Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Χημικών Μηχανικών

Διπλωματική Εργασία

Μέθοδος εμπλουτισμού για βελτίωση πρόβλεψης βιοδραστικότητας μικρών μορίων

Γεώργιος Χ. Κάργας

Επιβλέπων:

Σαρίμβεης Χαράλαμπος, Καθηγητής ΕΜΠ

Φεβρουάριος 2018

Πρώτα απ’ όλα, θέλω να ευχαριστήσω τον επιβλέποντα της διπλωματικής εργασίας μου, Καθηγητή κ. Χαράλαμπο Σαρίμβεη, για την πολύτιμη βοήθεια και καθοδήγησή του κατά τη διάρκεια της δουλειάς μου. Επίσης, είμαι ευγνώμων στον κ. Γεώργιο Δρακάκη για την προσεκτική ανάγνωση της εργασίας μου καθ’ όλη τη διάρκεια της συγγραφής αυτής, για τις πολύτιμες υποδείξεις του καθώς και για την πολύτιμη βοήθειά του στον προγραμματισμό με Python. Πάνω απ’ όλα, είμαι ευγνώμων στους γονείς μου, Χρήστο και Παρασκευή Κάργα για την ολόψυχη αγάπη και υποστήριξή τους όλα αυτά τα χρόνια. Αφιερώνω αυτή την εργασία στην μητέρα μου και στον πατέρα μου.

Γεώργιος Χ. Κάργας

Δηλώνω υπεύθυνα ότι η διπλωματική εργασία είναι εξ' ολοκλήρου δικό μου έργο και κανένα μέρος της δεν είναι αντιγραμμένο από έντυπες ή ηλεκτρονικές πηγές, μετάφραση από ξενόγλωσσες πηγές και αναπαραγωγή από εργασίες άλλων ερευνητών ή φοιτητών. Όπου έχω βασιστεί σε ιδέες ή κείμενα άλλων, έχω προσπαθήσει με όλες μου τις δυνάμεις να το προσδιορίσω σαφώς μέσα από την καλή χρήση αναφορών ακολουθώντας την ακαδημαϊκή δεοντολογία.

Σύνοψη

Η αύξηση της υπολογιστικής ισχύος και της ποσότητας των δεδομένων που καλείται να επεξεργαστεί ο ερευνητής έχουν συμβάλλει στην ολοένα αυξανόμενη προσέγγιση των υπολογιστικών μεθόδων για το σχεδιασμό φαρμάκων. Ως πρώτος στόχος των επιστημόνων που ασχολούνται με τη σχεδίαση φαρμάκων με τη βοήθεια ηλεκτρονικού υπολογιστή τίθεται η αποτελεσματική απεικόνιση των δομών κανονικών και παθολογικών μορίων τα οποία στη συνέχεια συγκρίνονται με παθογενή ένζυμα και ενεργούς υποδοχείς αντίστοιχα οπότε και καθορίζεται ο στόχος φαρμακευτικού σχεδιασμού. Επομένως, αν είναι γνωστή η δομή μιας πρωτεΐνης και ο τρόπος που ο υποδοχέας ή η ενεργός περιοχή της, δύναται να προσομοιωθεί το μοντέλο δράσης της, εξοικονομώντας τον χρόνο και το κόστος που θα απαιτούσαν αντίστοιχες πειραματικές δοκιμές.

Σε αυτή την εργασία, θα σκιαγραφηθεί μια επέκταση στις *in silico* προβλέψεις για τη βιοδραστικότητα, ενώ εισάγεται ένα εξαιρετικά ερμηνεύσιμο και επαναλήψιμο μοντέλο για προ-επεξεργασία των μορίων *in silico*, προβλέποντας τις πρωτεΐνες στις οποίες πραγματοποιείται συσχέτιση με τη νόσο της ψύχωσης. Αυτή η ροή εργασίας (workflow) είναι μεταβιβάσιμη και σε άλλες νόσους.

Λέξεις κλειδιά

Μηχανισμός δράσης: Βιολογική απόκριση που προέρχεται από την χορήγηση φαρμάκου, εξαιτίας της αλληλεπίδρασης του μορίου με συγκεκριμένες πρωτεΐνες στόχους. Ορίζει τις λειτουργικές αλλαγές σε μοριακό επίπεδο, σε αντίθεση με το 'τρόπο δράσης', ο οποίος αναφέρεται στις αλλαγές που παρατηρούνται σε κυτταρικό επίπεδο μετά τη χορήγηση μια ουσίας.

Χημειογενωμική: Πεδίο έρευνας που μελετά την αλληλεπίδραση μεταξύ μικρών μορίων και πρωτεϊνών στόχων σε μεγάλη κλίμακα π.χ. λαμβάνοντας υπόψη τη δραστηριότητα των πολλαπλών προσδεμάτων έναντι πολλαπλών στόχων άμεσα, σε αντίθεση με το να λαμβάνει υπόψη μόνο ξεχωριστές αλληλεπιδράσεις προσδέματος-στόχου).

Πρόβλεψη Στόχου: Οι *in silico* τεχνικές αναγνώρισης χρησιμοποιούνται για να συνάγουν την αλληλεπίδραση μεταξύ μικρού μορίου και πρωτεΐνης αξιοποιώντας τα διαθέσιμα δεδομένα βιοδραστηριότητας από τα προσδέματα.

Συστημική Βιολογία(Systems biology): Διεπιστημονική (βιολογική, μαθηματική, αναλυτική κλπ.) προσέγγιση η οποία αντλεί από την αναγνωρισμένη συνθετότητα των ζωντανών οργανισμών των οποίων η συμπεριφορά μπορεί να κατανοηθεί πλήρως μόνο λαμβάνοντας υπόψη όλες τις δυναμικές αλληλεπιδράσεις των μερών του βιολογικού συστήματος.

Abstract

The increase in computational power and the amount of data that a researcher is supposed to process have contributed to the ever-increasing approach of computational methods for drug design. The first objective of scientists involved in computer-aided drug design is to effectively depict the structures of normal and pathological molecules which are then compared to pathogenic enzymes and active receptors respectively, as the purpose of pharmaceutical design is set. Thus, if the structure of a protein is known and the way in which the receptor or its active region acts, the model of action can be simulated, saving the time and cost that would be required by corresponding experimental testing.

In this diploma thesis, an extension of in silico predictions for bioactivity will be outlined, while a highly interpretable and repeatable model for pre-processing of in silico molecules is introduced, predicting the proteins correlated with psychosis disease. This workflow is transferable to other diseases.

Keywords

Mechanism of action: Biological response resulting from drug administration because of the interaction of the molecule with specific target proteins. Defines the functional changes at a molecular level, as opposed to the "mode of action", which refers to the changes occurring at the cellular level after the administration of a substance.

Chemo Genomics: Research field studying the interaction between small molecules and protein targets on a large scale (i.e., taking account of the activity of multiple ligands against multiple targets at once, in contrast to only take into account individual ligand-target interactions).

Target prediction: *In silico* identification techniques are used to deduce the interaction between small molecule and protein utilizing the available data bioactivity of the ligands.

Systems biology: Interdisciplinary (biological, mathematical, analytical etc.) approach that draws on the acknowledged complexity of living organisms whose behaviour can be fully understood only by considering all dynamic interactions of the parts of a biological system.

Biological pathway: A set of interactions between protein and molecules which leads to a specific reaction of the cell, organ or organism.

Περιεχόμενα

1. Εισαγωγή

1.1 Ανακάλυψη φαρμάκων

Η ανακάλυψη νέων φαρμάκων προκύπτει ως αποτέλεσμα συστηματικών μελετών και έρευνας. Μέχρι τα μέσα του 19ου αιώνα η ανακάλυψη φαρμάκων βασιζόταν σε ατομική προσπάθεια,[1] ενώ σήμερα η ανακάλυψη φαρμάκων αποτελεί αποτέλεσμα ομαδικής εργασίας και της στενής συνεργασίας επιστημόνων από διάφορους τομείς, όπως η ιατρική, η βιοχημεία, η χημεία, η επιστήμη των υπολογιστών, η φαρμακολογία, η μικροβιολογία, η τοξικολογία, η φυσιολογία, η παθολογία.

Μετά από ποικίλες διαδικασίες, αυτό το οποίο συνήθως ανακαλύπτεται δεν είναι το κλινικά χρησιμοποιούμενο φάρμακο, αλλά μια βιοδραστική ένωση, που ονομάζεται ένωση-οδηγός (lead compound).[2] Η ένωση-οδηγός είναι μια πρωτότυπη ένωση που παρουσιάζει επιθυμητή βιολογική ή φαρμακολογική δραστικότητα ενώ στην πλειονότητα των περιπτώσεων παρουσιάζει και ανεπιθύμητα χαρακτηριστικά, όπως για παράδειγμα, υψηλή τοξικότητα, άλλες βιολογικές δράσεις πέραν της επιθυμητής, κακή διαλυτότητα ή προβλήματα μεταβολισμού. Μετά την ανακάλυψη, η δομή της ένωσης-οδηγού θα πρέπει να τροποποιηθεί ώστε να ενισχυθεί η επιθυμητή δραστικότητα και να εξαφανισθούν ή να ελαχιστοποιηθούν οι ανεπιθύμητες ιδιότητες της.

Για την ανακάλυψη ενός φαρμάκου ή ένωσης-οδηγού απαιτείται η ύπαρξη δοκιμασίας (assay) των διάφορων ενώσεων ως προς μια συγκεκριμένη βιολογική δραστικότητα, ώστε να είναι δυνατή η εκτίμηση της δραστικότητας αυτής. Μια τέτοια δοκιμασία (bioassay or screen) μας δίνει τη δυνατότητα να εκτιμήσουμε τη δραστικότητα μιας ένωσης, δηλαδή το ειδικό βιολογικό ή φαρμακολογικό αποτέλεσμα που προκαλεί, αλλά και τη δραστικότητα αυτής, συνήθως σε σύγκριση ως προς μια γνωστή ένωση αναφοράς.[3] Ο έλεγχος της δραστικότητας μιας ένωσης περιλαμβάνει *in vitro* δοκιμασίες, όπως την αναστολή ενός ενζύμου ή την πρόσδεση του σε έναν υποδοχέα, και *in vivo* δοκιμασίες, όπως τη μέτρηση μιας παραμέτρου σε ένα πειραματόζωο.[3]

Η ανακάλυψη φαρμάκων, προγενέστερα, βασιζόταν στις φαινοτυπικές ενδείξεις στο επίπεδο των οργανισμών, όπως η επίδραση των βοτάνων και άλλων φυσικών θεραπειών στους ανθρώπους. Σε αυτήν την περίπτωση, εξεταζόταν μόνο ο επιθυμητός φαινότυπος (δηλαδή η ίαση του ασθενούς), ενώ η επιλεκτική δράση των διαφόρων συστατικών του φαρμάκου στον οργανισμό παρέμενε ανεξερεύνητη.[4]

Η ασπιρίνη και άλλα παρόμοια φάρμακα, των οποίων η αποτελεσματικότητα έχει αποδειχθεί μέσω φαινοτυπικών ενδείξεων, ήταν διαθέσιμα στην αγορά για πολλά χρόνια δίχως να είχε επιβεβαιωθεί πλήρως ο τρόπος δράσης του φαρμάκου (Mode of Action,

MoA). Ομοίως, ακόμη και σήμερα, κάθε χρόνο μόνο ένας μικρός αριθμός φαρμάκων εγκρίνεται από την αμερικανική Υπηρεσία Τροφίμων και Φαρμάκων (US Food and Drug Administration - FDA), από τα οποία ένα μέρος αυτών έχουν γνωστό τρόπο δράσης.[5]

1.2 Μέθοδοι ελέγχου

Τα τελευταία χρόνια, στα πλαίσια της ανακάλυψης και εξέλιξης των φαρμάκων, η ιατρική έχει στραφεί στον κλάδο της μοριακής βιολογίας εστιάζοντας σε συγκεκριμένα γονιδιακά προϊόντα. Στόχος αποτελεί, η απομόνωση και ο προσδιορισμός των ουσιών αυτών που επηρεάζουν την ομαλή λειτουργία του κυττάρου ή του οργανισμού.[6] Τις πιο κοινές μεθόδους για την απεικόνιση και ταυτοποίηση των βιοδραστικών μικρών μορίων αποτελούν οι προσεγγίσεις που βασίζονται είτε στο φαινότυπο είτε στη στοχευμένη απεικόνιση μορίων.

Ο έλεγχος του φαινοτύπου (phenotypic screening) αποτελεί ένα είδος απεικόνισης που χρησιμοποιείται στη βιολογία και στην ανακάλυψη φαρμάκων (Phenotypic Drug Discovery-PDD) για την ταυτοποίηση ουσιών, όπως μικρά μόρια, πεπτίδια, RNAi, που διαμορφώνουν τον φαινότυπο ενός κυττάρου ή οργανισμού. Η απεικόνιση του φαινοτύπου στηρίζεται στις επιδράσεις, τους φαινοτύπους, που προκαλούν οι ουσίες στα κύτταρα, τους ιστούς ή σε έναν ολόκληρο οργανισμό. [6] Συγκεκριμένα, μετριέται in vitro η επίδραση των ουσιών σε μια καθαρή στοχευμένη πρωτεΐνη. Επομένως αυτές οι απεικονίσεις θα μπορούσαν να οδηγήσουν στην ταυτοποίηση ενός μορίου που διαμορφώνει μία ασθένεια στον φαινότυπο, δρώντας είτε σε μία θέση είτε σε πολλές θέσεις ταυτόχρονα.

Η προσέγγιση του φαινοτύπου αποτελεί μία μέθοδο όπου ο στόχος είναι συνδεδεμένος με μία σχετική με την ασθένεια βιολογική απόκριση, τα αντισώματα δεσμεύονται με φυσικό στόχο στα κύτταρα, μπορούν να ανακαλυφθούν παράλληλα αρκετοί στόχοι ή μηχανισμοί δράσης και τέλος προσφέρεται μεγαλύτερη δυνατότητα καινοτομίας με λιγότερο ανταγωνισμό. Αντίθετα, η μέθοδος αυτή[7] περιορίζεται στους ήδη γνωστούς στόχους ή στους ήδη γνωστούς μηχανισμούς δράσης, ενώ η ταυτοποίηση του στόχου αποτελεί απαιτητική διαδικασία όπου μπορεί να χρειαστεί ο συνδυασμός αυτής με την μέθοδο της στοχευμένης απεικόνισης.

Από το 1980 κι έπειτα, όπου σημειώθηκαν σημαντικές πρόοδοι στον κλάδο της μοριακής βιολογίας και της γονιδιωματικής, ο φαινοτυπικός έλεγχος αντικαταστάθηκε από τον έλεγχο καθορισμένων στόχων που εμπλέκονται στην ασθένεια. Αναπτύχθηκε, λοιπόν, η υπόθεση[6] πως ένας συγκεκριμένος βιολογικός στόχος καθορίζει την ασθένεια και τέθηκε το ζήτημα ανίχνευσης του. Με τον καθορισμό της ουσίας ή των ουσιών που προκαλούν την ασθένεια, πραγματοποιείται μια σειρά *in vivo* πειραμάτων ώστε να διαπιστωθεί η

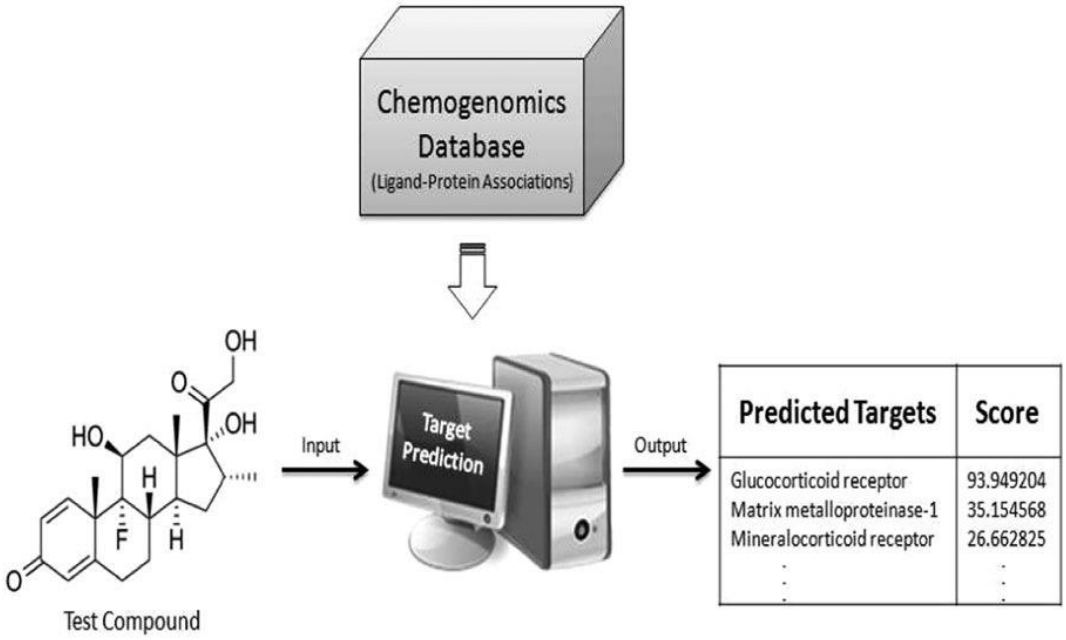
ακρίβεια της αρχικής υπόθεσης. Η μέθοδος αυτή αναφέρεται ως στοχευμένη απεικόνιση (Target-based Drug Discovery, TDD) και ακολουθείται από αρκετές εταιρείας έρευνας και ανάπτυξης φαρμάκων.

Η στοχευμένη προσέγγιση πλεονεκτεί σε σχέση με την αντίστοιχη φαινοτυπική καθώς είναι γνωστός ο μοριακός στόχος και ο μηχανισμός δράσης αυτού ενώ αποτελεί έναν απλό σχεδιασμό (High Throughput Screening, HTS).[7] Αντιθέτως, ενδέχεται ο στόχος να μην αποτελεί καθοριστικό παράγοντα της ασθένειας, ενώ οι βιβλιογραφικές μελέτες αυτού δεν παρουσιάζουν αναπαραγωγισιμότητα. Επίσης, ορισμένες φορές οι βιοχημικές και λειτουργικές αναφορές δεν συσχετίζονται και είναι γνωστός μονάχα ένας δραστικός μηχανισμός.

Κατά την τελευταία δεκαετία,[6] ωστόσο, οι υπεύθυνοι για την ανάπτυξη φαρμάκων προτείνουν συνδυασμό των δύο παραπάνω προσεγγίσεων δίχως αποκλειστική χρήση μόνο μίας εξ' αυτών, καθώς η πρώτη παρουσιάζει ελλείψεις ενώ η δεύτερη δεν ήταν από μόνη της αρκετή για τον πλήρη προσδιορισμό των καθοριστικών παραγόντων.

1.3 *In silico* μέθοδος πρόβλεψης στόχων

Οι μεγάλες φαρμακευτικές εταιρείες τα τελευταία χρόνια, βρίσκονται υπό την πίεση να μειώσουν τις ζημίες από πρόσφατες ή τις επικείμενες λήξεις των διπλωμάτων ευρεσιτεχνίας και τη μειωμένη παραγωγικότητα, γνωστή και ως «innovation gap»[8]. Αυτό το γεγονός συνεπάγεται πως το κόστος έρευνας δε μετατρέπεται σε κέρδος καθώς μόνο ένα μικρό ποσοστό νέων φαρμάκων εισέρχονται στην αγορά.[8]



Εικόνα 1.3-1: Επισκόπηση target-prediction.Η πληροφορία της ορφανής ένωσης δακτυλικών αποτυπωμάτων τροφοδοτείται εντός του αλγορίθμου, η οποία προβλέπει την πιθανότητα πρόσδεσης (score) στις πρωτεΐνες με βάση προηγούμενη γνώση. Η μέθοδος αυτή καθορίζει τη σχέση μεταξύ της ένωσης και των πρωτεϊνών-στόχων, συνδέοντάς το περαιτέρω με το MoA (αναπαράχθηκε από *Ravindranath et al.*)[9]

Ιστορικά, η ανακάλυψη νέων φαρμάκων βασίζεται στη χημεία και τη φαρμακολογία. Ωστόσο, με την έλευση των γονιδιωματικών επιστημών, η βιολογία καθιερώθηκε ως η βασική κινητήρια δύναμη, στην οποία βασίζεται η ανακάλυψη νέων φαρμάκων. Τα προγράμματα ανακάλυψης φαρμάκων, συνήθως, αρχίζουν με την ταυτοποίηση των κατάλληλων στόχων του φαρμάκου (drug targets). Στις περισσότερες περιπτώσεις,[2] αυτοί οι στόχοι είναι βιομόρια όπως πρωτεΐνες, ένζυμα και δίαυλοι ιόντων. Κατά τη σταδιακή επαλήθευση του στόχου, θα πρέπει να εξασφαλίζεται ένα επαρκές επίπεδο «εμπιστοσύνης» πως ο στόχος σχετίζεται με την υπό μελέτη ασθένεια και η επικείμενη διαμόρφωσή του θα οδηγήσει σε αποτελεσματική θεραπεία της νόσου.

Οι αρχικές προβλέψεις για την επαλήθευση του στόχου, συνήθως, λαμβάνονται *in vitro* και σε ζωικά μοντέλα, ενώ η ακριβής επαλήθευση πραγματοποιείται κατόπιν κλινικών δοκιμών σε ανθρώπους.[2]

Αφού επαληθευτεί ο στόχος, θα πρέπει να προσδιοριστούν οι ρυθμιστές του στόχου, οι οποίοι μπορεί να δρουν συναγωνιστικά ή ανταγωνιστικά στην περίπτωση των υποδοχέων, ως ενεργοποιητές ή αναστολείς των ενζύμων και των ιοντικών διαύλων. Αυτό το στάδιο της αρχικής ταυτοποίησης ξεκινά με το σχεδιασμό και την ανάπτυξη κατάλληλου ανιχνευτή για την παρακολούθηση του υπό μελέτη στόχου.[2]

Στη συνέχεια, η διαλογή υψηλής απόδοσης (High-throughput screening, HTS) εκθέτει το στόχο σε ένα μεγάλο αριθμό από χημικές ενώσεις (τυπικά της τάξεως των 10^5)[2] που έρχονται όλο και περισσότερο από υψηλής ταχύτητας παράλληλη και συνδυαστική σύνθεση. Οι δραστικές ενώσεις που επιδεικνύουν δοσοεξαρτώμενη ρύθμιση του στόχου, χαρακτηρίζονται ως επικεφαλείς ενώσεις όταν αποδεικνύεται ένας καθορισμένος βαθμός εκλεκτικότητας για τον υπό μελέτη στόχο ενώ τα πρώτα θετικά αποτελέσματα λαμβάνονται από ζωικά μοντέλα. Έπειτα οι επικεφαλείς ενώσεις βελτιστοποιούνται ως προς την ισχύ και την εκλεκτικότητά τους, ενώ καθορίζονται οι φυσικοχημικές τους ιδιότητες, οι φαρμακοκινητικές και τα χαρακτηριστικά ασφαλείας τους αξιολογούνται πριν επιλεχθούν για την ανάπτυξη φαρμάκων.[2]

Παρά του γεγονότος πως το μεγαλύτερο μέρος της διαδικασίας της πρώιμης φαρμακευτικής έρευνας βασίζεται κυρίως σε πειραματικές εργασίες στο εργαστήριο, ο ρόλος του υπολογιστή στην ανάπτυξη φαρμάκων αποκτά σταδιακά όλο και μεγαλύτερο ρόλο.[2]

Ο σχεδιασμός φαρμακευτικών ουσιών με τη βοήθεια ηλεκτρονικού υπολογιστή αποτελεί μια από τις βασικές μεθόδους ανακάλυψης νέων φαρμακοφόρων ουσιών. Πράγματι, τα τελευταία χρόνια, ο σχεδιασμός φαρμάκων με τη βοήθεια ηλεκτρονικού υπολογιστή (Computer Aided Drug Design, CADD) έχει εξελιχθεί σε έναν αρκετά υποσχόμενο και παράλληλα αρκετά ενδιαφέροντα επιστημονικό τομέα, με αρκετές εφαρμογές και σε βιομηχανικό επίπεδο.[10] Σε αυτή την κατεύθυνση βοήθησαν τόσο οι υπερσύγχρονοι με τεράστιες υπολογιστικές δυνατότητες ηλεκτρονικοί υπολογιστές όσο και η ανάπτυξη αρκετών πακέτων λογισμικού που υλοποιούν μία σειρά πειραματικών προσεγγίσεων πάνω στον τομέα της σύνθεσης χημικών ουσιών. Το σύνολο των βιοχημικών διεργασιών που πραγματοποιούνται κατά την επίδραση ενός φαρμάκου σε έναν ζωντανό οργανισμό, καθώς και ο λεπτομερής μηχανισμός των διαμορφώσεων ενός φαρμακευτικού μορίου στην πορεία του μέσα στον οργανισμό είναι αρκετά πολύπλοκες και όχι απόλυτα γνωστές στους επιστήμονες. Η δομή του κάθε μορίου καθορίζει σε σημαντικό βαθμό την αλληλεπίδραση του με άλλα μόρια και τη βιοχημική του δράση.[11]

Κατά την ανακάλυψη νέων φαρμάκων, προγενέστερα, εξεταζόταν μόνο ο επιθυμητός φαινότυπος (δηλαδή η ίαση του ασθενούς), ενώ η επιλεκτική δράση των διαφόρων συστατικών του φαρμάκου στον οργανισμό δεν διερευνούνταν. Τα μεταγενέστερα χρόνια, θεωρήθηκε πως κάθε φάρμακο αποτελείται από «magic bullets», δηλαδή κάθε φάρμακο δρα διαμορφώνοντας έναν στόχο για την αντίστοιχη ασθένεια (σχέση κλειδιού-κλειδαριάς), ενώ τα τελευταία χρόνια θεωρείται πως κάθε φάρμακο αποτελείται από πολλαπλούς στόχους. Επομένως για να μπορέσει ένα φάρμακο να εξασφαλίσει την επιθυμητή δραστικότητα στον άνθρωπο, θα πρέπει να προσδιοριστούν οι στόχοι που εξασφαλίζουν την αποτελεσματικότητα αλλά και να αποφευχθούν εκείνοι που δημιουργούν παρενέργειες.[4]

Με αυτή τη θεώρηση αλλάζει ο τρόπος αντίληψης της βιοδραστικότητας, μετασχηματίζοντάς την από μονοδιάστατη σε πολυδιάστατη. Με την χρήση μοντέλων πρόβλεψης των στόχων είναι δυνατή η πρόβλεψη της βιοδραστικότητας σε μικρότερο χρόνο με μεγαλύτερη απόδοση.[4]

Σήμερα, με τη χρήση υπολογιστικών μεθόδων και χάρις την πληθώρα βιβλιογραφικών δεδομένων βιοδραστικότητας ενώσεων[4], είναι δυνατός ο σχεδιασμός συμπλόκων που θα εξασφαλίζουν το επιθυμητό προφίλ βιοδραστικότητας, προβλέποντας παράλληλα δευτερεύουσες και εναλλακτικές βιοδραστικότητες σε πρώιμο στάδιο. Δημόσιες βάσεις δεδομένων όπως ChEMBL,[12] PubChem[13] και ChemBank,[14] συνεχώς αυξάνουν σε μέγεθος, με την ChEMBL να περιέχει δεδομένα για 700.000 μικρομόρια και 2,7 εκατομμύρια δεδομένα βιοδραστικότητας ενώσεων. Επομένως, γίνεται εφικτή η αντιστοίχιση του χημικού και βιολογικού υπόβαθρου στα εκάστοτε μοντέλα ώστε να εξαχθεί ένα βιολογικό φάσμα με σκοπό την πρόβλεψη της φαινοτυπικής δράσης των νέων μορίων.

Οι *in silico* μέθοδοι μπορούν να εκμεταλλευτούν την προηγούμενη γνώση των αλληλεπιδράσεων συμπλόκου-στόχου (ligand-target), που συλλέγονται για παράδειγμα από τη βιβλιογραφία ή τα διπλώματα ευρεσιτεχνίας, η οποία οργανώνεται στις χημειογενωμικές (chemogenomic) βάσεις δεδομένων. Στη συνέχεια, αυτά τα δεδομένα αναλύονται υπολογιστικά, προκειμένου να κάνουν προβλέψεις για τα νέα, μη δοκιμασμένα μόρια ή να προτείνουν νέες αλληλεπιδράσεις φαρμάκου-στόχου για τις ενώσεις που ήδη κυκλοφορούν. Με το συνδυασμό των σχέσεων δομής δραστικότητας (Structure Activity Relationships, SAR) και των μηχανικών μεθόδων, είναι δυνατό να εξερευνηθεί ένας μεγάλος χημειογενωμικός χώρος όπου δύναται να αναγνωριστεί ένας μηχανισμός δράσης όπου προηγουμένως ήταν αδύνατος.[15]

Τέτοιες μέθοδοι μπορούν να λειτουργήσουν συμπληρωματικά της *in vitro* ανίχνευσης και ως εκ τούτου, με την πληθώρα των πληροφοριών που αποθηκεύονται σε δημόσιες βάσεις δεδομένων (βλέπε Πίνακα 1.3-1), έχουν γίνει όλο και πιο σημαντικές στην ανακάλυψη φαρμάκων.

Πίνακας 1.3-1: Λίστα με τις βασικότερες βάσεις δεδομένων όπως αναπαράχθηκε από *Koutsoukas et al.*[4] το 2011

Βάση Δεδομένων	Δεδομένα
ChEMBL	>700χιλ. μικρομόρια, >2,7 εκατ. Δεδομένα βιοδραστικότητας
PubChem	250χιλ. ενώσεις, 2500 βιοαναλύσεις
WOMBAT	
DrugBank	4800 εγγραφές φαρμάκων, συμπεριλαμβανομένων: >1350 FDA-εγκεκριμένα φάρμακα και 123 FDA-εγκεκριμένων βιολογικών
ChemBank	Πληροφορίες για εκατοντάδες χιλιάδες μικρομόρια και εκατοντάδες βιοιατρικές αναλύσεις
Comparative Toxicogenomics Database (CTD)	6000 ενώσεις, 1,6 εκατ. δεδομένα χημειο-γονιδιακών ασθενειών
SuperTarget	500 φάρμακα, 2500 στόχους-πρωτεΐνες και 7300 αλληλεπιδράσεις φαρμάκου-στόχου
MATADOR	Χειροκίνητα σημιεώνονται ενώσεις από τη βάση δεδομένων SuperTarget
Therapeutic Target Database (TTD)	Περιέχει 1906 στόχους, συμπεριλαμβανομένων 358 επιτυχημένων και 251 υπό κλινικών δοκιμών, 43 που έχουν διακοπεί, 1254 ερευνητικών στόχων 5124 φάρμακα, συμπεριλαμβανομένων των 1511 εγκεκριμένων, 1118 κλινικών δοκιμών και 2331 πειραματικά φάρμακα
PubChem BioActivity	
BindingDB	>271χιλ. ενώσεις, >620χιλ. συγγενείς δέσμευσης 5526 πρωτεινικών στόχων
DrugPort (EBI)	1492 εγκεκριμένα φάρμακα και 1664 μοναδικών πρωτεινικών στόχων
Potential Drug Target Database (PDTD)	Περιέχει 1207 καταγραφές όπου οι 841 είναι γνωστοί και πιθανοί στόχοι φαρμάκων με δομές από την βάση δεδομένων Protein Data Bank
Promiscuous	> 20 εκατ. εκδόσεις, περιλαμβάνει >10 εκατ. πρωτεΐνες, >25 χιλ. ενώσεις, όπως επίσης και παρενέργειες φαρμάκων και διάφορα άλλα

	δεδομένα, όπως την οπτικοποίηση του δικτύου πρωτεϊνών
T3DB	Βάση δεδομένων τοξινών και τοξινών-στόχων, >2900 τοξίνες, >1300 τοξίνες-στόχους και >33000 συσχετίσεις τοξίνων-στόχων
Chem2Bio2RDF	RDF βάση δεδομένων με ενσωματωμένα χημικά, βιολογικά και φαινοτυπικά δεδομένα

Η μεγαλύτερη πρόκληση εμφανίζεται στο να ενσωματωθούν σωστά όλες οι πηγές των διαθέσιμων πληροφοριών (βιοχημικές, λειτουργικές, φαινοτυπικές, γενετικές και η διαδρομή), προκειμένου να είναι σε θέση να συνδέσουν τις ενώσεις με τους στόχους τους και, τελικά, την επίδρασή τους σε ένα βιολογικό σύστημα. Η πρόβλεψη του στόχου επιτρέπει την κάλυψη των κενών μεταξύ χημικής και βιολογικής γνώσης.

Η αναζήτηση ομοιότητας για πρόβλεψη στόχου με βάση τα προσδέματα θεωρείται η απλούστερη μορφή[4] στην *in silico* πρόβλεψη στόχου και έχει καθιερωθεί στις επιστημονικές εργασίες. Οι προβλέψεις βασίζονται στην αρχή της μοριακής ομοιότητας με τις αναγνωρισμένες βιοενεργές ενώσεις από τις χημειογενωμικές βάσεις δεδομένων. Η απλουστευτική φύση αυτών των μεθόδων σημαίνει ότι έχουν την ικανότητα να θεωρούν τη δομή ενός μορίου ως σύνολο, το οποίο παρεμποδίζει σημαντικά την δύναμη πρόβλεψης αυτών των μοντέλων. Επίσης, οι αλγόριθμοι εξόρυξης δεδομένων έχουν την ικανότητα να λαμβάνουν υπόψη πολλαπλούς συνδυασμούς από θραύσματα ενώσεων εφαρμόζοντας τεχνικές αναγνώρισης. Οι μέθοδοι αυτές είναι πλέον ελκυστικές για την πρόβλεψη στόχου εξαιτίας του ότι έχουν επιδείξει τις ικανότητές τους σε πιο αναλυτικές τελικές προβλέψεις. Ένα από τα πρώτα και πιο διαδεδομένα παραδείγματα εξόρυξης δεδομένων για την αποσαφήνιση στόχου είναι το λογισμικό Prediction of Activity Spectra for Substances (PASS)[16] το οποίο αφομοιώθηκε από τις βιοδραστηριότητες για περισσότερα από 270.000 ζεύγη προσδεμάτων ενώσεων. Οι συγγραφείς εφάρμοσαν το μοντέλο σε πολυεπίπεδες γειτονιές περιγραφέων ατόμων (Multilevel Neighborhoods of Atoms, MNA), παράγοντας προβλέψεις που βασίζονται στις εκτιμήσεις πιθανοτήτων του Bayes.[17]

Οι πιο απλές μέθοδοι πρόβλεψης του στόχου (target prediction) βασίζονται στην μελέτη των απλών μορίων, καθώς οι θεωρήσεις μελέτης του συμπλόκου (Ligand-based methods), βασίζονται στην αρχή των χημικών ομοιοτήτων.[4] Συγκεκριμένα, θεωρείται πως ενώσεις που παρουσιάζουν παρόμοιες χημικές δομές έχουν την τάση να παρουσιάζουν αντίστοιχα παρόμοια βιολογική δράση σε σχέση με το αντίθετο, παρόλο που οι διαφορετικές δομές της ίδιας ένωσης δρουν με διαφορετικό τρόπο με την πρωτεΐνη στόχο. Αυτές οι μέθοδοι στηρίζονται στην προγενέστερη γνώση της βιοδραστικότητας του συμπλόκου και των πρωτεϊνικών δομών, ώστε μέσα σε ελάχιστο χρονικό διάστημα να φιλτράρει όλο το μεγάλο όγκο των βάσεων δεδομένων και να

εντοπίσει τις πιο χημικά συνδεδεμένες ενώσεις που περιέχουν παρόμοιες δομές ώστε να συνδέονται με την αντίστοιχη πρωτεΐνη στόχο.[4]

Συμπληρωματικά, περαιτέρω μέθοδοι πρόβλεψης της βιοδραστικότητας των χημικών ενώσεων αποτελούν οι μέθοδοι εξόρυξης δεδομένων (data mining methods, DMM).[4] Οι μέθοδοι αυτές, σε αντίθεση με τις μεθόδους χημικών ομοιοτήτων που εντοπίζουν απλά πιθανούς στόχους με βάση την ομοιότητα των συμπλόκων των μοριακών υποδομών, εκτελούν αναγνώριση προτύπων στις χημειογενωμικές βάσεις δεδομένων και συμπλεγμάτων που παρουσιάζουν ομοιότητες στον πολυδιάστατο χώρο.[4]

Υπάρχουν πολλές κατηγορίες *in silico* μεθόδων οι οποίες βασίζονται στον τρόπο πρόσδεσης, με τις κυριότερες να παρουσιάζονται παρακάτω:[4]

- *In silico* πρόβλεψη στόχων βασισμένη σε απλή δομή.

Αυτές οι μέθοδοι βασίζονται σε προϋπάρχουσα γνώση των βιοενεργών προσδεμάτων και των πρωτεϊνικών δομών και μπορούν να ανασύρουν σε λίγο χρονικό διάστημα τεράστιες χημικές βιβλιοθήκες για να εντοπίσουν τις σχετικές χημικά δομές οι οποίες μοιράζονται υποδομικές ομοιότητες – και αλληλεπιδρούν- με παρόμοιες πρωτεΐνες στόχους.[4]

- Πρόβλεψη στόχου με βάση μεθόδους εξόρυξης δεδομένων.

Ένας ολοένα αναπτυσσόμενος τομέας που χρησιμοποιείται για την πρόβλεψη της βιοδραστικότητας βασίζεται στη μέθοδο εξόρυξης δεδομένων. Αυτές οι μέθοδοι εν αντιθέσει με τις μεθόδους χημικής ομοιότητας που εντοπίζουν απλά πιθανούς στόχους με βάση τις ομοιότητες πρόσδεσης των μοριακών υποδομών, εκτελούν αναγνώριση δομών στο χημειογενομικό χώρο και του συμπλέγματος των ενώσεων που παρουσιάζουν ομοιότητες στον πολυδιάστατο χώρο.[4] Αυτές οι μέθοδοι βασίζονται σε υποστηρικτικές διανυσματικές μηχανές (Support Vector Machines, SVMs) και πρότυπα γραμμικής παλινδρόμησης (Linear Regression Models, LRMs), πολλαπλές κατηγορίες του Bayes (Multiple-Category Bayesian) όπως εφαρμόζονται από τους Nidhi et al.[18] και την προσέγγιση Συνόλου Ομοιότητας (Similarity Ensemble Approach, SEA).[4]

Οι ταξινομητές Naïve Bayes (NB) αποτελούν μια δημοφιλή οικογένεια αλγόριθμων που χρησιμοποιούνται για την πρόβλεψη της βιοδραστικότητας των ενώσεων. Οι Nidhi et al.,[18] χρησιμοποίησαν ένα πολύ-ταξικό αλγόριθμο Naïve Bayes σε ένα σετ δεδομένων που αποτελείται από περισσότερες από 960 πρωτεΐνες στόχου που προέρχονται από τη βάση δεδομένων με το όνομα «World Of Molecular BioAcTivity» (WOMBAT)[19]. Ένας άλλος αλγόριθμος πρόβλεψης στόχου αναπτύχθηκε από τους Koutsoukas et al.,[15] και έχει την ικανότητα να προβλέπει τις σχέσεις δομικής δραστηριότητας (Structure Activity Relationships, SARs) για ορφανές ενώσεις χρησιμοποιώντας είτε τον τροποποιημένο Λαπλασιανό Naïve Bayes ταξινομητή είτε τον Parzen-Rosenblatt Window (PRW). Ο

αλγόριθμος εφαρμόστηκε σε πάνω από 155.000 ζεύγη προσδέματος-πρωτεϊνών από την βάση δεδομένων ChEMBL14, και περιέχει 894 διαφορετικές πρωτεΐνες-στόχους. Μετά τη συγκριτική κατάταξη των πειραμάτων βρέθηκε ότι ο PRW απέδωσε συνολικά καλύτερα σε σχέση με τον αλγόριθμο Naïve Bayes, επιτυγχάνοντας ανάκληση και ακρίβεια της τάξης του 66,6 και 63,3 % αντίστοιχα.[4]

- Χημειογενωμικές προσεγγίσεις

Αυτές οι μέθοδοι έχουν ως στόχο να αναλύσουν εκτενώς τη σχέση συμπλόκου-στόχου καθώς οι πρωτεΐνες στόχοι κατηγοριοποιούνται με βάση τις ομοιότητες των συμπλόκων έτσι ώστε κάποιος να μπορεί να διαπιστώσει τη δυνατότητα σύνδεσης του στόχου με το σύμπλοκο βασιζόμενος στην εκάστοτε δομή αυτών. Ως εκ τούτου, οι μέθοδοι στηρίζονται σε πρότυπα γραμμικής παλινδρόμησης όπως το drugCIPHER από τους Zhao et al.,[20] σε μεθόδους με βάση το αποτύπωμα πρωτεΐνης - προσδέματος (Protein–Ligand Fingerprint-based method, PLFP), σε τοπολογικούς περιγραφείς, που υποστηρίζουν τις συνθέσεις για να προβλέψουν την βιοδραστηριότητα των ενώσεων σε πρωτεϊνικούς στόχους, με τη χρήση λογισμικών όπως το Knime, Inte:Ligand PharmacophoreDB και το LigandScout.[4]

- Αποτυπώματα HTS στην πρόβλεψη στόχου.

Αυτές οι μέθοδοι περιλαμβάνουν μικρούς μοριακούς βιολογικούς περιγραφείς που ονομάζονται «έλεγχος υψηλής διεκπεραίωσης» (High-Throughput Screening - HTS). Για να συγκριθούν τα HTS αποτυπώματα των δύο ενώσεων, η μετρητική ομοιότητα θα πρέπει να λάβει υπόψη της την σπανιότητα των δεδομένων. Σύνθετα HTS αποτυπώματα συγκρίνονται χρησιμοποιώντας μια σταθμισμένη συσχέτιση Pearson για να δημιουργηθεί ένας πίνακας (Matrix) N-προς-N ομοιότητας. Αυτός ο πίνακας ομοιότητας μπορεί να χρησιμοποιηθεί για να δημιουργηθεί ένα δίκτυο όπου κάθε κόμβος (ένωση) συνδέεται εάν ο βαθμός ομοιότητας είναι πάνω από 0.85.[17]

- Ενδεχόμενη αξιοποίηση των δεδομένων βιοδραστηριότητας για σχεδιασμό φαρμάκων πολλαπλών στόχων: Εντοπίζοντας το σωστό στόχο.

Τα Βιολογικά δίκτυα είναι δομικά αρκετά σύνθετα, η πρόβλεψη του λειτουργικού αποτελέσματος των παρεμβάσεων ή των συνεπειών μεταλλάξεων που πρέπει αργότερα να αντιμετωπιστεί φαρμακολογικά είναι μη αμελητέα. Τα φάρμακα σχεδόν πάντα επιδρούν σε πάνω από έναν στόχους , έως κάποιο βαθμό, σαν συνέπεια των δομικών ομοιοτήτων μεταξύ του επιθυμητού στόχου και άλλων συγγενών πρωτεϊνών.[17]

1.4 Εφαρμογές της *in silico* πρόβλεψης στόχου

Ένας σημαντικός αριθμός εφαρμογών της *in silico* πρόβλεψης στόχων έχουν δημοσιευθεί πρόσφατα, που αφορούν από την διευκρίνιση του μηχανισμού δράσης των ενώσεων ως και την ανάλυση των ψευδών θετικών αποτελεσμάτων σε δοκιμασίες γονιδίου αναφοράς.

Για παράδειγμα, η «Συνολική Προσέγγιση Ομοιότητας» (Similarity Ensemble Approach, SEA), αποτελεί μία μέθοδο με την οποία διευκρινίστηκαν επιπρόσθετοι στόχοι φαρμάκων που βρίσκονται ήδη στην αγορά.[4] Υπολογίστηκαν οι ομοιότητες 3665 FDA-εγκεκριμένων και πειραματικών φαρμάκων έναντι εκατοντάδων στόχων με αποτέλεσμα την ανακάλυψη νέων διασυνδέσεων με πρωτεΐνες. Συγκεκριμένα προσδιορίστηκε η περίπτωση του ανταγωνισμού του β1 υποδοχέα από την φλουοξετίνη (fluoxetine) του Prozac, της αναστολής του διαβιβαστή σεροτονίνης (serotonin) από την ινφεπροδίλη (ifenprodil) του Vadilex και του ανταγωνισμού του H4 υποδοχέα από την δελαβιρδίνη (delavirdine) του Rescriptor.

Επίσης, οι Muller et al.[21] χρησιμοποιώντας υψηλής απόδοσης σάρωση, εξέτασαν 2.000 ενεργές θέσεις φαρμάκων από SC-PDB για τον εντοπισμό υποθετικών βιολογικών στόχων των παραγώγων του ικριώματος 1,3,5-τριαζεπαν-2,6-διόνη. Συγκεκριμένα, επικεντρώθηκαν σε πέντε μόρια εκπροσώπους αυτής της συνδυαστικής βιβλιοθήκης. Μεταξύ των πολλά υποσχόμενων βιολογικών στόχων που προσδιορίστηκαν και επαληθεύτηκαν in vitro ήταν η εκκρινόμενη φωσφολιπάση A2 (sPLA2).[4]

Παρακάτω αναλύονται οι δύο ειδικές εφαρμογές οι οποίες ισχύουν για αυτό το έργο, η πολυφαρμακολογία και η επαναστόχευση φαρμάκων.

1.4.1 Πολυφαρμακολογία

Η αντίληψη των «Magic Bullets»[22] (που επίσης αναφέρεται και ως η προσέγγιση «ένα φάρμακο-μία ασθένεια») αναφέρθηκε αρχικά από τον Ehrlich[23] και μέχρι πρόσφατα αποτελούσε τη βασική θεώρηση για το σχεδιασμό των φαρμάκων. Συγκεκριμένα, στηριζόταν στην παραγωγή μικρών μορίων με μεγάλη εκλεκτικότητα απέναντι σε ένα μόνο στόχο, ο οποίος θεωρείται υπεύθυνος ή σχετίζεται με την παθογένεια της ασθένειας. Η παραδοχή, με βάση την οποία εξελίχθηκε αυτή η φιλοσοφία, είναι πως τα παραγόμενα φάρμακα θα είναι ασφαλέστερα και πιο αποτελεσματικά καθώς θα αποφεύγονται οι πιθανές ανεπιθύμητες αντιδράσεις. Ιδιαίτερα, στις περιπτώσεις των διαταραχών του Κεντρικού Νευρικού Συστήματος (ΚΝΣ- Central Nervous System, CNS), όπως κατάθλιψη και σχιζοφρένεια, η φιλοσοφία των «magic bullets» αποδείχτηκε αναποτελεσματική.[4]

Το πρώτο μοντέλο λογισμικού που υποστήριζε την σκιαγράφηση των φασμάτων βιοδραστικότητας στηριζόμενο στην σύσταση των υπαρχόντων φαρμάκων ήταν το Ligand Profiler, διαθέσιμο από την Discovery Studio. Το πρόγραμμα πραγματοποιούσε απεικονίσεις μιας ομάδας ενώσεων σε μια σειρά φαρμακοφόρων μοντέλων που έχουν επιλεχθεί από τον χρήστη. Συγκεκριμένα, πραγματοποιείται μία μέτρηση σχετική με το πόσο καλά η ένωση αντιστοιχεί στη χημική συνάρτηση των ιδιοτήτων του φαρμάκου ενώ στα αποτελέσματα παρουσιάζεται η πιθανότητα αντιστοίχισης της κάθε ένωσης με τον εκάστοτε στόχο.[4]

Τα δεδομένα των φαρμάκων μπορούν να ανακτηθούν από ελεύθερες διαδικτυακές βάσεις δεδομένων όπως DrugBank,[24] BindingDB,[25] PDBBind[26] και PDTD.[27] Στη παρούσα διπλωματική εργασία ως βάση δεδομένων για την ανάκτηση των φαρμακευτικών δομών χρησιμοποιήθηκε η DrugBank ενώ ως λογισμικό πραγματοποίησης των μετρήσεων, το KNIME.[28]

1.4.2 Επαναστόχευση Φαρμάκων

Ο όρος επαναστόχευση φαρμάκων (drug repositioning, drug repurposing) αναφέρεται στην ανεύρεση νέων ασθενειών στις οποίες μπορούν να χρησιμοποιηθούν ήδη υπάρχοντα φάρμακα, με διαφορετικές αρχικές ενδείξεις.[29]

Η επαναστόχευση φαρμάκων έχει ήδη οδηγήσει σε ορισμένες σημαντικές επιτυχίες, όπως είναι η σιλδεναφίλη (κύρια δραστική ουσία του φαρμάκου Viagra της εταιρίας Pfizer), που αναπτύχθηκε αρχικά για την αντιμετώπιση της πνευμονικής υπέρτασης, ενώ πλέον χρησιμοποιείται για την θεραπεία της στυτικής δυσλειτουργίας.[29]

Ένα σημαντικό πλεονέκτημα της επαναστόχευσης φαρμάκων σε σχέση με τις παραδοσιακές μεθόδους ανάπτυξής τους είναι ότι τα επιλεγόμενα φάρμακα ήδη χρησιμοποιούνται, και, επομένως, έχουν περάσει ένα σημαντικό αριθμό ελέγχων τοξικότητας.[29]

Η επαναστόχευση φαρμάκων μεγιστοποιεί τη χρήση της υπάρχουσας γνώσης και δημιουργεί ελπίδες ότι θα αποτελέσει ένα σημαντικό συμπλήρωμα στις τρέχουσες τεχνικές ανάπτυξης φαρμάκων και αντιμετώπισης ασθενειών.[29]

1.5 Περιορισμοί της υπολογιστικής πρόβλεψης στόχων

Η σχεδίαση φαρμάκων με τη βοήθεια ηλεκτρονικού υπολογιστή περιλαμβάνει την ανακάλυψη νέων φαρμακοφόρων μορίων από αναζήτηση σε βάσεις χημικών πληροφοριών, τη βελτιστοποίηση των φαρμακοφόρων μορίων (με σκοπό την ελαχιστοποίηση των παρενεργειών όπως τοξικότητα), καθώς και τη σχεδίαση «εκ νέου» μορίων που μπορούν να προσδένονται σε συγκεκριμένους υποδοχείς για να λειτουργούν ως ανταγωνιστές ή αναστολείς. Παράλληλα η μοριακή σχεδίαση επιτρέπει την τρισδιάστατη αναπαράσταση των φαρμακοφόρων μορίων και τη μελέτη των φυσικοχημικών τους ιδιοτήτων.

Όμως παρόλη την εξέλιξη των τελευταίων χρόνων παραμένουν κάποια ανοικτά προβλήματα, τα οποία με την επίλυσή τους θα συμβάλλουν σημαντικά στην περαιτέρω εξέλιξη της *in silico* πρόβλεψης στόχου.

Μερικοί από τους περιορισμούς που συναντώνται κατά την υπολογιστική πρόβλεψη στόχων είναι:

1. η απουσία ενός γενικού και ενιαίου εργαλείου σχεδίασης μοριακών δομών που να περιλαμβάνει το σύνολο των βιολογικών μορίων,[30]
2. η αυξημένη υπολογιστική πολυπλοκότητα που εκφράζεται σε χρόνο και απαιτούμενους πόρους και η οποία αυξάνει εκθετικά με την αύξηση του μεγέθους του υπό εξέταση μορίου,[4]
3. η επιλογή του κατάλληλου μοντέλου αναπαράστασης (ανάλογα πάντα με το βιολογικό μόριο) και ο καθορισμός των κρίσιμων παραμέτρων που πρέπει να εξεταστούν ειδικότερα σε επίπεδο διανυσματικής γεωμετρίας,[4]
4. η αντιμετώπιση των σφαλμάτων στα δεδομένα εισόδου και η ανακατασκευή ενός τρισδιάστατου μοντέλου από ελλιπή ή λανθασμένα δεδομένα,[30]
5. η ταυτόχρονη αναπαράσταση ενός συνόλου φυσικοχημικών ιδιοτήτων με τρόπο που η πληροφορία να είναι κατανοητή και ερμηνεύσιμη από τον ερευνητή,[31]
6. το πρόβλημα των επικαλυπτόμενων περιοχών,[11]

1.6 Ψύχωση, αντιψυχωτικά χάπια και οι μηχανισμοί τους

1.6.1 Ψύχωση – Διαταραχή του Κεντρικού Νευρικού Συστήματος

Η διαταραχή του Κεντρικού Νευρικού Συστήματος (ΚΝΣ) αποτελεί μια ευρεία κατηγορία των συνθηκών υπό τις οποίες ο εγκέφαλος δεν λειτουργεί όπως θα έπρεπε,

περιορίζοντας την υγεία και την ικανότητα λειτουργίας αυτού. Η διαταραχή αυτή μπορεί να οφείλεται είτε σε μια κληρονομική μεταβολική διαταραχή, είτε να είναι αποτέλεσμα βλάβης από μία λοίμωξη, μία εκφυλιστική κατάσταση, ένα αγγειακό εγκεφαλικό επεισόδιο, όγκο στον εγκέφαλο ή κάποιο άλλο πρόβλημα, είτε να προκύψει από άγνωστους ή πολλούς παράγοντες.[32]

Η ψύχωση είναι μια χρόνια νόσος του εγκεφάλου που προσβάλλει περίπου το 1% των ανθρώπων παγκοσμίως. Ο προσφορότερος τρόπος για την κατανόηση των συμπτωμάτων της ψύχωσης είναι η προσεκτική μελέτη της υποκειμενικής προσωπικής εμπειρίας των ασθενών. Κανένα σύμπτωμα από μόνο του δεν αρκεί για τη διάγνωση της αρρώστιας. Οποιοδήποτε σύμπτωμα μπορεί να εμφανιστεί και σε άλλες ιατρικές καταστάσεις. Από την άλλη μεριά, κανένα σύμπτωμα δεν είναι αναγκαστικά παρόν σε όλους τους πάσχοντες. Ο κάθε ασθενής βιώνει το δικό του συνδυασμό συμπτωμάτων που μπορεί να περιλαμβάνει μεγαλύτερη ένταση κάποιων συμπτωμάτων και μικρότερη κάποιων άλλων. Δεν υπάρχει καμιά εργαστηριακή μέθοδος που να επιβεβαιώνει τη διάγνωση. Ο στόχος ορισμένων επιλεγμένων εργαστηριακών εξετάσεων είναι μόνο ο αποκλεισμός άλλων ιατρικών καταστάσεων που μπορεί να ευθύνονται για τα συμπτώματα.[33]

1.6.2 Θεραπεία, τρόπος δράσης αντιψυχωτικών φαρμάκων, παρενέργειες

Τα αντιψυχωτικά φάρμακα αποτελούν τον κορμό αντιμετώπισης των διαταραχών του Κεντρικού Νευρικού Συστήματος και ειδικότερα της ψύχωσης. Αναλυτικότερα δεν προσφέρουν ριζική θεραπεία με τον τρόπο που τα αντιβιοτικά θεραπεύουν κάποιες λοιμώξεις ή μια χειρουργική επέμβαση αποκαθιστά πλήρως ένα κάταγμα, αλλά ελέγχουν αποτελεσματικά τα συμπτώματα στο μεγαλύτερο μέρος των ασθενών και τους διευκολύνουν στην προσπάθειά τους να ζήσουν όσο το δυνατόν φυσιολογικότερη ζωή.

Επιπρόσθετα, τα αντιψυχωτικά φάρμακα μειώνουν δραστικά την πιθανότητα υποτροπής χωρίς όμως να την εξαλείφουν, ενώ δεν είναι εξίσου αποτελεσματικά για όλα τα συμπτώματα της αρρώστιας. Τα λεγόμενα θετικά συμπτώματα που είναι πιο θορυβώδη (παραληρητικές ιδέες, ψευδαισθήσεις) ανταποκρίνονται καλύτερα, ενώ τα αρνητικά συμπτώματα (έλλειψη πρωτοβουλίας και συναισθηματικής ανταπόκρισης, δυσκολίες στην προσοχή και τη μνήμη) που έχουν δυσμενέστερες επιπτώσεις στην καθημερινή λειτουργικότητα, λιγότερο καλά.

Τα αντιψυχωτικά φάρμακα διακρίνονται αδρά σε δύο μεγάλες ομάδες, τα πρώτης γενιάς ή κλασσικά ή τυπικά και τα δεύτερης γενιάς ή άτυπα. Ο βασικός μηχανισμός δράσης είναι κοινός και αφορά την μείωση της παθολογικά αυξημένης δραστηριότητας της ντοπαμίνης. Η κλοζαπίνη (Leponex) είναι αδιαμφισβήτητα το αποτελεσματικότερο αντιψυχωτικό, το μόνο που έχει επανειλημμένως αποδειχθεί αποτελεσματικό στην ανθεκτική σχιζοφρένεια δηλαδή σε συμπτωματολογία που δεν έχει ανταποκριθεί σε άλλα

φάρμακα. Μεταξύ των υπολοίπων ατύπων φαρμάκων, μόνο η ολανζαπίνη, η αμισουλπρίδη και η ρισπεριδόνη φαίνεται να υπερέχουν σε αποτελεσματικότητα των τυπικών φαρμάκων. Τα άλλα άτυπα αντιψυχωτικά έχουν παρόμοια αποτελεσματικότητα με τα παλαιότερα φάρμακα.[34]

Τα αντιψυχωτικά φάρμακα συγκαταλέγονται μεταξύ των ασφαλέστερων στην ιατρική. Παρ' όλα αυτά ανεπιθύμητες ενέργειες μπορεί να εμφανιστούν.

Οι συνηθέστερες ανεπιθύμητες ενέργειες των πρώτης γενιάς αντιψυχωτικών περιλαμβάνουν τις εξής:[34]

1. Εξωπυραμιδικά συμπτώματα δηλαδή ανεπιθύμητες δράσεις στις κινήσεις του σώματος. Η οξεία δυστονία είναι μια ακούσια επίμονη και ενδεχομένως επώδυνη σύσπαση κάποιων μυών συνήθως του λαιμού, της γλώσσας ή των ματιών που προκαλεί στροφή του κεφαλιού, ή των ματιών ή δυσκολία στην ομιλία ή την κατάποση. Συμπτώματα που μοιάζουν με αυτά της νόσου Πάρκινσον είναι η δυσκαμψία των μυών και η γενικότερη επιβράδυνση των κινήσεων καθώς και τρόμος (τρέμουλο) κυρίως στα χέρια. Η ακαθισία είναι ένα δυσάρεστο υποκειμενικό αίσθημα ανησυχίας που συνοδεύεται από τάση για διαρκή κίνηση. Η όψιμη δυσκινησία μπορεί να εμφανιστεί μετά από μήνες ή χρόνια αντιψυχωτικής αγωγής και αφορά ακούσιες κινήσεις κυρίως στην περιοχή του προσώπου και του λαιμού. Το κακόηθες σύνδρομο νευροληπτικών είναι μια σπάνια αλλά επικίνδυνη ανεπιθύμητη ενέργεια που χαρακτηρίζεται από σύγχυση, υπνηλία ή διέγερση, αλλαγές στην αρτηριακή πίεση και την καρδιακή συχνότητα, υψηλό πυρετό και έντονη δυσκαμψία.

2. Αντιχολινεργικές ανεπιθύμητες ενέργειες που περιλαμβάνουν ξηροστομία, δυσκοιλιότητα, δυσκολία στην ούρηση, θαμπή όραση, ταχυκαρδία και δυσκολίες στη συγκέντρωση και τη μνήμη.

3. Καταστολή, υπνηλία και υπόταση.

4. Αύξηση της ορμόνης προλακτίνη με ενδεχόμενες συνέπειες διαταραχές της έμμηνης ρύσης που αν παραταθούν χρονικά δημιουργούν κίνδυνο οστεοπόρωσης και σεξουαλική δυσλειτουργία.

5. Αύξηση του βάρους.

Τα δεύτερης γενιάς αντιψυχωτικά φάρμακα εμφανίζουν έντονη ετερογένεια με το καθένα να έχει διακριτά χαρακτηριστικά. Σε σχέση με τα πρώτης γενιάς, τα δεύτερης γενιάς συνδέονται με μικρότερα ποσοστά εξωπυραμιδικών παρενεργειών και αύξησης της προλακτίνης. Ενώ εμφανίζουν μεγαλύτερα ποσοστά αύξησης του βάρους, του σακχάρου

και των λιπιδίων του αίματος. Όμως, υπάρχουν έντονες διαφορές μεταξύ τους καθώς η κλοζαπίνη και η ολανζαπίνη συνδέονται με σημαντική αύξηση βάρους και επίδραση στις μεταβολικές παραμέτρους ενώ η ρισπεριδόνη και η κετιαπίνη με μέτριες επιδράσεις και η ζιπρασιδόνη και αριπιπραζόλη με ακόμα μικρότερες επιδράσεις.[35]

1.7 Επεκτάσεις της πρόβλεψης στόχων στην παρούσα εργασία

Στην παρούσα διπλωματική εργασία πραγματοποιήθηκε μία μορφή επέκτασης της πρόβλεψης στόχων στον τομέα της ψύχωσης καθώς τέθηκε ως στόχος η συσχέτιση ορισμένων πρωτεϊνών - στόχων με την ασθένεια της ψύχωσης μέσω της χρήσης αντίστοιχου λογισμικού για τον υπολογισμό αυτής. Καταρχάς, χρησιμοποιήθηκαν μικρά μόρια από την διαδικτυακή βάση δεδομένων της DrugBank. Η συγκεκριμένη βάση δεδομένων επιλέχθηκε καθώς περιέχει περισσότερες από 4100 εγγραφές φαρμάκων με τις οποίες συνδέονται 14000 τουλάχιστον πρωτεΐνες ή στόχοι των φαρμάκων.[36] Στη συνέχεια με τη χρήση του μοντέλου πρόβλεψης στόχων των *Koutsoukas et. al,*[4] πραγματοποιήθηκε μελέτη των πιθανών πρωτεϊνών στόχων που συνδέονται με την ψύχωση και με τον περαιτέρω εμπλουτισμό των εξαχθέντων αποτελεσμάτων προσδιορίστηκαν οι πρωτεΐνες με το μεγαλύτερο δείκτη πρόσδεσης με την ψύχωση. Συγκεκριμένα ο εμπλουτισμός πραγματοποιήθηκε με παραλλαγή της μεθόδου των *Liggi, Drakakis, Koutsoukas et. al* [37] όπου τα μόρια με καλύτερο estimation score ήταν πιο κοντά στο 1 και με χειρότερο πιο κοντά στο 0, επιλέχθηκαν τα μόρια με estimation score τουλάχιστον 0,99. Τέλος, τα επιλεχθέντα μόρια μέσω βιβλιογραφικής αναζήτησης συσχετίστηκαν με την ασθένεια της ψύχωσης.

2. Θεωρητικό Υπόβαθρο

2.1 Μοριακή Αναπαράσταση

2.1.1 SMILES

Πριν τον ορισμό της γλώσσας SMILES, είναι σημαντικό να οριστεί το φυσικό μοντέλο στο οποίο στηρίζεται, το μοντέλο των δεσμών χημικού σθένους, το οποίο χρησιμοποιεί ένα μαθηματικό γράφημα για να αναπαραστήσει ένα μόριο.[38]

Στο χημικό γράφημα οι κόμβοι είναι άτομα, και οι άκρες είναι ημι-σταθεροί δεσμοί που μπορεί να είναι μονοί, διπλοί ή τριπλοί ανάλογα με τους κανόνες της θεωρίας του δεσμού σθένους. Αυτό το απλό νοητικό μοντέλο παρουσιάζει μικρή ομοιότητα με την υφιστάμενη κβαντομηχανική πραγματικότητα των ηλεκτρονίων, των πρωτονίων και νετρονίων, παρ' όλα αυτά έχει αποδειχθεί ως μια αξιοσημείωτα χρήσιμη προσέγγιση για τον τρόπο συμπεριφοράς των ατόμων όταν βρίσκονται σε κοντινή απόσταση το ένα με το άλλο. Ωστόσο, το μοντέλο του δεσμού σθένους αποτελεί μια ατελή αναπαράσταση της μοριακής δομής και η γλώσσα SMILES έχει κληρονομήσει αυτές τις ατέλειες. Οι χημικοί δεσμοί είναι συχνά ταυτομερικοί, αρωματικοί ή διπολικοί και όχι τακτικά και ακέραια πολλαπλοί. Οι απεντοπισμένοι δεσμοί, οι κεντρικοί δεσμοί, οι δεσμοί υδρογόνου και άλλες δια-ατομικές δυνάμεις που χαρακτηρίζονται από κβαντομηχανική περιγραφή δεν ακολουθούν το μοντέλο δεσμού σθένους.[38]

Η SMILES αναπτύχθηκε αρχικά ως κατοχύρωση της Daylight Chemical Information Systems. Από τα τέλη της δεκαετίας του 1980, η SMILES αποτελεί ευρέως αποδεκτό πρότυπο για την ανταλλαγή των μοριακών δομών. Πολλά ανεξάρτητα πακέτα λογισμικού SMILES είναι γραμμένα σε C, C++, Java, Python, LISP, και ίσως ακόμα σε FORTRAN.[38]

Η SMILES είναι μια απλή αλλά περιεκτική χημική γλώσσα στην οποία τα μόρια και οι αντιδράσεις προσδιορίζονται χρησιμοποιώντας χαρακτήρες ASCII που αναπαριστούν άτομα και σύμβολα δεσμών. Η SMILES περιλαμβάνει τις ίδιες πληροφορίες που βρίσκονται σε έναν εκτεταμένο πίνακα συνδέσεων αλλά με αρκετά πλεονεκτήματα. Μια συμβολοσειρά SMILES είναι κατανοητή από τον άνθρωπο, πολύ συμπυκνωμένη και αν κανονικοποιηθεί αναπαριστά μια μοναδική συμβολοσειρά που μπορεί να χρησιμοποιηθεί ως καθολικό αναγνωριστικό για μια συγκεκριμένη χημική δομή. Επιπλέον, μια χημικά σωστή και κατανοητή απεικόνιση μπορεί να γίνει από μια οποιαδήποτε συμβολοσειρά SMILES είτε συμβολίζει μόριο είτε αντίδραση.[38]

Πίνακας 2.1-1: Γενικό παράδειγμα από δομή σε smiles[39]

Αιθανόλη	CCO
Οξικό οξύ	CC(=O)O
Κυκλοεξάνιο	C1CCCCC1
Πυριδίνη	c1cnccc1
Trans-2-βουτένιο	C/C=C/C
L-αλανίνη	N[COOH](C)C(=O)O
Χλωριούχο νάτριο	[Na+].[Cl-]
Αντίδραση μετατόπισης	C=CCBr>>C=CCI

Η ανάπτυξη της SMILES ξεκίνησε από τον David Weininger στα τέλη τη δεκαετίας του 80 χρησιμοποιώντας την έννοια του γραφήματος με κόμβους να συμβολίζουν τα άτομα και τις άκρες ως δεσμούς να αναπαριστούν ένα μόριο. Οι παρενθέσεις χρησιμοποιούνται για να συμβολίσουν τα σημεία διακλάδωσης και οι αριθμητικές επισημάνσεις προσδιορίζουν τα σημεία σύνδεσης των δακτυλίων. Η βασική γραμματική της SMILES περιλαμβάνει ακόμα ισοτοπικές πληροφορίες, ρυθμίσεις για τους διπλούς δεσμούς και χειρομορφία όπου οδήγησε σε αυτό που είναι γνωστό ως ισομερική SMILES.[40]

Από τη σύλληψή της, η SMILES έχει τροποποιηθεί και επεκταθεί για να συμπεριλάβει όχι μόνο τα νέα χαρακτηριστικά αλλά και δυο επιπλέον χημικές γλώσσες: την SMARTS, μια επέκταση της SMILES που επιτρέπει τον προσδιορισμό των μοριακών δομών και των ιδιοτήτων τους με αναζήτηση των επιμέρους δομών με τα επίπεδα ακρίβειας να ποικίλουν και την SMIRKS, μια περιορισμένη έκδοση της reaction SMARTS που περιλαμβάνει τις αλλαγές σε επίπεδο ατόμων-δεσμών που ορίζουν τις τυπικές αντιδράσεις.[39]

2.1.2 SMARTS

Η αναζήτηση υποδομής, η διαδικασία εύρεσης μιας συγκεκριμένης διάταξης σε ένα μόριο, είναι από τις σημαντικότερες εργασίες για τους υπολογιστές στη χημεία. Χρησιμοποιείται σχεδόν σε κάθε εφαρμογή που χρησιμοποιεί την ψηφιακή αναπαράσταση ενός μορίου, την απεικόνιση του (για να τονίσει μια συγκεκριμένη λειτουργική ομάδα), όπως στον σχεδιασμό των φαρμάκων (αναζητώντας σε όμοιες βάσεις δεδομένων για παρόμοιες δομές και δράση) και στην αναλυτική χημεία (αναζητώντας για δομές που είχαν από πριν χαρακτηριστεί και συγκρίνοντας τα δεδομένα με αυτά μιας άγνωστης δομής).

Η γλώσσα SMARTS επιτρέπει να καθοριστούν οι υποδομές χρησιμοποιώντας κανόνες που είναι απευθείας επεκτάσεις της SMILES. Για παράδειγμα, για να αναζητηθούν, σε μια βάση δεδομένων, οι δομές που περιέχουν φαινόλες, θα χρησιμοποιηθεί η SMARTS συμβολοσειρά [OH]c1ccccc1, η οποία είναι γνωστή στους εξοικειωμένους με την SMILES. Πράγματι, σχεδόν όλες οι προδιαγραφές SMILES είναι έγκυροι SMARTS στόχοι. Χρησιμοποιώντας την SMARTS, μπορούν να πραγματοποιηθούν ευέλικτες και αποτελεσματικές αναζητήσεις των προδιαγραφών των υποδομών με τρόπο που να έχουν νόημα για τους χημικούς.[41]

Στην γλώσσα SMILES, υπάρχουν δύο θεμελιώδεις τύποι συμβόλων: τα άτομα και οι δεσμοί. Αντίστοιχα, στην SMARTS χρησιμοποιούνται σύμβολα δεσμών και ατόμων για να περιγράψουν ένα γράφημα. Ωστόσο, στην SMARTS οι επισημάνσεις για ττους κόμβους και τις άκρες του γραφήματος επεκτείνονται ώστε να περιλαμβάνουν λογικούς τελεστές και ειδικά σύμβολα για άτομα και δεσμούς. Αυτοί οι τελεστές επιτρέπουν στους δεσμούς και στα άτομα της SMARTS να είναι πιο γενικοί. Για παράδειγμα, το SMARTS ατομικό σύμβολο [C, N] είναι ένα άτομο που μπορεί να είναι αλειφατικό C ή αλειφατικό N, το SMARTS σύμβολο για το δεσμό ~ (tilde) ταιριάζει σε κάθε δεσμό.[41]

2.1.3 Συσχέτιση γλώσσας SMARTS - SMILES

Όλες οι εκφράσεις SMILES είναι επίσης έγκυρες εκφράσεις SMARTS, αλλά η σημασιολογία αλλάζει διότι η SMILES περιγράφει μόρια ενώ η SMARTS περιγράφει σχήματα. Το μόριο που αναπαρίσταται από μια συμβολοσειρά SMILES ταιριάζει συνήθως με την ίδια συμβολοσειρά που χρησιμοποιείται στην SMARTS.[42]

Η SMILES ερμηνεύεται ως ένα μόριο, και είναι το μόριο που προκύπτει (και όχι η συμβολοσειρά SMILES) και υπόκειται σε αναζήτηση. Με τον ίδιο τρόπο, η SMARTS ερμηνεύεται ως σχήμα, σε αυτό το σχήμα (και όχι στην συμβολοσειρά SMARTS) αντιστοιχίζονται μόρια. Για παράδειγμα, στη SMILES το "C1=CC=CC=C1"

(κυκλοεξατριένιο) ερμηνεύεται ως μόριο βενζολίου. Το μόριο αυτό αντιστοιχίζεται από το SMARTS c1ccccc1, το οποίο ερμηνεύεται ως ένα σχήμα με “έξι αρωματικά άτομα άνθρακα σε δακτύλιο”. Η SMARTS "C1=CC=CC=C1" δημιουργεί ένα σχήμα ("έξι αλειφατικά άτομα άνθρακα σε ένα δακτύλιο με εναλλασσόμενους μονούς και διπλούς δεσμούς") το οποίο δεν αντιστοιχίζεται σε βενζόλιο αλλά με το μη-αρωματικό φαινυλικό κατιόν με SMILES C1=CC=CC=[CH+]1.[42]

Όταν τα άτομα προσδιορίζονται χωρίς αγκύλες στη SMILES, χρησιμοποιούνται προεπιλεγμένες τιμές. Στην SMARTS οι απροσδιόριστες ιδιότητες δεν ορίζονται ως μέρος του σχήματος. Για παράδειγμα, στην SMILES το O σημαίνει ένα αλειφατικό άτομο οξυγόνου με μηδενικό φορτίο και δύο άτομα υδρογόνου, δηλαδή το νερό. Στην SMARTS, η ίδια έκφραση σημαίνει κάθε αλειφατικό άτομο οξυγόνου ανεξάρτητα από το φορτίο, το αριθμό των ατόμων υδρογόνου κλπ., θα αντιστοιχεί στο νερό, αλλά επίσης και στην αιθανόλη, στην ακετόνη, στο μοριακό οξυγόνο, στα ιόντα οξωνίου. Προσδιορίζοντας τα όρια του [OH2] περιορίζεται το σχήμα στο να αντιστοιχεί μόνο στο νερό, ο οποίος αποτελεί τον προσδιορισμό της SMILES για το νερό.[42]

Επίσης, υπάρχουν κάποιοι αναχρονισμοί στους περισσότερους ερμηνευτές της SMILES που μπορούν να οδηγήσουν σε σύγχυση. Κάποιοι ερμηνευτές της SMILES επιτρέπουν «κρυφά» άτομα υδρογόνου να προστεθούν ως φανερά σαν συντόμευση στην εισαγωγή δεδομένων. Για παράδειγμα, Η SMILES για 1H-πυρρόλιο είναι "[nH]1cccc1" η οποία αντιστοιχίζεται ως SMARTS από το "n1cccc1".[42]

Τέλος, οι περισσότερες εκφράσεις SMARTS δεν είναι έγκυρες εκφράσεις SMILES. Η συμβολοσειρά "cOc" είναι μια έγκυρη SMARTS, η οποία αντιστοιχίζεται με ένα αλειφατικό οξυγόνο που συνδέεται με δύο αρωματικά άτομα άνθρακα ως μέρος ενός μεγαλύτερου μορίου(π.χ. του διφαυνυλικού αιθέρα). Ωστόσο, "cOc" δεν περιγράφει καθ’αυτό το μόριο επομένως δεν αποτελεί έγκυρη SMILES.[41]

2.1.4 Η μορφή SDF

Ίσως η πιο διαδεδομένη αναπαράσταση 2D και 3D μορίου μικρής χημικής δομής είναι αυτή που χρησιμοποιείται στη μορφή αρχείων SDF.[43] Η αναπαράσταση SDF περιέχει τρεις συντεταγμένες, έναν πίνακα συνδέσεων που περιγράφει τις πληροφορίες για τους δεσμούς και τη στερεοχημεία. Σύνθετες χημικές αντιδράσεις και ενδιάμεσες αντιδράσεις μπορούν επίσης να αναπαρασταθούν σε ένα αρχείο SDF. Αυτή η μορφή αρχείων χρησιμοποιείται ευρέως για την ανταλλαγή δεδομένων μικρών μοριακών δομών μεταξύ βάσεων δεδομένων χημειο-πληροφορικής.[44]

2.2 Μοριακό αποτύπωμα

Από τις αρχές της χημειο-πληροφορικής, υπήρχε η «διαμάχη» για το εάν είναι ανώτεροι οι δισδιάστατοι ή τρισδιάστατοι περιγραφείς και μέθοδοι. Η συζήτηση αυτή συνεχίζεται, και εξαρτάται από τις περιπτώσεις των ερευνών, με διαφορετικά συμπεράσματα κάθε φορά. Για τα αποτυπώματα, τα οποία είναι συμβολοσειρές από Bit των μοριακών δομών και ιδιοτήτων ή διαστάσεις των κωδικοποιημένων περιγραφέων έχει, επίσης, μελετηθεί επισταμένα, και σε αυτή την περίπτωση μπορούν να εξαχθούν κάποια σαφή συμπεράσματα: τα δισδιάστατα αποτυπώματα αποτελούν συχνά ισχυρά εργαλεία αναζήτησης, και ακόμα και η απλή αναζήτηση συμβολοσειράς και η αρίθμηση διανυσμάτων ατόμων αναγνωρίζει με επιτυχία τις ενεργές ενώσεις. Δεν προκαλεί λοιπόν έκπληξη πως η αναζήτηση δισδιάστατης ομοιότητας εξακολουθεί να είναι θέμα έρευνας της χημειο-πληροφορικής. Τα δισδιάστατα αποτυπώματα τελευταίας τεχνολογίας περιλαμβάνουν διαδρομές κατακερματισμένης συνδεσιμότητας, και αποτυπώματα δομικά και βασισμένα σε λεξικό και πολυεπίπεδο ατομικό περιβάλλον. Σε πολλές εκδόσεις και για ιστορικούς λόγους, τα daylight αποτυπώματα χρησιμοποιούνται σαν σταθερά για την συγκριτική αξιολόγηση (benchmarking). Επιστημονικά, είναι δύσκολο να δεχθεί κανείς οποιοδήποτε δισδιάστατο αποτύπωμα ως σταθερά για την αναζήτηση ομοιότητας.[45]

Αρχικά, τα δισδιάστατα αποτυπώματα αναπτύχθηκαν για την αναζήτηση ομοιότητας χρησιμοποιώντας απλά πρότυπα μορίων, όμως ανεξάρτητες έρευνες έχουν δείξει ότι η απόδοση της αναζήτησης ενισχύεται όταν χρησιμοποιούνται πολλαπλές ενώσεις αναφοράς. Κατά προτίμηση, όλα τα πρότυπα πρέπει να είναι γνωστά αντιδραστήρια, όμως ακόμα και μόρια τα οποία είναι όμοια σε μια απλή ένωση αναφοράς μπορούν να συμπεριληφθούν σε μια αρχική αναζήτηση ομοιότητας, ανεξάρτητα από την δραστικότητά τους. Αυτή η διαδικασία είναι γνωστή ως ‘turbo’ αναζήτηση ομοιότητας. Πρόσφατες έρευνες για την αύξηση της απόδοσης στην αναζήτηση του αποτυπώματος χρησιμοποιούν πολλαπλές αναφορές σε ενώσεις και είναι περισσότερο επικεντρωμένες σε στρατηγικές είτε με βάση την κλίμακα είτε με βάση το μέσο όρο. Επίσης, πιο συγκεκριμένα, στην αξιολόγηση των συστημάτων εναλλακτικής βαθμολόγησης, χρησιμοποιούνται μέθοδοι όπως του «πλησιέστερου γείτονα» και της συγχώνευσης δεδομένων.[45]

Στη συγχώνευση δεδομένων και στις μεθόδους του πλησιέστερου γείτονα, οι τιμές ομοιότητας καθορίζονται ξεχωριστά για κάθε διαθέσιμη ένωση αναφοράς. Για κάθε βάση δεδομένων των ενώσεων, ο βαθμός ομοιότητας υπολογίζεται είτε ως ο μέσος όρος ομοιότητας έναντι ενός προκαθορισμένου αριθμού των πλησιέστερων γειτόνων στο σετ αναφοράς είτε ως το μέγιστο.[45] Η τελευταία προσέγγιση ορίζεται ως ο k-«πλησιέστερος γείτονας» , k-NN. Ωστόσο, οι μέθοδοι του «πλησιέστερου γείτονα», και συγκεκριμένα η 1-NN, έχουν συχνά το μειονέκτημα ότι η ικανότητά τους να αναγνωρίσουν ενεργές

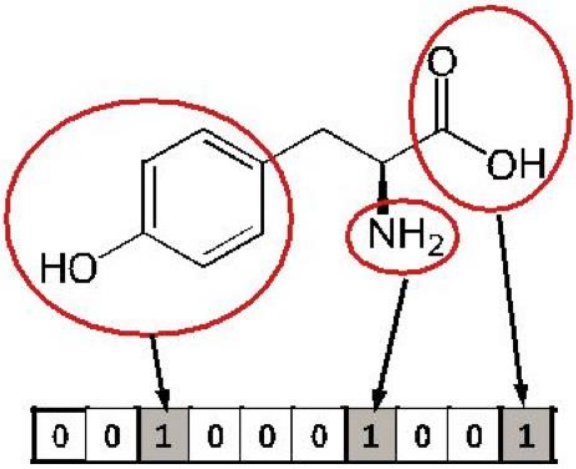
ενώσεις που είναι δομικά διαφορετικές είναι μειωμένη σε σχέση με τις μεθόδους που χρησιμοποιούν συνολικές πληροφορίες από πολλαπλές ενώσεις.[45]

Οι θέσεις αποτυπώματος bit που διαφέρουν σαν ζεύγη των υπό δοκιμή μορίων καθορίζονται ως εισαγωγή για μια συνάρτηση πυρήνα (kernel function) ώστε να εξαχθούν συναρτήσεις πυκνότητας πιθανότατα για ήδη γνωστές ενεργές και ανενεργές ενώσεις. Αυτές οι συναρτήσεις πυκνότητας χρησιμοποιούνται στη συνέχεια για να εκτιμήσουμε την πιθανότητα του εάν ένα μόριο είναι ενεργό, βασιζόμενοι στις ρυθμίσεις αποτυπώματος. Στους υπολογισμούς των στοιχείων αναφοράς (benchmark calculations) BKD υπερτερεί συγκριτικά με άλλες μεθόδους που βασίζονται στα αποτυπώματα.[45]

2.2.1 Τύποι μοριακού αποτυπώματος

Υπάρχουν αρκετοί τύποι μοριακών αποτυπωμάτων που βασίζονται στη μέθοδο με την οποία μια μοριακή αναπαράσταση μετατρέπεται σε συμβολοσειρά bit. Οι περισσότερες μέθοδοι χρησιμοποιούν μόνο το δισδιάστατο μοριακό γράφημα και συνεπώς ονομάζονται δισδιάστατα αποτυπώματα (2D fingerprints). Ωστόσο, μερικές μέθοδοι καταφέρνουν να αποθηκεύσουν τρισδιάστατες πληροφορίες, με πιο γνωστά τα φαρμακοφόρα αποτυπώματα. Οι κύριες προσεγγίσεις είναι υποδομές-κλειδιά, τοπολογικά ή με βάση τη διαδρομή αποτυπώματα (topological or print-based) και κυκλικά αποτυπώματα (circular fingerprints).[46]

Τα αποτυπώματα με βάση τις υποδομές κλειδιά θέτουν τα bits από τις συμβολοσειρές bit με βάση την παρουσία της ένωσης σε συγκεκριμένες υποδομές ή με βάση χαρακτηριστικά από μια δεδομένη λίστα με δομικά κλειδιά. Αυτό συνήθως σημαίνει ότι αυτά τα αποτυπώματα είναι περισσότερο χρήσιμα όταν χρησιμοποιούνται με μόρια που είναι πιθανόν να είναι κυρίως καλυμμένα από τα δεδομένα δομικά κλειδιά, αλλά όχι τόσο όταν τα μόρια είναι απίθανο να περιέχουν τα δομικά κλειδιά, καθώς τα χαρακτηριστικά τους δεν αναπαριστώνται. Ο αριθμός των bits καθορίζεται από των αριθμό των δομικών κλειδιών και κάθε bit σχετίζεται με την παρουσία ή την απουσία ενός απλού δεδομένου χαρακτηριστικού στο μόριο, κάτι το οποίο δεν συμβαίνει σε άλλους (μαρκαρισμένους) τύπους αποτυπωμάτων.[46]



Σχήμα 2.2-1: Αναπαράσταση ενός υποθετικού τοπολογικού αποτυπώματος 10-bit, με τρία bit καθώς οι υποδομές που αντιπροσωπεύουν υπάρχουν στο μόριο (σε κύκλο). (αναπαράχθηκε από *Cereto-Massagué, A. et al.*)[47]

Υπάρχουν δύο συγκρούσεις bit (bit collisions), οι οποίες είναι bits που τίθενται για παραπάνω από ένα θραύσμα, αυτές είναι πιθανά αποτυπώματα με μειωμένο αριθμό bits. Στο σχήμα φαίνονται μόνο τα θραύσματα και τα bits για ένα αρχικό άτομο. Για το πλήρες αποτύπωμα, αυτή η διαδικασία θα μπορούσε να εκτελεστεί για κάθε άτομο στο μόριο. Τα κυκλικά αποτυπώματα ακολουθούν μια παρόμοια προσέγγιση, αλλά με τη δημιουργία αποτυπωμάτων στην ακτίνα του αρχικού ατόμου αντί για γραμμικά θραύσματα.[47]

Δύο δημοφιλή δισδιάστατα αποτυπώματα με βάση τη δομή αποτελούν το αποτύπωμα MACCS (Molecular Access System)[48] και το αποτύπωμα BCI.[49]

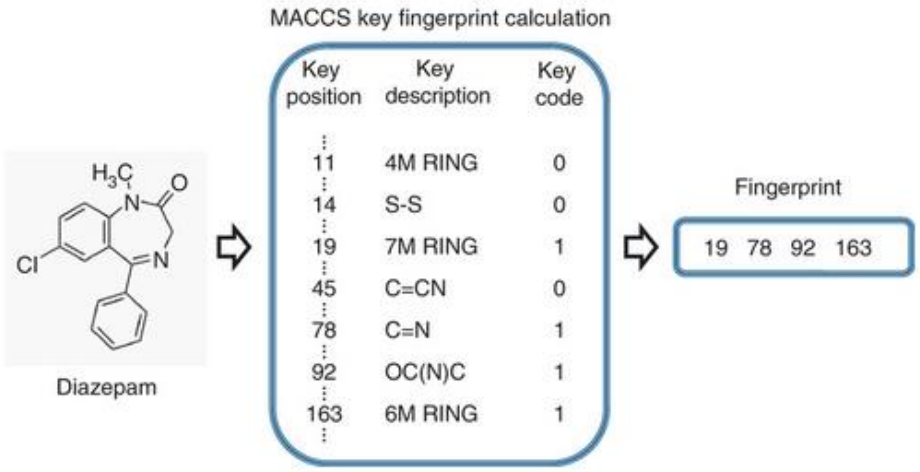
2.2.2 Αποτύπωμα MACCS

Η χρήση δισδιάστατων αποτυπωμάτων για την αναπαράσταση των μοριακών δομών αποτελείται από την παραγωγή ενός bit διανύσματος που κωδικοποιεί την παρουσία ή την απουσία διαφορετικών υποδομών ή φαρμακευτικών χαρακτηριστικών σε κάθε θέση bit. Ένα από τα πιο συνηθισμένα δισδιάστατα αποτυπώματα είναι το αποτύπωμα MAACS (Molecular Access System, το οποίο ήταν από τα πρώτα λογισμικά που διαχειρίστηκε μόρια σε υπολογιστή, και αναπτύχθηκε από την MDL).[50]

Το αποτύπωμα MACCS χρησιμοποιεί ένα σετ δομικών χαρακτηριστικών που χρησιμοποιείται για να κωδικοποιήσει το μόριο σε δυαδική αναπαράσταση. Η ακόλουθη εικόνα δείχνει ένα παράδειγμα κωδικοποίησης του μορίου του diazepam σε

αναπαράσταση MACCS. Η εικόνα δείχνει μόνο ένα μέρος του συνόλου των δομικών χαρακτηριστικών.[50]

Οι key positions(πρώτη στήλη) στην παρακάτω εικόνα αντιστοιχούν σε ένα χημικό χαρακτηριστικό, το οποίο περιγράφεται από το key description (δεύτερη στήλη). Το key code είναι μια δυαδική τιμή η οποία σχετίζεται με την παρουσία ή την απουσία χημικού χαρακτηριστικού: είναι 1 εάν το χημικό χαρακτηριστικό βρίσκεται στο μόριο, διαφορετικά είναι 0. Από τα key codes μπορούμε να πάρουμε το τελικό αποτύπωμα: μόνο τα key positions που κωδικοποιούν ένα θραύσμα που είναι παρόν σε ένα μόριο χρησιμοποιούνται ώστε να μειωθεί η διασπορά του key code vector.[50]



Εικόνα 2.2-1: Η διαδικασία παραγωγής αποτυπώματος MACCS φαίνεται στο παράδειγμα με το μόριο του Diazepam.[50]

2.2.3 Αποτύπωμα εκτεταμένης συνδεσιμότητας

Τα αποτυπώματα εκτεταμένης συνδεσιμότητας (Extended Connectivity Fingerprint, ECFPs) είναι κυκλικά τοπολογικά αποτυπώματα σχεδιασμένα για μοριακό χαρακτηρισμό, αναζήτηση ομοιότητας και μοντελοποίηση δομής-αντίδρασης (structure-activity modeling). Επιπρόσθετα βρίσκονται ανάμεσα στα πιο δημοφιλή εργαλεία αναζήτησης στην ανακάλυψη φαρμάκων και έχουν μεγάλη ποικιλία εφαρμογών.[51]

Η αρχική εφαρμογή των ECFPs ήταν στο πεδίο της ρομποτικής σάρωσης υψηλής απόδοσης (high-throughout screening). Στην αξιολόγηση των αποτελεσμάτων των HTS, τα ECFPs χρησιμοποιούνται ευρέως για την ανάλυση ψευδών θετικών και/ή ψευδών αρνητικών απαντήσεων. Επιπλέον, τα ECFPs συχνά εφαρμόζονται στις μελέτες

εικονικής διαλογής με βάση τις προσδέσεις (ligand-based virtual screening studies) για να διακρίνουν τα ενεργά από τα ανενεργά. Διεξοδικές μελέτες αποκάλυψαν ότι αυτά τα κυκλικά αποτυπώματα είναι τυπικά ανάμεσα στα περισσότερο αποδοτικά εργαλεία αναζήτησης.[52]

Άλλα πεδία φαρμακολογικής έρευνας που σχετίζονται με την αναζήτηση ομοιότητας, περιλαμβάνουν τη χημική ομαδοποίηση (chemical clustering) και την ανάλυση της βιβλιοθήκης των ενώσεων, οι οποίες αξιοποιούν με επιτυχία τις πλούσιες πληροφορίες που κωδικοποιούνται σε αυτά τα αποτυπώματα.[51]

Πέρα από την αναζήτηση ομοιότητας, τα ECFPs είναι κατάλληλα για την αναγνώριση της παρουσίας ή της απουσίας ιδιαίτερων υποδομών. Συνεπώς χρησιμοποιούνται συχνά στην δημιουργία μοντέλων QSAR και QSPR κατά τη διάρκεια της φάσης βελτιστοποίησης, συμπεριλαμβάνοντας την πρόβλεψη των ιδιοτήτων ADMET.[52]

Οι κύριες ιδιότητες των ECFPs είναι οι εξής:[51]

- Αναπαριστούν μοριακές δομές μέσω των γειτνιαζόντων κυκλικών ατόμων.
- Μπορούν να υπολογιστούν πολύ γρήγορα.
- Τα χαρακτηριστικά τους αναπαριστούν την παρουσία ορισμένων υποδομών.
- Δεν είναι προκαθορισμένα και μπορούν να αναπαριστούν ένα μεγάλο αριθμό διαφορετικών μοριακών χαρακτηριστικών (συμπεριλαμβανομένων και των στερεοχημικών πληροφοριών).
- Είναι σχεδιασμένα για να αναπαριστούν και την παρουσία και την απουσία της λειτουργικότητας, αφού και οι δυο είναι σημαντικές για την ανάλυση της μοριακής δραστηριότητας.
- Η μέθοδος παραγωγής τους μπορεί να οριστεί με ευελιξία ώστε να παράγει ποικίλους τύπους κυκλικών αποτυπωμάτων για διάφορες εφαρμογές.[52]

Οι τρεις κύριες παράμετροι των ECFPs είναι η μέγιστη διάμετρος, το μήκος του αποτυπώματος και οι καταμετρήσεις του αναγνωριστικού:[52]

1. Η Διάμετρος (Diameter)

Η παράμετρος αυτή καθορίζει τη μέγιστη διάμετρο των κυκλικών γειτνιαζόντων που λαμβάνεται υπόψη για κάθε άτομο. Η προεπιλεγμένη διάμετρος είναι 4. Αυτή είναι μια κύρια παράμετρος των ECFPs, η οποία ελέγχει τον αριθμό και το μέγιστο μέγεθος των θεωρούμενων γειτνιαζόντων ατόμων, και συνεπώς ελέγχει το μήκος της λίστας αναπαράστασης των αναγνωριστικών, όπως επίσης και τον αριθμό των bits “1” στο καθορισμένο μήκος της αναπαράστασης της συμβολοσειράς.

Τα ECFPs ξεχωρίζουν συνήθως από αυτή την παράμετρο. Για παράδειγμα, το ECFP_4 σημαίνει ότι η μέγιστη διάμετρος τίθεται στο 4, ECFP_6 σημαίνει διάμετρο 6.

Η κατάλληλη τιμή για τη μέγιστη διάμετρο εξαρτάται από την επιθυμητή εφαρμογή. Σύμφωνα με τους *Rogers and Hahn,*[53] η διάμετρος 4 είναι συνήθως επαρκής για αναζήτηση ομοιότητας (similarity searching) και ομαδοποίηση (clustering), ενώ οι μέθοδοι εκμάθησης της δραστηριότητας συχνά ωφελούνται από την μεγαλύτερη δομική λεπτομέρεια χρησιμοποιώντας μεγαλύτερο όριο, για παράδειγμα 6 ή 8.

2. Το Μήκος

Αυτή η παράμετρος καθορίζει το μήκος της αναπαράστασης συμβολοσειράς bit. Το προεπιλεγμένο μήκος είναι 1024. Μεγαλύτερο μήκος μειώνει την πιθανότητα της σύγκρουσης bit και επομένως μειώνει την απώλεια πληροφοριών. Ωστόσο, η διαχείριση μεγαλύτερων αποτυπωμάτων απαιτεί περισσότερο υπολογιστικό χρόνο και χωρητικότητα μνήμης.

3. Οι Καταμετρήσεις (Counts)

Αυτή η παράμετρος ελέγχει το εάν τα παραγόμενα ακέραια αναγνωριστικά (generated integer identifiers) είναι αποθηκευμένα με βάση την αρίθμηση των εμφανίσεών τους ή κάθε αναγνωριστικό συγκρατείται μόνο μια φορά ανεξάρτητα από τον αριθμό των αντίστοιχων υποδομικών χαρακτηριστικών στο εισαγόμενο μόριο (input molecule)**.** Η προεπιλεγμένη απάντηση είναι ‘όχι’, που σημαίνει ότι το κάθε αναγνωριστικό αποθηκεύεται μόνο μια φορά.

Οι πρώτες δύο παράμετροι διαδραματίζουν παρόμοιο ρόλο στο μέγιστο μήκος του σχήματος και στις παραμέτρους του μήκους του αποτυπώματος, ενώ έχουν παρόμοια αποτελέσματα στο περιεχόμενο των πληροφοριών, το χρόνο παραγωγής και στον απαιτούμενο αποθηκευτικό χώρο των αποτυπωμάτων.[52]

2.2.4 Molprint 2D

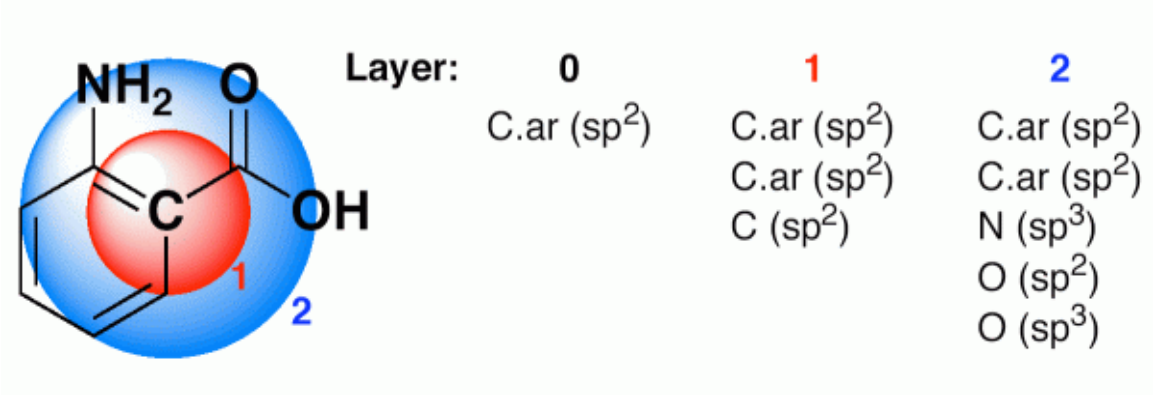
MolPrint (γνωστά και ως MolPrint 2D) περιγραφείς (descriptors)[54] [55] είναι ένα συγκεκριμένο είδος κυκλικών δακτυλικών αποτυπωμάτων που χρησιμοποιούν τους ατομικούς τύπους Sybyl MOL2. Πιο συγκεκριμένα, βασίζονται σε μετρήσεις των MOL2 ατομικών τύπων γύρω από κάθε βαρύ άτομο του μορίου. Σε αντίθεση με τα διαρθρωτικά (structural) κλειδιά δεν αντλούν στοιχεία από ένα περιορισμένο σύνολο των διαρθρωτικών θραυσμάτων (όπως τα κλειδιά MACCS). Αντίθετα απαριθμούν όλα τα περιβάλλοντα άτομα που είναι παρόντα σε ένα μόριο. Οι MolPrint 2D περιγραφείς είναι

παρόμοιοι με τα SciTegic's (Pipeline Pilot) εκτεταμένης-συνδεσιμότητας αποτυπώματα (ECFP), αλλά τα χαρακτηριστικά των Molprint 2D δεν κατακερματίζονται.[51] Η εφαρμογή των MolPrint 2D που χρησιμοποιούνται σε OCHEM χρησιμοποιεί τους ατομικούς τύπους κυριολεκτικά όπως εμφανίζονται σε ένα αρχείο MOL2, δηλαδή, ένας αρωματικός άνθρακας κωδικοποιείται ως "C.ar" , ένα sp2-υβριδοποιημένο άτομο οξυγόνου ως "O.2", κ.λπ.

Για κάθε βαρύ άτομο όλα τα γειτονικά άτομα σε ένα δεδομένο αριθμό δεσμών συμπίπτουν και κωδικοποιούνται ως μία συμβολοσειρά (string). Μια τέτοια συμβολοσειρά ξεκινά πάντα με το βαρύ άτομο C στο κέντρο του χαρακτηριστικού, που ακολουθείται από τριάδες της μορφής D-T-N, όπου D είναι η απόσταση σε δεσμούς από το κεντρικό άτομο (D σε {1, 2, ...}), T ο τύπος του ατόμου (T είναι ένας έγκυρος Sybyl MOL2 ατομικός τύπος), και N είναι ο αριθμός των τομικών τύπων T που μπορούν να βρεθούν σε απόσταση D από το κεντρικό άτομο C. Το κεντρικό άτομο και όλες οι τριάδες διαχωρίζονται με ερωτηματικά. Συνολικά, έχουν ως αποτέλεσμα μία συμβολοσειρά της μορφής: C;D-T-N;D-T-N;D-T-N;… Στην πράξη, διαπιστώθηκε ότι οι τιμές για το D μέχρι 3 θα πρέπει να εξετάζονται για την παραγωγή περιγραφέων, με D=2 το πλέον κοινώς χρησιμοποιούμενο. Όσο υψηλότερη είναι η τιμή του D, τόσο πιο συγκεκριμένες γίνονται οι λειτουργίες από τη φύση της κατασκευής τους.[56]

Η δυαδική φύση των περιγραφέων καθιστά τους MolPrint περιγραφείς ως πιο δεκτικούς σε ορισμένους τύπους μεθόδων μοντελοποίησης (όπως Bayes ή μεθόδους k-NN), περισσότερο από ότι για παράδειγμα σε μοντέλα νευρωνικών δικτύων. Τα παραγώμενα μοντέλα είναι σχετικά εύκολο να ερμηνευτούν, αφού το κάθε χαρακτηριστικό αντιστοιχεί σε περίπου μία λειτουργική ομάδα (αν και χωρίς σαφείς πληροφορίες σχετικά με τη σειρά των δεσμών μεταξύ των ατόμων).[56]

Οι MolPrint περιγραφείς έχουν χρησιμοποιηθεί με επιτυχία σε εικονική διαλογή (virtual screening)[57] και πρόβλεψη προσδέματος-στόχου[18] όπου έχει δειχθεί ότι μπορούν να συλλάβουν ένα μεγάλο ποσό των πληροφοριών που σχετίζουν τη μοριακή δομή με τη βιοδραστικότητα έναντι ενός στόχου πρωτεΐνης.



Εικόνα 2.2-2: Ο περιγραφέας MolPrint 2D βασίζεται σε απαριθμήσεις ατομικών τύπων γύρω από κάθε βαρύ άτομο του μορίου. Ο περιγραφέας που χρησιμοποιείται εδώ χρησιμοποιεί ατομικούς τύπους mol2. [58–60]

Στην εικόνα 2.2-2 παρουσιάζεται ένα παράδειγμα μοριακής απεικόνισης χρησιμοποιώντας έναν περιγραφέα MolPrint 2D. Στο **επίπεδο 0** απεικονίζεται ο αρωματικός άνθρακας (τροχιάς sp^2). Στο **επίπεδο 1** απεικονίζονται δύο αρωματικοί άνθρακες (τροχιάς sp^2) και ένας αλειφατικός άνθρακας (τροχιάς sp^2). Τέλος, στο **επίπεδο 2** απεικονίζονται δύο αρωματικοί άνθρακες (τροχιάς sp^2), ένα άτομο αζώτου (τροχιάς sp^3), ένα άτομο οξυγόνου (τροχιάς sp^2) και ένα άτομο οξυγόνου (τροχιάς sp^3).[57]

2.3 Αλγόριθμοι Μηχανικής Μάθησης

Η συνάφεια του πρωτεϊνικού προσδέματος αποτελεί τον καθοριστικό παράγοντας για αρκετές ζωτικές διαδικασίες, όπως την κυτταρική σήμανση, τη γονιδιακή ρύθμιση, το μεταβολισμό και την ανοσία, η οποία εξαρτάται από την δέσμευση σε κάποιο μόριο υποστρώματος. Η ακριβής πρόβλεψη των σχέσεων συνάφειας σε μεγάλα σετ συμπλεγμάτων πρωτεϊνικών προσδεμάτων παραμένει ένα από τα πιο δύσκολα και άλυτα προβλήματα της υπολογιστικής βιομοριακής επιστήμης, με εφαρμογές στην ανακάλυψη νέων φαρμάκων, την χημική και δομική βιολογία.

Η υπολογιστική μοριακή πρόσδεση (computational molecular docking) περιλαμβάνει τις πόζες πρόσδεσης δεκάδων χιλιάδων έως εκατομμυρίων υποψήφιων προσδεμάτων σε μια τοποθεσία πρόσδεσης υποδοχέα πρωτεΐνης στόχου ενώ χρησιμοποιώντας μια κατάλληλη συνάρτηση βαθμολόγησης, αξιολογεί την συνάφεια κάθε υποψήφιου προσδέματος ώστε να ταυτοποιήσει τους κορυφαίους υποψηφίους αναστολείς πρωτεϊνών.[61]

Επομένως, χρησιμοποιείται μια συνάρτηση βαθμολόγησης ώστε να βαθμολογήσει, να κατατάξει και να αναγνωρίσει οδηγούς φαρμάκων, καθώς η πιστότητα με την οποία προβλέπει τη συνάφεια ενός υποψήφιου προσδέματος για μια τοποθεσία πρωτεϊνικής πρόσδεσης και η υπολογιστική της ικανότητα διαδραματίζουν σημαντικό ρόλο στην ακρίβεια της αποδοτικότητας της εικονικής διαλογής. Παρά τις επισταμένες προσπάθειες

σε αυτό το πεδίο, μέχρι στιγμής δεν υπάρχει μια γενικά αποδεκτή συνάρτηση βαθμολόγησης που να υπερτερεί έναντι των υπολοίπων. Επομένως, στη συνέχεια εξετάζεται μια ποικιλία νέων συναρτήσεων βαθμολόγησης, χρησιμοποιώντας διαφορετικές προσεγγίσεις του machine learning (ML) σε συνδυασμό με ένα σύνολο διαφορετικών χαρακτηριστικών των συμπλεγμάτων των πρωτεϊνικών προσδεμάτων με σκοπό την σημαντική βελτίωση της ακρίβειας της συνάρτησης βαθμολόγησης σε σύγκριση με τις υπάρχουσες συμβατικές συναρτήσεις.[61]

2.3.1 Ταξινομητής Naïve Bayes (Naïve Bayes Classifier)

Σε πολλές εφαρμογές η σχέση μεταξύ του συνόλου των χαρακτηριστικών γνωρισμάτων και της μεταβλητής της κατηγορίας είναι μη-ντετερμινιστική. Με άλλα λόγια, η ετικέτα κατηγορίας μιας εγγραφής του test set δεν μπορεί να προβλεφθεί με απόλυτη βεβαιότητα, ακόμα κι αν τα χαρακτηριστικά της γνωρίσματα είναι ίδια με μερικά μιας έγγραφης του training set. Η κατάσταση μπορεί να χειροτερέψει εξαετίας δεδομένων που έχουν υποστεί κάποιο θόρυβο και μπορεί να έχουν παραμορφωθεί ή σε περίπτωση ύπαρξης συντελεστών που επηρεάζουν την ταξινόμηση και δεν έχουν ληφθεί υπόψη στην ανάλυση.

Οι Bayesian Classifiers, προσεγγίζουν ως τρόπο επίλυσης των προβλημάτων ταξινόμησης βασισμένο στις πιθανολογικές σχέσεις μεταξύ του συνόλου των χαρακτηριστικών γνωρισμάτων και της μεταβλητής της κλάσης – κατηγορίας. Χαρακτηριστικά γνωρίσματα των Bayesians μεθόδων εκμάθησης:

- Κάθε παρατηρούμενο παράδειγμα εκμάθησης μπορεί να αυξήσει ή να μειώσει δραματικά την εκτιμώμενη πιθανότητα ότι μια υπόθεση – πρόβλεψη είναι σωστή.
- Η προγενέστερη γνώση μπορεί να συνδυαστεί με τα παρατηρηθέντα στοιχεία για να καθορίσει την τελική πιθανότητα μιας υπόθεσης.
- Οι Μπεϋσιανοί μέθοδοι μπορούν να προσαρμόσουν τις υποθέσεις που κάνουν τις πιθανολογικές προβλέψεις.
- Οι νέες περιπτώσεις μπορούν να ταξινομηθούν με το συνδυασμό των προβλέψεων των πολλαπλών υποθέσεων, που καθορίζονται από τις πιθανότητές τους.
- Μπορούν να παρέχουν πρότυπα για την λήψης της βέλτιστης απόφασης κατά την οποία μπορούν να μετρηθούν και άλλες πρακτικές μέθοδοι.

Ένας βασικός ταξινομητής Naïve Bayes υπολογίζει την class-conditional πιθανότητα θεωρώντας ότι τα χαρακτηριστικά γνωρίσματα είναι υπό όρους ανεξάρτητα μεταξύ τους, δεδομένης της ετικέτας κατηγορίας y. Η υπό όρους υποθετική ανεξαρτησία μπορεί να δηλωθεί τυπικά ως εξής:[62]

$$P(X \mid Y = y) = \prod_{i=1}^{d} P(X_i \mid Y = y)$$

όπου κάθε σύνολο χαρακτηριστικών γνωρισμάτων $X = \{X_1, X_2, ..., X_d\}$, έχει d χαρακτηριστικά γνωρίσματα για κάθε εγγραφή.

Με την υπό όρους υποθετική ανεξαρτησία, αντί να υπολογίζουμε την classconditional πιθανότητα για κάθε συνδυασμό του X, υπολογίζουμε μόνο την πιθανότητα του κάθε Xi, δοσμένου του Y. Η τελευταία προσέγγιση – μέθοδος είναι πιο πρακτική γιατί δεν απαιτεί ένα πολύ μεγάλο training set για να πετύχει μια καλή εκτίμηση της πιθανότητας.[62]

Για να ταξινομήσουμε μία δοκιμαστική εγγραφή, ο Naïve Bayes ταξινομητής υπολογίζει τη μεταγενέστερη πιθανότητα για κάθε κατηγορία Y:

$$P(X \mid Y = y) = \frac{P(Y)\prod_{i=1}^{d} P(X_i \mid Y)}{P(X)}$$

Μιας και το P(X) είναι προσαρμοσμένο για κάθε Y, το βέλτιστο είναι να επιλεχθεί η κατηγορία που μεγιστοποιεί τον αριθμητή $P(Y)\prod_{i=1}^{d} P(X_i \mid Y)$.

2.3.2 Δέντρα Απόφασης

Τα δέντρα απόφασης (Decision trees) αποτελούν αντικείμενο διεξοδικής μελέτης καθώς αποτελούν εργαλεία στην ανακάλυψη γνώσης και στα συστήματα στήριξης αποφάσεων.

Η λειτουργία ενός δένδρου απόφασης είναι σχετικά απλή, καθώς ένα πρόβλημα ταξινόμησης μπορεί να επιλυθεί κάνοντας μια σειρά από σωστά δομημένες ερωτήσεις σχετικά με το κάθε ένα από τα χαρακτηριστικά γνωρίσματα κάθε εγγραφής του test set. Κάθε φορά που λαμβάνεται μια απάντηση, αυτόματα μία νέα καθορισμένη ερώτηση πραγματοποιείται, η οποία μόλις απαντηθεί άλλη μία καθορισμένη επακόλουθη έως ότου εξαχθεί το επιθυμητό συμπέρασμα. Η σειρά των ερωτήσεων και οι δυνατές απαντήσεις τους μπορούν να οργανωθούν υπό την μορφή ενός δένδρου απόφασης, το οποίο απεικονίζει μια ιεραρχική δομή που αποτελείται από κόμβους και κατευθυνόμενες ακμές.[63]

Στα δέντρα αποφάσεων, τα φύλλα αναπαριστούν τις ταξινομήσεις και τα κλαδιά αναπαριστούν τους διαχωρισμούς με βάση τα χαρακτηριστικά που οδηγούν σε ταξινομήσεις. Αυτά τα δέντρα προσεγγίζουν τις συναρτήσεις στόχων διακριτών τιμών ενώ αποτελούν μια πρακτική μέθοδος που χρησιμοποιείται ευρέως στα επαγωγικά συμπεράσματα.[63]

Τα δέντρα αποφάσεων έχουν προσφέρει πολλά στην ανακάλυψη γνώσης και στα συστήματα στήριξης αποφάσεων εξαιτίας του φυσικού και κατανοητού τρόπου να ταξινομούν με σχηματική διαδικασία μέσω μιας αλληλουχίας ερωτήσεων. Οι αλγόριθμοι για την κατασκευή των δέντρων αποφάσεων χρησιμοποιούν συχνά ευρετική μέθοδο, η οποία καταλήγει σε σύντομα δέντρα. Η εύρεση του συντομότερου δέντρου αποφάσεων αποτελεί ένα δύσκολο πρόβλημα βελτιστοποίησης.[63]

2.3.3 Επαγωγικοί κανόνες

Ο επαγωγικός κανόνας αποτελεί μια από τις πιο σημαντικές τεχνικές στην μηχανική εκμάθηση. Εφόσον οι κανονικότητες που βρίσκονται στα δεδομένα συχνά εκφράζονται σύμφωνα με κανόνες, ο επαγωγικός κανόνας αποτελεί θεμελιώδες εργαλείο στην εξόρυξη δεδομένων. Συνήθως οι κανόνες είναι εκφράσεις με την παρακάτω μορφή:[64]

$$if\,(attribute-1, value-1)\ and\ (attribute-2, value-2)\ and \ldots and \\ (attribute-n, value-n)\ then\ (decision, value)$$

Ορισμένα συστήματα επαγωγής κανόνων οδηγούν σε περισσότερους σύνθετους κανόνες, στους οποίους οι τιμές των ιδιοτήτων μπορούν να εκφραστούν με άρνηση κάποιων τιμών ή από μια υποομάδα τιμών στον τομέα των ιδιοτήτων.[64]

Τα δεδομένα από τα οποία επάγονται οι κανόνες παρουσιάζονται συνήθως σε μορφή παρόμοια με πίνακα όπου οι περιπτώσεις (ή παραδείγματα) είναι ετικέτες (ή ονόματα) για τις σειρές και οι μεταβλητές επισημαίνονται ως ιδιότητες και απόφαση.[64]

2.3.4 Τυχαία Δάση

Τα τυχαία δάση (Random Forests) αποτελούν μία εξελιγμένη τεχνική μηχανικής μάθησης ταξινόμησης, η οποία είναι αλληλένδετη με την μέθοδο των Δέντρων Απόφασης. Τα τυχαία δάση είναι ένα σύνολο από ταξινομητές οι οποίοι δημιουργούν παράλληλα πολλά δέντρα αποφάσεων οπού κάθε κόμβος του δέντρου είναι ένα τυχαίο υποσύνολο των εξεταζόμενων χαρακτηριστικών.

Συγκεκριμένα, τα RFs είναι ένας συνδυασμός από δέντρα απόφασης, τέτοια ώστε κάθε δέντρο να εξαρτάται από τις τιμές που λαμβάνονται τυχαία από ένα πίνακα και έχουν την ίδια κατανομή για κάθε δέντρο. Η ιδέα των αυξανόμενων συνόλων δέντρων και της απόφασης της κατηγοριοποίησης βάση ψηφίσματος έχει βελτιώσει την ακρίβεια της κατηγοριοποίησης. Το σφάλμα γενίκευσης των τυχαίων δασών αυξάνεται και φτάνει ένα όριο καθώς το πλήθος των δέντρων στο δάσος μεγαλώνει.[65]

Η εφαρμογή των τυχαίων δασών στην πρόβλεψη των χρονοσειρών παραμένει αρκετά περιορισμένη, παρά την υψηλή τους απόδοση στην κατηγοριοποίηση και την ικανότητα τους για γενίκευση σε δεδομένα το οποία δεν έχουν χρησιμοποιηθεί κατά την φάση εκπαίδευσης τους. Δυστυχώς η υψηλή τους ικανότητα γενίκευσης και ακρίβειας συνοδεύεται από το τίμημα της περιορισμένης ερμηνευσιμότητας. Επιπλέον η απόδοση τους εξαρτάται κατά πολύ από την κατάλληλη ρύθμιση των παραμέτρων τους και την επιλογή ενός καλού υποσυνόλου χαρακτηριστικών που θα χρησιμοποιηθεί ως είσοδος.

2.3.5 Γραμμική Παλινδρόμηση

Η ανάλυση παλινδρόμησης (regression analysis) είναι μια καθοδηγούμενη τεχνική η οποία γενικεύει ένα σύνολο αριθμητικών δεδομένων δημιουργώντας ένα μαθηματικό μοντέλο που συσχετίζει ένα ή περισσότερα χαρακτηριστικά εισόδου με ένα χαρακτηριστικό εξόδου. Με τη γραμμική παλινδρόμηση (linear regression), πραγματοποιείται μοντελοποίηση των μεταβολών μιας εξαρτημένης μεταβλητής ως γραμμικό συνδυασμό ενός ή περισσοτέρων ανεξάρτητων μεταβλητών. Η εξίσωση της γραμμικής παλινδρόμησης έχει τη μορφή:

$$Y = f(X_1, X_2, \ldots X_k) + \varepsilon = \beta_o + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

Όπου X_1, X_2, … X_k είναι ανεξάρτητες μεταβλητές και $\beta_o, \beta_1 \ldots \beta_k$ είναι σταθερές και ε είναι το σφάλμα. Η συνάρτηση $f(X_1, X_2, \ldots X_k)$ αντιπροσωπεύει την εξαρτημένη μεταβλητή και συχνά γράφεται απλώς ως Y. Γενικά, η γραμμική παλινδρόμηση ενδείκνυται όταν η σχέση μεταξύ των εξαρτημένων και των ανεξαρτήτων μεταβλητών είναι σχεδόν γραμμική.[66]

2.3.6 Μη γραμμική Παλινδρόμηση

Η μη γραμμική παλινδρόμηση (Nonlinear regression) είναι μια μέθοδος εύρεσης ενός μη γραμμικού μοντέλου της σχέσης μεταξύ εξαρτημένης μεταβλητής και ενός σετ ανεξάρτητων μεταβλητών. Σε αντίθεση με την παραδοσιακή γραμμική παλινδρόμηση, η

οποία περιορίζεται στην εκτίμηση γραμμικών μοντέλων, η μη γραμμική παλινδρόμηση μπορεί να εκτιμήσει μοντέλα με αυθαίρετες σχέσεις μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών. Αυτό επιτυγχάνεται χρησιμοποιώντας αλγορίθμους επαναληπτικής εκτίμησης (iterative estimation algorithms).[67]

2.3.7 Μηχανές Διανυσμάτων Υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines, SVM) είναι μια μέθοδος που ανήκει στην κατηγορία των μοντέλων της επιβλεπόμενης μάθησης και σήμερα θεωρείται καινοτόμος στο πεδίο της μηχανικής μάθησης. Οι μηχανές διανυσμάτων υποστήριξης προσπαθούν να εντοπίσουν το βέλτιστο υπερεπίπεδο το οποίο μπορεί να ταξινομήσει τα σημεία σε κατηγορίες και το οποίο παρουσιάζει τη μεγαλύτερη απόσταση μεταξύ των σημείων των διαφορετικών κατηγοριών. Είναι κατάλληλοι, όταν ο χώρος αναζήτησης είναι πολύπλοκος ή πολύ μεγάλος.

Οι μηχανές διανυσμάτων υποστήριξης ανήκουν στο σύνολο των αλγόριθμων με επιβλεπόμενη μάθηση που μπορούν να χρησιμοποιηθούν για κατηγοριοποίηση ή μη-γραμμική παλινδρόμηση. Τα SVMs αντιπροσωπεύουν μια επέκταση των μη γραμμικών μοντέλων του γενικού αλγορίθμου που αναπτύχτηκε από τον Vapnik.[68]Αναπτυχτήκαν σε ένα πολύ ενεργό ερευνητικό τομέα και έχουν ως τώρα εφαρμοστεί σε πολλά επιστημονικά προβλήματα. Επιπλέον τα SVMs δεν είναι περιορισμένα στην ακτινωτή συνάρτηση βάσης, αλλά μπορούν να έχουν μια επιλογή από μια ποικιλία συναρτήσεων πυρήνα.

Αρχικά, τα SVMs μοντέλα προορίζονταν για την ταξινόμηση γραμμικών διαχωριζόμενων αντικειμένων. Για οποιαδήποτε γραμμικώς διαχωριζόμενα σύνολα από δύο κλάσεις τα SVMs είναι ικανά να βρουν τα βέλτιστα υπερεπίπεδα τα οποία τις διαχωρίζουν και ταυτόχρονα έχουν την μεγαλύτερη απόσταση μεταξύ των δυο υπερεπιπέδων.

Τα SVMs έχουν επίσης την δυνατότητα να διαχωρίσουν αντικείμενα τα οποία είναι μη γραμμικώς διαχωρίσιμα. Σε αυτή την περίπτωση, οι συντεταγμένες των αντικειμένων αντιστοιχίζονται στο χώρο, χρησιμοποιώντας μη γραμμικές συναρτήσεις. Ο χώρος στον οποίο κάθε αντικείμενο αντιστοιχίζεται είναι χώρος υψηλού αριθμού διαστάσεων στον οποίο δύο κλάσεις μπορούν να διαχωριστούν με ένα γραμμικό διαχωριστή.

Τα SVMs είναι κυρίως τεχνικές κατηγοριοποίησης οι οποίες ωστόσο έχουν επεκταθεί με σκοπό την εφαρμογή τους σε προβλήματα μη-γραμμικής παλινδρόμησης, όπως πρόβλεψη χρηματοοικονομικών χρονοσειρών. Αυτό έγινε εφικτό από την εισαγωγή της

ε-ευαίσθητης συνάρτησης απωλειών από τον Vapnik, η οποία καθιέρωσε τα διανύσματα μη-γραμμικής παλινδρόμησης (Support Vector Regression).

Τα SVMs και τα SVRs έχουν ήδη εφαρμοστεί σε πολλές εφαρμογές πρόβλεψης και κατηγοριοποίησης. Παρά το ισχυρό θεωρητικό τους υποβάθρου και των υποσχόμενων πειραματικών τους αποτελεσμάτων δεν έχουν ωστόσο καταφέρει να ξεπεράσουν τα εμπόδια των διαστάσεων και της ρύθμισης των παραμέτρων τους. Ακόμη σε μερικές περιπτώσεις απλά νευρωνικά δίκτυα τα έχουν ξεπεράσει σε απόδοση.[69]

2.4 Μετρήσεις αξιολόγησης της επίδοσης

Η αξιολόγηση της επίδοσης στα συστήματα που βασίζονται στην μηχανική εκμάθηση πραγματοποιούνται περισσότερο πειραματικά παρά αναλυτικά. Για να αξιολογηθεί αναλυτικά ένα τυπικό μοντέλο προδιαγραφών για το πρόβλημα είναι απαραίτητο να είναι διαθέσιμο το ίδιο το σύστημα. Αυτό είναι αρκετά δύσκολο και εγγενώς μη εφαρμόσιμο για τις μηχανές, οι οποίες είναι μη γραμμικές και διαφέρουν χρονικά.

Η πειραματική αξιολόγηση ενός μοντέλου που βασίζεται στη μηχανική εκμάθηση εκτελείται σύμφωνα με μετρήσεις αξιολόγησης της επίδοσης όπως είναι οι Balanced Accuracy, F-Score, Precision/Recall, R^2. Αφού λοιπόν υπάρχουν αρκετές μέθοδοι μέτρησης που μπορούν να χρησιμοποιηθούν για αξιολόγηση, είναι εξαιρετικά δύσκολο να συγκριθούν τα τρέχοντα αποτελέσματα έρευνας με προηγούμενες εργασίες εκτός και αν το προηγούμενο πείραμα έγινε από έναν ερευνητή κάτω από τις ίδιες συνθήκες. Η εύρεση μιας κοινής μέτρησης επίδοσης μπορεί να απλοποιήσει αυτή τη σύγκριση, αλλά ακόμα οι ερευνητές δεν έχουν συμφωνήσει σε κάτι τέτοιο. Πειραματικές μελέτες έχουν δείξει ότι μόνο ένα μικρό ποσοστό δομοστοιχείων λογισμικού (software modules) προκαλεί σφάλματα σε συστήματα λογισμικού. Επομένως, η πλειοψηφία των δομοστοιχείων λογισμικού αναπαρίσταται με μη εσφαλμένες επισημάνσεις (labels) και οι υπόλοιπες μαρκάρονται με ετικέτες σφάλματος κατά τη διαδικασία της φάσης μοντελοποίησης. Αυτά τα είδη σετ δεδομένων ονομάζονται imbalanced / unbalanced / skewed και υπάρχουν διαφορετικές μέθοδοι μέτρησης για να εκτιμηθεί η επίδοση των τεχνικών πρόβλεψης σφάλματος σε μη ισορροπημένα σετ δεδομένων.[70]

Σύμφωνα με πειραματικές μελέτες, η πλειοψηφία των δομοστοιχείων λογισμικού δεν προκαλεί σφάλματα σε συστήματα λογισμικού, καθώς τα δομοστοιχεία με σφάλματα αποτελούν το 20% του συνόλου των δομοστοιχείων. Εάν χωριστούν τα δομοστοιχεία σε δύο διαφορετικούς τύπους, με σφάλματα και χωρίς, η πλειοψηφία τους θα ανήκει στην τάξη αυτών που δεν περιέχουν σφάλματα και τα υπόλοιπα θα ανήκουν στην τάξη με τα σφάλματα. Επομένως, τα σετ δεδομένων που χρησιμοποιούνται στα λογισμικά πρόβλεψης σφάλματος είναι μη-ισορροπημένα. Η παράμετρος της ακρίβειας δεν μπορεί να χρησιμοποιηθεί για την αξιολόγηση της απόδοσης των μη-ισορροπημένων σετ

δεδομένων. Για παράδειγμα, ένας συνηθισμένος αλγόριθμος ο οποίος μαρκάρει κάθε δομοστοιχείο ως στοιχείο χωρίς σφάλμα, μπορεί να έχει 90% ακρίβεια εάν τα δομοστοιχεία με σφάλματα είναι 10%. Επομένως, οι ερευνητές χρησιμοποιούν διαφορετικές μεθόδους μέτρησης για την επικύρωση των μοντέλων πρόβλεψης των σφαλμάτων στα λογισμικά.[70]

2.4.1 Accuracy, Balanced Accuracy και R^2

Η επαλήθευση του μοντέλου για τους αλγόριθμους μηχανικής εκμάθησης θα πρέπει να διασφαλίζει ότι τα δεδομένα μετασχηματίστηκαν σωστά σύμφωνα με το μοντέλο, ενώ το μοντέλο αναπαριστά το σύστημα με μια αποδεκτή ακρίβεια. Υπάρχουν διάφορες τεχνικές επαλήθευσης από τις οποίες η πιο γνωστή είναι η τεχνική N-fold cross-validation, όπου διαιρεί τα δεδομένα σε N μέρη, από τα οποία καθένα αποτελείται από ίσο αριθμό μερών από τα αρχικά δεδομένα. Για κάθε μέρος, πραγματοποιούνται (N-1) αριθμοί δοκιμών, ενώ το τεστ επαναλαμβάνεται M φορές με τυχαία σειρά κάθε φορά καθώς συγκεκριμένες σειρές δύναται να βελτιώσουν ή να υποβαθμίσουν το αποτέλεσμα αισθητά.

Πίνακας 2.4-1: Πίνακας ενδεχομένων[70]

	NO (Prediction)	**YES (Prediction)**
NO (Actual)	True Negative (TN) A	False Positive (FP) B
YES (Actual)	False Negative (FN) C	True Positive (TP) D

Οι στήλες αντιπροσωπεύουν τα πιθανά αποτελέσματα και οι γραμμές την πραγματική κατηγορία των ετικετών. Τα εσφαλμένα δομοστοιχεία παρουσιάζονται με την ετικέτα YES και τα μη εσφαλμένα δομοστοιχεία με την ετικέτα NO. Γι αυτό τα στοιχεία της διαγωνίου (TN, TP) στον παραπάνω πίνακα αντιπροσωπεύουν τις σωστές προβλέψεις ενώ τα υπόλοιπα στοιχεία (FN, FP) αντιπροσωπεύουν τις εσφαλμένες προβλέψεις. Για

παράδειγμα, εάν ένα δομοστοιχείο έχει προβλεφθεί ως εσφαλμένο (YES) παρόλο που είναι ένα μη εσφαλμένο δομοστοιχείο (NO), τοποθετείται στο B κελί του πίνακα και ο αριθμός των δειγμάτων στο κελί B αυξάνεται κατά 1. Έπειτα από M*N δοκιμές υπολογίζονται οι τιμές A, B, C και D, από τις οποίες υπολογίζεται η μέτρηση αξιολόγησης της επίδοσης.[70]

Ως ακρίβεια (accuracy) ορίζεται η αναλογία όλων των προβλέψεων που ήταν σωστές:

$$Accuracy = \frac{\#\sigma\omega\sigma\tau\acute{\omega}\nu\ \pi\rho o\beta\lambda\acute{\epsilon}\psi\epsilon\omega\nu}{\#\pi\rho o\beta\lambda\acute{\epsilon}\psi\epsilon\omega\nu}$$

Για ένα πρόβλημα ταξινόμησης δύο κλάσεων, η ποσότητα εκφράζεται με την βοήθεια του πλήθους των σωστών και εσφαλμένων ταξινομήσεων ενός συστήματος ως εξής:

$$Accuracy = \frac{A + D}{(A + B + C + D)}$$

Η Ισορροπημένη Ακρίβεια (Balanced Accuracy) υπολογίζεται ως ο μέσος όρος της αναλογίας των αληθών απαντήσεων κάθε κατηγορίας ξεχωριστά.[71]

$$Balanced\ Accuracy = \frac{\frac{A}{(A + B)} + \frac{D}{(C + D)}}{2}$$

Τα δύο μεγέθη που αναφέρονται παραπάνω, η ακρίβεια και η ισορροπημένη ακρίβεια, αποτελούν μετρητικές ταξινόμησης οι οποίες χρησιμοποιούνται ευρέως για τη μέτρηση επίδοσης πολλών εργασιών. Συμπληρωματικά, αναφέρεται παρακάτω μία μετρητική ανάλυσης παλινδρόμησης, η R^2.

Το R^2 υπολογίζει την ισχύ συσχετισμού μεταξύ του αριθμού των πραγματικών και των προβλεπόμενων σφαλμάτων. Πραγματικά, παρουσιάζει το ποσοστό της μεταβλητότητας της προβλεπόμενης μεταβλητής που αντιπροσωπεύει το εκάστοτε μοντέλο πρόβλεψης. Η R^2 εξίσωση παρουσιάζεται παρακάτω:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Όπου y_i ο πραγματικός αριθμός των σφαλμάτων, $\hat{y}_i$ ο προβλεπόμενος αριθμός των σφαλμάτων, $\overline{y}$ ο μέσος αριθμός των σφαλμάτων.

Γενικά, όσο υψηλότερη είναι η τιμή του R^2 τόσο καλύτερα το προβλεπόμενο μοντέλο προσεγγίζει τα πραγματικά δεδομένα. Η τιμή του R^2 κυμαίνεται μεταξύ των αριθμών -1 και 1.

- Για R^2=+1, υπάρχει τέλεια θετική συσχέτιση μεταξύ των δύο τιμών.
- Για R^2=0, δεν υπάρχει καμία (γραμμική) συσχέτιση μεταξύ των δύο τιμών.
- Για R^2= -1, υπάρχει τέλεια αρνητική συσχέτιση μεταξύ των δύο τιμών.

Όταν ο συντελεστής συσχέτισης είναι κοντά στο −1 ή στο +1 η γραμμική συσχέτιση των δύο τιμών είναι ισχυρή ενώ όταν είναι κοντά στο 0 οι τιμές είναι πρακτικά ασυσχέτιστες.[72]

2.4.2 Ακρίβεια και Ανάκληση

Τα δύο πιο συχνά και βασικά μέτρα ως προς την αποτελεσματικότητα ανάκτησης πληροφοριών αποτελούν η ακρίβεια και η ανάκληση.

Η ακρίβεια (precision) αποτελεί μέτρο της συνάφειας των εξαχθέντων αποτελεσμάτων. Δηλαδή, σύμφωνα με τον πίνακα 2.4-1, ορίζεται η ακρίβεια ως ο λόγος των αληθών θετικών αποτελεσμάτων προς το σύνολο των θετικών αποτελεσμάτων.

$$Precision = \frac{A}{A + B}$$

Ως ανάκληση (recall – degree of completeness) ορίζεται η δεσμευμένη πιθανότητα αν ένα στιγμιότυπο ανήκει σε μια κλάση, έστω c, αυτή να αναγνωριστεί σωστά από τον ταξινομητή:

$$Recall = \frac{\#\sigma\omega\sigma\tau\acute{\omega}\nu\ \pi\rho o\beta\lambda\acute{\epsilon}\psi\epsilon\omega\nu\ \kappa\lambda\acute{\alpha}\sigma\eta\varsigma\ c}{\#\delta\epsilon\delta o\mu\acute{\epsilon}\nu\omega\nu\ \kappa\lambda\acute{\alpha}\sigma\eta\varsigma\ c}$$

Επίσης, η ανάκληση αποτελεί το μέτρο των συναφών αποτελεσμάτων που εξήχθησαν. Για ένα πρόβλημα ταξινόμησης δύο κλάσεων προκύπτει:

$$Recall_P = \frac{D}{D + C}$$

$$Recall_N = \frac{A}{A + B}$$

Συγκεκριμένα, οι δυο μετρικές αυτές μέθοδοι χρησιμοποιούνται συχνά για την αναγνώριση του ονόματος των γονιδίων και των πρωτεϊνών. Μετρώντας τους δύο τύπους των λαθών που μπορούν να γίνουν κατά την διάρκεια της αναγνώρισης του ονόματος του γονιδίου, η ακρίβεια αξιολογεί τον αριθμό των φορών που μια λέξη ή φράση αναγνωρίζεται εσφαλμένα ως όνομα πρωτεΐνης ή γονιδίου.

Στη περίπτωση των γονιδίων, η ακρίβεια ορίζεται ως ο αριθμός των ορθά αναγνωρισμένων ονομάτων γονιδίων ή πρωτεϊνών διαιρούμενος από τον συνολικό αριθμό των αναγνωρισμένων γονιδίων ή πρωτεϊνών. Αυτός ο αριθμός επηρεάζεται δε, από την ομωνυμία. Ομωνυμία έχουμε όταν μια και μόνο λέξη ή φράση μπορεί να αναφέρεται σε αρκετά διαφορετικά γονίδια και ακόμα μπορεί να αναφέρεται σε έννοιες που δεν αφορούν γονίδια. Για παράδειγμα, το PSA μπορεί να αναφέρεται στο Prostate Specific Antigen, στην πρωτεΐνη S (alpha) ή στον Poultry Science Association. Αρκετές προσεγγίσεις έχουν προταθεί για την επίλυση της αμφισημίας μεταξύ λέξης και νοήματος οι οποίες είναι αρκετά αποτελεσματικές.

Από την άλλη μεριά, η ανάκληση (Recall) ορίζεται ως ο συνολικός αριθμός σωστά αναγνωρισμένων γονιδίων ή ονομάτων πρωτεϊνών διαιρούμενος με τον συνολικό αριθμό των ονομάτων γονιδίων και πρωτεϊνών που πραγματικά υπάρχουν στο κείμενο. Η ανάκληση είναι συνήθως 100% εξαιτίας των συνωνύμων της που δεν είναι παρόντα στον θησαυρό αλλά και εξαιτίας των ορθογραφικών παραλλαγών. Οι Tuason et al.[73] αναφέρουν μια ανάκληση της τάξης του 36.2% χρησιμοποιώντας μια προσέγγιση που βασίζεται σε λεξικό για να αναγνωρίσουν τα ονόματα των γονιδίων σε 45.000 περιλήψεις που σχετίζονται με τα γονίδια των ποντικών. Μια ανάλυση 200 περιλήψεων έδειξε ότι το 30% των περιπτώσεων όπου το γονίδιο δεν ανιχνεύεται αποδίδεται στο γεγονός ότι το όνομα του γονιδίου δεν αναφέρεται στην περίληψη. Ένα 51% αυτών των περιπτώσεων προέκυψε από ‘απλές παραλλαγές ονομάτων’. Για να αντιμετωπιστούν αυτές τις παραλλαγές στα ονόματα, έχουν προταθεί η τυχαία αντιστοίχιση της κατά προσέγγιση συμβολοσειράς και η πιθανολογική παραγωγή μεταβλητών. Ωστόσο, τα ονόματα των γονιδίων με μικρές παραλλαγές μπορούν να συμβολίζουν διαφορετικά γονίδια, όπως cyclin-dependent kinase inhibitor 1A και cyclin-dependent kinase inhibitor 1B. Είναι λοιπόν προτιμότερο να ληφθούν υπόψη παραλλαγές που συμμορφώνονται με προκαθορισμένους κανόνες.

2.4.3 Μέτρηση F

Στην πράξη οι δύο παραπάνω μετρικές δεν μπορούν να εκτιμηθούν χωριστά, καθώς παρέχουν μια αλληλοσυμπληρούμενη εικόνα της αποτελεσματικότητας ενός ταξινομητή. Ένα μέτρο που τα συνδυάζει είναι η συνάρτηση F.

Η παραδοσιακή F-measure ή balanced F-score ορίζεται λαμβάνοντας υπόψη την ακρίβεια (precision) και την ανάκληση (recall), καθώς αποτελεί τον αρμονικό μέσο (harmonic mean) της ανάκλησης και ακρίβειας.[70]

$$F - measure = \frac{2 * (Recall * Precision)}{Recall + Precision}$$

2.4.4 Συντελεστής συσχέτισης Matthew's

Ο συντελεστής συσχέτισης Matthew's (Matthew's Correlation Coefficient, MCC) χρησιμοποιείται στη μηχανική μάθηση ως μέτρο αξιολόγησης της δυαδικής ταξινόμησης. Ο συντελεστής λαμβάνει υπόψη αληθινές και ψευδής, θετικές και αρνητικές μετρήσεις, ενώ γενικά θεωρείται ως ένα ισορροπημένο μέτρο ταξινόμησης ακόμη και αν οι κατηγορίες αποτελούνται από πολύ διαφορετικά μεγέθη. Ο MCC λαμβάνει τιμές συσχέτισης μεταξύ -1 και +1, όπου συντελεστής ίσος με 1 αποτελεί μια τέλεια πρόβλεψη, ίσος με 0 κατά μέσο όρο τυχαία πρόβλεψη και ίσος με -1 αντιπροσωπεύει αντίστροφη πρόβλεψη. Συγκεκριμένα, η αναλυτική σχέση προσδιορισμού του MCC παρουσιάζεται ακολούθως:[74]

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

Ενώ σύμφωνα με τα στοιχεία του πίνακα 2.4-1 , η παραπάνω σχέση μετασχηματίζεται ως εξής:

$$MCC = \frac{D \cdot A - B \cdot C}{\sqrt{(D + C)(A + B)(D + B)(A + C)}}$$

2.5 Μέθοδοι Επικύρωσης

Στην εξόρυξη δεδομένων η πιο γνωστή μέθοδος είναι η ταξινόμηση προτύπων (classification) ενώ υπάρχουν πάρα πολλοί αλγόριθμοι κατηγοριοποίησης. Η αποτελεσματικότητα ενός αλγορίθμου μπορεί να εκτιμηθεί βάση της ακρίβειας (accuracy) της κατηγοριοποίησης. Η εκτίμηση της ακρίβειας κατηγοριοποίησης αποτελεί πολύ σημαντικό ζήτημα καθώς δείχνει την δυνατότητα του αλγορίθμου να ανταποκριθεί σε δεδομένα που δε του έχουν καταχωρηθεί.

Η ποιότητα των μοντέλων εξετάζεται με την εκτίμηση του σφάλματος γενίκευσης, δηλαδή την ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης. Με βάση την απόδοση του κάθε ταξινομητή στα διαθέσιμα δεδομένα προκύπτει το εάν ένα μοντέλο καλύπτει καλά το σύνολο των δεδομένων λειτουργίας του. Κατά την μάθηση του νευρωνικού δικτύου μέσα από ένα σύνολο δεδομένων εκπαίδευσης, το νευρωνικό δίκτυο μαθαίνει τις εσωτερικές παραμέτρους του για μια συνάρτηση αντιστοίχησης της κάθε εισόδου στην εκτιμώμενη έξοδο. Σε αυτή τη μάθηση υπάρχουν δύο σφάλματα, το σφάλμα εκπαίδευσης (trainError) και το σφάλμα ελέγχου (testError).

Το σφάλμα εκπαίδευσης που δείχνει πόσο καλά προσεγγίζει το μοντέλο τα Ntrain το πλήθος παραδειγμάτων εκπαίδευσης με τα οποία εκπαιδεύτηκε και είναι αυτό που μειώνεται κατά τη διάρκεια εκπαίδευσης, και το σφάλμα ελέγχου που δείχνει πόσο καλά γενικεύει το μοντέλο σε Ntest το πλήθος νέα παραδείγματα ελέγχου.

Οι πιο κάτω τεχνικές που θα αναφερθούν επιτρέπουν την καλύτερη χρήση των δεδομένων για εκπαίδευση (training), εκτίμηση απόδοσης (testing),επιλογή μοντέλου (model selection).

2.5.1 Διασταυρωμένη επικύρωση

Η διασταυρωμένη επικύρωση (Cross-validation), αποτελεί ένα μοντέλο τεχνικής επαλήθευσης. Σε ένα πρόβλημα πρόβλεψης, δίνεται συνήθως ένα σύνολο γνωστών δεδομένων με το οποίο η λειτουργεί η εκπαίδευση, καθώς και ένα σύνολο αγνώστων δεδομένων με βάση τα οποία ελέγχεται το μοντέλο. Ο στόχος της διασταυρωμένης επικύρωσης είναι να καθορίσει ένα σύνολο δεδομένων ώστε να εξετάσει το μοντέλο στη φάση της εκπαίδευσης, προκειμένου να δώσει μια εικόνα για το πώς το μοντέλο θα γενικευθεί σε ένα ανεξάρτητο σύνολο δεδομένων. Ένας γύρος διασταυρωμένης επικύρωσης περιλαμβάνει τον διαχωρισμό ενός δείγματος δεδομένων σε συμπληρωματικά υποσύνολα, την εκτέλεση της διαδικασίας μοντελοποίησης σε ένα υποσύνολο και την επικύρωση της ανάλυσης από το άλλο υποσύνολο. Για να μειωθεί η μεταβλητότητα, εκτελούνται πολλαπλοί γύροι διασταυρούμενης επικύρωσης

χρησιμοποιώντας διαφορετικά χωρίσματα και τα αποτελέσματα της επαλήθευσης συνυπολογίζονται.[75]

Στην διασταυρωμένη επικύρωση k-υποσυνόλων, το αρχικό δείγμα χωρίζεται τυχαία σε k επιμέρους σύνολα ίσου μεγέθους. Από τα επιμέρους δείγματα k, ένα δείγμα διατηρείται ως δείγμα επικύρωσης για τη δοκιμή του μοντέλου, και τα υπόλοιπα k - 1 δείγματα χρησιμοποιούνται ως δεδομένα εκπαίδευσης. Η διαδικασία διασταυρούμενης επικύρωσης κατόπιν επαναλαμβάνεται k φορές, με κάθε ένα από τα επιμέρους δείγματα k να χρησιμοποιείται ακριβώς μια φορά ως δεδομένο επικύρωσης. Τα αποτελέσματα από τα k-υποσύνολα μπορεί στη συνέχεια να χρησιμοποιηθούν συνολικά για την παραγωγή μιας ενιαίας εκτίμησης.[75]

Το πλεονέκτημα αυτής της μεθόδου είναι ότι όλες οι παρατηρήσεις χρησιμοποιούνται τόσο ως σύνολα εκπαίδευσης όσο και ως σύνολα επικύρωσης, και κάθε παρατήρηση χρησιμοποιείται σε σύνολο επικύρωσης ακριβώς μια φορά. Στη διασταυρωμένη επικύρωση k-υποσυνόλων, τα υποσύνολα επιλέγονται έτσι ώστε η μέση τιμή απόκρισης να είναι περίπου ίση σε όλα τα υποσύνολα.[75]

Στις μεθόδους cross validation περιλαμβάνονται η μέθοδος Random Subsampling, η μέθοδος K-Fold Cross-Validation και Bootstrap.[75]

Τυχαία δειγματοληψία (Random Subsampling)

Στη μέθοδο της τυχαίας δειγματοληψίας το σύνολο ελέγχου επιλέγεται ως τυχαίο δείγμα. Εφαρμόζεται τυχαία δειγματοληψία χωρίς επανατοποθέτηση και επιλέγονται Ν πρότυπα για το σύνολο ελέγχου. Τα εναπομείναντα πρότυπα σχηματίζουν το σύνολο εκπαίδευσης. Με αυτήν την ενέργεια μειώνεται η επιρροή που μπορεί να επιφέρει η κατανομή των στιγμιότυπων στο σύνολο δεδομένων. Η προηγούμενη διαδικασία επαναλαμβάνεται κ φορές ώστε να επιτευχθεί η μεγαλύτερη δυνατή μείωση της επιρροής.

K-fold cross-validation

Θεωρείται μια από τις πλέον αξιόπιστές μεθόδους για την αποτίμηση της ακρίβειας κατηγοριοποιητών. Στην περίπτωση αυτήν εφαρμόζεται ένας πιο συστηματικός τρόπος ώστε να ληφθούν τυχαία δείγματα για το σχηματισμό του συνόλου ελέγχου.

Εάν το σύνολο δεδομένων αποτελείται από Μ πρότυπα και έχει οριστεί ως αριθμός επαναλήψεων το k, τότε χωρίζεται το σύνολο σε k ισομεγέθη τμήματα μεγέθους Μ/k το καθένα. Στην i-οστή επανάληψη, το i-οστό τμήμα λειτουργεί ως σύνολο ελέγχου, ενώ τα υπόλοιπα k – 1 τμήματα αποτελούν το σύνολο εκπαίδευσης.

Για μεγάλο K η εκτίμηση του σφάλματος είναι αρκετά ακριβής αλλά με μεγάλες αποκλίσεις. Για μικρό K μειώνεται το πλήθος των πειραμάτων και το υπολογιστικό κόστος και το εκτιμώμενο σφάλμα θα είναι μεγαλύτερο από το πραγματικό αλλά με μικρότερες αποκλίσεις.

Η Leave-one-out είναι ειδική περίπτωση της K-Fold Cross Validation. Για ένα σύνολο δεδομένων με K παραδείγματα εκτελούνται K πειράματα. Κάθε παράδειγμα αφήνετε με την σειρά του έξω από το σύνολο εκπαίδευσης, και ο αλγόριθμος στην συνέχεια εκπαιδεύεται στα υπόλοιπα K-1. Δηλαδή χρησιμοποιούνται K-1 δείγματα για εκπαίδευση και ένα μόνο για έλεγχο. Η παραλλαγή αυτή χρησιμοποιείται μόνο σε μικρά σύνολα δεδομένων.

Η μέθοδος αυτή έχει ως πλεονέκτημα το ότι αποφεύγετε η τυχαία δειγματοληψία. Επίσης χρησιμοποιείται το μέγιστο δυνατό ποσό δεδομένων για εκπαίδευση. Μειονέκτημα αυτής αποτελεί το υψηλό υπολογιστικό κόστος.

2.5.2 Επικύρωση διαχωρισμένου δείγματος

Με την επικύρωση διαχωρισμένου δείγματος (Split-Sample Validation), το μοντέλο παράγεται χρησιμοποιώντας ένα training δείγμα κατόπιν εξέτασης σε ένα παρακρατημένο δείγμα. Μπορεί να προσδιοριστεί το μέγεθος ενός training δείγματος, ώστε να εκφράζεται ως ποσοστό του ολικού μεγέθους του δείγματος, ή σαν μια μεταβλητή που χωρίζει το δείγμα σε άλλα δείγματα (training και testing).

Εάν χρησιμοποιηθεί μια μεταβλητή για να οριστούν τα training και testing δείγματα, οι περιπτώσεις που έχουν ως μεταβλητή τιμή 1 αποδίδονται στο training δείγμα και όλες οι άλλες στο testing. Η μεταβλητή όμως δεν μπορεί να είναι η εξαρτημένη ή η μεταβλητή βαρύτητας ή η μεταβλητή επιρροής, δεν μπορεί να είναι επίσης μια αναγκαστική ανεξάρτητη μεταβλητή. Μπορούν να παρουσιαστούν τα χαρακτηριστικά και για τα training και για τα testing δείγματα ή μόνο για το testing.

Η επικύρωση διαχωρισμένου δείγματος θα πρέπει να χρησιμοποιείται με προσοχή σε αρχεία με λίγα δεδομένα (αρχεία δεδομένων με μικρό αριθμό περιπτώσεων). Μικρά μεγέθη σε training δείγματα είναι πιθανόν να αποδώσουν μοντέλα χαμηλής σημαντικότητας αφού δεν θα υπάρχουν αρκετές περιπτώσεις σε κάποιες κατηγορίες ώστε να μεγαλώσουν αρκετά το δέντρο μας.[67]

2.5.3 External Validation

Σε ορισμένες περιπτώσεις τίθεται ως στόχος της επικύρωσης, η επίδειξη ικανοποιητικής απόδοσης για ένα εξωτερικό σύνολο δεδομένων συγκριτικά με εκείνο βάση του οποίου είχε σχεδιαστεί ο αρχικός αλγόριθμος του μοντέλου. Δηλαδή αφορά δεδομένα που δεν έχει ξαναδεί ο αλγόριθμος και το μοντέλο που έχει σχεδιαστεί.[76]

Αντίστοιχα με το split validation όπου για παράδειγμα από τα 10 στοιχεία του dataset πραγματοποιείται train στα 7 και test στα 3, στην external validation πραγματοποιείται αντίστοιχα train και στα 10 στοιχεία ώστε να χτιστεί το μοντέλο και test σε οποιοδήποτε άλλο εξωτερικό dataset. Ορισμένοι φτιάχνουν «ψευδο-εξωτερικά» dataset, δηλαδή για το προαναφερθέν παράδειγμα, θα έπαιρναν περίπου 3 από τα 10 στοιχεία και θα τα έβγαζαν από το dataset, ώστε τώρα το dataset να περιέχει 7 στοιχεία, και έτσι το train/test γίνεται ίδιο με το split validation.

Σε ένα παράδειγμα της βιβλιογραφίας, πραγματοποιείται διερεύνηση των προγνωστικών μοντέλων επαλήθευσης κάποιας ασθένειας, όπου διερευνάται η εκδήλωση συμπτωμάτων του ασθενούς σε σχέση με τον ασθενή και τα χαρακτηριστικά της ασθένειας. Αφού ο στόχος της επαλήθευσης είναι να επιδείξει ικανοποιητική απόδοση για τους ασθενείς από έναν διαφορετικό πληθυσμό από το πρωτότυπο, είναι σαφώς επιθυμητό να αξιολογηθεί ένα μοντέλο για τα νέα δεδομένα που συλλέγονται από ένα κατάλληλο πληθυσμό ασθενών με διαφορετικό κέντρο δράσης. Σημαντικά θέματα σχεδιασμού, όπως η επιλογή του δείγματος και το μέγεθος του δείγματος έχουν σε μεγάλο βαθμό παραμεληθεί στη βιβλιογραφία. Η εξωτερική αξιολόγηση μπορεί να βασίζεται σε αναδρομικά δεδομένα και έτσι είναι βιώσιμη για την επικύρωση των μοντέλων επιβίωσης που χρειάζονται μακρά παρακολούθηση.[77]

2.6 Μέθοδοι εμπλουτισμού

Το τελικό στάδιο πολλών πρωτεϊνικών, γενετικών και μεταβολικών αναλύσεων είναι η παραγωγή μιας λίστας από «ενδιαφέροντα» βιομόρια. Χαρακτηριστικά παραδείγματα αυτών περιλαμβάνουν λίστες γονιδίων που κατατάσσονται από διαφορική ή συνέκφραση και έχουν ερευνηθεί σε πειράματα μικροσυστοιχίας. Συγκεκριμένα, εντοπίζονται λίστες από μονά νουκλεοτίδια πολυμορφισμού (Single-Nucleotide Polymorphism, SNP) που περιέχουν γονίδια που έχουν καταταχθεί από π-τιμές καθορίζονται από γενετικό συσχετισμό σε ένα φαινότυπο ενδιαφέροντος μέσα από μια μελέτη που αφορά όλο το γονιδίωμα, καθορίζονται επίσης από υπολογιστικά παραγόμενες λίστες θεωρούμενης

μεταγραφής ή στόχων mRNA που ορίζονται από πιθανότητες. Δυστυχώς τέτοιες λίστες κατάταξης είναι κενές δομής και στερούνται πλαισίου εφαρμογής .

Είναι δύσκολο να καθοριστεί ο τρόπος με τον οποίο τα γονίδια και τα πρωτεϊνικά τους προϊόντα αλληλεπιδρούν μεταξύ τους ή επηρεάζουν τις βιολογικές διεργασίες καθώς και ποια μπορεί να είναι η «φυσιολογική» τους συμπεριφορά μόνο με την απλή επανεξέτασή τους. Η συμπεριφορά αυτή αλλάζει κατά τη διάρκεια ασθένειας, διαταραχής ή θεραπείας. Οι χειροκίνητες αναζητήσεις γονίδιο-γονίδιο, ιδιαίτερα σε μεγάλες λίστες γονιδίων είναι εργασίες υπερβολικά κοπιώδεις και συχνά ακατόρθωτες. Με τον ίδιο τρόπο, οι λίστες κατάταξης των γονιδίων συνεισφέρουν ελάχιστα στην αντιγραφή της πολύπλοκης πραγματικότητας της βιολογίας, όπου τα γονίδια και οι πρωτεΐνες συνεργάζονται σε ομάδες που αλληλεπιδρούν με πολύπλοκο τρόπο για να δημιουργήσουν λειτουργικά συστήματα. Η εστίαση σε μια συλλογή από ενδιαφέροντα γονίδια και πρωτεΐνες σαν σύνολο δεν είναι μόνο βιολογικά πιο εύληπτο αλλά έχει και την τάση να αυξάνει τη στατιστική του ισχύ και να μειώνει τη διαστατικότητα. Η κατανόηση της λειτουργικής σημαντικότητας από τέτοιες λίστες με γονίδια αν και υπερβολική είναι ωστόσο μια κρίσιμη εργασία.

Ο εμπλουτισμός (που μερικές φορές αποκαλείται pathway analysis)[78] έχει εξελιχθεί σε δευτερεύουσα ανάλυση σε συλλογές γονιδίων και αναγνωρίζεται από υψηλής απόδοσης γενωμικές μεθόδους εξαιτίας της ικανότητάς του να παρέχει μια πολύτιμη ματιά στις συνολικές βιολογικές λειτουργίες που υπόκεινται σε μια λίστα γονιδίων. Με τη συστηματική χαρτογράφηση γονιδίων και πρωτεϊνών, στα συσχετισμένα βιολογικά σχόλια (όπως είναι οι όροι γονιδιακής οντολογίας [GO][79] ή η pathway membership) και στη συνέχεια συγκρίνοντας την κατανομή των όρων μέσα σε ένα σετ γονιδίων ενδιαφέροντος με την κατανομή στο παρασκήνιο αυτών των όρων, η ανάλυση εμπλουτισμού μπορεί να εντοπίσει όρους οι οποίοι αναπαρίστανται με στατιστική υπερβολή ή μη επαρκώς μέσα στη λίστα ενδιαφέροντος.[80] Εννοείται ότι τέτοιοι εμπλουτισμένοι όροι περιγράφουν κάποια σημαντική υποκείμενη βιολογική διεργασία ή συμπεριφορά. Για παράδειγμα, εάν 10% των γονιδίων από τις λίστες είναι κινάσες, σε σύγκριση με το 1% των γονιδίων του ανθρώπινου γονιδιώματος (το πληθυσμιακό υπόβαθρο) , και χρησιμοποιώντας κοινές στατιστικές μεθόδους (π.χ. χ2, το ακριβές τεστ του Fisher, τη διωνυμική πιθανότητα, ή την υπεργεωμετρική κατανομή) είναι πιθανό να καθοριστεί πως οι κινάσες εμπλουτίζονται στη λίστα γονιδίων και συνεπώς έχουν σημαντικές λειτουργίες στη βιολογική μελέτη που διεξάγεται.[81]

Πίνακας 2.6-1: Λίστα εργαλείων εμπλουτισμού *(αναπαράχθηκε από Huang et al.)*[82]

Εργαλείο Εμπλουτισμού	Έτος Κυκλοφορίας	Εργαλείο Εμπλουτισμού	Έτος Κυκλοφορίας	Εργαλείο Εμπλουτισμού	Έτος Κυκλοφορίας
FunSpec	2002	L2L	2005	PAGE	2005
Onto-express	2002	WebGestalt	2005	T-profiler	2005
EASE	2003	BayGO	2006	FuncCluster	2006
FatiGO/FatiWise/FatiGO+	2003	eGOn/GeneTools	2006	FatiScan	2007
FuncAssociate	2003	Gene Class Expression	2006	FINA	2007
GARBAN	2003	GOALIE	2006	GAzer	2007
GeneMerge	2003	GOFFA	2006	GeneTrail	2007
GoMiner	2003	GOLEM	2006	MetaGP	2007
MAPPFinder	2003	JProGO	2006	Ontologizer	2004
CLENCH	2004	PageMan	2006	POSOC	2004
GO::TermFinder	2004	STEM	2006	topGO	2006
GOAL	2004	WEGO	2006	GO-2D	2007
GOArray	2004	EasyGO	2007	GENECODIS	2007
GOStat	2004	g:Profiler	2007	GOSim	2007
GoSurfer	2004	ProbCD	2007	PalS	2008
OntologyTraverser	2004	GOEAST	2008	ProfCom	2008
THEA	2004	GOHyperGAll	2008	GOTM	2004
BiNGO	2005	CatMap	2004	ermineJ	2005
FACT	2005	Godist	2004	DAVID	2003
gfinder	2005	GO-Mapper	2004	GOToolBox	2004
Gobar	2005	iGA	2004	ADGO	2006
GOCluster	2005	GSEA	2005	FunNet	2008
GOSSIP	2005	MEGO	2005		

2.6.1 Gene Set Enrichment Analysis

Η Gene Set Enrichment Analysis (GSEA) λαμβάνει υπόψη της πειράματα με προφίλ έκφρασης του γονιδιώματος από δείγματα που ανήκουν σε δύο τάξεις και επισημαίνονται ως 1 ή 2. Τα γονίδια κατατάσσονται με βάση τη συσχέτιση μεταξύ της έκφρασης και της διάκρισης της τάξης χρησιμοποιώντας οποιαδήποτε κατάλληλη μετρική μέθοδο.[83]

Δεδομένου ενός ορισμένου σετ γονιδίων S (π.χ. γονίδια που κωδικοποιούν προϊόντα σε μια μεταβολική οδό, που βρίσκονται στην ίδια κυτογενική ζώνη, ή έχουν κοινή

κατηγορία GO), ο σκοπός της GSEA είναι να καθοριστεί εάν τα μέλη του S είναι τυχαία κατανεμημένα σε όλο το L ή βρίσκονται κυρίως στην κορυφή ή στο κάτω μέρος. Αναμένεται ότι τα σετ που σχετίζονται με τη φαινοτυπική διάκριση θα φανούν στην τελευταία κατανομή.[83]

Μια γενική θεώρηση του GSEA που δείχνει τη μέθοδο.[83]

i) Μια έκφραση ενός σετ δεδομένων που κατανέμεται από μια συσχέτιση με ένα φαινότυπο , ο αντίστοιχος heat map, και τα «gene tags» δηλαδή η τοποθεσία γονιδίων από ένα σετ S μέσα στην καθορισμένη λίστα.
ii) Τμήμα του τρέχοντος αθροίσματος για το S στο σετ δεδομένων περιλαμβάνει και την τοποθεσία του μέγιστου βαθμού εμπλουτισμού (maximum enrichment score - ES) και το κορυφαίο υπο-σετ.[83]

Υπάρχουν τρία κύρια βήματα στην μέθοδο GSEA:

Βήμα 1: Υπολογισμός του βαθμού εμπλουτισμού.

Βήμα 2: Εκτίμηση του Επιπέδου Σημαντικότητας (Significance Level) του ES.

Βήμα 3: Προσαρμογή για την δοκιμασία Πολλαπλής Υπόθεσης (Adjustment for Multiple Hypothesis Testing).

Συγκεκριμένα, στο βήμα 1 υπολογίζεται ο βαθμός Εμπλουτισμού (Enrichment Score, ES) όλης της λίστας κατάταξης L που αντανακλά το βαθμό στον οποίο ένα σετ S παρουσιάζει έμφαση στα άκρα του. Το σκορ υπολογίζεται προχωρώντας προς τα κάτω στη λίστα L, αυξάνοντας το αθροιστικό στατιστικό στοιχείο όταν συναντάται ένα γονίδιο στο S και μειώνοντας το όταν δεν συναντάται γονίδιο στο S. Το μέγεθος της προσαύξησης εξαρτάται από τη συσχέτιση του γονιδίου με το φαινότυπο. Το σκορ εμπλουτισμού είναι η μέγιστη απόκλιση από το μηδέν όταν συναντάται σε τυχαία διαδρομή. Αντιστοιχεί, λοιπόν, σε μια σταθμισμένη στατιστική του είδους Kolmogorov–Smirnov.

Έπειτα στο βήμα 2, υπολογίζεται η στατιστική σημαντικότητα (ονομαστική τιμή P) για το ES χρησιμοποιώντας εμπειρική διαδικασία δοκιμής μετάθεσης με βάση το φαινότυπο, η οποία διατηρεί τη σύνθετη συσχέτιση των δεδομένων έκφρασης των γονιδίων. Συγκεκριμένα, μετατίθενται οι επισημάνσεις των φαινοτύπων και υπολογίζεται εκ νέου το ES του σετ γονιδίων των δεδομένων που έχουν μετατεθεί, ενώ ακολουθεί η παραγωγή μηδενικής κατανομής για το ES. Η εμπειρική, ονομαστική τιμή P των υπό παρατήρηση ES υπολογίζεται σε σχέση με αυτή τη μηδενική κατανομή. Είναι σημαντικό η μετάθεση των επισημάνσεων τάξης να διατηρεί τις συσχετίσεις μεταξύ των γονιδίων και, κατά συνέπεια, να παρέχει μια περισσότερο βιολογικά λογική αξιολόγηση που θα μπορούσε να εξαχθεί από τα γονίδια που μετατίθενται.

Τέλος στο βήμα 3, όταν αξιολογείται μια ολόκληρη βάση από σετ γονιδίων ρυθμίζεται το υπολογιζόμενο επίπεδο σημαντικότητας ώστε να αναλογεί σε δοκιμασίες πολλαπλής υπόθεσης. Αρχικά, κανονικοποιείται το ES για κάθε σετ γονιδίων που αναλογεί στο μέγεθος του σετ, εξάγοντας ένα κανονικοποιημένο βαθμό εντοπισμού (Normalized Enrichment Score, NES). Στη συνέχεια, ελέγχεται η αναλογία λανθασμένων αποδοχών υπολογίζοντας τον εσφαλμένο ρυθμό ανακάλυψης (False Discovery Rate, FDR) που ανταποκρίνεται στο κάθε NES και αναπαριστά ένα όνομα λανθασμένης αποδοχής. Υπολογίζεται, λοιπόν, με την σύγκριση των τελικών παρατηρούμενων και μηδενικών κατανομών για το NES.[83]

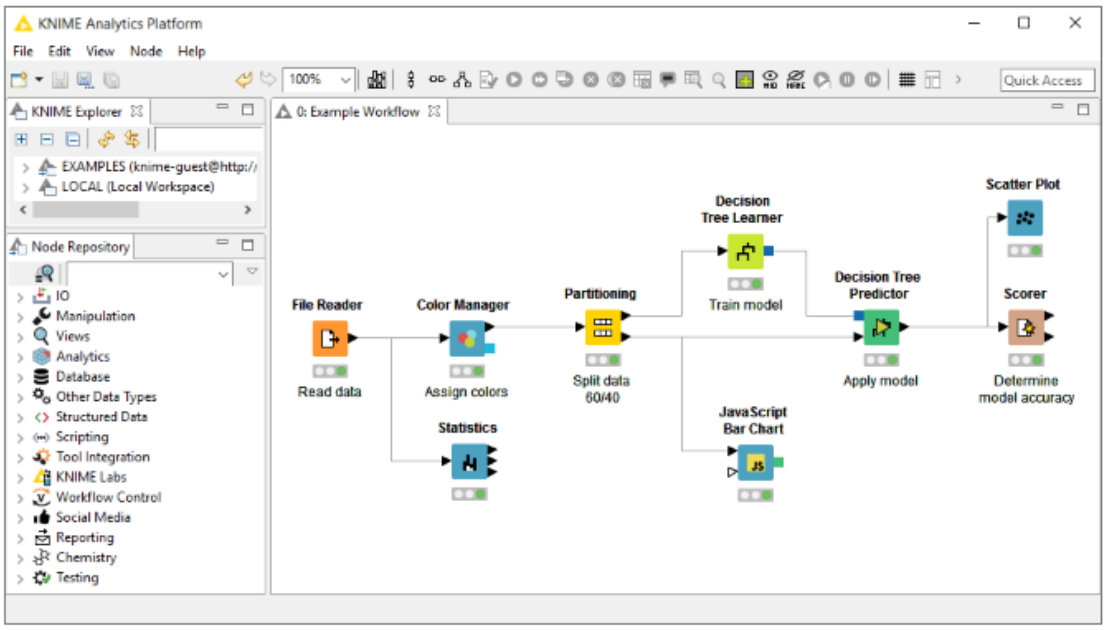
3. Υλικά και Μεθοδολογία

Στην παρούσα εργασία χρησιμοποιήθηκε το πρόγραμμα εξόρυξης δεδομένων KNIME και πιο συγκεκριμένα η έκδοση 3.2.1. Αναπτύχθηκε μια αλληλουχία (διάγραμμα ροής) από κόμβους με σκοπό την εξόρυξη ορισμένων πληροφοριών από ένα αρχείο τύπου sdf. Το συγκεκριμένο αρχείο που αντλήθηκε από τη βάση δεδομένων της DrugBank περιέχει όλες τις πληροφορίες για φαρμακευτικές ουσίες, όπως σε ποια τάξη ανήκει το κάθε φάρμακο, τα αποτυπώματα της κάθε ουσίας σε διαφορετικό format. Επίσης χρησιμοποιήθηκαν script της έκδοσης 2.7 της γλώσσας προγραμματισμού Python με σκοπό ορισμένες περαιτέρω εξορύξεις δεδομένων όπως η εύρεση και η καταγραφή μόνο των αντιψυχωτικών φαρμάκων καθώς και ο εμπλουτισμός (drug enrichment) μόνο των εγκεκριμένων φαρμάκων. Τέλος καταγράφηκαν οι πρωτεΐνες με υψηλότερο estimation score, που προέκυψε από τον εμπλουτισμό και έγινε βιβλιογραφική αναζήτηση καθώς και συσχέτιση των πρωτεϊνών αυτών με την ψύχωση.

3.1 Λογισμικό που χρησιμοποιήθηκε

3.1.1 KNIME

Το KNIME (the Konstanz Information Miner), χρησιμοποιείται για open-source ανάλυση δεδομένων, καθώς ενσωματώνει στοιχεία μηχανικής μάθησης και εξόρυξης δεδομένων. Η γραφική διεπαφή χρήστη επιτρέπει τη γρήγορη και εύκολη συναρμολόγηση των κόμβων για την προεπεξεργασία των δεδομένων, για την ανάλυση και μοντελοποίηση δεδομένων καθώς και για την οπτικοποίησή τους. Χρησιμοποιείται, κυρίως από το 2006,[28] στην φαρμακευτική έρευνα αλλά εξίσου και σε άλλους τομείς. Διατίθεται ως ελεύθερο λογισμικό στη σελίδα https://www.knime.org/.[28] Στην παρούσα εργασία χρησιμοποιήθηκε η έκδοση 3.3.2.



Εικόνα 3.1-1: Το εργαλείο – KNIME

3.1.2 Python

Η Python είναι μια δυναμική αντικειμενοστραφής γλώσσα προγραμματισμού που μπορεί να χρησιμοποιηθεί για πολλά είδη ανάπτυξης λογισμικού. Παρουσιάστηκε πρώτη φορά το 1991 από τον Guido van Rossum,[84] ένα Δανό προγραμματιστή. Προσφέρει ισχυρή υποστήριξη για ενσωμάτωση με άλλες γλώσσες και εργαλεία, ενώ διαθέτει ένα πλούσιο εύρος βιβλιοθηκών και διατίθεται ως ελεύθερο λογισμικό στην ιστοσελίδα https://www.python.org/.[84] Στην παρούσα εργασία χρησιμοποιήθηκε η έκδοση 2.7.

3.2 Πρόβλεψη Στόχου

3.2.1 Laplacian-modified Naïve Bayes

Ο Laplacian-Modified Naïve Bayes ταξινομητής που χρησιμοποιήθηκε στην παρούσα εργασία βασίζεται στην ανάλυση των Nigsch et al.[85] Έχοντας ως αρχή τον μπεϋσιανό κανόνα και θεωρώντας πως τα στοιχεία f_i είναι ανεξάρτητα, η $p(\omega_\alpha \mid x)$ μπορεί να υπολογιστεί από την παρακάτω σχέση:[15]

$$p(\omega_\alpha \mid x) = \frac{p(\omega_\alpha)}{p(x)}\left[\prod_{i=1}^{d}\frac{p(\omega_\alpha \mid f_i)}{p(\omega_\alpha)}p(f_i)\right]$$

Η κάθε κατηγορία ορίζεται ως εξής $p(\omega_\alpha) = N_{\omega_\alpha}/N$, όπου N_{ω_α} ο αριθμός των περιπτώσεων στην κατηγορία ω_α και N ο συνολικός αριθμός των περιπτώσεων. Αν ως $N^+{}_{i\omega_\alpha}$ οριστεί ο συνολικός αριθμός των εμφανίσεων του στοιχείου f_i στην κατηγορία ω_α και $N^+{}_i$ ο συνολικός των εμφανίσεων του στοιχείου f_i σε όλο το σετ, τότε η πιθανότητα της μη διορθωμένης κατηγορίας ενός στοιχείου ορίζεται από την παρκάτω σχέση:

$$p(\omega_\alpha \mid f_i) = \frac{N^+{}_{i\omega_\alpha}}{N^+{}_i}$$

Σύμφωνα με την παραπάνω σχέση, αν $N^+{}_{i\omega_\alpha} = 0$, τότε $p(\omega_\alpha \mid f_i) = 0$ και $p(\omega_\alpha \mid x) = 0$, ανεξάρτητο για όλα τα $p(\omega_\alpha \mid f_i)$. Η επιθυμητή επίλυση προσεγγίζεται καθώς το στοιχείο f_i εμφανίζεται σπάνια και ισχύει $p(\omega_\alpha \mid f_i) = p(\omega_\alpha)$. Ως εκ τούτου, αν ένα στοιχείο εξετάζεται D παραπάνω φορές, αναμένεται $D * p(\omega_\alpha)$ δείγματα να ανήκουν στην κατηγορία ω_α. Επομένως προσθέτοντας D παραπάνω δείγματα, η προηγούμενη σχέση μετασχηματίζεται ως ακολούθως:[15]

$$p'(\omega_\alpha \mid f_i) = \frac{N^+{}_{i\omega_\alpha} + D * p(\omega_\alpha)}{N^+{}_i + D}$$

Ενώ το όριο καθώς το $N^+{}_{i\omega_\alpha} \to 0$ και $N^+{}_i \to 0$ ορίζεται ως:

$$\lim_{N^+_{i\omega_\alpha}, N^+_i \to 0} p'(\omega_\alpha \mid f_i) = \lim_{N^+_{i\omega_\alpha}, N^+_i \to 0} \frac{N^+_{i\omega_\alpha} + D * p(\omega_\alpha)}{N^+_i + D} = p(\omega_\alpha)$$

Κατά τη διόρθωση Laplace, $D = p(\omega_\alpha)^{-1}$, ενώ ορίζεται η σχετική πιθανότητα ως εξής:

$$p_{rel}(\omega_\alpha \mid f_i) = \frac{p'(\omega_\alpha \mid f_i)}{p(\omega_\alpha)} = \frac{N^+_{i\omega_\alpha} + 1}{N^+_i * p(\omega_\alpha) + 1}$$

Ο αλγόριθμος μετασχηματίστηκε ώστε να αποφευχθούν αριθμητικά προβλήματα καθώς οι τιμές μικραίνουν. Επομένως, το score της κατηγορίας ω_α ενός νέου μορίου x υπολογίζεται από τη σχέση:[15]

$$S_{\omega_\alpha}(x) = \sum_i f_i \log\left[\frac{N^+_{i\omega_\alpha} + 1}{N^+_i * p(\omega_\alpha) + 1}\right] + \log \frac{\prod_{i=1}^{d} p(f_i)}{p(x)}$$

Όπου το f_i αντιπροσωπεύει τη διωνυμική τιμή του στοιχείου για το μόριο x.

3.2.2 Class-specific score cut-offs

Η θέσπιση ορίου κατώτερης τιμής στις τιμές που προκύπτουν από την εφαρμογή της τροποποιημένης Laplacian Naïve Bayes αποτελεί μία μέθοδο αρχικής ταξινόμησης και επιλεκτικής διαλογής των εξαχθέντων αποτελεσμάτων. Συγκεκριμένα, οι *Nguyen et al.* στη μελέτη τους, η οποία αφορά τη διαφορική επιλογή ενώσεων από τις βιβλιοθήκες διαλογής υψηλής απόδοσης, πραγματοποίησαν τις αναλύσεις τους και εξήγαγαν αποτελέσματα εφαρμόζοντας όρια κατωφλίου.

Με τη θέσπιση ορίου κατώτερης τιμής, πραγματοποιείται διαλογή των τιμών με απώτερο στόχο τη βελτίωση του προφίλ της ανάλυσης βιοδραστικότητας. Οι *Drakakis et al.*[86] επισημαίνουν τα σφάλματα στα οποία μπορεί να οδηγήσει η χρήση τέτοιων ορίων ενώ προτείνουν συγκεκριμένους τρόπους για την πραγματοποίηση cut-offs ομάδων συγκεκριμένων πρωτεϊνών με αξιολόγηση της απόδοσής τους.

Τέλος, τα παράγωγα των cut-offs έχουν χρησιμοποιηθεί σε πρόσφατες δημοσιεύσεις ώστε να βελτιωθεί η ανάλυση του προφίλ βιοδραστικότητας, καθώς και ο υπολογισμός εμπλουτισμού σε μια Xenopus laevis χημική γενετική απεικόνιση.[86]

3.3 Enrichment Calculation

Ο υπολογισμός του εμπλουτισμού πραγματοποιείται ώστε να αποφευχθούν όσο το δυνατόν περισσότερο είτε οι υπερβολικές προβλέψεις είτε οι υποτιμήσεις αυτών, οι οποίες οφείλονται στην μεροληψία των δεδομένων που χρησιμοποιούνται. Για το σκοπό αυτό δημιουργήθηκε μια ποικιλόμορφη βιβλιοθήκη στο υπόβαθρο η οποία καλύπτει ένα μεγάλο φάσμα χημικού χώρου.

Αναλυτικά στη μελέτη των *Liggi et al.*,[37] ανακτήθηκαν δεδομένα εγκεκριμένων φαρμάκων από την DrugBank, παρόμοια μόρια φαρμάκων από την PubChem, ανθρώπινοι μεταβολίτες από την HDMB, φυσικά προϊόντα από την ZINC και υπολογιστικά παραχθέντες ενώσεις από την GDB13. Συνολικά συλλέχθηκαν 194849 μόρια με μοριακό βάρος να κυμαίνεται μεταξύ 100 και 900, από τα οποία τελικά επιλέχθηκαν τα 194433 μόρια.

Η τιμή του Estimation Score περιγράφει το κλάσμα των τυχαίων δειγμάτων των οποίων η συχνότητα για ένα δεδομένο στόχο/οδό είναι πάνω από την παρατηρούμενη ένα. Υπολογίζεται μετρώντας πόσες φορές η συχνότητα της πρόβλεψης για το στόχο/οδό στα προαναφερθέντα τυχαία σύνολα δεδομένων είναι μεγαλύτερη ή ίση με το ένα στο σύνολο των δεδομένων δοκιμής $(R_i \geq F_t)$. Η απόλυτη συχνότητα (C) που λαμβάνεται στη συνέχεια διαιρείται με τον συνολικό αριθμό των τυχαίων δεδομένων n αποδίδοντας μια τιμή μεταξύ 0 και 1.[37]

$$Estimation\ Score = \frac{C}{n}$$

Το ES αποτελεί την πρώτη μέθοδο κατάταξης των στόχων και των οδών με βάση τον εμπλουτισμό, καθώς εκτιμάται η στατιστική συνάφεια των προβλεπόμενων στόχων και μονοπατιών. Συγκεκριμένα, καθώς η τιμή του ES προσεγγίζει την τιμή 1, η συχνότητα πρόβλεψης στα τυχαία σύνολα δεδομένων είναι υψηλότερη από εκείνη που παρατηρήθηκε στο σετ δοκιμής, επομένως δεν φαίνεται να είναι σημαντικός ο στόχος. Αντιθέτως, καθώς η τιμή του ES προσεγγίζει την τιμή 0, εμφανίζεται ισχυρή συσχέτιση.[37]

Η αναλογία πιθανοτήτων (Odds Ratio) για έναν συγκεκριμένο στόχο / μονοπάτι υπολογίζεται συγκρίνοντας τη συχνότητα της πρόβλεψης στο σετ δοκιμών (F_t) και τη συχνότητα πρόβλεψης στην κατανομή υποβάθρου (F_b), λαμβάνοντας υπόψη τον συνολικό αριθμό προβλέψεων σε κάθε σετ (N).[37]

$$Odds\ Ratio = \frac{F_t / N_t}{F_b / N_b}$$

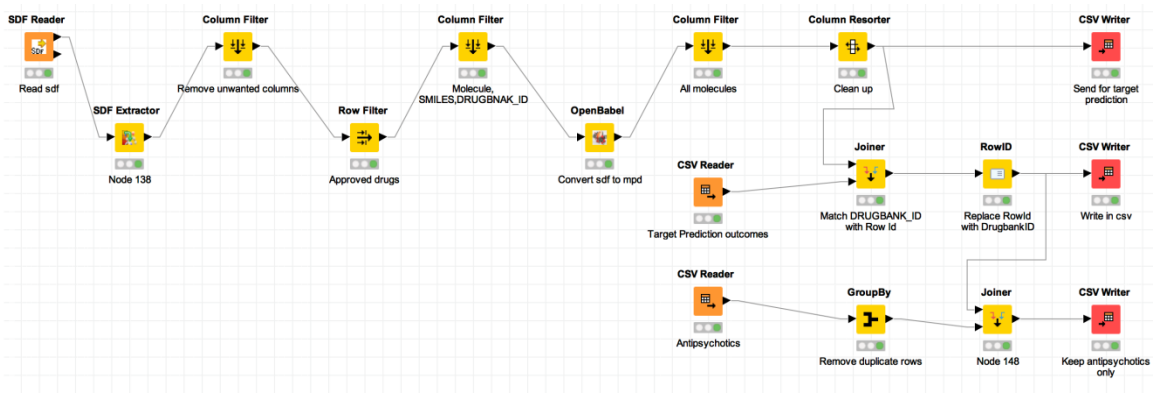
Ο μέσος όρος είναι παρόμοιος με την αναλογία πιθανοτήτων, αλλά σε μεγαλύτερη κλίμακα από τη στιγμή που 10.000 (n) σύνολα δεδομένων χρησιμοποιούνται για τον υπολογισμό. Αυτά τα σύνολα δεδομένων έχουν τον ίδιο αριθμό ενώσεων του συνόλου δοκιμών και παράγονται τυχαία από την κατανομή του υποβάθρου. Ο μέσος όρος για έναν συγκεκριμένο στόχο / μονοπάτι υπολογίζεται κατόπιν σύγκρισης της συχνότητας πρόβλεψης σε κάθε τυχαίο σύνολο δεδομένων (R_i) με τη συχνότητα πρόβλεψης στο σύνολο δεδομένων δοκιμής: [37]

$$Average\ Ratio = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{F_t}$$

Στην παρούσα διπλωματική εργασία, ο εμπλουτισμός πραγματοποιήθηκε με παραλλαγή της μεθόδου των *Liggi, Drakakis, Koutsoukas et. Al* όπου τα μόρια με καλύτερο estimation score ήταν πιο κοντά στο 1 και με χειρότερο πιο κοντά στο 0, επιλέχθηκαν τα μόρια με estimation score τουλάχιστον 0,99.

3.4 Μεθοδολογία

Στο πρώτο μέρος του υπολογιστικού μέρους της παρούσας διπλωματικής εργασίας, ανακτήθηκαν όλα τα εγκεκριμένα μόρια από τη βάση δεδομένων DrugBank[24] και αποθηκεύτηκαν ως αρχείο τύπου .sdf προς περαιτέρω επεξεργασία στο πρόγραμμα Knime, ενώ απορρίφθηκαν τα μόρια με χαρακτηρισμό ως παράνομα, προς διερεύνηση, αποσυρμένα και πειραματικά. Στη συνέχεια, πραγματοποιήθηκαν διαδοχικά φιλτραρίσματα ώστε να μελετηθούν μόνο τα ID, τα αποτυπώματα smiles και η μορφοποίηση σε κυκλικά αποτυπώματα Molprint2D των εγκεκριμένων φαρμάκων. Από την πληθώρα των πληροφοριών πραγματοποιήθηκε «καθαρισμός» αυτών μέσω της πρόβλεψης στόχου[4] και δημιουργήθηκε script στην Python ώστε να συγκρατηθούν τα φάρμακα της DrugBank που ήταν καταχωρημένα ως αντιψυχωτικοί παράγοντες (antipsychotics agents).



Σχήμα 3.4-1:Διάγραμμα ροής του Knime, του πρώτου μέρους της εργασίας.

Στον Πίνακα 3.4-1 παρατίθεται ο κώδικας της Python που χρησιμοποιήθηκε ώστε να συγκρατηθούν τα φάρμακα της DrugBank που ήταν καταχωρημένα ως αντιψυχωτικοί παράγοντες (antipsychotics agents). Οι προτάσεις οι οποίες είναι χρωματισμένες με πράσινο χρώμα αποτελούν τα σχόλια του εκάστοτε κώδικα και χρησιμοποιήθηκαν τα αρθρώματα (modules) : pandas, numpy.

Πίνακας 3.4-1:Script 1 της Python. Το script διαβάζει από το αρχείο drugbank.xml (το οποίο περιέχει όλα τα φάρμακα της βάσης δεδομένων DrugBank) και με συγκρίσεις μεταξύ των id της DrugBank και των id των αντιψυχωτικών τα καταγράφει.

```
#! /bin/python2

drugbank_file_path = 'drugbank.xml'
category_to_search = 'Antipsychotic Agents'
path_of_file_to_write = 'id_list.txt'
id_list = []
# first read the file
with open (drugbank_file_path) as inputfile:
    # each line is an element of the 'data' list
    data = inputfile.readlines()

# now parse the read file for the drugs we need
for index, element in enumerate(data):
    # new drugs always start with '<drug ',
    # also make sure that the next element is the id one
    # we need this since there are some <drugbank-id primary="true">DBSALTXXXXXXX</drugbank-id>
    # lines that need to be avoided
    if '<drug ' in element and '<drugbank-id primary="true">' in data[index+1]:
        # now get the id (it is the immediate next element, but do a customary check for it)
        drug_id = data[index+1].lstrip('<drugbank-id primary="true">').rstrip('</drugbank-id>\n')
```

```
    if  category_to_search in element and '<category>' in data[index+1]:
        id_list.append(drug_id)
    # if we match the category, save the id
    #elif category_to_search in element: #
    #    id_list.append(drug_id)



# finally write the list to file
with open(path_of_file_to_write, 'w') as file_to_write:
    for drug_id in id_list:
      file_to_write.write("%s\n" % drug_id)
```

Στη συνέχεια, με το παρακάτω script συλλέχθηκαν τα δεδομένα από το πρώτο script επιλογής και συγκρίθηκαν με τα αποτελέσματα από το target prediction. Μέσω αυτής της επαλήθευσης, ο όγκος των προς ανάλυση δεδομένων περιορίστηκε σε 36 αντιψυχωτικούς παράγοντες, στους οποίους θα ακολουθήσει περαιτέρω μελέτη.

Πίνακας 3.4-2: Script 2 της Python. Σε αυτό το script από το αρχείο drugbank-approved.csv (το οποίο περιέχει όλα τα εγκεκριμένα φάρμακα), κρατήθηκαν μόνο τα εγκεκριμένα αντιψυχωτικά. Έπειτα καταγράφηκαν στο αρχείο matched_id_list.txt

```
#! /bin/python2
import random

drugbank_file_path = 'drugbank-approved.csv'
category_to_search = 'Antipsychotic Agents'
path_of_file_to_write = 'matched_id_list.csv'
matched_id_list = []
other_id_list = []
drug_id = False
with open (drugbank_file_path) as inputfile:
    data = inputfile.readlines()

for index, element in enumerate(data):
    if '<drug ' in element and '<drugbank-approved-id primary="true">' in data[index+1]:
        if drug_id and drug_id not in matched_id_list: other_id_list.append(drug_id)
        drug_id = data[index+1].lstrip('<drugbank-approved-id primary="true">').rstrip('</drugbank-approved-id>\n')
    elif category_to_search in element:
        matched_id_list.append(drug_id)

# finally write the list to file
with open(path_of_file_to_write, 'w') as file_to_write:
    for drug_id in matched_id_list:
        file_to_write.write("%s\n" % drug_id)

sets = []
while (len(other_id_list) > 38):
    new_set = []
    for drug_id in range (38):
        choose = random.randint(0,len(other_id_list))
        new_set.append(other_id_list.pop(choose-1))
    sets.append(new_set)
```

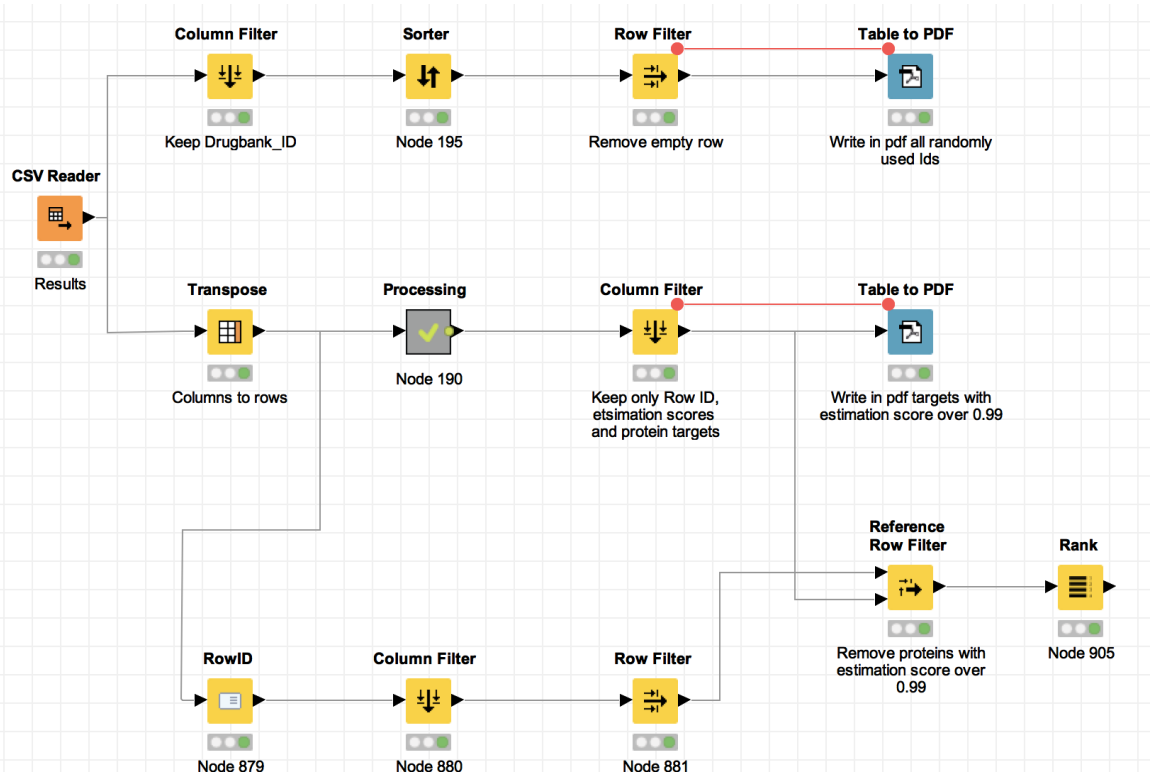
```
with open(path_of_file_to_write, 'w') as file_to_write:
    for set_index, drug_set in enumerate(sets):
        for drug_id in drug_set:
            file_to_write.write("%s\n" % drug_id)
        file_to_write.write('===== set %s \n' % str(set_index+1))
```

Στο δεύτερο μέρος της υπολογιστικής εργασίας, πραγματοποιήθηκε το κομμάτι του εμπλουτισμού,[87] με χρήση του script που αναφέρεται στον Πίνακα 3.4-3. Στο συγκεκριμένο script δημιουργήθηκαν διακόσια datasets των τριανταέξι ids το καθένα (το κάθε dataset έπρεπε να έχει ίδιο αριθμό ids με αυτό των αντιψυχωτικών). Για κάθε ένα από τα datasets έγινε σύγκριση με το dataset των αντιψυχωτικών, όσον αφορά στην κάθε στήλη (πρωτεΐνη στόχο). Σε κάθε σύγκριση έχοντας έναν μετρητή, κρατήθηκε το dataset που υπερίσχυε (όπως περιγράφεται στον υπολογισμό του εμπλουτισμού). Στη συνέχεια διαιρέθηκε το πόσες φορές εμφανίστηκε το dataset που θέλαμε με το συνολικό αριθμό των datasets, έτσι ώστε να ολοκληρωθεί το επόμενο στάδιο της εργασίας που ήταν ο εμπλουτισμός (drug enrichment). Έπειτα με τυχαίο τρόπο συγκρίθηκαν τα estimation scores των datasets όλων των φαρμάκων με τα τα estimation scores του dataset των αντιψυχωτικών, κρατώντας τα estimation scores για κάθε dataset. Οι πρωτεΐνες στόχοι με estimation score εγγύτερα στο 1 θεωρήθηκε πως προσδένουν καλύτερα, ενώ εκείνες με τιμές που προσεγγίζουν την τιμή 0 δεν παρουσιάζουν την επιθυμητή πρόσδεση. Με βάση τους υπολογισμούς που πραγματοποιήθηκαν μεγαλύτερες πιθανότητες πρόσδεσης στις πρωτεΐνες στόχους είχαν τα datasets με score εγγύτερα στο 1, γι' αυτό συγκρατήθηκαν προς μελέτη οι πρωτεΐνες στόχοι με estimation scores μεγαλύτερα από 0,99.

Πίνακας 3.4-4: Script 3 της Python. Αρχικά δημιουργήθηκαν με τυχαίο τρόπο και συγκρίθηκαν τα διακόσια datasets όλων των φαρμάκων με το dataset των αντιψυχωτικών και κρατήθηκαν τα estimation scores του κάθε dataset. Τέλος, τα dasatets αυτά καταγράφηκαν στο αρχείο results.csv

```
#! /bin/python2


import pandas
import numpy

number_of_desired_samples = 200
antipsychotics_file = '200.csv'
target_prediction_file = '256.csv'
database_file = 'drugbank-approved-1 .csv'
output_file = 'results.csv'

database = pandas.read_csv(database_file)
antipsychotics_dataframe = pandas.read_csv(antipsychotics_file, header = None, names = ['antipsychotics'])
target_prediction = pandas.read_csv(target_prediction_file)

# create a Series of the approved agents for comparing
approved_antipsychotics = antipsychotics_dataframe.antipsychotics.values

# the length of the randomly created samples will equal the antipsycho dataset
approved_agents_num = len(approved_antipsychotics)
```

```
# create a dataframe keeping only the antipsychotics
approved_antipsychotics_dataframe = database.loc[database.ID.isin(approved_antipsychotics)]

# remove the antipsychotic agents from the db file to avoid using antipsycho agents in the random samples
for antipshychotic in approved_antipsychotics:
    database = database.loc[database.ID != antipshychotic]

# create number_of_desired_samples sets of random agents. each sample will have agents
# equal to the number of antiphsychotic agents
sets_list = []
for sample_no in range(number_of_desired_samples):
    sets_list.append(database.sample(approved_agents_num))

# just store one row per set. This row will be the sum of all rows & the resulting row will be a pandas Series
sets_sum_list = []
for random_set in sets_list:
    sets_sum_list.append(random_set.sum())

# do the same for the approved_antipsychotics
approved_antipsychotics_sums = approved_antipsychotics_dataframe.sum()

# finally concatenate the sums into one dataframe. this will be the final dataframe
results = approved_antipsychotics_sums
for set_sum in sets_sum_list:
    results = pandas.concat([results,set_sum], axis=1)
results = results.transpose()
results.index = numpy.arange(len(results))


# do a copy so as not to lose the results original file
comparison_dataframe = results.copy()

# compare each value of each row of the dataframe to the value of the antipsychotic sample. results will be
True/False
for comparison_candidate in comparison_dataframe[1:].iterrows():
    comparison_dataframe.loc[comparison_candidate[0]] = (comparison_dataframe.iloc[0] >
comparison_dataframe.loc[comparison_candidate[0]])

# if the random sample wins, set the value = 0, else 1
comparison_dataframe = comparison_dataframe.applymap(lambda x: 1 if x else 0)

# for each column get the average value and append it to a list
estimation_score = []
for column in comparison_dataframe:
    if column != 'ID':
        estimation_score.append(comparison_dataframe[column].mean())

# convert the estimation_score list to a pandas dataframe object in order to add it to the results Dataframe
estimation_score.insert(0,numpy.nan)
estimation_score = pandas.DataFrame(estimation_score).transpose()
estimation_score.columns = results.columns
estimation_score.index = ['estimations']
# do the final addition
results = pandas.concat([results,estimation_score])
# export to csv
results.to_csv(output_file)
```



Σχήμα 3.4-2: Διάγραμμα ροής του Knime, του δεύτερου μέρους της εργασίας.

4. Αποτελέσματα και συζήτηση

4.1 Αποτελέσματα Διπλωματικής εργασίας

Από το λογισμικό Knime εξήχθησαν τελικά οι πρωτεΐνες στόχοι με τα προβλεπόμενα σκορ πρόσδεσής τους. Με βάση τους υπολογισμούς που πραγματοποιήθηκαν μεγαλύτερες πιθανότητες πρόσδεσης στις πρωτεΐνες στόχους είχαν τα datasets με score εγγύτερα στο 1. Συγκεκριμένα, επιλέχθηκαν εκείνες με estimation score μεγαλύτερο του 0,99, οπότε προέκυψε ο ακόλουθος πίνακας.

Πίνακας 4.1-1: Αποτελέσματα, πρωτεϊνών στόχων με τα προβλεπόμενα σκορ πρόσδεσης

Estimations	Protein targets
1	5-hydroxytryptamine receptor 1A
1	5-hydroxytryptamine receptor 1F
1	5-hydroxytryptamine receptor 2A
1	5-hydroxytryptamine receptor 2B
1	5-hydroxytryptamine receptor 2C
1	5-hydroxytryptamine receptor 3A
0.995	5-hydroxytryptamine receptor 5A
1	5-hydroxytryptamine receptor 6
1	5-hydroxytryptamine receptor 7
1	Sodium-dependent serotonin transporter
1	Alpha-1D adrenergic receptor
1	Alpha-2A adrenergic receptor
1	Alpha-2B adrenergic receptor
1	Alpha-2C adrenergic receptor
1	Histamine H1 receptor
1	Histamine H2 receptor
1	Histamine H4 receptor
1	Multidrug resistance-associated protein 1
0.99	Multidrug resistance protein 1
1	Cytochrome P450 2D6
1	D(1A) dopamine receptor
1	D(1B) dopamine receptor
1	D(2) dopamine receptor
1	D(3) dopamine receptor
1	D(4) dopamine receptor
1	Sodium-dependent dopamine transporter
1	Dual specificity mitogen-activated protein kinase kinase 1
1	Tyrosine-protein kinase Fyn
1	Tyrosine-protein kinase YES
0.99	Serine/threonine-protein kinase PLK3
1	Fibroblast growth factor receptor 1
1	Sodium- and chloride-dependent glycine transporter 1
1	Muscarinic acetylcholine receptor M1
1	Muscarinic acetylcholine receptor M2

0.995	Muscarinic acetylcholine receptor M3
1	Muscarinic acetylcholine receptor M4
1	Muscarinic acetylcholine receptor M5
1	Sodium-dependent noradrenaline transporter
1	Orexin receptor type 2
1	Sigma non-opioid intracellular receptor 1
1	Potassium voltage-gated channel subfamily H member 2
1	Voltage-dependent T-type calcium channel subunit alpha-1G
0.995	Bcl2 antagonist of cell death
1	NADPH oxidase 1

- 5-hydroxytryptamine receptors

Οι 5-hydroxytryptamine (5-HT) υποδοχείς αποτελούν τους νευρώνες σεροτονίνης και ταξινομούνται στο νευρικό σύστημα στην περιοχή του εγκεφαλικού φλοιού, τον ιππόκαμπο, το διάφραγμα και την αμυγδαλή. Οι περιοχές αυτές του εγκεφάλου αντιπροσωπεύουν περιοχές που σχετίζονται με διάφορους τομείς της γνωστικής λειτουργίας. [88] Η 5-HT νεύρωση είναι διάχυτη με εκτείνεις διακλαδώσεις, οι οποίες φαίνεται να επηρεάζουν εκτενέστερο μέρος νευρώνων του εγκεφάλου από αυτό που αναφέρθηκε αρχικά. Οι σεροτονίνες σημαίνονται από 14 υποκατηγορίες υποδοχέων με διαφορετικές λειτουργικές και μεταβιβαστικές ιδιότητες, οι οποίες ταξινομούνται σε 7 κύριες ομάδες τις 1, 2, 3, 4 , 5, 6 και με υποκατηγορίες αντίστοιχα a, b και c.[89]

Ορισμένοι από αυτούς τους υποδοχείς ενδέχεται να διαδραματίσουν σημαντικό ρόλο κατά τη δημιουργία ψυχιατρικών ασθενειών.

Καταρχάς, ο υποδοχέας σεροτονίνης 1A (5-HT1A) αποτελεί τον κύριο μεσολαβητή για τη σηματοδότηση της ενεργότητας των σεροτονινών στο κεντρικό νευρικό σύστημα. Ο υποδοχέας αυτός έχει συσχετιστεί με τη γνωστική λειτουργία[88] και την κατάθλιψη[90] ενώ ο υποδοχέας 5-HT_{1F} έχει συσχετιστεί με την ημικρανία.[91]

Η οικογένεια υποδοχέων 5-HT_2, διαπιστώνεται πως εμφανίζει ένα ευρύ φάσμα παθολογίας, καθώς οι υποδοχείς αυτοί συσχετίζονται με ινώσεις και φλεγμονές, με το καρδιαγγειακό σύστημα, τον καρκίνο, το σκελετικό σύστημα και με διαταραχές του κεντρικού νευρικού συστήματος, ενώ δεν παραλείπεται εκτενέστερη αναφορά σε μελλοντική έρευνα αυτών στον τομέα του αιμοποιητικού συστήματος, της Amyotrophic lateral sclerosis και της παχυσαρκίας με το μεταβολικό σύνδρομο.[92]

Συγκεκριμένα, ως προς τις διαταραχές του κεντρικού νευρικού συστήματος, οι υποδοχείς 5-HT_2 εμφανίζουν συσχέτιση με την κατάθλιψη, τη ψύχωση, τον εθισμό, τη σίτιση και την παρορμητικότητα. Για παράδειγμα, ο πολυμορφισμός του γονιδίου HTR2A το οποίο

κωδικοποιείται για τον 5-HT_{2A} υποδοχέα, έχει συσχετιστεί με μείζονα καταθλιπτική ασθένεια (Major Depressive illness, MDD)[93] και με τη δραστικότητα των θεραπειών με αντικαταθλιπτικά φάρμακα.[94] Επίσης, οι 5-HT_{2C} και 5-HT_{2B} υποδοχείς διαμορφώνουν αντιψυχωτικές αποκρίσεις, ενώ άτυπα αντιψυχωτικά που συμπεριφέρονται ανταγωνιστικά ως προς τον 5-HT_{2A} υποδοχέα μπορεί να παρουσιάζουν αποτελεσματικότητα έναντι των αρνητικών συμπτωμάτων, δίχως να προκαλούν περαιτέρω διαταραχές.[95]

Επιπρόσθετα, οι παραλλαγές του γονιδίου HTR3A οι οποίες κωδικοποιούνται ως υπομονάδα Α του υποδοχέα 5-HT3, βρέθηκε πως σχετίζονται με τη διπολική διαταραχή και την σχιζοφρένεια.[96]

Αντίστοιχα, ο πολυμορφισμός του γονιδίου 5-HT5A που κωδικοποιείται ως ο υποδοχέας 5-HT_{5A}, βρέθηκε πως σχετίζεται με την σχιζοφρένεια[97] και μείζονες διαταραχές όπως την ψύχωση,[98] ενώ έχουν πραγματοποιηθεί εκτενείς μελέτες όπου τελικά συσχετίζονται οι 5-HT_6 και 5-HT_7 υποδοχείς με την σχιζοφρένεια.[99]

- Sodium-dependent serotonin transporter

Ο μεταφορέας σεροτονίνης (serotonin transporter, SERT ή 5-HTT) είναι γνωστός και ως νάτριο-εξαρτώμενος μεταφορέας σεροτονίνης (Sodium-dependent serotonin transporter) και αποτελεί μια ανθρώπινη πρωτεΐνη η οποία κωδικοποιείται από το γονίδιο SLC6A4.

Η δυσλειτουργική σηματοδότηση της σεροτονίνης έχει συνδεθεί με την παθογένεια του αυτισμού, των ψυχαναγκαστικών διαταραχών,[100] των διαταραχών στη διάθεση και με την σχιζοφρένεια, ενώ η υπολειτουργία αυτής σχετίζεται με την παθογένεια κατάθλιψης,[101] άγχους, ψυχαναγκαστικών διαταραχών (obsessive compulsive disorder, OCD) και ψύχωση. Γι αυτόν το λόγο, το άγχος και η κατάθλιψη αντιμετωπίζονται με SSRI's τα οποία αναστέλλουν την μεταφορά σεροτονίνης (5-HTT), ενώ η ψύχωση ελέγχεται με φάρμακα τα οποία εμποδίζουν τους υποδοχείς σεροτονίνης και ντοπαμίνης.[102] Συμπληρωματικά, ο γενετικός πολυμορφισμός και η επιγενετική διαταραχή του 5-HTT εμπλέκεται στην παθογένεια ψυχικών ασθενειών,[103] με χαρακτηριστικό παράδειγμα τον πολυμορφισμό του 5-HTT σε 5-HTTLPR.[102]

- Alpha Adrenergic receptors

Οι αδρενεργικοί υποδοχείς, στους οποίους συμπεριλαμβάνονται οι α- , οι β και οι υποδοχείς ντοπαμίνης, ανήκουν σε μια ευρεία οικογένεια, πρωτεϊνης G, επτά διαμεμβρανικών υποδοχέων.

Οι άλφα υποδοχείς ευρίσκονται στα λεία μυϊκά κύτταρα των αγγείων, στις προσυναπτικές νευρικές απολήξεις στους αεραγωγούς, και στους υποβλεννογόνιους αδένες, όπου μπορεί να εισφέρουν στη θερμική εξισορρόπηση του εισπνεόμενου αέρα.

Οι α1-υποδοχείς αποτελούν τους κλασσικούς μετασυναπτικούς υποδοχείς των μυϊκών κυττάρων των αγγείων και μεταφέρουν την αγγειοσυσπαστική δράση της νοραδρεναλίνης. Οι α2-υποδοχείς των νευρώνων εντοπίζονται, κυρίως, προσυναπτικά και αποτελούν το κατασταλτικό σκέλος του μηχανισμού ρύθμισης, μέσω του οποίου, η νοραδρεναλίνη αυτοκαταστέλλει την απελευθέρωση της στις νευρικές απολήξεις του συμπαθητικού συστήματος. Η υποδιαίρεση των α-αδρενεργικών υποδοχέων σε α1 και α2 υπότυπους έχει δείξει ότι μόνο οι εκλεκτικοί α1-αδρενεργικοί αποκλειστές είναι δυνητικά χρήσιμοι.[104]

Η πιο λεπτομερής υποδιαίρεση των α1-αδρενεργικών υποδοχέων σε α1A, α1B και 1D υπότυπους έχει οδηγήσει στην ανάπτυξη περισσότερο ή λιγότερο εκλεκτικών ανταγωνιστών με βάση αυτούς τους υποπληθυσμούς υποδοχέων. Ωστόσο, τέτοιοι νέοι παράγοντες δεν έχουν επιδείξει ρεαλιστική βελτίωση στην αντιυπερτασική φαρμακευτική θεραπεία. Για τους εκλεκτικούς ανταγωνιστές των α1A-αδρενεργικών υποδοχέων, όπως η ταμσουλοσίνη, η αλφουζοσίνη και η τεραζοσίνη, έχει δειχθεί μια μέτρια εκλεκτικότητα για τους α1-αδρενεργικούς υποδοχείς στον προστάτη και αυτά τα φάρμακα χρησιμοποιούνται σήμερα για την βελτίωση της ροής των ούρων σε ασθενείς με καλοήθη υπερπλασία του προστάτη.[104]

Επομένως οι α1 υποδοχείς σχετίζονται με την αγγειοσύσπαση, την αύξηση της περιφερικής αντίστασης, την αύξηση της αρτηριακής πίεσης, τη μυδρίαση, την επίταση της σύγκλεισης έσω σφιγκτήρα ουροδόχου κύστης. Επίσης, οι α2 υποδοχείς σχετίζονται με την αναστολή απελευθέρωσης νορεπινεφρίνης και την αναστολή απελευθέρωσης ινσουλίνης.

Για τον α2 αδρενεργικό υποδοχέα έχουν ταυτοποιηθεί τρία διακριτά γονίδια τα οποία διακρίνονται αντίστοιχα ως α2A, α2B και α2C.

Ο αδρενεργικός υποδοχέας α2C (Alpha-2C Adrenergic receptor, ADRA2C), βρίσκεται στο χρωμόσωμα 4p16.3 και αποτελεί ένα υποψήφιο γονίδιο για τη σχιζοφρένεια[105] καθώς δεσμεύει την κλοζαπίνη (clozapine),[106] ένα άτυπο νευροληπτικό χρήσιμο για την ανθεκτική θεραπεία της σχιζοφρένειας. Επιπλέον, ο ADRA2C δεσμεύει την κλονιδίνη η οποία συνταγογραφείται για τρεις ψυχιατρικές ασθένειες.[107] [108]

Επομένως, φάρμακα που δρουν μέσω του α2C υποδοχέα παρουσιάζουν θεραπευτικές ιδιότητες σε διαταραχές που σχετίζονται με αυξημένες ξαφνικές κρίσεις, όπως η σχιζοφρένεια, η διαταραχή ελλειμματικής προσοχής, η διαταραχή μετατραυματικού στρες αλλά και τις διαταραχές που παρουσιάζονται κατά την απόσυρση ενός φαρμάκου σε ασθενή με χρόνια χρήση. Εκτός από τον α2C υποδοχέα, ο α2A υποδοχέας διαδραματίζει εξίσου σημαντικό ρόλο στη διαμόρφωση καταστάσεων συμπεριφοράς, καθώς μπορεί να συμβάλλει με προστατευτικό ρόλο σε ορισμένες μορφές κατάθλιψης και άγχους.[109]

Από την άλλη μεριά, ο α2B υποδοχέας εντοπίζεται στις λείες μυϊκές ίνες των αγγείων, στους νεφρούς και σε μικρό ποσοστό στον εγκέφαλο, με δεδομένα υπέρ της συμμετοχής του στην παθογένεια της παχυσαρκίας.[110] Επίσης, αποτελεί στόχο του φαρμάκου Asenapine το οποίο χρησιμοποιείται στη θεραπεία της ψύχωσης, της σχιζοφρένειας και τη διπολικής διαταραχής.

- Histamine receptor

Στον εγκέφαλο, οι επιδράσεις της ισταμίνης εκφράζονται μέσω τεσσάρων ειδών υποδοχέων (H1, H2, H3, H4), οι οποίοι έχουν οριστεί με τη βοήθεια λειτουργικών δοκιμασιών με τον μετέπειτα σχεδιασμό των επιλεκτικών αγωνιστών και ανταγωνιστών και την κλωνοποίηση των γονιδίων τους.[111] Και οι τέσσερος ανήκουν στην οικογένεια των υποδοχέων με επτά διαμεμβρανικές περιοχές και συνδέεται με γουανυλυλίου νουκλεοτίδια ευαίσθητες G-πρωτεΐνες.

Ο αποκλεισμός των υποδοχέων H1 στον εγκέφαλο πιθανώς σχετίζεται με κατασταλτικές, παχυντικές και σπασμολυτικές ιδιότητες πολλών φαρμάκων που εμφανίζουν υψηλή συσχέτιση με τον H1 υποδοχέα. Δεδομένα που ελήφθησαν από H1 υποδοχείς ποντικών στηρίζουν σε μεγάλο βαθμό το ρόλο του H1 υποδοχέα ως προς τη διέγερση και τη γνωστική λειτουργία,[112] το άγχος και την επιθετικότητα,[113] την αλγαισθησία,[114] την αντιεπιληπτική συμπεριφορά και τη ρύθμιση της πρόσληψης τροφής[115] και του σωματικού βάρους.[116]

Επίσης, ο διεισδυτικός εγκεφαλικός ανταγωνιστής του H2 υποδοχέα, η ζολαντιδίνη, έχει χρησιμοποιηθεί για να διερευνηθεί η εμπλοκή των υποδοχέων H2 σε διάφορες νευροχημικές και συμπεριφορικές αποκρίσεις.[117] Ένας αριθμός τρικυκλικών αντικαταθλιπτικών αποτελεί έναν πολύ ισχυρό αναστολέα του H2 υποδοχέα, όπου είναι συνδεδεμένη η αδενυλυλοκυκλάση επί των μεμβρανών του εγκεφάλου, ενώ δεν ενδείκνυται για την προετοιμασία άθικτων κυττάρων. Επιπλέον, η ιδέα πως τα αντικαταθλιπτικά αντλούν την κλινική αποτελεσματικότητα τους από τον αποκλεισμό

του εγκεφαλικού υποδοχέα Η2 φαίνεται απίθανη καθώς αντίστοιχος αποκλεισμός δεν παρατηρήθηκε μετά από χρόνια θεραπείες.[118]

Τέλος, όσον αφορά τον Η4 υποδοχέα, η δομή του γονιδίου του, η κωδική αλληλουχία και η φαρμακευτική του δράση σχετίζεται σαφώς με τις αντίστοιχες του Η3 υποδοχέα. Οι ανταγωνιστές που έχουν σχεδιασθεί μέχρι σήμερα παρουσιάζουν κυρίως αντιφλεγμονώδεις ιδιότητες, ενώ το υψηλότερο επίπεδο έκφρασης του δεν παρουσιάστηκε στον εγκέφαλο αλλά στον αιμοποιητικό ιστό και στα αιμοποιητικά κύτταρα (μυελό των οστών, σπλήνα, λευκοκύτταρα).[118] Επομένως με αυτόν τον τρόπο ορίζεται ο κύριος ρόλος του στο ανοσοποιητικό σύστημα, με εφαρμογές στο άσθμα, τις αλλεργίες, τον χρόνιο κνησμό και τις αυτοάνοσες νόσους. Οι Η4 υποδοχείς ρυθμίζουν τη μετανάστευση των ηωσινόφιλων κυττάρων και πραγματοποιούν επιλεκτική πρόσληψη των μαστοκυττάρων τα οποία με τη σειρά τους οδηγούν στη διεύρυνση των ανοσολογικών αποκρίσεων της ισταμίνης και τελικά σε χρόνια φλεγμονή.[119] Ο πρωτεϊνικός στόχος Η4 αποτελεί στόχο του αντιψυχωτικού φαρμάκου Clozapine δίχως όμως να έχει φαρμακευτική δράση επί του στόχου.

- Multidrug resistance-associated protein 1 (MRP1/ABCC1)

Η πολυανθεκτική πρωτεΐνη αντίστασης 1 (MRP1/ABCC1) είναι ένας ATP-προσδεδεμένος (ABC) διαμεμβρανικός μεταφορέας με σημαντική κλινική σημασία που προσδίδει ανθεκτικότητα στα φάρμακα σε καρκινικά κύτταρα μειώνοντας τη συσσώρευση του φαρμάκου με ενεργό εκροή. Η MRP1 βρέθηκε ότι παρέχει έλλειψη ανταπόκρισης σε ανθρακυκλίνες, Vinca αλκαλοειδή και στις επιποδοφυλλοτοξίνες. Η πρωτεΐνη MRP1 σχετίζεται με τον καρκίνο[120] παρέχοντας προστασία στον υγιή ιστό από τα αντικαρκινικά φάρμακα,[121] με την επιληψία,[122] με τη νόσο HIV[122] και με τη νόσο Parkinson και Alzheimer.[123]

- Multidrug resistance protein 1 (MDR1)

Το γονίδιο MDR1 (ABCB1) εντοπίζεται στο χρωμόσωμα 7q21.1 και αποτελείται από 28 εξόνια. Η P-γλυκοπρωτεΐνη (P-glycoprotein 1, Pgp), που κωδικοποιείται από το γονίδιο αυτό, είναι μια εξαρτώμενη από ATP αντλία εκροής με ευρεία εξειδίκευση υποστρώματος. Η Pgp βρίσκεται σε μια πληθώρα ιστών υπεύθυνων για την απορρόφηση και απέκκριση ουσιών από το σώμα και είναι έτσι σε θέση να ελέγχει τη διαθεσιμότητα ενός μεγάλου αριθμού φαρμάκων.[124]

Εντοπίζεται κυρίως στα φυσιολογικά επιθηλιακά κύτταρα των επινεφριδίων, των νεφρών, του ήπατος, του παχέος εντέρου, της νήστιδας, του παγκρέατος, των τριχοειδών ενδοθηλιακών κυττάρων στους όρχεις καθώς και στον αιματοεγκεφαλικό φραγμό[125]. Εκφράζεται, επίσης, στα κορυφαία επιθηλιακά κύτταρα του λεπτού και του παχέος εντέρου, καθώς και των νεφρικών σωληναρίων. Μελέτες έχουν δείξει την ύπαρξη υψηλών επιπέδων της πρωτεΐνης στον πλακούντα και το ενδομήτριο εγκύων γυναικών[126]. Το πρότυπο έκφρασης της MDR1 υποδεικνύει ότι υπό φυσιολογικές συνθήκες η κύρια λειτουργία της είναι η προστασία των ζωτικών οργάνων στα οποία εντοπίζεται από ξενοβιώτες. Ιδιαίτερα αυξημένα επίπεδα της P γλυκοπρωτεΐνης εντοπίζονται σε ανθεκτικά στην χημειοθεραπεία καρκινικά κύτταρα, ειδικά αν προέρχονται από κύτταρα που εκφράζουν και στην φυσιολογική κατάστασή τους το μόριο αυτό.[127]

Σύμφωνα με μελέτες, το γονίδιο MDR1 σχετίζεται με τη σχιζοφρένεια[128] και τις διαταραχές της διάθεσης όπως την κατάθλιψη,[129] ενώ έχει πραγματοποιηθεί μελέτη συσχέτισής του με φάρμακα για το κεντρικό νευρικό σύστημα.[130]

- Cytochrome P450 2D6 (CYP2D6)

Το ηπατικό ένζυμο CYP2D6 του κυτοχρώματος P450 εμπλέκεται στον μεταβολισμό περίπου 25% των φαρμάκων που χρησιμοποιούνται σήμερα στην κλινική πράξη.[131] Εκτός από την μελέτη των πολυμορφισμών του κάθε μεταβολικού ενζύμου, η κατηγοριοποίηση των ατόμων ως «κανονικοί», «αργοί» και «γρήγοροι μεταβολιστές» μπορεί να γίνει και με μια λειτουργική δοκιμή (που καλείται και μελέτη φαινοτύπου) για να αποτιμηθεί το επίπεδο δραστικότητας ενός συγκεκριμένου ενζύμου, το οποίο εμπλέκεται στο μεταβολισμό μιας ομάδας φαρμάκων.

Το CYP2D6 αποτελεί την πιο καλά μελετημένη ισομορφή P450 με πολυμορφισμούς, που επιδρούν στο μεταβολισμό φαρμάκων, είναι υπεύθυνο για την οξείδωση πάνω από 70 διαφορετικών φαρμάκων και έχουν ταυτοποιηθεί τουλάχιστον 95 αλληλόμορφά του. Υποστρώματα του CYP2D6 είναι αντιαρρυθμικά όπως flecainide, mexiletine, propafenone, αντικαταθλιπτικά όπως amitriptyline, aroxetine, venlafaxine, fluoxetine, αντιψυχωσικά όπως clorpromazine, haloperidol, thoridazine, β-αναστολείς όπως labetalol, timolol, propanolol, pindolol, metoprolol) και αναλγητικά φάρμακα όπως codeine, fentanyl, meperidine, oxycodone.[132]

- Dopamine (DA) receptors

Το ντοπαμινεργικό σύστημα αποτελείται από διάφορες ομάδες κυττάρων, οι οποίες γενικά αναφέρονται ως A8 έως A15 και βρίσκεται στο μέσο του εγκεφάλου και τον υποθάλαμο.[133]

Η ντοπαμίνη (DA) αλληλεπιδρά με πέντε διαφορετικούς υποδοχείς, οι οποίοι επισημαίνονται ως D1 έως D5 και υποδιαιρούνται σε δύο κύριες υποκατηγορίες: την D1 (που αποτελείται από τους υποδοχείς D1 και D5) και την D2 (που αποτελείται από τους υποδοχείς D2, D3 και D4).

Οι υποδοχείς ντοπαμίνης D1, D2, D3, D4 και D5 κωδικοποιούνται το γονίδιο DRD1, DRD2, DRD3, DRD4 και DRD5 αντίστοιχα. Τα DRD1 και DRD5 δεν έχουν εσώνια, ενώ τα DRD2, DRD3 και DRD4 περιέχουν έξι, πέντε και τρία εσώνια στις κωδικές περιοχές τους, αντίστοιχα, παρέχοντας παραλλαγές ματίσματος.[134] Το γονίδιο DRD1 βρίσκεται στο χρωμόσωμα 5q35.1, το DRD2 βρίσκεται στο χρωμόσωμα 11q23.1, το DRD3 βρίσκεται στο χρωμόσωμα 3q13.3, το DRD4 βρίσκεται στο 11p15.5 και το DRD5 βρίσκεται στο χρωμόσωμα 4p16.1.

Όλοι οι DA υποδοχείς είναι συζευγμένοι G-πρωτεΐνης υποδοχείς (GPCRs), με την κατηγορία D1 να πραγματοποιεί σύζευξη με το Gs (και ενδεχομένως με το Gq) και την κατηγορία D2 να πραγματοποιεί σύζευξη με το Gi. Οι περισσότεροι από τους DA υποδοχείς βρίσκονται συγκεντρωμένοι εντός των βασικών γαγγλίων, ειδικά το ραχιαίο και κοιλιακό ραβδωτό σώμα, με τον D1 να βρίσκεται και στον προμετωπιαίο φλοιό. Επιπλέον, ο D3 βρίσκεται στο Island of Calleja, ο υποδοχέας D4 στην αμυγδαλή και ο υποδοχέα D5 στον ιππόκαμπο.[135] [136]

Η DA είναι ίσως ένα από τους πιο μελετημένους νευροδιαβιβαστές και είναι γνωστό πως διαδραματίζει σημαντικό ρόλο στην κίνηση και την συντονισμένη κίνηση, στα κίνητρα και την ανταμοιβή, καθώς και στη γνωστική λειτουργία. Επομένως, ενδεχόμενες αλλοιώσεις του ντοπαμινεργικού συστήματος εμπλέκονται σε αρκετές διαταραχές, όπως η νόσος του Parkinson, η εξάρτηση από τα ναρκωτικά και τον εθισμό[137] και την σχιζοφρένεια.[138]

Σύμφωνα με μελέτες το DRD1 έχει συσχετιστεί με την διπολική διαταραχή,[139] με το σύνδρομο της ελλειμματικής προσοχής/υπερκινητικότητας (Attention-Deficit Hyperactivity Disorder - ADHD)[140] και με τον αυτισμό,[141] ενώ γενετικές πολυμορφίες στις λειτουργικές περιοχές του DRD1 διαδραματίζουν ένα ρυθμιστικό ρόλο ως προς το ενδεχόμενο εμφάνισης σχιζοφρένειας ή ψύχωσης.[142]

Επίσης, ο D2 υποδοχέας, σύμφωνα με μελέτες, εμφανίζεται μεγάλη συσχέτιση με την εμφάνιση ψύχωσης[143] και σχιζοφρένειας ενώ νεότερες μελέτες δείχνουν πως από την κατηγορία των D2 υποδοχέων, ο υποδοχέας που χαρακτηρίζεται ως D2High είναι

εκείνος που τελικά που διαδραματίζει σημαντικότερο ρόλο στην εμφάνιση της νόσου, έναντι του D2Low.[144]

Επιπλέον, ο D3 υποδοχέας έχει συσχετιστεί άμεσα με την ψύχωση και την σχιζοφρένεια,[145] ενώ ο πολυμορφισμός αυτού έχει συσχετιστεί με τον αυτισμό[146] και την μονοπολική διαταραχή.[147] Τέλος, οι υποδοχείς D4 και D5(D1B) έχουν συσχετιστεί με την σχιζοφρένεια και το ADHD.[148] [149]

Ασυνήθιστα υψηλές ντοπαμινεργικές μεταδόσεις έχουν συνδεθεί με την ψύχωση και τη σχιζοφρένεια. Και η αύξηση, αλλά και η μείωση της ντοπαμίνης επηρεάζουν τους ψυχωσικούς και σχιζοφρενείς.

Αντιψυχωσικά φάρμακα ενεργούν κυρίως ως ανταγωνιστές της ντοπαμίνης, αναστέλλοντας τα επίπεδα των υποδοχέων ντοπαμίνης. Τα τυπικά αντιψυχωτικά συνηθέστερα δρουν επί των D2 υποδοχέων, ενώ τα άτυπα φάρμακα ενεργούν, επίσης, σε D1, D3 και D4 υποδοχείς. Οι αμφεταμίνες, η μεθαμφεταμίνη και η κοκαΐνη, που μπορεί να αυξήσουν τα επίπεδα της ντοπαμίνης κατά δέκα φορές, μπορεί να προκαλέσουν προσωρινά ψύχωση. Ωστόσο, πολλά μη-ντοπαμινεργικά φάρμακα μπορεί να προκαλέσουν οξεία και χρόνια ψύχωση.

- Sodium-dependent dopamine transporter (DAT)

Ο μεταφορέας της ντοπαμίνης (DAT) είναι μέλος μιας μεγάλης οικογένειας Na^+/Cl^- - εξαρτώμενων νευρωνικών μεταφορέων της πλασματικής μεμβράνης, η οποία επίσης περιλαμβάνει το μεταφορέα της νορεπινεφρίνης (NET) και της σεροτονίνης (SERT) καθώς και τους μεταφορείς των αμινοξέων: γ-αμινοβουτυρικό οξύ (GAT), γλυκίνη, προλίνη, ταυρίνη, κρεατίνη και μπεταϊνη.[150]

Ο DAT θεωρείται ένας ειδικός μάρτυρας για τους ντοπαμινεργικούς νευρώνες καθώς εκφράζεται αποκλειστικά και μόνο στους νευρώνες που συνθέτουν DA. Ο DAT όσο και ο NET έχει αποδειχθεί ότι μπορούν να μεταφέρουν ντοπαμίνη καθώς και νορεπινεφρίνη, δηλαδή ο καθένας τους μεταφέρει εκτός από το δικό του υπόστρωμα και το υπόστρωμα του άλλου.[151]

Ο μεταφορέας της ντοπαμίνης είναι στόχος πολλών νοοτρόπων και ψυχοτρόπων φαρμάκων, η δε συμμετοχή των ντοπαμινεργικών νευρώνων στο κύκλωμα των κοιλιακών βασικών γαγγλίων, αλλά και στη μεσοφλοιϊκή ντοπαμινεργική προβολή, καθιστούν το μεταφορέα της ντοπαμίνης ένα πολύ ενδιαφέροντα στόχο φαρμακολογικών χειρισμών για την ίαση ασθενειών όπως η σχιζοφρένεια, το Parkinson και η κατάθλιψη καθώς και για την απεξάρτηση από χημικές ουσίες.[152]

- Dual specificity mitogen-activated protein kinase kinase 1 (MAP2K1)

Η διπλής εξειδίκευσης ενεργοποιημένη από μιτογόνα πρωτεϊνικών κινασών κινάση 1 (MAP2K1), που ονομάζεται επίσης MEK1, βρίσκεται στο χρωμόσωμα 15q22.1-q22.33 και κωδικοποιεί μια πρωτεϊνική κινάση MEK1, που είναι γνωστός ως ο μετέπειτα στόχος του RAF και είναι ανάντι του ERK1 και ERK2. Οι μεταλλάξεις της MAP2K1 συμβαίνουν σχεδόν πάντα στα εξόνια 2 και 3 και οι περισσότερες προκαλούν συνεχή ενεργοποίηση της κινάσης MAP2K1. Οι μεταλλάξεις της MAP2K1 έχουν εμπλακεί με την εμφάνιση αρκετών ανθρώπινων καρκίνων συμπεριλαμβανομένων των μελανωμάτων, του πνεύμονα, των αδενοκαρκινωμάτων του παχέος εντέρου και της λευχαιμίας τριχωτών κυττάρων.[153] Επιπρόσθετα, σύμφωνα με μελέτες, η μεθυλίωση του DNA του γονιδίου MEK1 (ενός υποκινητή CpG νησίδας που βρίσκεται περίπου 30 kb ανοδικά του γονιδίου που κωδικοποιεί τη MEK1), συσχετίστηκε με την εφόρου ζωής χρήση αντιψυχωτικών φαρμάκων.[154]

- Tyrosine-protein kinase Fyn

Η πρωτεϊνική κινάση τυροσίνης Fyn, αποτελεί μέλος της οικογένειας των Src κινάσων και εκφράζεται έντονα στον εγκεφαλικό ιστό και τα κύτταρα του αίματος. Η Fyn συμμετέχει στην ανάπτυξη του εγκεφάλου, στη συναπτική διαβίβαση μέσω της φωσφορυλίωσης των υπομονάδων των υποδοχέων NMDA και στη ρύθμιση της συναισθηματικής συμπεριφοράς. Η Fyn απαιτείται για την μεταγωγή σήματος στους ραβδωτούς νευρώνες η οποία ενεργοποιείται από την αλοπεριδόλη. Επίσης, ασθενείς που πάσχουν από σχιζοφρένεια εμφανίζουν ανωμαλίες στην Fyn.[155] [156]

- Tyrosine-protein kinase YES

Η πρωτεϊνική κινάση τυροσίνης YES, αποτελεί μέλος της οικογένειας των Src κινάσων, ενώ συμμετέχει στη γονιμοποίηση, τη διαφοροποίηση και την αγγειακή συστολή.[157] Στη βιβλιογραφία συσχετίζεται κυρίως με διάφορες μορφές καρκίνου.[158]

- Serine/Threonine-protein kinase PLK3

Οι Polo-like κινάσες (Polo-Like Kinases, PLK) ανήκουν στην οικογένεια των κινάσων σερίνης και θρεονίνης, ενώ διαδραματίζουν πολλαπλούς ρόλους στον κυτταρικό κύκλο.

Η PLK3 κινάση εκφράζεται σε όλο τον κυτταρικό κύκλο και αυξάνεται από G1 σε μίτωση. Η έκφρασή της ρυθμίζεται αυξητικά σε εντόνως πολλαπλασιαστικούς όγκους των ωοθηκών και του καρκίνου του μαστού και συσχετίζεται με τις δυσμενέστερες προγνώσεις. Επιπρόσθετα, στη ρύθμιση της μίτωσης, η PLK3 πιστεύεται πως εμπλέκεται στον κατακερματισμό Golgi κατά τη διάρκεια του κυτταρικού κύκλου και στην αντιμετώπιση ενδεχόμενης βλάβης του DNA.[159] Επίσης, η κινάση PLK3 συσχετίζεται με νευροεκφυλιστικές διαταραχές,[160] όπως τη νόσο του Parkinson.[161]

- Fibroblast growth factor receptor 1 (FGFR1)

Οι αυξητικοί παράγοντες των ινοβλαστών (FGFs) στα θηλαστικά σχηματίζουν μια οικογένεια 21 συγγενικών πρωτεϊνών (FGF1-21) που συνδέονται σε τουλάχιστον 4 τύπους υποδοχέων FGF (FGFR1-4) ρυθμίζοντας κρίσιμες βιολογικές διαδικασίες όπως η κυτταρική αύξηση και διαφοροποίηση. Οι FGFs είναι μονομερή πολυπεπτίδια, τα οποία διμερίζονται και συνδέονται σε δυο υποδοχείς, επάγοντας το διμερισμό τους. Για το διμερισμό και την ενεργοποίηση των υποδοχέων FGF-Rs απαιτείται και η πρόσδεση της ηπαρίνης στην εξωκυτταρική περιοχή των υποδοχέων.

Μεταλλάξεις του υποδοχέα FGFR1 οδηγούν σε νεφρική δυσπλασία και αγενεσία.[162] [163] Επίσης, ο υποδοχέας FGFR1 ανευρίσκεται επίσης μεταλλαγμένος στο σύνδρομο Kallmann, όπου έχει διαπιστωθεί ότι η σύνδεσή του με τους FGFs διαμεσολαβείται από την ανοσμίνη και τις πρωτεογλυκάνες HSPGs.[164]

Ακόμα, ο FGFR1 συσχετίζεται με διάφορες μορφές καρκίνου[165] όπως τον καρκίνο του μαστού,[166] με το σύνδρομο Pfeiffer[167] και με διαταραχές του κεντρικού νευρικού συστήματος, όπως την κατάθλιψη,[168] τη διπολική διαταραχή και την σχιζοφρένεια.[169]

- Sodium- and Chloride-dependent glycine transporter 1 (GlyT-1)

Ο νάτριο- και χλωριούχο-εξαρτώμενος διαβιβαστής γλυκίνης 1 (Sodium- and Chloride-dependent glycine transporter 1, GlyT-1) αποτελεί μια πρωτεΐνη η οποία κωδικοποιείται στο γονίδιο SCL6A9. Μελέτες υβριδισμού έχουν αποκαλύψει ότι η GlyT-1 εκφράζεται ευρέως σε όλον τον εγκέφαλο. Επίσης η GlyT-1 συσχετίζεται με διαταραχές λόγω χρήσης μεθαμφεταμίνης,[170] με την ψύχωση[171] και την σχιζοφρένεια.[172]

- Muscarinic acetylcholine receptor (mAChR)

Οι μουσκαρινικοί υποδοχείς (Muscarinic acetylcholine receptor , mAChR) παρουσιάζουν δομικές διαφορές σε σχέση με τους νικοτινικούς υποδοχείς και ανήκουν στη μεγάλη οικογένεια των G-συζευγμένων με πρωτεΐνη υποδοχέων. Υπάρχουν πέντε υποκατηγορίες μουσκαρινικών υποδοχέων, οι M1, M2, M3, M4 και M5. Οι υποδοχείς M1, M3 και M5 πραγματοποιούν θετική σύζευξη με την G πρωτεΐνη Gq/11 και επακολουθεί ενεργοποίηση της φωσφολιπάσης Cb και κινητοποίηση του ασβεστίου. Οι μουσκαρινικοί υποδοχείς M2 και M4 αναστέλλουν την αδενυλική κυκλάση και το σχηματισμό cAMP καθώς αλληλεπιδρούν με τους ιοντικούς διαύλους.

Οι μουσκαρινικοί υποδοχείς διεγείρονται από την μουσκαρίνη και την ακετυλοχολίνη και δεσμεύονται από την ατροπίνη. Οι μουσκαρινικοί υποδοχείς βρίσκονται στο κεντρικό νευρικό σύστημα και στο περιφερικό νευρικό σύστημα, στην καρδιά, τους πνεύμονες, την άνω γαστρεντερική οδό και στους ιδρωτοποιούς αδένες.

Επίσης, οι μουσκαρινικοί υποδοχείς M1, M2, M3, M4 και M5 έχουν συσχετιστεί με ψυχωτικές διαταραχές.[173] Συγκεκριμένα, οι M1 και M4 υποδοχείς έχουν συσχετιστεί με τη νόσο του Alzheimer και με την σχιζοφρένεια,[174] ενώ δεν εντοπίστηκαν μεταβολές στη δομή των υποδοχέων M2 και M3 υποκείμενων που έπασχαν από σχιζοφρένεια,[175] παρόλα αυτά αποτελεί στόχο των αντιψυχωτικών φαρμάκων. Από την έλλειψη των υποδοχέων M5, είναι πιθανό να δημιουργηθούν ελλείμματα συμπεριφοράς καθώς η ενεργοποίηση τους είναι γνωστό πως διεγείρει την απελευθέρωση της ντοπαμίνης στον επικλινή πυρήνα. Επομένως, ο μουσκαρινικός υποδοχέας M5 εμπλέκεται στην ρύθμιση αρκετά σημαντικών φαρμακολογικών λειτουργιών και της συμπεριφοράς.[176]

- Sodium- dependent noradrenaline transporter (NAT1)

Ο νάτριο- εξαρτώμενος διαβιβαστής νοραδρεναλίνης (Sodium- dependent noradrenaline transporter, NAT1), γνωστός και ως διαβιβαστής νορεπινεφρίνης (NET), κωδικοποιείται στο γονίδιο SLC6A2.

Η Νοραδρεναλίνη είναι μια μονοαμίνη-νευροδιαβιβαστής που παράγεται από τους νοραδρενεργικούς νευρώνες. Το κυτταρικό σώμα τους βρίσκεται κυρίως στον υπομέλα τόπο το οποίο νευροανατομικά εντοπίζεται στη γέφυρα του εγκεφαλικού στελέχους. Από εκεί οι νευράξονες των νοραδρενεργικών νευρώνων φέρονται σε διάφορες περιοχές του εγκεφάλου και του νωτιαίου μυελού. Επιπλέον, η νοραδρεναλίνη εκκρίνεται σε νευροδραστικές συνάψεις με περιφερικούς ιστούς-στόχους από τους μεταγαγγλιακούς νευρώνες του συμπαθητικού αυτόνομου νευρικού συστήματος. Η νοραδρεναλίνη δεσμεύεται από ειδικούς υποδοχείς, είτε στον τελικό νευράξονα (προσυναπτικοί), είτε

στον μετασυναπτικό νευρώνα ή ιστό (μετασυναπτικοί). Έτσι φέρονται εις πέρας μεταβολές στη νευρωνική και ιστική λειτουργία, που εκδηλώνονται με αλλαγές σε διάφορες ανθρώπινες λειτουργίες.

Ο νευροδιαβιβαστής νοραδρεναλίνης διαδραματίζει σημαντικό ρόλο στην ανθρώπινη φυσιολογία και παθολογία, καθώς εμπλέκεται στη ρύθμιση της διάθεσης, τη ρύθμιση του ύπνου, την έκφραση της συμπεριφοράς και την σε γενικό βαθμό επαγρύπνηση και εγρήγορση.[177] Επίσης, έχει συσχετιστεί η δράση του με τη λειτουργία των αντικαταθλιπτικών φαρμάκων[178] και φαρμάκων για τη νόσο του Parkinson.[179]

Γενικότερα, τα γονίδια μονοαμινικών νευροδιαβιβαστών έχουν προταθεί ως υποψήφια γονίδια στην παθογένεση διαφόρων νευρολογικών και ψυχιατρικών διαταραχών, συμπεριλαμβανομένης της κατάθλιψης, της σχιζοφρένειας, του άγχους και της κατάχρησης ουσιών.[180]

- Orexin receptor type 2 (OX2)

Ο υποδοχέας ορεξίνης τύπου 2 (Orexin receptor type 2, Ox2R ή OX2), γνωστός και ως υποδοχέας υποκρετίνης τύπου 2 (Hypocretin Receptor Type 2, HCRTR2), είναι μια πρωτεΐνη που κωδικοποιείται από το γονίδιο HCRTR2.

Ο OX2 είναι ένας υποδοχέας συζευγμένος με G-πρωτεϊνη η οποία εκφράζεται αποκλειστικά στον εγκέφαλο, ενώ δεσμεύει τόσο νευροπεπτίδια ορεξίνης Α όσο και Β. Επίσης, ο OX2 σχετίζεται με τον κεντρικό μηχανισμό ανάδρασης που ρυθμίζει τη διατροφική συμπεριφορά, ενώ έχει συσχετιστεί με διαταραχές του ύπνου, όπως η ναρκοληψία.[181]

Οι υποκρετίνες συμμετέχουν στη λειτουργία των νευροενδοκρίνων και των αντιδράσεων του στρες (διεγείρουν τον υποθάλαμο-υπόφυση-επινεφρίδια), στην κεντρική αντίληψη του πόνου (αναλγητικό), στις αυτόνομες λειτουργίες και στις ισχαιμικές εγκεφαλικές αλλαγές. Ως εκ τούτου, μεταβληθέντα συστήματα υποκρετίνης μπορεί να συμμετέχουν σε ορισμένες νευρολογικές και ψυχιατρικές διαταραχές, γι αυτό και πραγματοποιούνται μελέτες ώστε να διαπιστωθεί το ποσοστό συσχέτισής τους με διαταραχές όπως η σχιζοφρένεια και η κατάθλιψη.[181]

- Sigma non-opioid intracellular receptor 1 (SIGMAR 1)

Ο Sigma μη οπιοειδής ενδοκυτταρικός υποδοχέας 1 (Sigma non-opioid intracellular receptor 1, SIGMAR 1) κωδικοποιείται από το γονίδιο SIGMAR 1, το οποίο κωδικοποιεί

μια πρωτεΐνη υποδοχέα που αλληλεπιδρά με μία ποικιλία ψυχωσεομιμητικών ουσιών, συμπεριλαμβανομένων της κοκαΐνης και των αμφεταμινών. Ο υποδοχέας συμβάλλει στις κυτταρικές λειτουργίες των διαφόρων ιστών που σχετίζονται με το ενδοκρινικό, το ανοσοποιητικό και το νευρικό σύστημα.

Ο SIGMAR 1 υποδοχέας έχει συσχετιστεί με την καρδιαγγειακή λειτουργία, τη σχιζοφρένεια,[182] [183] τη νόσο Huntington,[184] την κλινική κατάθλιψη,[185] τις συνέπειες της κατάχρησης κοκαΐνης και τον καρκίνο.

- Potassium voltage-gated channel subfamily H member 2 (KCNH2)

Το τασεοελεγχόμενο κανάλι καλίου της υποοικογένειας Η μέλους 2 (Potassium voltage-gated channel subfamily H member 2, KCNH2) είναι ένα γονίδιο που κωδικοποιεί πρωτεΐνες.

Η οργανωμένη νευρωνική πυροδότηση είναι ζωτικής σημασίας για την επεξεργασία του φλοιού η οποία διαταράσσεται στην σχιζοφρένεια. Στον ιππόκαμπο ενός σχιζοφρενούς, η έκφραση του KCNH2-3.1 είναι 2,5 φορές μεγαλύτερη από την έκφραση του KCNH2-1A. Μια ανάλυση 5 σετ κλινικών δεδομένων απέδειξαν συσχέτιση των SNPs των KCNH2 με την σχιζοφρένεια.[186]

- Voltage-dependent T-type calcium channel subunit alpha-1G (CACNA1G)

Οι τασο-εξαρτώμενοι δίαυλοι ασβεστίου τύπου Τ της υποομάδας άλφα-1G (Voltage-dependent T-type calcium channel subunit alpha-1G, CACNA1G ή Cav3.1) αποτελούν γονίδια πρωτεϊνικής κωδικοποίησης ενώ κωδικοποιούνται ως γονίδιο CACNA1G.

Οι τασο-εξαρτώμενοι δίαυλοι ασβεστίου (Voltage-sensitive calcium channels, VSCC) διαμεσολαβούν την είσοδο των ιόντων ασβεστίου στα κύτταρα και συμμετέχουν επίσης σε μια ποικιλία διεργασιών όπως την σύσπαση των μυών, την απελευθέρωση ορμόνης ή νευροδιαβιβαστών, τη γονιδιακή έκφραση, την κυτταρική κινητικότητα, την κυτταρική διαίρεση και τον κυτταρικό θάνατο . Η ισόμορφη άλφα-1G προκαλεί ρεύματα ασβεστίου τύπου Τ. Οι δίαυλοι ασβεστίου τύπου Τ ανήκουν στη «χαμηλής τάσης ενεργοποιημένη» ομάδα. Το ρεύμα τύπου Τ μπλοκάρεται από χαμηλές συγκεντρώσεις νικελίου, ενώ όλα τα ρεύματα ασβεστίου μπλοκάρονται από υψηλές συγκεντρώσεις νικελίου ή καδμίου. Οι δίαυλοι τύπου Τ εξυπηρετούν ρυθμιστικές λειτουργίες στους δύο κεντρικούς νευρώνες και στα καρδιακά κομβικά κύτταρα ενώ υποστηρίζουν τη σήμανση του ασβεστίου στα εκκριτικά κύτταρα και τον αγγειακό λείο μυ. Επίσης, μπορούν να εμπλέκονται στη

ρύθμιση της πυροδότησης των νευρώνων που είναι σημαντικοί για την επεξεργασία πληροφοριών, καθώς και σε διεργασίες κυτταρικής ανάπτυξης.

Γενικότερα, οι δίαυλοι ασβεστίου τύπου Τ συσχετίζεται με παθοφυσιολογικές καταστάσεις όπως την επιληψία,[187] τον αυτισμό, την υπέρταση, την κολπική μαρμαρυγή,[188] τη συγγενή καρδιακή ανεπάρκεια,[189] τον πόνο,[187] την ψύχωση και τον καρκίνο.[190]

- Bcl2 antagonist of cell death (Bcl-2)

Ανασταλτικά αποπτωτικά γονίδια ονομάζονται τα γονίδια των οποίων η έκφραση αναστέλλει τον προγραμματισμένο κυτταρικό θάνατο. Στην κατηγορία αυτή συμπεριλαμβάνονται γονίδια που κωδικοποιούν μέλη της οικογένειας πρωτεϊνών Bcl-2. Η οικογένεια των πρωτεϊνών Bcl-2 έχει κεντρικό ρόλο στη ρύθμιση της απόπτωσης και συσχετίζεται με την παθογένεια πολλών παθήσεων.

Η πρωτεΐνη Bcl-2 είναι ένα μέλος μιας πρωτεϊνικής οικογένειας, τα μέλη της οποίας συμμετέχουν στη ρύθμιση του αποπτωτικού προγράμματος στα κύτταρα των θηλαστικών. Προς το παρόν, τουλάχιστον 25 μέλη της οικογένειας Bcl-2 είναι γνωστά, τα οποία μπορούν να αναστέλλουν ή να ενεργοποιούν την απόπτωση. Η ομοιόσταση διατηρείται ελέγχοντας την ποσότητα των ενεργών προ- και αντι-αποπτωτικών μελών της οικογένειας Bcl-2. Ερεθίσματα όπως η καταστροφή του DNA, οδηγούν σε αύξηση της έκφρασης των προ-αποπτωτικών πρωτεϊνών Bcl-2. Αυτό καταστρέφει τη λεπτή ισορροπία ανάμεσα στις προ- και αντι-αποπτωτικές Bcl-2 οδηγώντας στην απόπτωση.[191]

Μέλη της Bcl-2 οικογένειας γονιδιακών πρωτεϊνών τα οποία προάγουν τον προγραμματισμένο κυτταρικό θάνατο περιλαμβάνουν τις πρωτεΐνες bad και bax. Αντίθετα, η έκφραση των γονιδιακών πρωτεϊνών bcl-2 και bcl-xL καταστέλλει το πρόγραμμα απόπτωσης.

Βλάβη στο γονίδιο Bcl-2 έχει ταυτοποιηθεί ως η αιτία ενός αριθμού καρκίνων[192], συμπεριλαμβανομένου του μελανώματος, του καρκίνου του μαστού, του προστάτη, της χρόνιας λεμφοκυτταρικής λευχαιμίας[193] και του καρκίνου του πνεύμονα. Επίσης, η βλάβη αυτή αποτελεί μία πιθανή αιτία της σχιζοφρένειας[155] και της αυτοανοσίας,[194] ενώ αποτελεί ένα παράγοντα αντοχής σε θεραπείες του καρκίνου.

- NADPH oxidase 1 (NADPH)

Η οξειδάση NADPH (υδρογονωμένο φωσφορικό νικοτιναμιδο-αδενινο δινουκλεοτίδιο, nicotinamide adenine dinucleotide phosphate-oxidase, NADPH) αποτελεί ένα συγκρότημα ένζυμων δεσμευμένων σε μεμβράνη που αντικρίζουν τον εξωκυτταρικό χώρο. Επίσης, μπορεί να βρεθεί στην πλασματική μεμβράνη, καθώς και στις μεμβράνες των φαγοσώματων που χρησιμοποιούνται από ουδετερόφιλα λευκά αιμοσφαίρια. Οι ισόμορφες ενώσεις της ορίζονται ως NOX1, NOX2, NOX3 και NOX4.

Μια μελέτη αναλύει τη δράση της οξειδάσης NADPH σε κεταμίνη με επαγόμενη απώλεια νευρωνικών παρβαλβουμίνης και έκφραση GAD67. Παρόμοια απώλεια παρατηρείται στη σχιζοφρένεια, με αποτελέσματα η οξειδάση NADPH να διαδραματίζει ρόλο στην παθοφυσιολογία της νόσου. [195] [196]

4.2 Συζήτηση αποτελεσμάτων

Συνοψίζοντας, στους παρακάτω πίνακες παρατίθενται οι πρωτεΐνες-στόχοι κατηγοριοποιημένες ανάλογα με τη δράση τους σύμφωνα με μελέτες που εξήχθησαν από τη βιβλιογραφική αναζήτηση. Επομένως, παρατίθενται τρεις κατηγορίες και συγκεκριμένα η πρώτη αφορά τις πρωτεΐνες που σχετίζονται με τη ψύχωση και τη σχιζοφρένεια, η δεύτερη αφορά εκείνες που σχετίζονται με άλλου είδους διαταραχές του ΚΝΣ και η τελευταία αφορά εκείνες τις πρωτεΐνες που τουλάχιστον από τη βιβλιογραφική αναζήτηση δε βρέθηκαν να συνδέονται με τη ψύχωση και γενικότερες διαταραχές του ΚΝΣ αλλά με άλλες νόσους.

Πίνακας 4.2-1: Πρωτεΐνες- στόχοι οι οποίες σύμφωνα με τη βιβλιογραφία σχετίζονται με την ψύχωση/σχιζοφρένεια

Συμβολισμός	Πρωτεΐνες στόχοι
HTR1A	5-hydroxytryptamine receptor 1A
HTR2A	5-hydroxytryptamine receptor 2A
HTR2B	5-hydroxytryptamine receptor 2B
HTR2C	5-hydroxytryptamine receptor 2C
HTR3A	5-hydroxytryptamine receptor 3A
HTR5A	5-hydroxytryptamine receptor 5A
HTR6	5-hydroxytryptamine receptor 6
HTR7	5-hydroxytryptamine receptor 7

SERT	Sodium-dependent serotonin transporter
ADRA1D	Alpha-1D adrenergic receptor
ADRA2B	Alpha-2B adrenergic receptor
ADRA2C	Alpha-2C adrenergic receptor
HRH1	Histamine H1 receptor
HRH2	Histamine H2 receptor
HRH4	Histamine H4 receptor
MDR1	Multidrug resistance protein 1
CYP2D6	Cytochrome P450 2D6
DRD1A	D(1A) dopamine receptor
DRD1B	D(1B) dopamine receptor
DRD2	D(2) dopamine receptor
DRD3	D(3) dopamine receptor
DRD4	D(4) dopamine receptor
DAT	Sodium-dependent dopamine transporter
MAP2K1	Dual specificity mitogen-activated protein kinase kinase 1
FYN	Tyrosine-protein kinase Fyn
FGFR1	Fibroblast growth factor receptor 1
GlyT-1	Sodium- and chloride-dependent glycine transporter 1
mAChR1	Muscarinic acetylcholine receptor M1
mAChR2	Muscarinic acetylcholine receptor M2
mAChR3	Muscarinic acetylcholine receptor M3
mAChR4	Muscarinic acetylcholine receptor M4
NAT1	Sodium-dependent noradrenaline transporter
OX2	Orexin receptor type 2

SIGMAR1	Sigma non-opioid intracellular receptor 1
KCNH2	Potassium voltage-gated channel subfamily H member 2
CACNA1G	Voltage-dependent T-type calcium channel subunit alpha-1G
BCL2	Bcl2 antagonist of cell death
NADPH	NADPH oxidase 1

Πίνακας 4.2-2: Πρωτεΐνες- στόχοι οι οποίες σύμφωνα με τη βιβλιογραφία σχετίζονται με άλλες διαταραχές του ΚΝΣ

Συμβολισμός	**Πρωτεΐνες στόχοι**	**Διαταραχές ΚΝΣ**
ADRA2A	Alpha-2A adrenergic receptor	Κατάθλιψη, άγχος
ABCC1	Multidrug resistance-associated protein 1	Parkinson, Alzheimer
PLK3	Serine/threonine-protein kinase PLK3	Νευροεκφυλιστικές, Parkinson
mAChR5	Muscarinic acetylcholine receptor M5	Συμπεριφορά

Πίνακας 4.2-3: Πρωτεΐνες- στόχοι οι οποίες σύμφωνα με τη βιβλιογραφία δεν σχετίζονται με την ψύχωση/σχιζοφρένεια ή άλλες διαταραχές του ΚΝΣ

Συμβολισμός	**Πρωτεΐνες στόχοι**	**Πάθηση**
HTR1F	5-hydroxytryptamine receptor 1F	Ημικρανία
YES	Tyrosine-protein kinase Yes	Καρκίνο

Επομένως, οι πρωτεΐνες-στόχοι HTR1F και YES δεν βρέθηκε μέσω βιβλιογραφικής αναζήτησης, να συσχετίζονται με την ψύχωση ή άλλες διαταραχές του κεντρικού νευρικού συστήματος. [197] [155]

5. Συμπεράσματα

Στην παρούσα διπλωματική εργασία, σκιαγραφήθηκε μια επέκταση στις in silico προβλέψεις για το σχεδιασμό φαρμάκων της ψύχωσης. Στόχος ήταν να εισαχθεί επαναλήψιμο μοντέλο για προ-επεξεργασία των μορίων *in silico*, ώστε να προβλεφθούν οι πρωτεΐνες-στόχοι, οι μεταξύ τους αλληλεπιδράσεις καθώς και η σύνδεσή τους με τη φαινοτυπική τους επίδραση.

Αναλυτικότερα, χρησιμοποιήθηκαν μικρά μόρια από την διαδικτυακή βάση δεδομένων της DrugBank και με τη χρήση του μοντέλου πρόβλεψης στόχων των *Koutsoukas et. al,*[4] πραγματοποιήθηκε μελέτη των πιθανών πρωτεϊνών στόχων που συνδέονται με την ψύχωση και με τον περαιτέρω εμπλουτισμό των εξαχθέντων αποτελεσμάτων προσδιορίστηκαν οι πρωτεΐνες με το μεγαλύτερο δείκτη πρόσδεσης με την ψύχωση. Συγκεκριμένα ο εμπλουτισμός πραγματοποιήθηκε με παραλλαγή της μεθόδου των *Liggi, Drakakis, Koutsoukas et. al.*[37] όπου τα μόρια με καλύτερο estimation score ήταν πιο κοντά στο 1 και με χειρότερο πιο κοντά στο 0, επιλέχθηκαν τα μόρια με estimation score τουλάχιστον 0,99.

Επομένως, επιλέχθηκαν συνολικά 44 πρωτεΐνες-στόχοι στις οποίες πραγματοποιήθηκε βιβλιογραφική αναζήτηση της συσχέτισής τους με τη ψύχωση αλλά και με άλλες διαταραχές του κεντρικού νευρικού συστήματος. Από τις 44 πρωτεΐνες-στόχους, μόλις 2 πρωτεΐνες δεν βρέθηκε να συσχετίζονται με τη ψύχωση, η HTR1F και η YES.

Οι δύο αυτές πρωτεΐνες σχετίζονται με τη θεραπεία ή την αλληλεπίδραση με άλλες νόσους όπως την αγγειακή λειτουργία του εγκεφάλου και τη λευχαιμία αντίστοιχα. Οι υπόλοιπες 42 πρωτεΐνες είτε αποτελούν άμεσες πρωτεΐνες – στόχους της νόσου είτε αποτελούν στόχους δράσης των αντιψυχωτικών φαρμάκων.

Στη συνέχεια, κρίνεται σκόπιμη η περαιτέρω μελέτη των δύο αυτών πρωτεϊνών ως προς το μηχανισμό δράσης, ώστε να επαληθευτεί η σύνδεση τους με τη νόσο της ψύχωσης. Προτείνεται, επιπλέον, η διερεύνησή τους και η αντιστοίχιση τους με μόρια που έχουν την ίδια βιοδραστικότητα, τα οποία όμως δεν είναι αντιψυχωτικά, ώστε να πραγματοποιηθεί εκ νέου επαναστόχευση.

Τα παραπάνω συμπεράσματα αποτελούν υποσχόμενο εδάφιο μελέτης, τα οποία μπορούν να αξιοποιηθούν σε νέα εργασία.

Τέλος, η ροή εργασίας (workflow) που ακολουθήθηκε στην παρούσα διπλωματική εργασία είναι μεταβιβάσιμη σε περαιτέρω κατηγορίες φαρμάκων ή ασθενειών και δύναται να χρησιμοποιηθεί για το γενικότερο σχεδιασμό φαρμάκων.

6. Βιβλιογραφία

1. Schedler, D. J. A. Drug Discovery: A History (Sneader, Walter). *J. Chem. Educ.* **83,** 215 (2006).

2. Terstappen, G. C. & Reggiani, A. In silico research in drug discovery. *Trends Pharmacol. Sci.* **22,** 23–26 (2001).

3. Carroll, P. M., Dougherty, B., Ross-Macdonald, P., Browman, K. & FitzGerald, K. Model systems in drug discovery: Chemical genetics meets genomics. *Pharmacol. Ther.* **99,** 183–220 (2003).

4. Koutsoukas, A. *et al.* From in silico target prediction to multi-target drug design: Current databases, methods and applications. *Journal of Proteomics* **74,** 2554–2574 (2011).

5. Eggert, U. S. The why and how of phenotypic small-molecule screens. *Nat. Chem. Biol.* **9,** 206–9 (2013).

6. Kotz, J. Phenotypic screening, take two. *Sci. Exch.* **5,** 1–3 (2012).

7. Gonzalez-Munoz, A. L., Minter, R. R. & Rust, S. J. Phenotypic screening: The future of antibody discovery. *Drug Discov. Today* **21,** 150–156 (2016).

8. Khanna, I. Drug discovery in pharmaceutical industry: Productivity challenges and trends. *Drug Discov. Today* **17,** 1088–1102 (2012).

9. Chavan Ravindranath, A. *et al.* Connecting Gene Expression Data from Connectivity Map and In Silico Target Predictions For Small Molecule Mechanism-of-Action Analysis. *Mol. BioSyst.* **11,** 86–96 (2014).

10. Mezey, P. G. Computer Aided Drug Design: Some Fundamental Aspects. *J. Mol. Model.* **6,** 150–157 (2000).

11. Εισαγωγή στη Βιοπληροφορική Β’ Μέρος.

12. ChEMBL. Available at: https://www.ebi.ac.uk/chembl/. (Accessed: 9th May 2017)

13. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, B. S. The PubChem Project. *Nucleic Acids Research* 44(D1):D1202-13 (2016). doi:10.1093/nar/gkv951

14. ChemBank - Login. Available at: http://chembank.broadinstitute.org/. (Accessed: 9th May 2017)

15. Koutsoukas, A. *et al.* In silico target predictions: Defining a benchmarking data set

and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model.* **53,** 1957–1966 (2013).

16. Lagunin, A., Stepanchikova, A., Filimonov, D. & Poroikov, V. PASS: prediction of activity spectra for biologically active substances. *Bioinformatics* **16,** 747–8 (2000).

17. Sriram, D. Medicinal Chemistry. **6,** 712 (2010).

18. Nidhi, Glick, M., Davies, J. W. & Jenkins, J. L. Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **46,** 1124–1133 (2006).

19. Olah, M. *et al.* WOMBAT: World of Molecular Bioactivity. in *Chemoinformatics in Drug Discovery* **23,** 221–239 (Wiley-VCH Verlag GmbH & Co. KGaA, 2005).

20. Zhao, S. & Li, S. Network-based relating pharmacological and genomic spaces for drug target identification. *PLoS One* **5,** (2010).

21. Muller, P. *et al.* In silico-guided target identification of a scaffold-focused library: 1,3,5-Triazepan-2,6-diones as novel phospholipase A2 inhibitors. *J. Med. Chem.* **49,** 6768–6778 (2006).

22. Drews, J. Case histories, magic bullets and the state of drug discovery. *Nat. Rev. Drug Discov.* **5,** 635–640 (2006).

23. Winau, F., Westphal, O. & Winau, R. Paul Ehrlich - In search of the magic bullet. *Microbes and Infection* **6,** 786–789 (2004).

24. DrugBank. Available at: https://www.drugbank.ca/. (Accessed: 9th May 2017)

25. Binding Database. (2001). Available at: www.bindingdb.org. (Accessed: 9th May 2017)

26. Liu, Z. *et al.* Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **50,** 302–309 (2017).

27. PDTD -- Potential Drug Target Database. Available at: http://www.dddc.ac.cn/pdtd/. (Accessed: 9th May 2017)

28. KNIME. KNIME | Open for Innovation. (2016). Available at: https://www.knime.org/. (Accessed: 9th May 2017)

29. Novac, N. Challenges and opportunities of drug repositioning. *Trends in Pharmacological Sciences* **34,** 267–272 (2013).

30. Sacan, A., Ekins, S. & Kortagere, S. Applications and limitations of in silico models in drug discovery. *Methods Mol. Biol.* **910,** 87–124 (2012).

31. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34,** D668–D672 (2006).

32. Central Nervous System Disease Definition. Available at: http://www.neuromodulation.com/central-nervous-system-disease-definition.

33. Σχιζοφρένεια — Δημήτρης Γούσης. Available at: http://www.dgousis.gr/διαταραχές/σχιζοφρένεια/. (Accessed: 23rd April 2016)

34. Jeffrey A. Lieberman, M. Treatment of Schizophrenia: The Current State of the Art. *Medscape Education Psychiatry & mental Health* (2011). Available at: http://www.medscape.org/viewarticle/753207. (Accessed: 23rd April 2016)

35. Monti, J. M., Torterolo, P. & Pandi Perumal, S. R. The effects of second generation antipsychotic drugs on sleep variables in healthy subjects and patients with schizophrenia. *Sleep Med. Rev.* 1–8 (2016). doi:10.1016/j.smrv.2016.05.002

36. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34,** D668-72 (2006).

37. Liggi, S. *et al.* Extending in silico mechanism-of-action analysis by annotating targets with pathways: application to cellular cytotoxicity readouts. *Future Med. Chem.* **6,** 2029–56 (2014).

38. James, C. A. OpenSMILES specification. 32 (2007). Available at: http://www.opensmiles.org/opensmiles.pdf. (Accessed: 23rd April 2016)

39. Daylight>Cheminformatics. Available at: http://www.daylight.com/smiles/index.html. (Accessed: 23rd April 2016)

40. James, C. A. OpenSMILES specification. 32 (2007).

41. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. 1–7 (2012). Available at: papers3://publication/uuid/C2639E7A-B7E5-4664-884E-5F3DDFE92FDB. (Accessed: 23rd April 2016)

42. James, C., Weininger, D. & Delaney, J. Daylight Theory Manual version 4.9; Daylight Chemical Information Systems, Inc.: Mission Viejo, CA;http://www.daylight.com/dayhtml/doc/theory/index.html. *Inc. St. Fe, N. Mex* (2008).

43. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Model.* **32,** 244–255 (1992).

44. Westbrook, J. & Berman, H. M. Ontologies for three-dimensional molecular structure. Available at: http://what-when-how.com/bioinformatics/ontologies-for-three-dimensional-molecular-structure-bioinformatics/. (Accessed: 23rd April

2016)

45. Hert, J. *et al.* Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **44,** 1177–1185 (2004).

46. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71,** 58–63 (2015).

47. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71,** 58–63 (2015).

48. Cereto-Massagu??, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71,** 58–63 (2015).

49. Barnard, J. M. & Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Model.* **37,** 141–142 (1997).

50. MDL's MACCS fingerprint. Available at: https://docs.chemaxon.com/display/CD/MDL's+MACCS+fingerprint. (Accessed: 23rd April 2016)

51. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50,** 742–754 (2010).

52. Chemical Fingerprints: Extended Connectivity Fingerprint (ECFP). *Chemical Fingerprints Documentation* Available at: https://docs.chemaxon.com/pages/viewpage.action?pageId=14483752. (Accessed: 23rd April 2016)

53. Drogers. David Rogers - ECFP Manuscript. (2010).

54. Bender, A., Mussa, H. Y., Glen, R. C. & Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Na??ve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **44,** 170–178 (2004).

55. Bender, A., Mussa, H. Y., Glen, R. C. & Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **44,** 1708–1718 (2004).

56. Eckert, H. & Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **12,** 225–233 (2007).

57. Glen, R. C. *et al.* Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **9,** 199–204 (2006).

58. Bender, A. & Glen, R. C. Molecular similarity: a key technique in molecular

informatics. *Org. Biomol. Chem.* **2,** 3204–18 (2004).

59. Bender, A., Mussa, H. Y., Glen, R. C. & Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Na??ve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **44,** 170–178 (2004).

60. Willett, P., Barnard, J. M. & Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **38,** 983–996 (1998).

61. Ashtawy, H. M. & Mahapatra, N. R. A comparative assessment of ranking accuracies of conventional and machine-learning-based scoring functions for protein-ligand binding affinity prediction. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **9,** 1301–1313 (2012).

62. Langley, P. & Sage, S. Induction of Selective Bayesian Classifiers. *Proc. Tenth Int. Conf. Uncertain. Artif. Intell.* 399–406 (1994). doi:10.1016/B978-1-55860-332-5.50055-9

63. Cha, S. & Tappert, C. A Genetic Algorithm for Constructing Compact Binary Decision Trees. *Entropy* **1,** 1–13 (2009).

64. Grzymala-busse, J. W. Rule Induction. in *Data Mining and Knowledge Discovery Handbook* 277–294 (2005). doi:10.1007/978-0-387-09823-4_13

65. Breiman, L. Random forests. *Mach. Learn.* **45,** 5–32 (2001).

66. Hand, D. J. *Principles of data mining*. *Drug Safety* **30,** (2007).

67. Ibm. IBM Knowledge Center. 2015 (2014).

68. Cherkassky, V. The Nature Of Statistical Learning Theory~. *IEEE Trans. Neural Networks* **8,** 1564–1564 (1997).

69. Peng, J. X., Rafferty, K. & Ferguson, S. Building support vector machines in the context of regularized least squares. *Neurocomputing* **211,** 129–142 (2016).

70. Catal, C. Performance evaluation metrics for software fault prediction studies. *Acta Polytech. Hungarica* **9,** 193–206 (2012).

71. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The balanced accuracy and its posterior distribution. in *Proceedings - International Conference on Pattern Recognition* 3121–3124 (IEEE, 2010). doi:10.1109/ICPR.2010.764

72. Tomaszewski, P., Lundberg, L. & Grahn, H. The accuracy of early fault prediction in modified code. *Proc. Fifth Conf. Softw. Eng. Res. Pract. Sweden* 57–63 (2005).

73. Tuason, O., Chen, L., Liu, H., Blake, J. A. & Friedman, C. Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pacific Symp.*

Biocomput. Pacific Symp. Biocomput. 238–249 (2004).

74. Lund, O., Nielsen, M., Lundegaard, C., Kesmir, C. & Brunak, S. Immunological Bioinformatics. *Methods* 312 (2005). doi:10.1093/bioinformatics/bth100

75. Schneider, J. Cross Validation. Available at: http://www.cs.cmu.edu/~schneide/tut5/node42.html. (Accessed: 19th December 2015)

76. Bleeker, S. E. *et al.* External validation is necessary in prediction research: A clinical example. *J. Clin. Epidemiol.* **56,** 826–832 (2003).

77. Altman, D. G. & Royston, P. What do we mean by validating a prognistic model? *Stat. Med.* **19,** 453–473 (2000).

78. Curtis, R. K., Orešič, M. & Vidal-Puig, A. Pathways to the analysis of microarray data. *Trends Biotechnol.* **23,** 429–435 (2005).

79. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

80. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2008).

81. Huang, D. W., Lempicki, R. a & Sherman, B. T. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).

82. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).

83. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102,** 15545–15550 (2005).

84. Rossum, G. Van & Drake, F. L. Python. *Milt. Q.* **42,** 270–272 (2008).

85. Nigsch, F., Bender, A., Jenkins, J. L. & Mitchell, J. B. O. Ligand-target prediction using winnow and naive bayesian algorithms and the implications of overall performance statistics. *J. Chem. Inf. Model.* **48,** 2313–2325 (2008).

86. Drakakis, G. *et al.* Comparing Global and Local Likelihood Score Thresholds in Multiclass Laplacian-Modified Naive Bayes Protein Target Prediction. *Comb. Chem. High Throughput Screen.* **18,** 323–330 (2015).

87. Liggi1, S., Drakakis1, G., Koutsoukas1, AlexiosCortes–Ciriano2, IsidroPatricia

Martínez–Alonso3, 4Malliavin2, Thérèse EAdrian Velazquez-Campoy3, 4, 5Brewerton6, Suzanne CBodkin6, Michael JEvans6, David AGlen1, Robert C, José Alberto Carrodeguas3, 4 & & Andreas Bender*, 1. Extending in silico mechanism-of- action analysis by annotating targets with pathways: application to cellular cytotoxicity readouts. **6,** 2029–2055 (2014).

88. Ögren, S. O. *et al.* The role of 5-HT1A receptors in learning and memory. *Behav. Brain Res.* **195,** 54–77 (2008).

89. Hoyer, D., Hannon, J. P. & Martin, G. R. Molecular, pharmacological and functional diversity of 5-HT receptors. *Pharmacol. Biochem. Behav.* **71,** 533–554 (2002).

90. Savitz, J., Lucki, I. & Drevets, W. C. 5-HT1A receptor function in major depressive disorder. *Prog. Neurobiol.* **88,** 17–31 (2009).

91. Mitsikostas, D. D. & Tfelt-Hansen, P. Targeting to 5-HT1F receptor subtype for migraine treatment: lessons from the past, implications for the future. *Cent Nerv Syst Agents Med Chem* **12,** 241–249 (2012).

92. Di Giovanni, G. & De Deurwaerdère, P. New therapeutic opportunities for 5-HT2C receptor ligands in neuropsychiatric disorders. *Pharmacol. Ther.* **157,** 125–162 (2015).

93. Christiansen, L. *et al.* Candidate Gene Polymorphisms in the Serotonergic Pathway: Influence on Depression Symptomatology in an Elderly Population. *Biol. Psychiatry* **61,** 223–230 (2007).

94. McMahon, F. J. *et al.* Variation in the gene encoding the serotonin 2A receptor is associated with outcome of antidepressant treatment. *Am. J. Hum. Genet.* **78,** 804–814 (2006).

95. Meltzer, H. Y. Update on typical and atypical antipsychotic drugs. *Annu. Rev. Med.* **64,** 393–406 (2013).

96. Niesler, B., Kapeller, J., Hammer, C. & Rappold, G. Serotonin type 3 receptor genes: HTR3A,B,C,D,E. *Pharmacogenoimcs* **9,** 501–504 (2008).

97. Khorana, N. *et al.* Binding of tetrahydrocarboline derivatives at human 5-HT5A receptors. *J. Med. Chem.* **46,** 3930–3937 (2003).

98. Birkett, J. T. *et al.* Association analysis of the 5-HT5A gene in depression, psychosis and antipsychotic response. *Neuroreport* **11,** 2017–2020 (2000).

99. East, S. Z., Burnet, P. W. J., Kerwin, R. W. & Harrison, P. J. An RT-PCR study of 5-HT6 and 5-HT7 receptor mRNAs in the hippocampal formation and prefrontal cortex in schizophrenia. *Schizophr. Res.* **57,** 15–26 (2002).

100. Szeszko, P. R. *et al.* Amygdala volume reductions in pediatric patients with obsessive-compulsive disorder treated with paroxetine: preliminary findings. *Neuropsychopharmacology* **29,** 826–832 (2004).

101. Werner, F.-M. & CoveÃ±as, R. Classical Neurotransmitters and Neuropeptides Involved in Major Depression in a Multi-neurotransmitter System: A Focus on Antidepressant Drugs. *Curr. Med. Chem.* **20,** 4853–4858 (2013).

102. Abdolmaleky, H. M. *et al.* DNA hypermethylation of serotonin transporter gene promoter in drug na??ve patients with schizophrenia. *Schizophr. Res.* **152,** 373–380 (2014).

103. Serretti, A. *et al.* Serotonin transporter gene (5-HTTLPR) and major psychoses. *Mol. Psychiatry* **7,** 95–9 (2002).

104. Ioanninamed.gr - α-αδρενεργικοί ανταγωνιστές. Available at: https://www.ioanninamed.gr/topics/common-disease/66-hypertension/721--adrenergic-antagonists. (Accessed: 15th January 2017)

105. Feng, J. *et al.* An in-frame deletion in the alpha(2C) adrenergic receptor is common in African--Americans. *Mol. Psychiatry* **6,** 168–172 (2001).

106. Fitton, A. & Heel, R. C. Clozapine. A review of its pharmacological properties, and therapeutic use in schizophrenia. *Drugs* **40,** 722–47 (1990).

107. Leckman, J. F. *et al.* Clonidine treatment of Gilles de la Tourette's syndrome. *Arch.Gen.Psychiatry* **48,** 324–328 (1991).

108. Reichow, B., Volkmar, F. R. & Bloch, M. H. Systematic review and meta-analysis of pharmacological treatment of the symptoms of attention-deficit/hyperactivity disorder in children with pervasive developmental disorders. *J. Autism Dev. Disord.* **43,** 2435–2441 (2013).

109. Philipp, M., Brede, M., Hein, L. & Physiological, L. H. Physiological significance of alpha2-adrenergic receptor subtype diversity: one receptor is not enough. *Am. J. Physiol. - Regul. Integr. Comp. Physiol.* **283,** 287–295 (2002).

110. MacDonald, E., Kobilka, B. K. & Scheinin, M. Gene targeting - Homing in on α2-adrenoceptor-subtype function. *Trends Pharmacol. Sci.* **18,** 211–219 (1997).

111. Hill, S. J. *et al.* International Union of Pharmacology. XIII. Classification of Histamine Receptors. *Pharmacol. Rev.* **49,** 253–278 (1997).

112. Lin, L. *et al.* Measurement of hypocretin/orexin content in the mouse brain using an enzyme immunoassay: The effect of circadian time, age and genetic background. *Peptides* **23,** 2203–2211 (2002).

113. Yanai, K. *et al.* Behavioural characterization and amounts of brain monoamines

and their metabolites in mice lacking histamine h1 receptors. *Neuroscience* **87,** 479–487 (1998).

114. Deliu, E. *et al.* Mechanisms of G protein-coupled estrogen receptor-mediated spinal nociception. *J. Pain* **13,** 742–754 (2012).

115. Yoshizawa, M. *et al.* Increased Brain Histamine H1 Receptor Binding in Patients with Anorexia Nervosa. *Biol. Psychiatry* **65,** 329–335 (2009).

116. Vehof, J. *et al.* Association of genetic variants of the histamine H1 and muscarinic M3 receptors with BMI and HbA1c values in patients on antipsychotic medication. *Psychopharmacology (Berl).* **216,** 257–265 (2011).

117. Mori, T., Narita, M., Onodera, K. & Suzuki, T. Involvement of histaminergic system in the discriminative stimulus effects of morphine. *Eur. J. Pharmacol.* **491,** 169–172 (2004).

118. Arrang, J. M. Histamine and Schizophrenia. *Int. Rev. Neurobiol.* **78,** 247–287 (2007).

119. Zampeli, E. & Tiligada, E. The role of histamine H 4 receptor in immune and inflammatory disorders. *Br. J. Pharmacol.* **157,** 24–33 (2009).

120. Munoz, M., Henderson, M., Haber, M. & Norris, M. Role of the MRP1/ABCC1 multidrug transporter protein in cancer. *IUBMB Life* **59,** 752–757 (2007).

121. Borst, P., Evers, R., Kool, M. & Wijnholds, J. A family of drug transporters: the multidrug resistance-associated proteins. *J. Natl. Cancer Inst.* **92,** 1295–1302 (2000).

122. Dallas, S., Miller, D. S. & Bendayan, R. Multidrug Resistance-Associated Proteins : Expression and Function in the Central Nervous System. **58,** 140–161 (2006).

123. Pahnke, J., Langer, O. & Krohn, M. Alzheimer's and ABC transporters - new opportunities for diagnostics and treatment. *Neurobiol. Dis.* **72,** 54–60 (2014).

124. Kunjachan, S., Rychlik, B., Storm, G., Kiessling, F. & Lammers, T. Multidrug resistance: Physiological principles and nanomedical solutions. *Adv. Drug Deliv. Rev.* **65,** 1852–1865 (2013).

125. Tatsuta, T., Naito, M., Oh-hara, T., Sugawara, I. & Tsuruo, T. Functional involvement of P-glycoprotein in blood-brain barrier. *J. Biol. Chem.* **267,** 20383–20391 (1992).

126. Anger, G. J., Cressman, A. M. & Piquette-Miller, M. Expression of ABC efflux transporters in placenta from women with insulin-managed diabetes. *PLoS One* **7,** e35027 (2012).

127. Fojo, A. T. *et al.* Expression of a multidrug-resistance gene in human tumors and tissues. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 265–9 (1987).

128. Tovilla-Zarate, C. A. *et al.* Association study between the MDR1 gene and clinical characteristics in schizophrenia. *Rev. Bras. Psiquiatr.* **36,** 227–232 (2014).

129. Qian, W. *et al.* MDR1 gene polymorphism in Japanese patients with schizophrenia and mood disorders including depression. *Biol. Pharm. Bull.* **29,** 2446–50 (2006).

130. Girardin, F. Membrane transporter proteins: A challenge for CNS drug development. *Dialogues Clin. Neurosci.* **8,** 311–321 (2006).

131. Ingelman-Sundberg, M., Sim, S. C., Gomez, A. & Rodriguez-Antona, C. Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol. Ther.* **116,** 496–526 (2007).

132. Owen, R. P., Sangkuhl, K., Klein, T. E. & Altman, R. B. Cytochrome P450 2D6. *Pharmacogenet. Genomics* **19,** 559–62 (2009).

133. Fuxe, K. Evidence for the existence of monoamine neurons in the central nervous system. *Zeitschrift für Zellforsch. und Mikroskopisch* **596,** 573–596 (1965).

134. Gingrich, J. a & Caron, M. G. Recent advances in the molecular biology of dopamine receptors. *Annu. Rev. Neurosci.* **16,** 299–321 (1993).

135. Rommelfanger, K. S. & Wichmann, T. Extrastriatal dopaminergic circuits of the basal ganglia. *Front. Neuroanat.* **4,** 1–17 (2010).

136. Undieh, A. S. Pharmacology of signaling induced by dopamine D1-like receptor activation. *Pharmacol. Ther.* **128,** 37–60 (2010).

137. Koob, G. F. & Volkow, N. D. Neurocircuitry of addiction. *Neuropsychopharmacology* **35,** 217–238 (2010).

138. Maia, T. V. & Frank, M. J. An Integrative Perspective on the Role of Dopamine in Schizophrenia. *Biol. Psychiatry* **81,** 52–66 (2017).

139. Shi, J. *et al.* Neurotransmission and bipolar disorder: a systematic family-based association study. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **147B,** 1270–7 (2008).

140. Heijtz, R. D., Kolb, B. & Forssberg, H. Motor inhibitory role of dopamine D1 receptors: Implications for ADHD. *Physiol. Behav.* **92,** 155–160 (2007).

141. Hettinger, J. A., Liu, X., Schwartz, C. E., Michaelis, R. C. & Holden, J. J. A. A DRD1 haplotype is associated with risk for autism spectrum disorders in male-only affected sib-pair families. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **147,** 628–636 (2008).

142. Andreou, D. *et al.* Associations between a locus downstream DRD1 gene and cerebrospinal fluid dopamine metabolite concentrations in psychosis. *Neurosci. Lett.* **619,** 126–130 (2016).

143. Simpson, E. H. & Kellendonk, C. Insights About Striatal Circuit Function and Schizophrenia From a Mouse Model of Dopamine D2 Receptor Upregulation. *Biol. Psychiatry* **81,** 21–30 (2017).

144. Seeman, P. Are dopamine D2 receptors out of control in psychosis? *Prog. Neuro-Psychopharmacology Biol. Psychiatry* **46,** 146–152 (2013).

145. Richtand, N. M., Woods, S. C., Berger, S. P. & Strakowski, S. M. D3 dopamine receptor, behavioral sensitization, and psychosis. *Neurosci. Biobehav. Rev.* **25,** 427–443 (2001).

146. de Krom, M. *et al. A Common Variant in DRD3 Receptor Is Associated with Autism Spectrum Disorder*. *Biological Psychiatry* **65,** (2009).

147. Dikeos, D. G. *et al.* Association between the dopamine D3 receptor gene locus (DRD3) and unipolar affective disorder. *Psychiatr. Genet.* **9,** 189–195 (1999).

148. Lauzon, N. M. & Laviolette, S. R. Dopamine D4-receptor modulation of cortical neuronal network activity and emotional processing: Implications for neuropsychiatric disorders. *Behav. Brain Res.* **208,** 12–22 (2010).

149. Kim, B.-N. *et al.* Shorter dinucleotide repeat length in the DRD5 gene is associated with attention deficit hyperactivity disorder. *Psychiatr. Genet.* **19,** 57 (2009).

150. Torres, G. E., Gainetdinov, R. R. & Caron, M. G. Plasma membrane monoamine transporters: structure, regulation and function. *Nat. Rev. Neurosci.* **4,** 13–25 (2003).

151. Liang, N. Y. & Rutledge, C. O. Evidence for carrier-mediated efflux of dopamine from corpus striatum. *Biochem. Pharmacol.* **31,** 2479–2484 (1982).

152. McHugh, P. C. & Buckley, D. A. The Structure and Function of the Dopamine Transporter and its Role in CNS Diseases. in *Vitamins and Hormones* **98,** 339–369 (2015).

153. Marks, J. L. *et al.* Novel MEK1 mutation identified by mutational analysis of epidermal growth factor receptor signaling pathway genes in lung adenocarcinoma. *Cancer Res.* **68,** 5524–5528 (2008).

154. Mill, J. *et al.* Epigenomic Profiling Reveals DNA-Methylation Changes Associated with Major Psychosis. *Am. J. Hum. Genet.* **82,** 696–711 (2008).

155. Cacabelos, R. & Martínez-Bouza, R. Genomics and Pharmacogenomics of

Schizophrenia. *CNS Neurosci. Ther.* **17,** 541–565 (2011).

156. Rybakowski, J., Borkowska, A., Dmitrzak-Weglarz, M., Skibinska, M. & Hauser, J. Association Between Polymorphisms of Drd1 and Fyn Genes and the Results of Wisconsin Test in Schizophrenia. *Schizophr. Res.* **102,**

157. Anguita, E. & Villalobo, A. Src-family tyrosine kinases and the Ca2+ signal. *Biochim. Biophys. Acta - Mol. Cell Res.* (2016). doi:10.1016/j.bbamcr.2016.10.022

158. Kinase Associated Diseases | Cell Signaling Technology. Available at: https://www.cellsignal.com/common/content/content.jsp?id=science-tables-kinase-disease. (Accessed: 23rd January 2017)

159. Advances in Phosphotransferases (Alcohol Group Acceptor) Research and ... - Βιβλία Google. Available at: https://books.google.gr/books?id=R_Yy-dyUpF0C&pg=PT264&lpg=PT264&dq=κινάσης+plk3+disease&source=bl&ots=LySq4_j4vx&sig=4LpEAYj0-MdMNTh7kpqn1LQDK3A&hl=el&sa=X&sqi=2&ved=0ahUKEwimvMnOu9nRAhUFVBQKHaUnCzQQ6AEIOTAE#v=onepage&q. (Accessed: 24th January 2017)

160. Donald J. Zack, D. S. W. Identification of molecular pathways and methods of use thereof for treating retinal neurodegeneration and other neurodegenerative disorders. (2013). doi:US20150030572 A1

161. Dzamko, N., Zhou, J., Huang, Y. & Halliday, G. M. Parkinson's disease-implicated kinases in the brain; insights into disease pathogenesis. *Front. Mol. Neurosci.* **7,** 57 (2014).

162. Sims-Lucas, S. *et al.* Fgfr1 and the IIIc isoform of Fgfr2 play critical roles in the metanephric mesenchyme mediating early inductive events in kidney development. *Dev. Dyn.* **240,** 240–249 (2011).

163. Poladia, D. P. *et al.* Role of fibroblast growth factor receptors 1 and 2 in the metanephric mesenchyme. *Dev. Biol.* **291,** 325–339 (2006).

164. Dodé, C. *et al.* Loss-of-function mutations in FGFR1 cause autosomal dominant Kallmann syndrome. *Nat. Genet.* **33,** 463–465 (2003).

165. Ahmad, I., Iwata, T. & Leung, H. Y. Mechanisms of FGFR-mediated carcinogenesis. *Biochim. Biophys. Acta - Mol. Cell Res.* **1823,** 850–860 (2012).

166. Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10,** 529–541 (2006).

167. Chokdeemboon, C. *et al.* FGFR1 and FGFR2 mutations in Pfeiffer syndrome. *J. Craniofac. Surg.* **24,** 150–2 (2013).

168. Borroto-Escuela, D. O., Tarakanov, A. O. & Fuxe, K. FGFR1-5-HT1A Heteroreceptor Complexes: Implications for Understanding and Treating Major Depression. *Trends Neurosci.* **39,** 5–15 (2016).

169. Gaughran, F., Payne, J., Sedgwick, P. M., Cotter, D. & Berry, M. Hippocampal FGF-2 and FGFR1 mRNA expression in major depression, schizophrenia and bipolar disorder. *Brain Res. Bull.* **70,** 221–227 (2006).

170. Morita, Y. *et al.* The glycine transporter 1 gene (GLYT1) is associated with methamphetamine-use disorder. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **147,** 54–58 (2008).

171. Manahan-Vaughan, D., Wildförster, V. & Thomsen, C. Rescue of hippocampal LTP and learning deficits in a rat model of psychosis by inhibition of glycine transporter-1 (GlyT1). *Eur. J. Neurosci.* **28,** 1342–1350 (2008).

172. Javitt, D. C. Glycine Transport Inhibitors and the Treatment of Schizophrenia. *Biol. Psychiatry* **63,** 6–8 (2008).

173. Keefe, R. S. E. & Harvey, P. D. *Novel Antischizophrenia Treatments*. *Handbook of Experimental Pharmachology* **213,** (2012).

174. Foster, D. J., Choi, D. L., Jeffrey Conn, P. & Rook, J. M. Activation of M1 and M4 muscarinic receptors as potential treatments for Alzheimer's disease and schizophrenia. *Neuropsychiatr. Dis. Treat.* **10,** 183–191 (2014).

175. Scarr, E., Keriakous, D., Crossland, N. & Dean, B. No change in cortical muscarinic M2, M3 receptors or [35S] GTP??S binding in schizophrenia. *Life Sci.* **78,** 1231–1237 (2006).

176. Yamada, M. *et al.* Novel insights into M5 muscarinic acetylcholine receptor function by the use of gene targeting technology. *Life Sci.* **74,** 345–353 (2003).

177. Zhou, J. Norepinephrine transporter inhibitors and their therapeutic potential. *Drugs Future* **29,** 1235–1244 (2004).

178. Haenisch, B. & Bönisch, H. Depression and antidepressants: Insights from knockout of dopamine, serotonin or noradrenaline re-uptake transporters. *Pharmacol. Ther.* **129,** 352–368 (2011).

179. Sommerauer, C., Rebernik, P., Reither, H., Nanoff, C. & Pifl, C. The noradrenaline transporter as site of action for the anti-Parkinson drug amantadine. *Neuropharmacology* **62,** 1708–1716 (2012).

180. Borowsky, B. & Hoffman, B. J. Neurotransmitter transporters: molecular biology, function, and regulation. *Int. Rev. Neurobiol.* **38,** 139–99 (1995).

181. Nishino, S. The hypocretin/orexin system in health and disease. *Biol. Psychiatry*

54, 87–95 (2003).

182. Takizawa, R. *et al.* Association between sigma-1 receptor gene polymorphism and prefrontal hemodynamic response induced by cognitive activation in schizophrenia. *Prog. Neuro-Psychopharmacology Biol. Psychiatry* **33,** 491–498 (2009).

183. Hashimoto, K. Activation of sigma-1 receptor chaperone in the treatment of neuropsychiatric diseases and its clinical implication. *J. Pharmacol. Sci.* **127,** 6–9 (2015).

184. Ryskamp, D. *et al.* The sigma-1 receptor mediates the beneficial effects of pridopidine in a mouse model of Huntington disease. *Neurobiol. Dis.* **97,** 46–59 (2017).

185. Zhang, B. *et al.* Sigma-1 receptor deficiency reduces GABAergic inhibition in the basolateral amygdala leading to LTD impairment and depressive-like behaviors. *Neuropharmacology* **116,** 387–398 (2017).

186. Huffaker, S. J. *et al.* A primate-specific, brain isoform of KCNH2 affects cortical physiology, cognition, neuronal repolarization and risk of schizophrenia. *Nat. Med.* **15,** 509–18 (2009).

187. Nelson, M. T., Todorovic, S. M. & Perez-Reyes, E. The role of T-type calcium channels in epilepsy and pain. *Curr. Pharm. Des.* **12,** 2189–97 (2006).

188. Fry, C. H., Sui, G. & Wu, C. T-type Ca2+ channels in non-vascular smooth muscles. *Cell Calcium* **40,** 231–239 (2006).

189. Vassort, G., Talavera, K. & Alvarez, J. L. Role of T-type Ca2+ channels in the heart. *Cell Calcium* **40,** 205–220 (2006).

190. Iftinca, M. C. Neuronal T-type calcium channels: what's new? Iftinca: T-type channel regulation. *J. Med. Life* **4,** 126–38 (2011).

191. Soane, L., Siegel, Z. T., Schuh, R. A. & Fiskum, G. Postnatal developmental regulation of Bcl-2 family proteins in brain mitochondria. *J. Neurosci. Res.* **86,** 1267–1276 (2008).

192. Vaux, D. L., Cory, S. & Adams, J. M. Bcl-2 gene promotes haemopoietic cell survival and cooperates with c-myc to immortalize pre-B cells. *Nature* **335,** 440–442 (1988).

193. Otake, Y. *et al.* induces stabilization of bcl2 mRNA Overexpression of nucleolin in chronic lymphocytic leukemia cells induces stabilization of bcl2 mRNA. *Hematology* **109,** 3069–3075 (2009).

194. Li, A., Ojogho, O. & Escher, A. Saving Death: Apoptosis for Intervention in

Transplantation and Autoimmunity. *Clin. Dev. Immunol.* **13,** 273–282 (2006).

195. Wang, X., Pinto-Duarte, A., Sejnowski, T. J. & Behrens, M. M. How Nox2-containing NADPH oxidase affects cortical circuits in the NMDA receptor antagonist model of schizophrenia. *Antioxid. Redox Signal.* **18,** 1444–62 (2013).

196. Sorce, S. *et al.* The NADPH Oxidase NOX2 Controls Glutamate Release: A Novel Mechanism Involved in Psychosis-Like Ketamine Responses. *J. Neurosci.* **30,** 11317–11325 (2010).

197. Gene Set - Psychotic Disorders. Available at: http://amp.pharm.mssm.edu/Harmonizome/gene_set/Psychotic+Disorders/CTD+Gene-Disease+Associations. (Accessed: 29th January 2017)