυπώματα ECFP για το CYP2A6. Αντίθετα, το CYP2B6 απαιτούσε τον συνδυασμό αποτυπωμάτων αλληλεπίδρασης, μοριακών περιγραφέων και αποτυπωμάτων Graph προς βέλτιστη απόδοση και επαρκή γενίκευση. Το CYP2C19 είχε παρομοίως βέλτιστη απόδοση, χρησιμοποιώντας, μόνο, μοριακούς περιγραφείς και αποτυπώματα γράφου. Πιο συγκεκριμένα, παρά την καλή ακρίβεια πρόβλεψης, τόσο στα σετ εκπαίδευσης, όσο και στα σετ ελέγχου, η συσχέτιση Matthews μειώθηκε από 0,88 στα δεδομένα εκπαίδευσης σε 0,43 στα δεδομένα ελέγχου. Επιπλέον, αντίστοιχη απόδοση παρατηρήθηκε στις ισομορφές CYP2C9, CYP2D6 και CYP3A4, με συσχέτιση Matthews στα δεδομένα εκπαίδευσης ίση με 0,91, 0,65 και 0,70, αντίστοιχα. Ωστόσο, η συσχέτιση Matthews στο σετ ελέγχου έδειξε

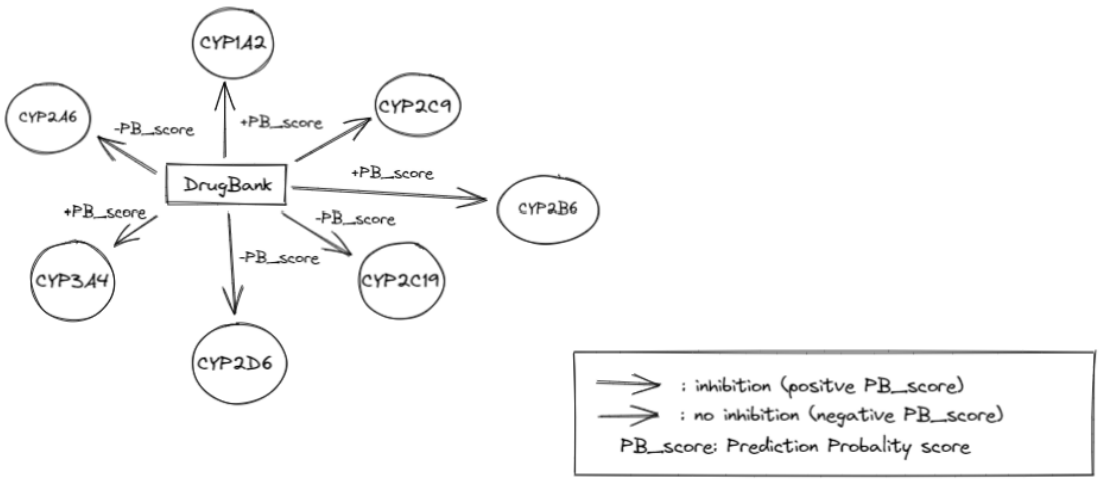
μείωση (66% για το CYP2C9, 18% για το CYP2D6 και 24% για το CYP3A4). Οι ισομορφές CYP2C9 και CYP2D6 χρησιμοποίησαν αποτυπώματα αλληλεπίδρασης, σε συνδυασμό με μοριακούς περιγραφείς και αποτυπώματα ECFP ή MACCS, αντίστοιχα, ενώ στο CYP3A4 χρησιμοποιούνται, μόνο, μοριακοί περιγραφείς και αποτυπώματα PubChem.
Στην Εικόνα 4 δίδονται τα 20 πιο σημαντικά χαρακτηριστικά και η σημαντικότητά τους κατά μέση τιμή, με βάση τον βέλτιστο τύπο χαρακτηριστικών για κάθε ισομορφή.



Εικόνα 4. Τα 20 πρώτα πιο σημαντικά χαρακτηριστικά ταξινομούνται με βάση τη μέση σημασία τους για τον βέλτιστο τύπο χαρακτηριστικών κάθε ισομορφής CYP450.

3.2 Εισαγωγή προβλέψεων στο γράφο

Τα αποτελέσματα των προβλέψεων, που προέκυψαν για κάθε χημικό μόριο, σε σχέση με την αντίστοιχη ισομορφή του κυτοχρώματος CYP450 (1A2, 2A6, 2B6, 2C9, 2C19, 2D6 και 3A4), εισήχθησαν στο γνωσιακό γράφο, ώστε να επανατοποθετεί/επαναστοχεύει, λαμβάνοντας υπόψιν τη δυνητική τοξικότητα. Τα δεδομένα, κατά την εισαγωγή τους στον γράφο, έχουν την παρακάτω δομή (Εικόνα 5).



Εικόνα 5. Διάγραμμα εισαγωγής των προβλέψεων για κάθε χημικό μόριο και ισομορφή του κυτοχρώματος CYP450 (1A2, 2A6, 2B6, 2C9, 2C19, 2D6 και 3A4) στον γράφο

3.3 Πρόβλεψη αλληλεπίδρασης φαρμάκου-πρωτεΐνης μέσω μηχανικής μάθησης σε γράφο

Από τα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν, με βάση τις δύο διαφορετικές τεχνικές εξαγωγής χαρακτηριστικών, λάβαμε τα αποτελέσματα που συνοψίζονται ακόλουθα.

	RF	SVM	KNN
Accuracy	80.80	81.88	81.38
Kappa	-2.17	-0.01	2.87
Sens	0.00	16.67	34.09
Spec	0.00	0.06	3.24
Precision	0.00	0.06	3.24
Recall	98.62	99.93	98.62
F1	*NA*	0.12	5.92
Matthews	-0.05	0.00	0.06
AUC	0.71	0.61	0.56

	RF	SVM	knn
Accuracy	80.65	81.83	81.38
Kappa	-2.44	-0.18	2.53
Sens	0.00	0.00	33.06
Spec	0.00	0.00	2.95
Precision	0.00	0.00	2.95
Recall	98.44	99.89	98.68
F1	*NA*	*NA*	5.42
Matthews	-0.05	-0.01	0.05
AUC	0.70	0.62	0.55

Πίνακας 2. Απόδοση των ταξινομητών στο σετ εκπαίδευσης (αριστερά) και στο σετ εξωτερικού ελέγχου (δεξιά), που έχουν σχεδιαστεί με χαρακτηριστικά από τον αλγόριθμο Node2Vec.

Τα αποτελέσματα του Πίνακα 2 αφορούν την εξαγωγή χαρακτηριστικών με την μέθοδο Node2Vec. Όπως παρατηρείται, η μέθοδος Node2Vec, αν και έχει υψηλή ακρίβεια, φαίνεται πως προβλέπει σωστά, μόνο την μια κλάση, όπως προκύπτει από τις τιμές ευαισθησίας και ειδικότητας.

	RF	SVM	KNN
Accuracy	95.07	96.52	95.77
Kappa	83.22	87.84	85.58
Sens	87.22	94.43	89.15
Spec	85.25	85.84	87.20
Precision	85.25	85.84	87.20
Recall	97.24	98.88	97.66
F1	86.22	89.93	88.16
Matthews	0.83	0.88	0.86
AUC	0.96	0.97	0.97

	RF	SVM	knn
Accuracy	95.18	96.85	96.20
Kappa	83.83	89.09	87.13
Sens	86.16	94.15	89.78
Spec	87.40	88.05	89.13
Precision	87.40	88.05	89.13
Recall	96.90	98.79	97.76
F1	86.78	91.00	89.45
Matthews	0.84	0.89	0.87
AUC	0.97	0.98	0.98

Πίνακας 3. Απόδοση των ταξινομητών σ το σετ εκπαίδευσης (αριστερά) και στο εξωτερικού ελέγχου (δεξιά) οι έχουν σχεδιαστεί με χαρακτηριστικά από τον αλγόριθμο FastRP.

Στον Πίνακα 3, συνοψίζονται τα αποτελέσματα των ταξινομητών οι οποίοι εκπαιδεύτηκαν με τα χαρακτηριστικά που εξήχθησαν με την τεχνική FastRP. Όπως διαφαίνεται, επιτυγχάνεται πρόβλεψη με πολύ υψηλή ακρίβεια για την κλάση ενός ζεύγους φαρμάκου-πρωτεΐνης.

4. Μελλοντικές Προοπτικές

Κατά την ανάλυση των αποτελεσμάτων των μοντέλων μηχανικής μάθησης, με βάση το γράφο, παρατηρήθηκε ότι στο σετ δεδομένων έχουν εισαχθεί πέραν των χημικών μορίων, φυσικά προϊόντα ή μονοκλωνικά αντισώματα. Συνεπώς, απαιτείται επιμέλεια των δεδομένων εκπαίδευσης και στη συνέχεια να υλοποιηθεί η εν λόγω προσέγγιση με τα νέα δεδομένα.

Επιπλέον, έχει σχεδιαστεί η δοκιμή της μεθόδου πρόβλεψης σύνδεσης σε υποδίκτυα, όπως τα ακόλουθα:

- Φάρμακο – Φάρμακο
- Πρωτεΐνη – Πρωτεΐνη
- Φάρμακο – Ασθένεια
- Πρωτεΐνη – Ασθένεια

Ως προς τις μεθόδους μηχανικής μάθησης στα δεδομένα γράφων, επιπρόσθετα, θα εφαρμοστούν οι αντίστοιχες μέθοδοι βαθιάς μάθησης, οι οποίες θα συγκριθούν ως προς την ακρίβεια και την ικανότητα γενίκευσης. Τέλος, θα ενταχθούν δεδομένα δομικής βιολογίας με σκοπό τον εμπλουτισμό των δεδομένων.

5. Αναφορές

1. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Ridwan Amode M, et al. Ensembl 2021. Nucleic Acids Res. 2021;49(D1):D884–91.

2. Relling M V., Klein TE. CPIC: Clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. Clin Pharmacol Ther [Internet]. 2011;89(3):464–7. Available from: http://dx.doi.org/10.1038/clpt.2010.279/nature06264

3. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.

4. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7.

5. Sherry ST, Ward M, Sirotkin K. dbSNP - database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res. 1999;9(8):677–9.

6. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 2019 Jan 8;47(D1):D930–40.

7. Huang HY, Lin YCD, Cui S, Huang Y, Tang Y, Xu J, et al. MiRTarBase update 2022: An informative resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. 2022;50(D1):D222–30.

8. Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open Targets Platform: Supporting systematic drug-target identification and prioritisation. Nucleic Acids Res. 2021;49(D1):D1302–10.