ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

3η Έκθεση Προόδου

Φεβρουάριος 2023 – Νοέμβριος 2023

Εκπονούμενη Διατριβή:

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

Ονοματεπώνυμο Υποψήφιου Διδάκτορα:
Σωτήριος Ουζούνης

Αριθμός Μητρώου:
2003

Επιβλέπων Καθηγητής:
Καλατζής Ιωάννης, Καθηγητής

Τριμελής Συμβουλευτική Επιτροπή:
Καλατζής Ιωάννης, Καθηγητής
Κατσίλα Θεοδώρα, Κύρια Ερευνήτρια, Εθνικό Ίδρυμα Ερευνών
Ματσούκας Μίνωας-Τιμόθεος, Επίκουρος καθηγητής

1. Εισαγωγή

Στόχος της διατριβή αυτής είναι η ανάπτυξη μεθοδολογίας για την επαναστόχευση φαρμάκων με βέλτιστη ακρίβεια. Η μεθοδολογία αυτή είναι αμιγώς υπολογιστική, ανάγοντας την επαναστόχευση φαρμάκων σε ένα πρόβλημα ανάλυσης ετερογενών δεδομένων μεγάλου όγκου. Για την επίλυση του προβλήματος αυτού, η προσέγγιση η οποία ακολουθείται έχει ως κύριο άξονα τη συλλογή, ομογενοποίηση, ενοποίηση και ανάλυση των παρακάτω:

- βιοιατρικών δεδομένων από δημόσια αποθετήρια
- δεδομένων ομικών τεχνολογιών
- φαρμακογονιδιωματικών δεδομένων
- δεδομένων ευρεσιτεχνιών βιοδραστικών μορίων
- δεδομένων κλινικών δοκιμών
- δεδομένων δομικής βιολογίας

Η συλλογή και επεξεργασία των δεδομένων από δημόσια αποθετήρια, όπως αυτή περιγράφεται στην 1η και 2η έκθεση προόδου, αποσκοπεί στη μετουσίωση των δεδομένων σε γνώση. Η παραγόμενη αυτή γνώση δύναται, όχι μόνο να επισπεύσει, αλλά και να καθοδηγήσει τις εργαστηριακές προσεγγίσεις επαναστόχευσης φαρμάκων. Οι εργαστηριακές μελέτες αποτελούν αναπόσπαστο κομμάτι, τόσο της διαδικασίας ανάπτυξης φαρμάκων, όσο και της επαναστόχευσης αυτών.

Μια από τις πιο διαδομένες μεθόδους πειραματικής αξιολόγησης της δράσης ενός φαρμάκου είναι η εύρεση της μέσης ανασταλτικής συγκέντρωσης (IC_{50}). Η τιμή IC_{50} ενός χημικού μορίου-συνδέτη για μια πρωτεΐνη-στόχο ορίζεται ως η συγκέντρωση του πρώτου ικανή να οδηγήσει σε αναστολή της βιοδραστικότητας κατά 50% [1] και μπορεί να μοντελοποιηθεί ως μια σχέση παλινδρόμησης.

Αναπτύχθηκε, λοιπόν, μια μεθοδολογία βαθιάς μάθησης για την ημιποσοτική πρόβλεψη της τιμής IC_{50} ενός χημικού μορίου-συνδέτη. Πιο συγκεκριμένα, η πρόβλεψη της τιμής IC_{50} μοντελοποιήθηκε ως πρόβλημα ταξινόμησης πολλαπλών κλάσεων. Οι κλάσεις αυτές προέκυψαν, ορίζοντας διακριτά εύρη τιμών IC_{50}, με βάση την φαρμακολογική τους ερμηνεία. Ως μεταβλητές εισόδου χρησιμοποιήθηκαν μοριακοί περιγραφείς για το χημικό μόριο-προσδέτη, μοριακοί περιγραφείς πρωτεϊνών, καθώς και περιγραφείς αλληλεπιδράσεων μεταξύ προσδέτη-πρωτεΐνης. Οι περιγραφείς αλληλεπίδρασης εξήχθησαν από υπολογισμούς μοριακής πρόσδεσης. Εν συνεχεία, τα δεδομένα μοντελοποιήθηκαν από βαθιά νευρωνικά δίκτυα, καθώς επίσης και από κλασικούς αλγορίθμους μηχανικής μάθησης, ώστε να γίνει σύγκριση της απόδοσης των μοντέλων.

Η μεθοδολογία αυτή μπορεί να συμβάλει σημαντικά στην επανοστόχευση φαρμάκων, ειδικότερα σε ασθένειες, όπως ο σακχαρώδης διαβήτης. Το βαθύ νευρωνικό δίκτυο έρχεται να προσφέρει ακόμη ένα εργαλείο στη συνολική ροή εργασίας ανάλυσης 1D και 3D δεδομένων για την επαναστόχευση φαρμάκων στον σακχαρώδη διαβήτη.

2. Υλικά και Μέθοδοι

2.1. Λήψη δεδομένων και ορισμός κλάσεων

Τα δεδομένα που χρησιμοποιήθηκαν για την ανάπτυξη της μεθοδολογίας αυτής χωρίζονται σε 1D, 3D και πειραματικά αποτελέσματα. Τα 1D δεδομένα αφορούν τις δομές προσδέτη (χημικό μόριο) σε μορφή SMILES [2], δηλαδή σε μια μορφή, που κωδικοποιεί την χημική δομή σε μια αλληλουχία συμβολο-σειρών, ώστε να είναι αντιληπτή από τον υπολογιστή. Ως 1D δεδομένα ελήφθησαν, επίσης, οι αλληλουχίες αμινοξικών καταλοίπων σε μορφή FASTA. Τα 1D δεδομένα εξήχθησαν από το αποθετήριο ChEMBL [3].Τα 3D δεδομένα για τα φάρμακα ελήφθησαν από την πλατφόρμα DrugBank [4], ενώ επιλέχθηκαν μόνο, όσα

έχουν λάβει έγκριση ή βρίσκονται υπό διερεύνηση. Οι 3D δομές των πρωτεϊνών ελήφθησαν από το αποθετήριο Protein Data Bank [5]. Τα πειραματικά δεδομένα ελήφθησαν, επίσης, από την ChEMBL και αφορούν δοκιμασίες, οι οποίες ποσοτικοποιούν τη σχέση προσδέτη-πρωτεΐνης με την τιμή IC_{50}.

Με βάση τα πειραματικά δεδομένα που ελήφθησαν, κάθε ζευγάρι προσδέτη-πρωτεΐνης ταξινομήθηκε σε μια κλάση. Η ταξινόμηση στις κλάσης έγινε με βάση την ενδιάμεση τιμή IC_{50} που είχε κάθε ζεύγος. Επιπλέον, πριν την ταξινόμηση σε κλάσεις, το σύνολο των τιμών κανονικοποιήθηκε (κλίμακα nM). Οι κλάσεις που ορίστηκαν, καθώς και το σύνολο των ζευγαριών που φέρουν δίδεται παρακάτω:

- Κλάση 0: IC50 < 10, ισχυρή αναστολή, πλήθος ζευγαριών = 2.934
- Κλάση 1: 10 ≤ IC50 < 100, αναστολή, πλήθος ζευγαριών = 3.041
- Κλάση 2: 100 ≤ IC50 < 1,000, μέτρια αναστολή, πλήθος ζευγαριών = 3.859
- Κλάση 3: 1,000 ≤ IC50 < 10,000, χαμηλή αναστολή, πλήθος ζευγαριών = 6.558
- Κλάση 4: 10,000 ≤ IC50 < 100,000, καθόλου αναστολή, πλήθος ζευγαριών = 4.776

2.2. Προεπεξεργασία δεδομένων και παραγωγή περιγραφέων

Κατά την προεπεξεργασία των δεδομένων που ελήφθησαν, οι 1D προσδέτες μετατράπηκαν σε 3D, μέσω της παρακάτω ροής εργασίας:

- Αφαίρεση αλάτων από τα αρχεία SMILES
- Αφαίρεση διπλότυπων SMILES
- Προσθήκη υδρογόνων στα SMILES
- Μετατροπή των SMILES σε 3D SDF αρχεία
- Ελαχιστοποίηση ενέργειας στα παραγόμενα SDF αρχεία.
- Ορισμός των μορίων σε PH 7,4.

Η διαδικασία επεξεργασίας των χημικών μορίων που περιγράφεται παραπάνω έγινε με το λογισμικό OpenBabel [6]. Εν συνεχεία, έχοντας επεξεργαστεί τα χημικά μόρια, έγινε εξαγωγή 1D, 2D και 3D πειργραφέων (βλ. 2^{η} έκθεση προόδου) με τα λογισμικά Rcdk [7], RDkit [8] και Mordred[9].

Κατά την προεπεξεργασία των πρωτεϊνικών αλληλουχιών, η αρχική βιβλιοθήκη χωρίστηκε σε μεμονωμένα αρχεία FASTA, με το καθένα να αντιστοιχεί σε έναν συγκεκριμένο πρωτεϊνικό στόχο. Τα αρχεία αυτά φιλτραρίστηκαν, ώστε να διατηρηθούν μόνο οι πρωτεϊνικοί στόχοι για τους οποίους υπάρχουν πειραματικά δεδομένα. Έπειτα πραγματοποιήθηκε ποιοτικός έλεγχος σε κάθε αρχείο FASTA, χρησιμοποιώντας το πακέτο protr [10] για να εξαλειφθούν τυχόν περιπτώσεις εσφαλμένης μορφοποίησης. Στη συνέχεια, έγινε ο υπολογισμός των μοριακών περιγραφέων για τις πρωτεΐνες, χρησιμοποιώντας τη βιβλιοθήκη Rcpi [11]. Έτσι, δημιουργήθηκαν συνολικά 14 σύνολα δεδομένων με περιγραφείς πρωτεϊνών, που εμπίπτουν σε έξι κατηγορίες:

- Κατηγορία 1 - Pseudo-amino acid – 2 σετ περιγραφέων με 130 μεταβλητές συνολικά.
- Κατηγορία 2 - Quasi-sequence-order descriptors – 2 σετ περιγραφέων με 160 μεταβλητές συνολικά.
- Κατηγορία 3 - CTD descriptors – 3 σετ περιγραφέων με 147 μεταβλητές συνολικά.
- Κατηγορία 4 - Conjoint triad descriptors – 1 σετ περιγραφέων με 343 μεταβλητές.
- Κατηγορία 5 - Autocorrelation - 3 σετ περιγραφέων με 720 μεταβλητές συνολικά.
- Κατηγορία 6 - Amino acid composition – 3 σετ περιγραφέων με 8420 μεταβλητές συνολικά.

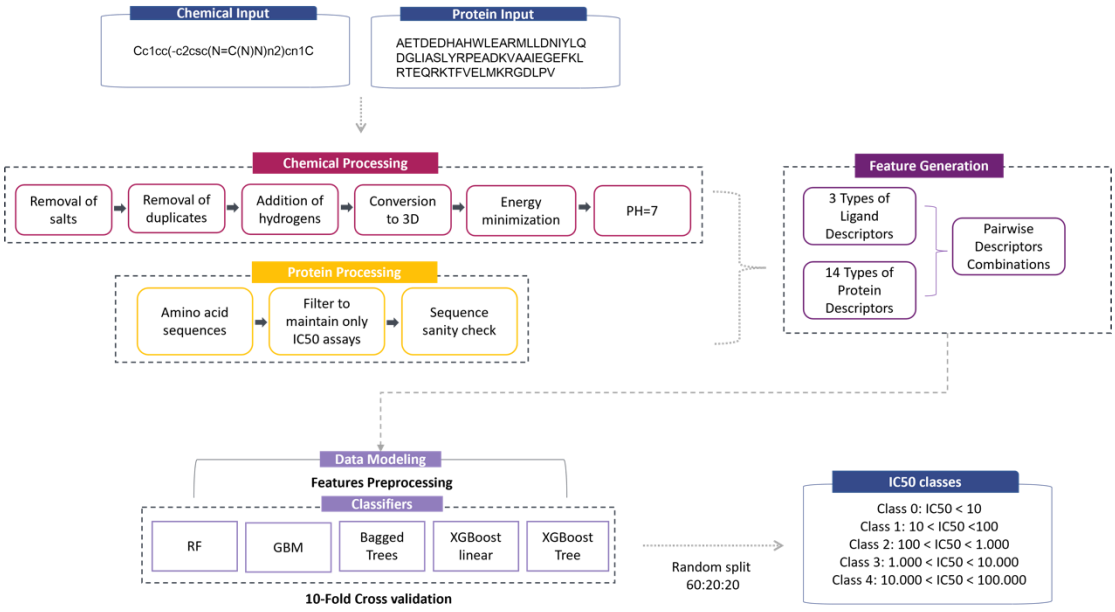
Για τις 3D δομές πρωτεϊνών που ελήφθησαν από την PDB ακολουθήθηκε η εξής διαδικασία προεπερξεργασίας:

- Αφαίρεση μορίων νερού
- Αφαίρεση συγκρυσταλλωμένων προσδετών
- Ορισμός μερικών φορτιών στα άτομα

Η παραπάνω ροή εργασίας υλοποιήθηκε με το λογισμικό Autodock Tools [12]. Στη συνέχεια, οι προσδέτες και πρωτεϊνικοί στόχοι χρησιμοποιήθηκαν για την εκτέλεση υπολογισμών μοριακής προσδέσης με χρήση του λογισμικού Autodock Vina [13], [14]. Από τα αποτελέσματα των υπολογισμών αυτών εξήχθησαν οι μοριακοί περιγραφείς αλληλεπίδρασης με χρήση του λογισμικού Open Drug Discovery Toolkit [15].

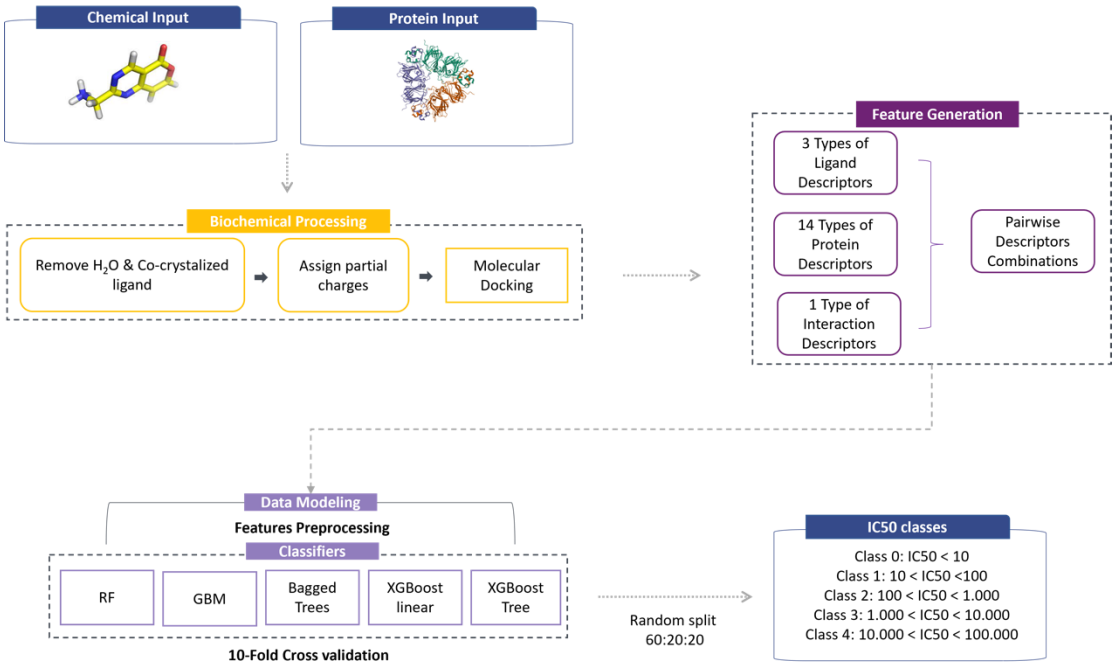
2.3. Μηχανική μάθηση - 1D & 3D δεδομένα

Αρχικά, τα δεδομένα χρησιμοποιήθηκαν για την εκπαίδευση αλγορίθμων μηχανικής μάθησης, ώστε να υπάρχει μια απόδοση ως προς την ακρίβεια διαχωρισμού των 5 κλάσεων. Για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης έγινε συνδυασμός, ανά δυο, των τριών σετ περιγραφέων των προσδετών και των 14 σετ περιγραφέων των πρωτεϊνών, οδηγώντας σε 3 × 14, δίνοντας 42 διαφορετικά δεδομένα εισόδου. Τα χαρακτηριστικά (περιγραφείς χημικών μορίων και πρωτεϊνών) φιλτραρίστηκαν, εξετάζοντας όσα έχουν χαμηλή διακύμανση και υψηλή συσχέτιση (>0,75). Επιπλέον, έγινε επιλογή χαρακτηριστικών με τη μέθοδο Recursive Feature Elimination [16]. Οι ταξινομητές που χρησιμοποιήθηκαν για τη μοντελοποιήση των δεδομένων είναι: 1) Random Forest, 2) XGBoost linear, 3) XGBoost tree, 4) Bootstrap Aggregating-Treebag και 5) Gradient Boosting Machines. Για την αξιολόγηση των αποτελεσμάτων, τα δεδομένα των ζευγών προσδέτη-πρωτεΐνης χωρίστηκαν σε 60:20:20 ως σετ εκπαίδευσης, επικύρωσης και εξωτερικού ελέγχου, αντίστοιχα. Η επιλογή υπερ-παραμέτρων για κάθε ταξινομητή έγινε με τη μέθοδο «10-fold cross-validation». Οι επιλεγμένες υπερ-παράμετροι χρησιμοποιήθηκαν για την εκπαίδευση ενός τελικού μοντέλου, η απόδοση του οποίου αξιολογήθηκε με βάση το εξωτερικό σετ ελέγχου. Η ροή εργασίας που ακολουθήθηκε αποδίδεται στην Εικόνα 1.



Εικόνα 1. Γραφική αναπαράσταση της ροής εργασίας για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης με τα 1D δεδομένα προσδέτη-πρωτεΐνης σε μορφή πίνακα.

Η αντίστοιχη διαδικασία ακολουθήθηκε και για τη δημιουργία των μοντέλων μηχανικής μάθησης, που εκπαιδεύτηκαν με 3D δεδομένα (περιγραφείς αλληλεπίδρασης προσδέτη-πρωτεΐνης). Εδώ, σε καθένα από τα 42 διαφορετικά σετ δεδομένων, προστέθηκαν οι περιγραφείς αλληλεπίδρασης. Στην Εικόνα 2 παρουσιάζεται γραφικά η διαδικασία που ακολουθήθηκε.



Εικόνα 2. Γραφική αναπαράσταση της ροής εργασίας για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης με τα 1D και 3D δεδομένα προσδέτη-πρωτεΐνης σε μορφή πίνακα.

2.4. Μετασχηματισμός περιγραφέων σε εικόνα

Τα χαρακτηριστικά (περιγραφείς προσδέτη/πρωτεΐνης/αλληλεπίδρασης) συναντώνται σε μορφή πίνακα, όπου κάθε γραμμή είναι ένα ζεύγος προσδέτη-πρωτεΐνης, το οποίο αναπαρίσταται από ένα διάνυσμα Ν-θέσεων, με κάθε τιμή του διανύσματος να βρίσκεται ανά στήλη πίνακα, δημιουργώντας έναν πίνακα Ν-στηλών. Στη συγκεκριμένη μεθοδολογία, κύριος σκοπός ήταν η αξιολόγηση της απόδοσης ενός μοντέλου βαθιάς μάθησης, το οποίο δέχεται ως είσοδο το διάνυσμα χαρακτηριστικών ενός ζεύγους προσδέτη–πρωτεΐνης σε μορφή εικόνας. Συνεπώς, το διάνυσμα των χαρακτηριστικών μετατρέπεται σε μια εικόνα, όπου κάθε εικονο-στοχείο έχει τιμή αντίστοιχη με αυτή του περιγραφέα, που αντιπροσωπεύει. Η μετατροπή αυτή έγινε με βάση τη βιβλιοθήκη IGTD [17], με σκοπό τη σύγκριση της απόδοσής της, σε σχέση με τους αλγορίθμους μηχανικής μάθησης, οι οποίοι εκπαιδεύτηκαν με τα χαρακτηριστικά σε μορφή διανύσματος. Η μέθοδος αυτή είναι ελάχιστα μελετημένη στο πεδίο της πρόβλεψης αλληλεπίδρασης προσδέτη-πρωτεΐνης και γι’ αυτό επιλέχθηκε [18]. Άρα, μετασχηματίζεται το πρόβλημα σε ταξινόμηση εικόνων, δίνοντας τη δυνατότητα αξιοποιήσης προ-εκπαιδευμένων συνελικτικών νευρωνικών δικτύων σε δεδομένα εικόνας. Με αυτόν τον τρόπο αίρεται ο περιορισμός του όγκου δεδομένων που απαιτείται για την εκπαίδευση ενός δικτύου βαθιάς μάθησης.

Για τη δημιουργία των εικόνων, στην περίπτωση των περιγραφέων των πρωτεϊνών, όλα τα χαρακτηριστικά (14 σύνολα δεδομένων) συνενώθηκαν σε έναν ενιαίο πίνακα. Στη συνέχεια, φιλτραρίστηκαν, απαλείφοντας μεταβλητές, με συσχέτιση μεγαλύτερη του 0,75. Αντίστοιχα, για τον μετασχηματισμό των περιγραφέων των προσδετών σε εικόνα, συνενώθηκαν και τα τρία σύνολα δεδομένων και αφαιρέθηκαν τα χαρακτηριστικά με μηδενική διακύμανση. Στην περίπτωση των περιγραφέων των αλληλεπιδράσεων, αρχικά έγινε συνένωση των περιγραφέων αυτών με τους περιγραφείς των πρωτεϊνών και προσδετών σε μια ενιαία μήτρα, η οποία μετασχηματίστηκε σε μια εικόνα 60 x 60 pixel εικονο-στοιχείων, ανά ζεύγος, συνδέτη-πρωτεΐνης. Κατά την μετατροπή των χαρακτηριστικών από την μορφή πίνακα σε εικόνες δημιουργήθηκαν δύο σύνολα εικόνων, χρησιμοποιώντας διαφορετική

μέθοδο για τον υπολογισμό των αποστάσεων κατά ζεύγη χαρακτηριστικών: ένα προέκυψε με βάση την Ευκλείδεια απόσταση και ένα, με βάση τον συντελεστή συσχέτισης Pearson.

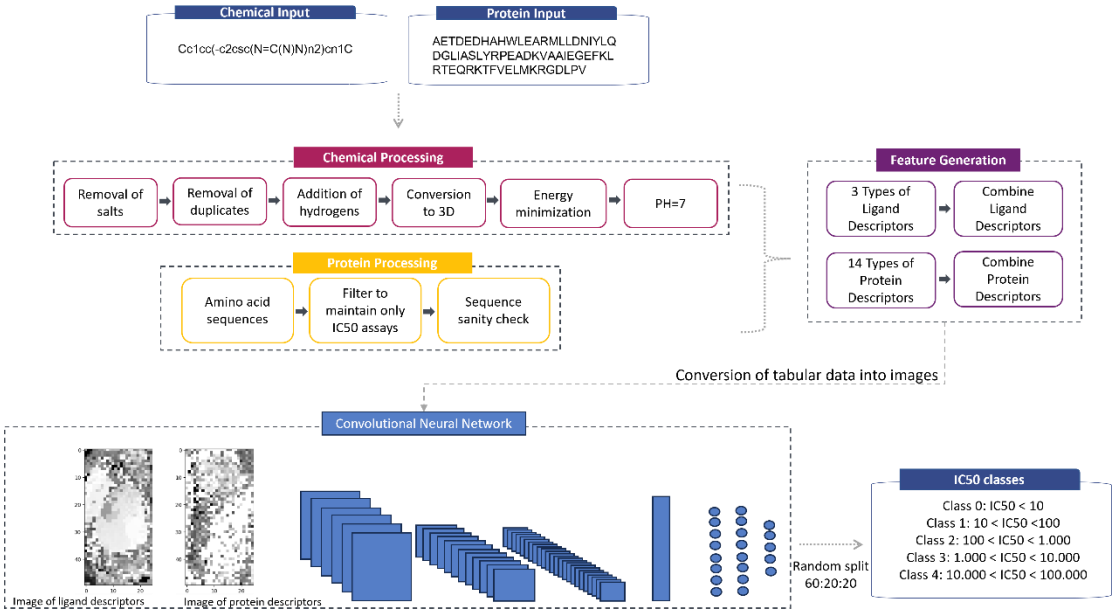
2.5. Βαθιά μάθηση - 1D & 3D δεδομένα

Για την εκπαίδευση νευρωνικών δικτύων βαθιάς μάθησης, με βάση τα 1D δεδομένα, ακολουθήθηκαν δυο ροές εργασίας. Η 1η βασίστηκε στον μετασχηματισμό ενός διανύσματος χαρακτηριστικών σε εικόνα 50 x 25 pixels για τον προσδέτη και την πρωτεΐνη, αντίστοιχα. Στη συνέχεια, οι δυο εικόνες – που, πλέον, φέρουν τους περιγραφείς των προσδετών σε μια εικόνα και τους περιγραφείς των πρωτεϊνών σε άλλη εικόνα - ενώθηκαν. Έτσι, σχηματίζεται μια εικόνα 50 x 50 pixels, η οποία εμπεριέχει τους περιγραφείς, τόσο των προσδετών, όσο και των πρωτεϊνών. Η εικόνα αυτή χρησιμοποιείται ως είσοδος για το νευρωνικό δίκτυο. Για τη 2η ροή εργασίας, κάθε διάνυσμα προσδέτη και πρωτεΐνης μετασχηματίστηκε σε μια εικόνα 35 x 35 pixels, αντίστοιχα. Η εικόνα με τους περιγραφείς των προσδετών τροφοδοτήθηκε σε ένα συνελικτικό δίκτυο και η εικόνα με τους περιγραφείς των πρωτεϊνώντροφοδοτήθηκε σε ένα άλλο συνελικτικό δίκτυο. Οι πίνακες χαρακτηριστικών, που προέκυψαν από τα δυο ξεχωριστά νευρωνικά δίκτυα, ενοποιήθηκαν και χρησιμοποιήθηκαν ως είσοδος σε ένα ενιαίο δίκτυο ταξινόμησης.

Για κάθε ροή εργασίας ακολουθήθηκε η ίδια μέθοδος για την ανάπτυξη των δικτύων. Αρχικά, επιλέχθηκαν τα μοντέλα: : EfficientNet, MobileNet, ResNet-50, ResNet-101, VGG16 και VGG19. Σε κάθε μοντέλο έγινε αρχικοποίηση των βαρών με βάση την εκπαίδευση των δικτύων στο σετ δεδομένων ImageNet [19], ενώ αναπτύχθηκε και ένα νευρωνικό δίκτυο με ειδικό σχεδιασμό. Η επιλογή των παραμέτρων έγινε με «grid search» σε ένα πεπερασμένο σύνολο τιμών, που επιλέχθηκε χειροκίνητα. Οι υπερ-παράμετροι που αξιολογήθηκαν ήταν οι εξής:

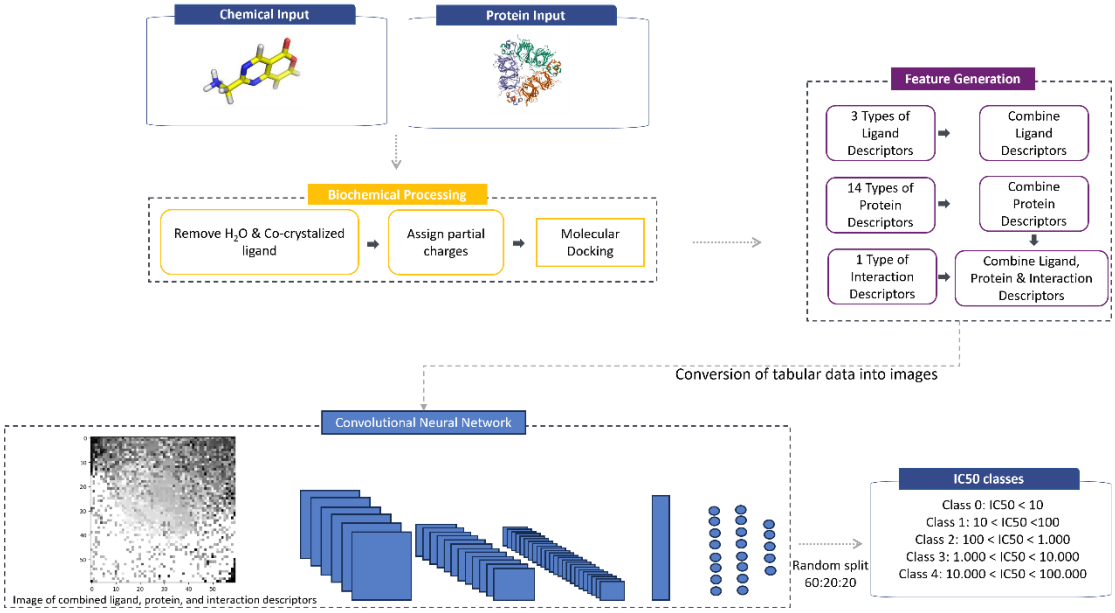
- learning rate
- batch size
- training epochs
- dropout rates

Επίσης, αξιολογήθηκε το πλήθος νευρώνων στα επίπεδα ταξινόμησης και το επίπεδο στο οποίο γίνεται εξαγωγή χαρακτηριστικών από την προ-εκπαιδευμένη συνελικτική βάση. Η μέθοδος αξιολόγησης των μοντέλων έγινε, αντίστοιχα, με αυτή στην μηχανική μάθηση, όπου τα δεδομένα χωρίστηκαν σε 60:20:20 ως σετ εκπαίδευσης, επικύρωσης και εξωτερικού ελέγχου, αντίστοιχα. Κατά την εκπαίδευση των δικτύων χρησιμοποιήθηκε πρόωρη διακοπή και μείωση κατά 0,1 του ρυθμού εκμάθησης, όταν η απώλεια επικύρωσης έφτανε σε κορεσμό. Αυτή η διαδικασία επιλέχθηκε για να αποφευχθεί η υπερ-προσαρμογή των μοντέλων στα δεδομένα. Η μεθοδολογία παρουσιάζεται στην Εικόνα 3.



Εικόνα 3. Γραφική αναπαράσταση της ροής εργασίας για την εκπαίδευση των αλγορίθμων βαθιάς μάθησης με τα 1D δεδομένα προσδέτη-πρωτεΐνης σε μορφή εικόνας.

Για την εκπαίδευση νευρωνικών δικύων βαθιάς μάθησης με βάση τα 3D δεδομένα ακολουθήθηκε η ίδια μεθοδολογία για την ανάπτυξη των δικτύων με τη διαφορά ότι σε αυτή την περίπτωση είχαμε μόνο μια ροή εργασίας. Όπως περιγράφεται και παραπάνω (ενότητα 2.4), οι περιγραφείς αλληλεπιδράσεων ενώθηκαν με τους περιγραφείς των πρωτεϊνών και προσδετών σε μια ενιαία μήτρα, η οποία μετασχηματίστηκε σε εικόνες 60 x 60 pixel, ανά ζεύγος συνδέτη-πρωτεΐνης. Οι εικόνες αυτές τροφοδοτήθηκαν στα νευρωνικά δίκτυα. Η προσέγγιση αυτή αναπαρίσταται στην Εικόνα 4.



Εικόνα 4. Γραφική αναπαράσταση της ροής εργασίας για την εκπαίδευση των αλγορίθμων βαθιάς μάθησης με τα 1D και τα 3D δεδομένα προσδέτη-πρωτεΐνης σε μορφή εικόνας.

3. Αποτελέσματα

3.1. Απόδοση μοντέλων μηχανικής & βαθιάς μάθησης

Στην περίπτωση των μοντέλων μηχανικής μάθησης, τα οποία εκπαιδεύτηκαν με τα 1D δεδομένα (τους περιγραφείς των πρωτεϊνών και προσδετών) βρέθηκε ότι βέλτιστος ταξινόμητης είναι τα Gradient Boosting Machines. Ο συγκεκριμένος αλγόριθμος χρησιμοποιώντας τους περιγραφείς από το rdkit και τους περιγραφείς των πρωτεϊνών CTDD αποδίδει: ισορροπημένη ακρίβεια: 54-67%, ακρίβεια: 28-56% και ανάκληση: 15-57% ανά κατηγορία στο σετ δοκιμής. Σημαντικό είναι να σημειώσουμε ότι σε αυτή την περίπτωση, μετά από την ανάλυση σημαντικότητας των μεταβλητών ανά μοντέλο, οι σημαντικές μεταβλητές ήταν μόνο περιγραφείς των πρωτεϊνών. Έτσι, φαίνεται ότι ο αλγόριθμος δε λαμβάνει υπόψη του τα χημικά μόρια (προσδέτες), αλλά μόνο την αμινοξική αλληλουχία.

Αντίστοιχα, στην περίπτωση των μοντέλων βαθιάς μάθησης, που χρησιμοποιούν τους 1D περιγραφείς, προέκυψε ότι τα μοντέλα EfficientNet, ResNet-50 και ResNet-101 απέτυχαν να γενικεύσουν, ενώ τα μοντέλα MobileNet, Vgg16 και Vgg19 απέδωσαν: συνολική ακρίβεια 45-48% για τα δεδομένα εκπαίδευσης, 43% για το σετ επικύρωσης και 42% για το σετ δοκιμών. Μεταξύ των μοντέλων μηχανικής και βαθιάς μάθησης, που χρησιμοποιούν δεδομένα 1D (χωρίς περιγραφείς αλληλεπίδρασης), παρατηρείται μια ελαφρά αύξηση στην ακρίβεια (1-3%).

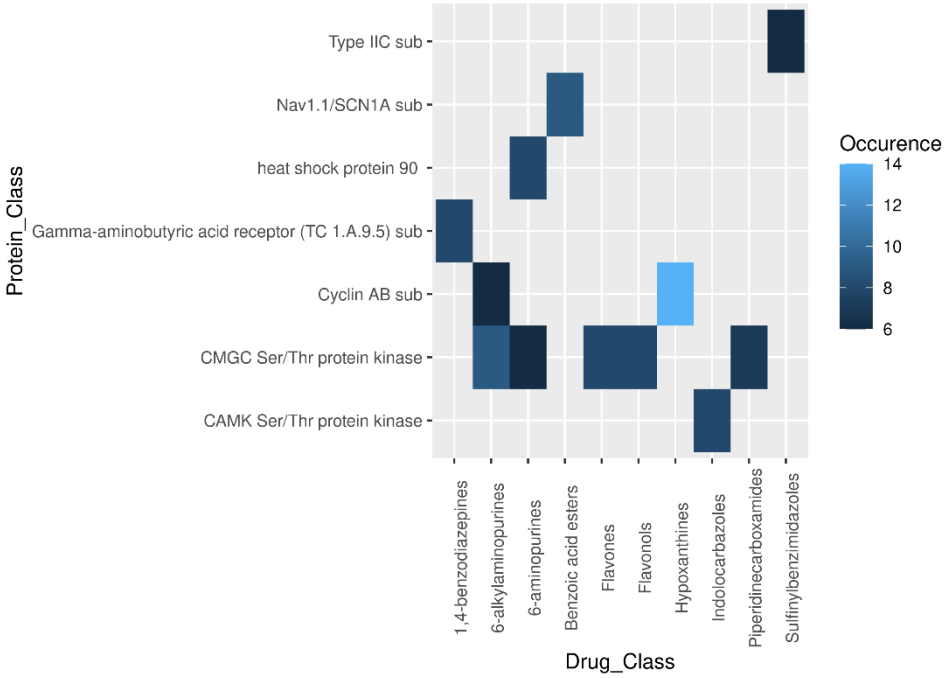
Για τα μοντέλα μηχανικής μάθησης, που χρησιμοποιούν και περιγραφείς αλληλεπίδρασης συνδέτη-πρωτεΐνης, δηλαδή εμπλουτίζονται από τα 3D δεδομένα, βρέθηκε ως καλύτερος ταξινομητής ο xgbTree. Σε αυτήν την περίπτωση, ο ταξινομητής απέδωσε καλύτερα, χρησιμοποιώντας τους περιγραφείς από το rdkit και τους περιγραφείς των πρωτεϊνών CTDD (ισορροπημένη ακρίβεια 73-82%, ακρίβεια 55-76% και ανάκληση 52-72% ανά κατηγορία), με αύξηση της απόδοσης, σε σχέση με τα 1D δεδομένα. Αντίστοιχη αύξηση παρατηρείται και στην περίπτωση των μοντέλων βαθιάς μάθησης με περιγραφείς αλληλεπίδρασης συνδέτη-πρωτεΐνης. Τα μοντέλα, σε αυτή την περίπτωση, παρουσιάζουν αύξηση στην ακρίβεια και επιτυγχάνουν 99% ακρίβεια κατά την εκπαίδευση, ωστόσο κατά την εξωτερική αξιολόγηση η απόδοση τους μειώνεται, υποδηλώνοντας την υπερ-προσαρμογή του μοντέλου στα δεδομένα. Στον Πίνακα 1 βρίσκονται, συνοπτικά, τα αποτελέσματα του βέλτιστου μοντέλου ανά τύπο δεδομένων.

Method	**Data Type**	**Descriptors**	**Overall train accuracy**	**Overall validation accuracy**	**Overall test accuracy**
ML	Tabular	Protein - Ligand	43%	42%	41%
DL	Image	Protein – Ligand	48%	42%	42%
ML	Tabular	Protein – Ligand – Interaction	65%	63%	61%
ML	Image	Protein – Ligand – Interaction	99%	60%	66%

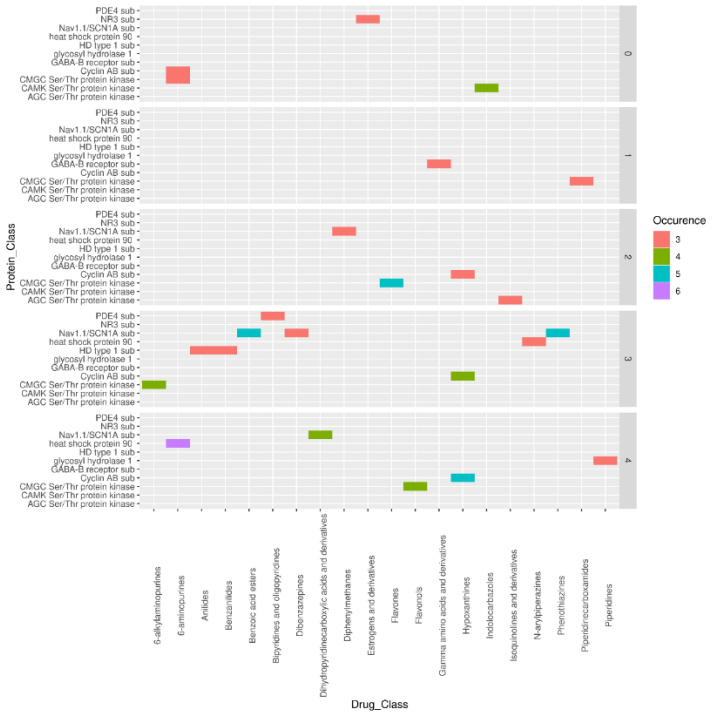
Πίνακας 1. Σύνοψη της συνολικής απόδοσης των διαφορετικών μεθόδων που χρησιμοποιούνται ανά δεδομένα εισόδου, δεδομένα εκπαίδευσης, επικύρωσης και εξωτερικού ελέγχου.

3.2. Απόδοση μοντέλων ανά οικογένειες προσδέτη-πρωτεΐνης

Οι προσδέτες και οι πρωτεΐνες κατηγοριοποιήθηκαν με βάση την οικογένεια στην οποία ανήκουν και έγινε εκ νέου ανάλυση της απόδοσης των μοντέλων, με βάση την ακρίβεια πρόβλεψης στα ζεύγη μεταξύ των οικογενειών. Ενδεικτικά αποτελέσματα της ανάλυσης αυτής βλέπουμε στις Εικόνες 5 και 6.



Εικόνα 5. Ζεύγη οικογένειας προσδέτη-πρωτεΐνης με ακρίβεια πρόβλεψης άνω του 90% στα δεδομένα εξωτερικού ελέγχου.



Εικόνα 6. Ζεύγη οικογένειας προσδέτη-πρωτεΐνης ανά κατηγορία IC_{50} με ακρίβεια πρόβλεψης άνω του 90% στα δεδομένα δοκιμής.

4. Μελλοντικές Προοπτικές

Με γνώμονα ότι η απόδοση και η αξιοπιστία των μοντέλων είναι χαμηλή, φαίνεται να ενθαρρύνεται ο περαιτέρω εμπλουτισμός με πειραματικά δεδομένα. Τα δεδομένα αυτά δύναται να ληφθούν από το αποθετήριο PDBbind [20], εξασφαλίζοντας περισσότερα δεδομένα εκπαίδευσης και μειώνοντας τον θόρυβο.

5. Ευχαριστίες

Αποδίδονται ευχαριστίες στο EuroHPC Joint Undertaking για την πρόσβαση στο EuroHPC supercomputer Karolina, hosted by IT4Innovations (Czech Republic), όπως επετεύχθη στο πλαίσιο της πρόσκλησης EuroHPC Development Access και στα μέλη της ομάδας έργου: Β. Παναγιωτόπουλο, ΜΤ Ματσούκα και Θ. Κατσίλα.

6. Βιβλιογραφία

[1] C. Yung-Chi and W. H. 1973. Prusoff. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem Pharmacol* 22(23): 3099–3108. doi: 10.1016/0006-2952(73)90196-2.

[2] D. Weininger. 2002. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28(1): 31–36. doi: 10.1021/ci00057a005.

[3] D. Mendez *et al.* 2019. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res* 47(D1): D930–D940. doi: 10.1093/nar/gky1075.

[4] D. S. Wishart *et al.* 2018. DrugBank 5.0: A major update to the DrugBank database for 2018,. *Nucleic Acids Res* 46(D1): D1074–D1082. doi: 10.1093/nar/gkx1037.

[5] H. M. Berman *et al.* 2000. The Protein Data Bank. *Nucleic Acids Res* 28(1): 235–242. doi: 10.1093/NAR/28.1.235.

[6] N. M. O'Boyle *et al.* 2011. Open Babel: An Open chemical toolbox,. *J Cheminform* 3(10). doi: 10.1186/1758-2946-3-33.

[7] R. Guha. 2007. Chemical Informatics Functionality in R. *J Stat Softw* 18(5):1–16. doi: 10.18637/JSS.V018.I05.

[8] G. Landrum *et al.* 2023. rdkit/rdkit: 2023_03_1 (Q1 2023) Release. Zenodo

[9] H. Moriwaki *et al.* 2018. Mordred: A molecular descriptor calculator. *J Cheminform* 10(1):1–14. doi: 10.1186/s13321-018-0258-y.

[10] N. Xia *et al.* 2015. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* 31(11): 1857–1859. doi: 10.1093/bioinformatics/btv042.

[11] D. S. Cao *et al.* 2015. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics* 31(2): 279–281. doi: 10.1093/bioinformatics/btu624.

[12] G. M. Morris *et al.* 2009. Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* 30(16): 2785–2791. doi: 10.1002/jcc.21256.

[13] O. Trott and A. J. Olson. 2009. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31(2):455-461. doi: 10.1002/jcc.21334.

[14] J. Eberhardt *et al.* 2021. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model* 61(8): 3891–3898. doi: 10.1021/acs.jcim.1c00203.

[15] M. Wójcikowski, P. Zielenkiewicz, and P. Siedlecki. 2015. Open Drug Discovery Toolkit (ODDT): A new open-source player in the drug discovery field. *J Cheminform* 7(26) . doi: 10.1186/s13321-015-0078-2.

[16] I. Guyon *et al.* 2002. Gene selection for cancer classification using support vector machines. *Mach Learn* 46(1–3): 389–422. doi: 10.1023/A:1012487302797/METRICS.

[17] Y. Zhu *et al.* 2021. Converting tabular data into images for deep learning with convolutional neural networks. *Sci Rep* 11(1): 1–11. doi: 10.1038/s41598-021-90923-y.

[18] R. Gorantla *et al.* 2023. From Proteins to Ligands: Decoding Deep Learning Methods for Binding Affinity Prediction. *J Chem Inf Model.* doi: 10.1021/acs.jcim.3c01208.

[19] J. Deng *et al.* 2009. ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA pp. 248–255. doi: 10.1109/cvpr.2009.5206848.

[20] R. Wang *et al.* 2005. The PDBbind Database: Methodologies and Updates. *J Med Chem* 48(12): 4111–4119. doi: 10.1021/jm048957q.