

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ - ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ**  
**ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ**  
**ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ**



**ΤΟΜΕΑΣ: ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ**  
**ΕΡΓΑΣΤΗΡΙΟ: MULTI-DIMENSIONAL DATA ANALYSIS AND**  
**KNOWLEDGE MANAGEMENT**

**Διπλωματική Εργασία**

του φοιτητή του Τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας  
Υπολογιστών της Πολυτεχνικής Σχολής του Πανεπιστημίου Πατρών

**ΜΗΝΑ ΧΑΜΑΜΤΖΟΓΛΟΥ του ΙΩΑΚΕΙΜ**

**ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ: 1020700**

Θέμα

**Εντοπισμός αλληλεπιδράσεων πρωτεΐνων μέσω συμπλήρωσης  
μητρώων και αλγορίθμων βαθιάς μάθησης**

Επιβλέπων

Καθηγητής Βασίλειος Μεγαλοοικονόμου

Αριθμός Διπλωματικής Εργασίας:

Πάτρα, Ιούνιος 2020

## ΠΙΣΤΟΠΟΙΗΣΗ

Πιστοποιείται ότι η διπλωματική εργασία με θέμα

Εντοπισμός αλληλεπιδράσεων πρωτεΐνών μέσω συμπλήρωσης  
μητρώων και αλγορίθμων βαθιάς μάθησης

του φοιτητή του Τμήματος Ηλεκτρολόγων Μηχανικών και  
Τεχνολογίας Υπολογιστών

Μηνά Χαμαμτζόγλου του Ιωακείμ

(Α.Μ.: 1020700)

παρουσιάτηκε δημόσια και εξετάστηκε στο τμήμα Ηλεκτρολόγων  
Μηχανικών και Τεχνολογίας Υπολογιστών στις

\_\_\_\_/\_\_\_\_/\_\_\_\_

Ο Επιβλέπων

Ο Διευθυντής του Τομέα

Βασίλειος Μεγαλοοικονόμου  
*Καθηγητής*

Βασίλειος Παλιούράς  
*Καθηγητής*

## **Στοιχεία διπλωματικής εργασίας**

**Θέμα: Εντοπισμός αλληλεπιδράσεων πρωτεΐνών μέσω συμπλήρωσης μητρώων και αλγορίθμων βαθιάς μάθησης**

**Φοιτητής: Μηνάς Χαμαμτζόγλου του Ιωακείμ**

**Ομάδα επίβλεψης  
Καθηγητής Βασίλειος Μεγαλοοικονόμου**

**Διδακτορικός Θωμάς Παπαστεργίου**

**Εργαστήριο  
Multi-Dimensional Data Analysis and Knowledge Management**

**Περίοδος εκπόνησης της εργασίας:  
Νοέμβριος 2019 - Ιούνιος 2020**

**Η εργασία αυτή γράφτηκε στο XΕΛΤΕΧ και χρησιμοποιήθηκε η γραμματοσειρά GFS Didot του Greek Font Society.**

## Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η μελέτη και η κατανόηση των πρωτεΐνικών αλληλεπιδράσεων (Protein Protein Interactions - PPIs), καθώς και η δημιουργία μοντέλων μηχανικής μάθησης (machine learning models) για τον εντοπισμό των PPIs.

Η αλληλεπίδραση μεταξύ πρωτεΐνών αποτελεί ένα αντικείμενο μεγάλης σημασίας για την κατανόηση της λειτουργίας των πρωτεΐνών και έναν από τους βασικούς στόχους της συστηματικής βιολογίας (systems biology) [1]. Αφού πρώτα παρουσιαστεί το θεωρητικό βιολογικό υπόβαθρο της εργασίας, γίνεται αναφορά στους τρόπους εντοπισμού των εν λόγω αλληλεπιδράσεων, τόσο πειραματικά όσο και υπολογιστικά, ενώ το κεφάλαιο ολοκληρώνεται με την επιγραμματική αναφορά στις μεγαλύτερες βάσεις δεδομένων πρωτεΐνικών αλληλεπιδράσεων.

Το δεύτερο μέρος της εργασίας αφορά την εισαγωγή στις βασικές έννοιες της μηχανικής μάθησης. Με τον όρο μηχανική μάθηση αναφερόμαστε σε μεθόδους ανάλυσης και επεξεργασίας δεδομένων με σκοπό την αυτοματοποίηση μοντέλων μηχανών και υπολογιστικών συστημάτων. Δίνονται ορισμοί και αναπτύσσονται έννοιες όπως τα νευρωνικά δίκτυα που θα αναπτυχθούν λεπτομερώς και στην υλοποίηση της εργασίας.

Έπειτα από την ολοκλήρωση του θεωρητικού πλαισίου της εργασίας, υλοποιείται η προσέγγιση της παρούσας διπλωματικής. Ξεκινάει με τον ορισμό του προβλήματος και την ανάλυση της εξαγωγής δεδομένων αλληλεπιδράσεων πρωτεΐνών. Έπειτα από τη δημιουργία του συνόλου δεδομένων, παρουσιάζονται τεχνικές συμπλήρωσης μητρώων (*matrix completion*), μαθηματικές μέθοδοι συμπλήρωσης μη-υπαρχόντων τιμών στο σύνολο των δεδομένων, με σκοπό την μείωση της διαστασιμότητας του προβλήματος και την βελτίωση της απόδοσης των αλγορίθμων μας. Παρουσιάζεται μια τεχνική παραγοντοποίησης μητρώου μέσω stochastic gradient descent (*matrix factorization with SGD*), καθώς και μια τεχνική τανηστικής αποδόμησης (*tensor decomposition*). Έπειτα από την επεξεργασία των δεδομένων ακολουθεί η ανάπτυξη νευρωνικών δικτύων για τον εντοπισμό αλληλεπιδράσεων. Παρουσιάζονται λεπτομερώς οι αρχιτεκτονικές που χρησιμοποιήθηκαν στα πλαίσια της εργασίας, ενώ συγκρίνεται η απόδοση των εν λόγω δομών με υπάρχοντα μοντέλα. Η εργασία ολοκληρώνεται με την παρουσίαση των μελλοντικών σκέψεων σχετικά με την εξέλιξη των μοντέλων και την βελτίωση των αποτελεσμάτων.

**Λέξεις-Κλειδιά:** Πρωτεΐνες, αλληλεπιδράσεις πρωτεΐνών, συμπλήρωση μητρώου, παραγοντοποίηση μητρώου, τανυστική αποδόμηση, μηχανική μάθηση, νευρωνικά δίκτυα.

## Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της διπλωματικής μου εργασίας κ. Βασίλειο Μεγαλοοικονόμου για την εμπιστοσύνη που μου έδειξε καθ' όλη τη διάρκεια εκπόνησης της εργασίας αλλά και την πολύτιμη καθοδήγησή του.

Στη συνέχεια, ευχαριστώ ιδιαιτέρως τον διδάκτορα Θωμά Παπαστεργίου για τις πολύτιμες συμβουλές αλλά και τις υποδείξεις του σχετικά με την κατεύθυνση της εργασίας, καθώς χωρίς αυτές η συγκεκριμένη εργασία δεν θα μπορούσε να ολοκληρωθεί. Παράλληλα, θα ήθελα να ευχαριστήσω όλα τα παιδιά του εργαστηρίου Multi-Dimensional Data Analysis and Knowledge Management για το όμορφο κλίμα και την ενθάρρυνση σε όλη την πορεία αυτής της εργασίας.

Εξαιρετικά ευγνώμων είμαι προς την οικογένειά μου, η οποία στήριξε όλες μου τις επιλογές και για τις θυσίες τους προκειμένου να ολοκληρώσω τις σπουδές μου.

Ευχαριστώ ιδιαίτερα τον Παναγιώτη και τον Βασίλη για τις όμορφες και εποικοδομητικές κουβέντες, προσφέροντάς μου συνεχώς νεα ερεθίσματα και βοηθώντας με να δω την εργασία με νεα οπτική. Τέλος, θα ήθελα να ευχαριστήσω όλους τους κοντινούς μου ανθρώπους, που στάθηκαν δίπλα μου τόσο στις εύκολες όσο και στις δύσκολες στιγμές.

---

# ΠΕΡΙΕΧΟΜΕΝΑ

<b>Κατάλογος σχημάτων</b>	<b>vii</b>
<b>Κατάλογος πινάκων</b>	<b>ix</b>
<b>1 Πρωτεΐνες</b>	<b>3</b>
1.1 Ορισμός, Λειτουργίες . . . . .	3
1.2 Αλληλεπίδραση πρωτεΐνης με πρωτεΐνη . . . . .	5
1.3 Μέθοδοι εντοπισμού αλληλεπιδράσεων πρωτεΐνης με πρωτεΐνη . . . . .	6
1.3.1 Πειραματικές μέθοδοι εντοπισμού . . . . .	7
1.3.2 Υπολογιστικές μέθοδοι εντοπισμού . . . . .	14
1.3.3 Βάσεις δεδομένων αλληλεπιδράσεων πρωτεΐνης με πρωτεΐνη . . . . .	19
<b>2 Μηχανική Μάθηση</b>	<b>22</b>
2.1 Εισαγωγή . . . . .	22
2.1.1 Ορισμός . . . . .	22
2.1.2 Ιστορική Αναδρομή . . . . .	23
2.2 Είδη μηχανικής μάθησης . . . . .	25
2.2.1 Επιβλεπόμενη Μάθηση . . . . .	26
2.2.2 Μη Επιβλεπόμενη Μάθηση . . . . .	27
2.2.3 Ενισχυτική Μάθηση . . . . .	28
2.3 Νευρωνικά Δίκτυα . . . . .	29
2.3.1 Νευρώνες . . . . .	29
2.3.2 Λειτουργία και Οργάνωση . . . . .	30
2.3.3 Αρχιτεκτονικές νευρωνικών δικτύων . . . . .	32
<b>3 Γλοποίηση</b>	<b>35</b>
3.1 Ορισμός προβλήματος . . . . .	36
3.2 Δεδομένα . . . . .	37

3.2.1	Ανάκτηση δεδομένων πρωτεϊγικών αλληλεπιδράσεων	37
3.2.2	Δημιουργία κομματιών επιφάνειας . . . . .	39
3.2.3	Ανάθεση κατηγορίας . . . . .	42
3.2.4	Χαρακτηριστικά δεδομένων . . . . .	44
3.2.5	Σύνοψη δεδομένων εισόδου . . . . .	48
3.3	Προεπεξεργασία . . . . .	50
3.3.1	Εργαλεία . . . . .	50
3.3.2	Διαδικασία . . . . .	50
3.4	Συμπλήρωση μητρώου . . . . .	53
3.4.1	Συμπλήρωση μητρώου μέσω SGD . . . . .	54
3.4.2	Παραγοντοποίηση μητρώου μέσω ταυυστικής αποδόμησης . . . . .	58
3.5	Εκπαίδευση . . . . .	63
3.5.1	Πλήρως διασυνδεδεμένη αρχιτεκτονική . . . . .	63
3.5.2	Συνελικτική αρχιτεκτονική . . . . .	65
<b>4</b>	<b>Αποτελέσματα</b>	<b>67</b>
4.1	Αποτελέσματα Εκπαίδευσης . . . . .	67
4.2	Μελλοντικές Εξελίξεις . . . . .	78
	<b>Βιβλιογραφία</b>	<b>79</b>

---

# ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

1.1 Δομή ενός αμινοξέος . . . . .	4
1.2 Δευτερογενής δομή πρωτεΐνων . . . . .	4
1.3 Μέθοδος Tap Tagging . . . . .	8
1.4 Μέθοδος Co-immunoprecipitation . . . . .	9
1.5 Απλοποιημένη αναπαράσταση PCAs . . . . .	10
1.6 Κύκλος μεθόδου Phage Display . . . . .	10
1.7 Μέθοδος Yeast Two-Hybrid . . . . .	12
1.8 Παράδειγμα Synthetic Lethality . . . . .	12
1.9 Παράδειγμα πρωτεΐνη με 3 domains ( <i>Pyruvate kinase</i> ) . . . . .	16
1.10 Παράδειγμα φυλογενετικού δέντρου . . . . .	17
1.11 Παράδειγμα φυλογενετικού προφίλ . . . . .	17
2.1 Δομή του Perceptron . . . . .	24
2.2 Support Vector Machines . . . . .	25
2.3 Επιβλεπόμενη Μάθηση . . . . .	26
2.4 Μη Επιβλεπόμενη Μάθηση . . . . .	27
2.5 Ενισχυτική Μάθηση . . . . .	28
2.6 Βιολογική Δομή Νευρώνα . . . . .	29
2.7 Μαθηματική δομή τεχνητού νευρώνα . . . . .	31
2.8 Feedforward αρχιτεκτονική . . . . .	32
2.9 Recurrent αρχιτεκτονική . . . . .	33
2.10 Convolutional αρχιτεκτονική . . . . .	34
2.11 General Adversarial Network αρχιτεκτονική . . . . .	34
3.1 Παράδειγμα επιφάνειας αλληλεπίδρασης σε PPI . . . . .	36
3.2 Εργαλείο PISA . . . . .	38
3.3 Εργαλείο PISCES . . . . .	39
3.4 Παράδειγμα χρήσης solvent angle . . . . .	41
3.5 Παράδειγμα unlabelled κομματιού . . . . .	43

3.6 Αριθμός ελλειπών τιμών . . . . .	51
3.7 Standardization . . . . .	52
3.8 Διακύμανση SGD . . . . .	55
3.9 Σχετικές απώλειες SGD . . . . .	58
3.10 Αποδόμηση ταυυστή τρίτης τάξης . . . . .	60
3.11 Απώλειες CP αποδόμησης . . . . .	62
3.12 10-fold cross validation . . . . .	64
3.13 Fully connected αρχιτεκτονική . . . . .	65
3.14 Max Pooling . . . . .	65
3.15 Flattening . . . . .	66
4.1 Ακρίβεια πλήρως διασυνδεδεμένου νευρωνικού δικτύου . . . . .	69
4.2 Απώλειες πλήρως διασυνδεδεμένου νευρωνικού δικτύου . . . . .	70
4.3 Ακρίβεια συνελικτικού νευρωνικού δικτύου . . . . .	70
4.4 Απώλειες συνελικτικού νευρωνικού δικτύου . . . . .	71
4.5 Τυπική Καμπύλη ROC . . . . .	73
4.6 Καμπύλη ROC πλήρως διασυνδεδεμένου νευρωνικού δικτύου . . . . .	74
4.7 Καμπύλη ROC συνελικτικού νευρωνικού δικτύου . . . . .	74
4.8 Καμπύλη Precision Recall πλήρως διασυνδεδεμένου νευρωνικού δικτύου . . . . .	75
4.9 Καμπύλη Precision Recall συνελικτικού νευρωνικού δικτύου	76

---

# ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1.1	Λειτουργίες πρωτεΐνων . . . . .	5
1.2	Τύποι Αλληλεπιδράσεων Πρωτεΐνης με Πρωτεΐνη . . . . .	6
1.3	Σύνοψη <i>in vitro</i> τεχνικών εντοπισμού PPIs . . . . .	13
1.4	Σύνοψη <i>in vivo</i> τεχνικών εντοπισμού PPIs . . . . .	14
1.5	Σύνοψη <i>in silico</i> τεχνικών εντοπισμού PPIs (1) . . . . .	18
1.6	Σύνοψη <i>in silico</i> τεχνικών εντοπισμού PPIs (2) . . . . .	19
3.1	Σύνοψη δεδομένων εισόδου . . . . .	49
3.2	Τυπική απόκλιση προ-επεξεργασμένων δεδομένων . . . . .	51
3.3	Συνάρτηση Απώλειας - Ορισμός, Παράγωγοι . . . . .	57
3.4	Δημοφιλείς τύποι $\beta_k$ [2] . . . . .	62
4.1	Χρόνοι σύγκλισης μεθόδων συμπλήρωσης μητρώων . . . . .	68
4.2	Χρόνοι σύγκλισης νευρωνικών δικτύων . . . . .	68
4.3	Μετρικές απόδοσης δυαδικής κατηγοριοποίησης . . . . .	72
4.4	Αποτελέσματα - Σύγκριση Μεθόδων . . . . .	77

---

# ΕΙΣΑΓΩΓΗ

Οι πρωτεΐνες αποτελούν μια κατηγορία μακρομορίων που ελέγχει όλες τις βιολογικές διαδικασίες μέσα σε ένα κύταρρο και είναι άρογκτα συνδεδεμένες με την υγεία των ανθρώπων. Παρ' ότι αρκετές πραγματοποιούν τις διεργασίες τους ως ανεξάρτητες οντότητες, η πλειοψηφία των πρωτεϊνών αλληλεπιδρά με άλλες πρωτεΐνες για την ορθή βιολογική δραστηριότητα. Επομένως, για την καλύτερη κατανόηση της λειτουργίας των πρωτεϊνών και συνεπώς και των οργανισμών, είναι σημαντικό ζήτημα ο εντοπισμός και η μελέτη των αλληλεπιδράσεων μεταξύ των πρωτεϊνών (γνωστές και ως *Protein Protein Interactions - PPI*). Αν και στη θεωρία φοίνεται ως μια απλή διαδικασία, πρακτικά αποτελεί ένα εξαιρετικά δύσκολο ζήτημα, διότι τα κύταρρα αντιδρούν σε μια πληθώρα ερεθισμάτων, καθιστώντας την έκφραση των πρωτεϊνικών διεργασιών μια δυναμική διαδικασία με πολλαπλές παραμέτρους. Παράλληλα, οι πρωτεΐνες που συμμετέχουν σε μια αλληλεπίδραση μπορεί να μην είναι πάντα ενεργοποιημένες ή εμφανείς (στην πραγματικότητα οι περισσότερες αλληλεπιδράσεις είναι προσωρινού χαρακτήρα και απαιτούν έναν αριθμό συνθηκών για να πραγματοποιηθούν).

Οι ίδιες οι αλληλεπιδράσεις αποτελούνται από εναν συνδυασμό υδροφοβικών δεσμών με δυνάμεις *Van der Waals* και γέφυρες αλάτων (*salt bridges*) σε συγκεκριμένες περιοχές (*binding domains*) σε κάθε πρωτεΐνη. Ο εντοπισμός τους γίνεται εργαστηριακά μέσω μιας σειράς βιοχημικών πειραμάτων, τόσο *in vivo* όσο και *in vitro* (π.χ. co-immunoprecipitation, TAP tagging, X-ray crystallography, Yeast two-hybrid κ.α.). Ωστόσο, ο τεράστιος αριθμός πιθανών αλληλεπιδράσεων σε συνδυασμό με το μεγάλο κόστος των παραπάνω πειραμάτων καθιστά αναγκαία τη δημιουργία αυτοματοποιημένων προσεγγίσεων για τον εντοπισμό ή/και την πρόβλεψη αλληλεπιδράσεων. Οι προσεγγίσεις αυτές, εκτελεσμένες σε υπολογιστή ή μέσω υπολογιστικής προσωμοίωσης, ονομάζονται μέθοδοι *in silico* και αποτελούν κομμάτι της βιοπληροφορικής (*bioinformatics*), τομέα που αφορά την

κατασκευή μεθόδων και εργαλείων λογισμικού για την κατανόηση και επεξεργασία βιολογικών δεδομένων. Η συγκεκριμένη εργασία ασχολείται με δομικές προσεγγίσεις *in silico* τεχνικών, οπου προβλέπεται αλληλεπιδραση μεταξύ δυο πρωτεΐνων με βάση τα δομικά τους χαρακτηριστικά. Ειδικότερα, ασχολείται με τη δημιουργία μοντέλων μηχανικής μάθησης για την πρόβλεψη αλληλεπιδράσεων μεταξύ πρωτεΐνων.

Η αύξηση της υπολογιστικής ισχύος έχει οδηγήσει τα τελευταία χρόνια στην ραγδαία ανάπτυξη του αντικειμένου της μηχανικής μάθησης, με τα μοντέλα μηχανικής μάθησης να αποδίδουν εξαιρετικά σε πολλούς τομείς όπως η υγεία, η βιολογία, τα οικονομικά κ.α. Στον τομέα της βιοπληροφορικής, η μηχανική μάθηση εφαρμόζεται σε τομείς τομείς όπως η γονιδιωματική (*genomics*), η εξέλιξη (*evolution*) και η βιολογία συστημάτων (*systems biology*). Πλέον, ο μεγάλος αριθμός βιολογικών δεδομένων δεν αποτελεί αναστατικό παράγοντα, αλλά το ερώτημα μετατοπίζεται στην κατανόηση του τεράστιου όγκου πληροφορίας και στην αξιοποίησή του. Ακολουθίες, δισδιάστατες και τρισδιάστατες δομές, καθώς και δίκτυα αλληλεπιδράσεων αποτελούν μόνο μερικά από τα δεδομένα ενδιαφέροντος. Όσον αφορά την πρόβλεψη αλληλεπιδράσεων πρωτεΐνων, η μηχανική μάθηση φαίνεται να έχει εξαιρετικά αποτελέσματα (Support Vector Machines [3], Random Forests [4], Convolutional Neural Networks [5], Bayesian networks [6] etc.).

Στην διπλωματική εργασία κατασκευάστηκαν νευρωνικά δίκτυα με σκοπό την πρόβλεψη αλληλεπιδράσεων μεταξύ πρωτεΐνων, ωστόσο ακολουθήθηκε μια νεα προσέγγιση όσον αφορά την επεξεργασία των δεδομένων. Ειδικότερα, εφαρμόστηκαν μέθοδοι συμπλήρωσης μητρώων (*matrix completion*), οπου συμπληρώθηκαν οι ελλιπείς τιμές των δεδομένων με βάση τις υπόλοιπες παρατηρίσιμες τιμές. Οι τεχνικές αυτές αποτελούν μια μορφή συνεργατικού φίλτραρισμάτος (*collaborative filtering*), κατά την οποία επιχειρούμε να εξάγουμε πληροφορία ή μοτίβα για τη συμπλήρωση τιμών συμπεριλαμβάνοντας πληροφορίες από πολλαπλές εγγραφές [7]. Οι μέθοδοι που χρησιμοποιήθηκαν ήταν η παραγοντοποίηση μητρώου μέσω *Stochastic Gradient Descent* (*Matrix Factorization with SGD*) και η τανηστική αποδόμηση *CP* (*CP Tensor Decomposition*), για τις οποίες γίνεται λεπτομερής ανάλυση στη συνέχεια.

Επειτα από την επεξεργασία των δεδομένων, εκπαιδεύτηκαν μοντέλα νευρωνικών δικτύων, και συγκεκριμένα ένα πλήρως διασυνδεδεμένο νευρωνικό δίκτυο (*fully connected neural network*) και ένα συνελικτικό νευρωνικό δίκτυο (*convolutional neural network*), ενώ αξιολογήθηκε η απόδοση τους σε σύγκριση με αντίστοιχες δημοσιευμένες εργασίες. Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων ήταν: Accuracy, Precision, Sensitivity, Specificity, F-score και Matthews Correlation Coefficient (MCC).

---

# ΚΕΦΑΛΑΙΟ 1

---

## ΠΡΩΤΕΪΝΕΣ

Στο κεφάλαιο αυτό παρουσιάζεται η θεωρία γύρω από το αντικείμενο των πρωτεΐνων και των πρωτεϊνικών αλληλεπιδράσεων (γνωστές ως Protein - Protein Interactions). Αρχικά, δίνεται ένας σύντομος ορισμός των πρωτεΐνων, που αποτελούν τα δομικά και λειτουργικά στοιχεία των οργανισμών. Στη συνέχεια, γίνεται μια εισαγωγή στο αντικείμενο των PPIs και αναφέρονται οι κατηγορίες αλληλεπιδράσεων. Έπειτα, αναπτύσσονται οι μέθοδοι εντοπισμού αλληλεπιδράσεων πρωτεΐνων τόσο σε πειραματικό στάδιο όσο και υπολογιστικά με τη βοήθεια υπολογιστών. Το κεφάλαιο κλείνει με μια σύντομη αναφορά στις βάσεις δεδομένων που έχουν δημιουργηθεί για την καταγραφή, αποθήκευση και ανάλυση των πρωτεϊνικών αλληλεπιδράσεων.

### 1.1 Ορισμός, Λειτουργίες

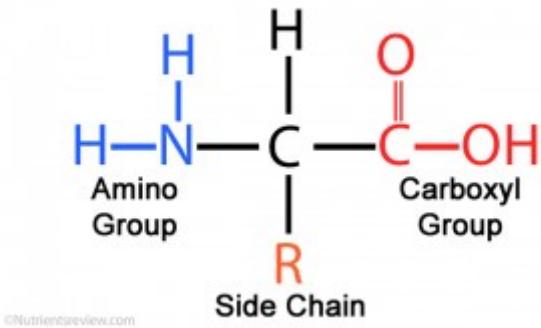
Οι πρωτεΐνες αποτελούν μεγάλα και πολύπλοκα μόρια που παίζουν καθοριστικό ρόλο στις διαδικασίες λειτουργίας και αναπαραγωγής ενός οργανισμού [8]. Στην πρωτοταγή δομή τους, αποτελούνται από μια ή περισσότερες αλυσίδες αμινοξέων (amino acid residues). Αμινοξέα ονομάζονται τα μόρια που αποτελούνται από 3 χημικές ομάδες και ένα άτομο υδρογόνου τα οποία συνδέονται μέσω ομοιοπολικών δεσμών με το ίδιο άτομο άνθρακα. Οι 3 αυτές χημικές ομάδες είναι:

1. μια καρβονική ομάδα RCOOH
2. μια αμινομάδα NH<sub>2</sub> και
3. μια πλευρική αλυσίδα μεταβλητού μεγέθους (συμβολίζεται ως R )

Υπάρχουν 20 διαφορετικοί τύποι αμινοξέων που μπορούν να συνδυαστούν για τη δημιουργία πρωτεΐνων. Η αλληλουχία των αμινοξέων κα-

θορίζει την 3-D δομή της πρωτεΐνης καθώς και τη λειτουργία της [9].

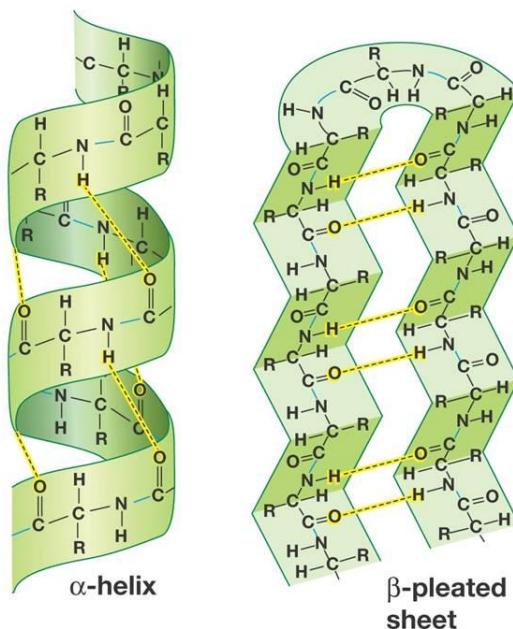
## Amino Acid Structure



Σχήμα 1.1: Δομή ενός αμινοξέος

Η δευτεροταγής δομή των πρωτεΐνων αφορά τη γενικότερη 3-D αρχιτεκτονική των τοπικών στοιχείων της. Οι συνηθέστερες δευτερογενής δομές είναι:

1. η "α-έλικα" δεξιόστροφη ( $\alpha$ -helix) και
2. η "β-πτυχωτή επιφάνεια" ( $\beta$ -sheets)



Σχήμα 1.2: Δευτερογενής δομή πρωτεΐνων

Όσον αφορά τις λειτουργίες των πρωτεΐνων, αυτές συναντώνται σε όλες τις διαστάσεις της κυτταρικής ζωής. Ορισμένες από αυτές παρουσιάζονται στον παρακάτω πίνακα:

Λειτουργίες Πρωτεΐνών	
Λειτουργία	Περιγραφή
Άμυνα	Πρωτεΐνες, όπως τα αντισώματα, δένονται πάνω σε άγνωστα σωματίδια (π.χ. ιοί, βακτήρια, προκειμένου να προστατεύσουν τον οργανισμό
Ένζυμο	Τα ένζυμα υλοποιούν σχεδόν όλες τις χημικές αντιδράσεις των κυττάρων. Ταυτόχρονα βοηθούν στη δημιουργία νέων μορίων διαβάζοντας τη γενετική πληροφορία που βρίσκεται αποθηκευμένη στο DNA.
Μεταφορά	Διάφορες πρωτεΐνες, όπως κάποια είδη ορμονών, μεταδίδουν σήματα για να συντονίσουν βιολογικές διεργασίες μεταξύ διαφορετικών κυττάρων, ιστών ή/και οργάνων.
Δομικό Στοιχείο	Παρέχουν δομή και στήριξη στα κύτταρα. Μακροσκοπικά, επιτρέπουν την κίνηση του οργανισμού.
Αποθήκευση	Δένουν και μεταφέρουν άτομα και μικρομόρια δια μέσω των κυττάρων σε όλο τον οργανισμό

Πίνακας 1.1: Λειτουργίες πρωτεΐνών

## 1.2 Αλληλεπίδραση πρωτεΐνης με πρωτεΐνη

Για να εκτελέσουν τις λειτουργίες τους μέσα στα κύτταρα, οι πρωτεΐνες σπανίως λειτουργούν ως απομονωμένες οντότητες. Αντιθέτως, πάνω από το 80% των πρωτεΐνών εκτελούν τις λειτουργίες τους σε ομάδες[10], με την μέση πρωτεΐνη να αλληλεπιδρά με 3 έως 10 άλλες πρωτεΐνες[11]. Μέσω της δημιουργίας ομάδων, οι πρωτεΐνες χειρίζονται ένα ευρύ φάσμα βιολογικών διεργασιών, μεταξύ των οποίων ο έλεγχος του μεταβολισμού και της ανάπτυξης των κυττάρων και η κυτταρική αλληλεπίδραση[1]. Επομένως, ένα σημαντικό βήμα για την κατανόηση των μοριακών σχέσεων μεταξύ των οργανισμών είναι η "χαρτογράφηση" των φυσικών αλληλεπιδράσεων πρωτεΐνης με πρωτεΐνη (PPIs).

Ως αλληλεπίδραση αναφερόμαστε στις συγκεκριμένες φυσικές επαφές μεταξύ ζεύγων πρωτεΐνών που συμβαίνουν μέσω επιλεκτικής μοριακής σύνδεσης σε ένα αυστηρά συγκεκριμένο βιολογικό πλαίσιο [12]. Η σύνδεση αυτή προκύπτει από μια πληθώρα βιοχημικών γεγονότων, μεταξύ

των οποίων λόγω ηλεκτροστατικών δυνάμεων, δεσμών υδρογόνου και λόγω υδροφοβικότητας. Οι πρωτεΐνες συνδέονται μεταξύ τους μέσω περιοχών τους που είναι τόσο γεωμετρικά όσο και φυσιοχημικά συμπληρωματικές μεταξύ τους, οπότε και προκύπτουν ενεργειακά ευνοϊκές αλληλεπιδράσεις. Στον παρακάτω πίνακα παρουσιάζονται τα είδη των αλληλεπιδράσεων που αναπτύσσουν οι πρωτεΐνες μεταξύ τους:

Τύποι Αλληλεπιδράσεων Πρωτεΐνης με Πρωτεΐνη		
Με βάση την αλληλεπίδραση	Ομο-ολιγομερής (Homooligomeric)	Ετερο-ολιγομερής (Heterooligomeric)
Με βάση την ευστάθειά τους	Υποχρεωτικές (obligate)	Μη υποχρεωτικές (non obligate)
Με βάση το προσδόκιμο ζωής	Εφήμερες (Transient)	Σταθερές (Permanent)

Πίνακας 1.2: Τύποι Αλληλεπιδράσεων Πρωτεΐνης με Πρωτεΐνη

Ειδικότερα, ομοολιγομερείς ονομάζονται οι αλληλεπιδράσεις μεταξύ ομοίων αλυσίδων ενώ ετεροολιγομέρεις αυτές μεταξύ διαφορετικών αλυσίδων. Επιπλέον διαφοροποίηση παρατηρείται όσον αφορά την ευστάθεια των δεσμών, με τις αλληλεπιδράσεις να διαχρίνονται σε υποχρεωτικές, όπου τα πρωτομερή (δομικές μονάδες ολιγομερών πρωτεΐνών) δεν μπορούν να υπάρξουν ως ευσταθείς δομές *in vivo*, και μη υποχρεωτικές. Οι αλληλεπιδράσεις διαφοροποιούνται και όσον αφορά το προσδόκιμο ζωής της σύνδεσης. Σταθερές ή μόνιμες είναι αλληλεπιδράσεις με υψηλή ευστάθεια και μόνιμο χαρακτήρα, ενώ οι εφήμερες αλληλεπιδράσεις αφορούν συνδέσεις και αποσυνδέσεις *in vivo*. Ταυτόχρονα, υπάρχει ένας περαιτέρω διαχωρισμός των εφήμερων αλληλεπιδράσεων σε αδύναμες (weak transient), όπου η ισορροπία συνεχώς χάνεται και ανακτάται, και ισχυρές (strong transient), όπου απαιτείται κάποια μοριακή διατάραξη για να χαθεί η ισορροπία[13]. Αξίζει να σημειωθεί ότι οι περισσότερες PPIs δεν εντάσσονται σε κάποια διακριτή κατηγορία. Παραδείγματος χάριν, η ευστάθεια δεν αποτελεί μια συγκεκριμένη διαφοροποίηση αλλά ένα φάσμα όπου αλληλεπιδράσεις αλλάζουν ανάλογα με τις φυσιολογικές και περιβαλλοντικές συνθήκες.

### 1.3 Μέθοδοι εντοπισμού αλληλεπιδράσεων πρωτεΐνης με πρωτεΐνη

Η δημιουργία ομάδων πρωτεΐνών (*protein complexes*) που προκύπτει από τις αλληλεπιδράσεις αυτών αυξάνει τον αριθμό των λειτουργικών στοιχείων και οδηγεί σε μεγαλύτερη ποικιλία πρωτεΐνών. Ωστόσο, ο πιθανός αριθμός συνδυασμών που μπορούν να δημιουργηθούν μέσω αλληλεπιδράσεων είναι τεράστιος, ενώ δεν είναι δυνατή η εξαγωγή γνώσης

σχετικά με τη δημιουργία αυτών καθαρά από γονιδιωματική πληροφορία (*genomic information*) [14].

Παράλληλα, παρουσιάζονται δυσκολίες τόσο στην κατανόηση όσο και στην έρευνα των PPIs. Αυτό συμβαίνει διότι το αντικείμενο των PPIs είναι ένα διεπιστημονικό και εξειδικευμένο πεδίο που εκτείνεται από τη βιολογία (κατανόηση πρωτεΐνών) μέχρι τη φυσική (κινητικές ιδιότητες PPIs) και τη χημεία (χημικές ιδιότητες αλληλεπιδράσεων), με αποτέλεσμα μια εμπειριστατωμένη μελέτη να απαιτεί τη συνεργασία και την ενσωμάτωση γνώσης και μεθόδων των παραπάνω πεδίων.

### 1.3.1 Πειραματικές μέθοδοι εντοπισμού

Οι πρώτες μέδιοδοι εντοπισμού αφορούν την πειραματική έρευνα σχετικά με τους συνδυασμούς που μπορούν πραγματικά να δημιουργήσουν ομάδες, καθώς και τη μελέτη χαρακτηριστικών, όπως η χημική συγγένεια και οι κινητικές ιδιότητες των PPIs, προκειμένου να κατανοηθούν πλήρως οι πολύπλοκες βιολογικές διεργασίες στις οποίες συμμετέχουν. Συνοπτικά, οι μέθοδοι πειραματικού εντοπισμού PPIs χωρίζονται σε 2 τύπους:

- μέθοδοι *in vitro*
- μέθοδοι *in vivo*

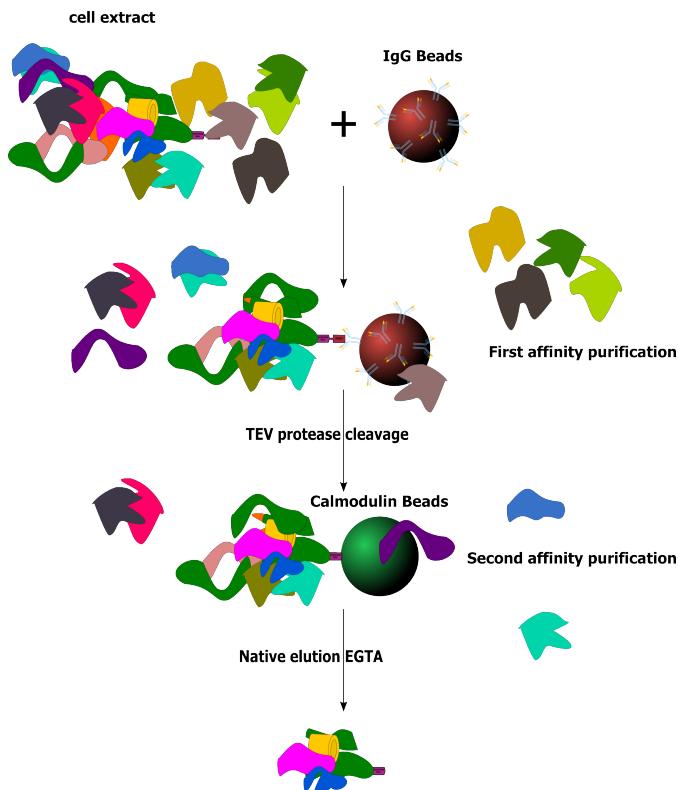
Κάθε μια από τις μεθόδους που παρουσιάζονται στη συνέχεια έχουν τα προτερήματά τους καθώς και τις αδυναμίες τους, ειδικά σε σχέση με την ευαισθησία (*sensitivity*) και την εξειδίκευση (*specificity*) τους.

#### 1.3.1.1 Μέθοδοι *in vitro*

Στις μεθόδους *in vitro*, μια συγκεκριμένη διαδικασία εκτελείται σε ένα ελεγχόμενο περιβάλλον εξωτερικά του ζωντανού οργανισμού. Οι συνηθέστερες *in vitro* μέθοδοι για τον εντοπισμό PPIs είναι οι tandem affinity purification, affinity chromatography, coimmunoprecipitation, protein arrays, protein fragment complementation, phage display, X-ray crystallography and NMR spectroscopy.

*Tandem affinity purification*, γνωστή και ως μέθοδος TAP tagging, είναι μια μέθοδος κάθαρσης βασιζόμενη στην ανοσοκαθίζηση, όπου ο σκοπός είναι η εξαγωγή από ένα κύτταρο μόνο μιας συγκεκριμένης πρωτεΐνης από όλες τις πρωτεΐνες με τις οποίες αλληλεπιδρούσε. Αναπτύχθηκε από τους Gavin et al. αρχικά για την ανάλυση των αλληλεπιδράσεων της μαγιάς [15]. Βασίζεται στον διπλό χαρακτηρισμό (tagging) της πρωτεΐνης ενδιαφέροντος σε επίπεδο χρωμοσωμάτων, ακολουθούμενο από μια διαδικασία κάθαρσης δυο βημάτων. Οι πρωτεΐνες που έπειτα από την κάθαρση παραμένουν συνδεδεμένες μπορούν στη συνέχεια να εντοπιστούν και να επεξεργαστούν (συνήθως ακολουθεί mass spectrometry (MS) επεξεργασία για τον υπολογισμό του λόγου μάζας προς φορτίο των ιόντων), με αποτέλεσμα τον εντοπισμό του βαθμού συμμετοχής της πρωτεΐνης ενδιαφέροντος

στην αλληλεπίδραση. Η μεθόδος TAP tagging επιτρέπει την αναγνώριση μις πληθώρας αλληλεπιδράσεων ενώ επιτρέπει τον έλεγχο δραστηριότητας μονομερών και πολυμερών πρωτεΐνων *in vivo*.

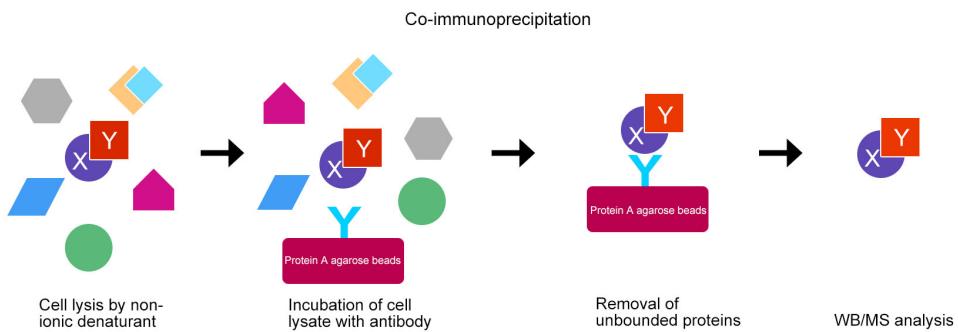


Σχήμα 1.3: Μέθοδος Tap Tagging

*Affinity Chromatography:* Μια τεχνική επιλεκτικής κάθαρσης ενός μορίου ή μιας ομάδας μορίων από πολύπλοκες δομές βασισμένη σε εξαιρετικά συγκεκριμένες αλληλεπιδράσεις μεταξύ δύο μορίων[16]. Ανήκει στις εργαστηριακές μεθόδους χρωματογραφίας και στα πλεονεκτήματά της παρουσιάζεται η γρήγορη αντιδρασιμότητα, η δυνατότητα εντοπισμού ακόμη και των ασθενέστερων αλληλεπιδράσεων ενώ ελέγχει ισόποσα κάθε πρωτεΐνη δείγματος όσον αφορά την αλληλεπίδραση της με την πρωτεΐνη ενδιαφέροντος. Ωστόσο, εμφανίζονται πολλές false positive αλληλεπιδράσεις που δεν υπάρχουν στο κυτταρικό σύστημα και αυτός ειναι ο κύριος λόγος που η εν λόγω τεχνική δεν μπορεί να χρησιμοποιηθεί εξ ολοκλήρου για τον έλεγχο και την αξιολόγηση PPIs.

*Co-immunoprecipitation:* Είναι μια ευρέως διαδεδομένη μέθοδος για την αναγνώριση PPIs, ειδικά όταν εκτελείται σε ενδογενείς πρωτεΐνες. Αποτελεί μια επέκταση της ανοσοκαθίζησης, κατά την οποία με τη χρήση εξειδικευμένων για την πρωτεΐνη ενδιαφέροντος αντισωμάτων είναι δυνατή η απομόνωση της συγκεκριμένης πρωτεΐνης, καθώς και άλλων μακρομορίων που παραμένουν μετά την αντίδραση συνδεδεμένα με αυτήν, από τον υπό-

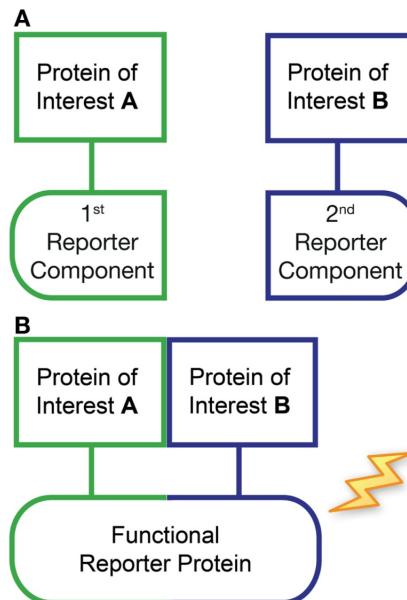
λοιπό κυτταρικό ιστό[17]. Η διαφορά με τις immunoprecipitation μεθόδους βρίσκεται στο αντικείμενο του πειράματος, που δεν ειναι πλεον η μελέτη της πρωτεΐνης ενδιαφέροντος αλλα η μελέτη των μακρομορίων με τα οποία αλληλεπιδρά.



Σχήμα 1.4: Μέθοδος Co-immunoprecipitation

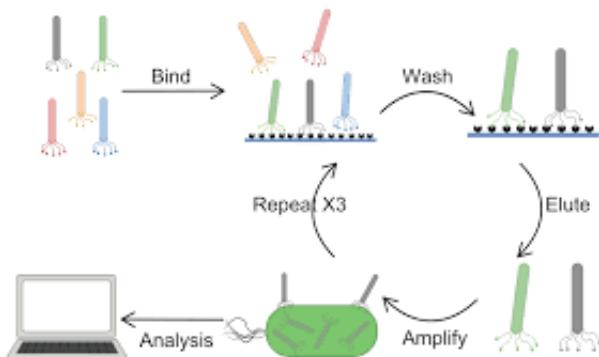
Οι μικροσυστοιχίες πρωτεΐνων (protein microarrays) αποτελούν μια από τις πιο εξελισσόμενες τεχνικές όσον αφορά τον εντοπισμό πρωτεΐνων, καθώς και την διερεύνηση των λειτουργιών και των αλληλεπιδράσεών τους. Μια μικροσυστοιχία πρωτεΐνων είναι ένα κομμάτι γυαλί, πάνω στο οποίο τοποθετούνται μόρια πρωτεΐνων σε ορισμένες αποστάσεις μεταξύ τους[18]. Τα πλεονεκτήματα τους βρίσκονται στην δυνατότητα παράλληλης και αυτοματοποιημένης ανάλυσης πολλαπλών μικροσυστοιχιών με μεγάλη ακρίβεια.

*Protein-fragment complementation assays:* Αποτελούν μια οικογένεια πειραμάτων για τον εντοπισμό PPIs. Στα PCAs, οι πρωτεΐνες αλληλεπιδρασης (οπου ονομάζονται δόλωμα-bait και λεία-prey) συνδέονται μέσω ομοιοπολικών δεσμών με μια τρίτη πρωτεΐνη, που δρα ως reporter. Οι αλληλεπιδράσεις μεταξύ των πρωτεΐνων ενδιαφέροντος φέρνουν τα αντίστοιχα μέρη της πρωτεΐνης reporter κοντά, προκειμένου να δημιουργηθεί μια λειτουργική πρωτεΐνη reporter, η οποία μπορεί να μετρηθεί. Μέσω των PCAs μπορούν να εντοπιστούν PPIs μεταξύ πρωτεΐνων ανεξαρτήτως βάρου σε ενδογενές επίπεδο[19].



Σχήμα 1.5: Απλοποιημένη αναπαράσταση PCAs

*Phage Display:* Αποτελεί μια εργαστηριακή μέθοδο για την μελέτη των PPIs που κάνει χρήση βακτηριοφάγων (αλλιώς φάγων, δηλαδή ιων που μολύνουν βακτήρια) για να συνδέσουν πρωτεΐνες μεσω της γενετική πληροφορία που τις κωδικοποιεί [20]. Η διαδικασία περιλαμβάνει την εισαγωγή γενετικής πληροφορίας μιας πρωτεΐνης ενδιαφέροντος σε έναν φάγο, με αποτέλεσμα να "προβάλλεται" η πρωτεΐνη στο εξωτερικό του φάγου ενώ η γενετική πληροφορία να παραμένει στο εσωτερικό του. Αυτοί οι φάγοι, που ονομάζονται και *display phages*, προβάλλονται πάνω σε άλλες πρωτεΐνες προκειμένου να εντοπιστούν πιθανές αλληλεπιδράσεις.



Σχήμα 1.6: Κύκλος μεθόδου Phage Display

Η κρυσταλλογραφία με ακτίνες X (*X-ray crystallography*) είναι μια μορφή μικροσκοπίας υψηλής ευχρίνειας, με σκοπό την οπτικοποίηση της ατομικής και μοριακής δομής των πρωτεΐνων, επιτρέποντας περαιτέρω κα-

τανόηση των λειτουργιών τους. Βασίζεται στη διάθλαση ακτίνων X, ύστερα από πρόσκρουσή τους στην κρυσταλλική δομή των πρωτεϊνών, προς συγκεκριμένες κατεύθυνσεις. Με την τεχνική αυτή είναι δυνατός ο σχεδιασμός καινοτόμων φαρμάκων που στοχεύουν συγκεκριμένες πρωτεΐνες [21].

Τέλος, τα τελευταία χρόνια παρουσιάζεται έντονο ερευνητικό ενδιαφέρον στην ανάλυση PPIs μέσω *nuclear magnetic resonance spectroscopy* (NMR). Η βάση της NMR spectroscopy είναι ότι οι μαγνητικά ενεργοί κυτταρικοί πυρήνες, όταν προσανατολίσονται υπό ένα ισχυρό μαγνητικό πεδίο, απορροφούν ηλεκτρομαγνητική ακτινοβολία σε χαρακτηριστικές συχνότητες που εξαρτώνται από το χημικό τους περιβάλλον. Εξετάζοντας τα διαφορετικά επίπεδα απορρόφησης ακτινοβολίας, οι ερευνητές μπορούν να αποκτήσουν εικόνα για την ηλεκτρονιακή δομή των μορίων και την λειτουργία τους [22].

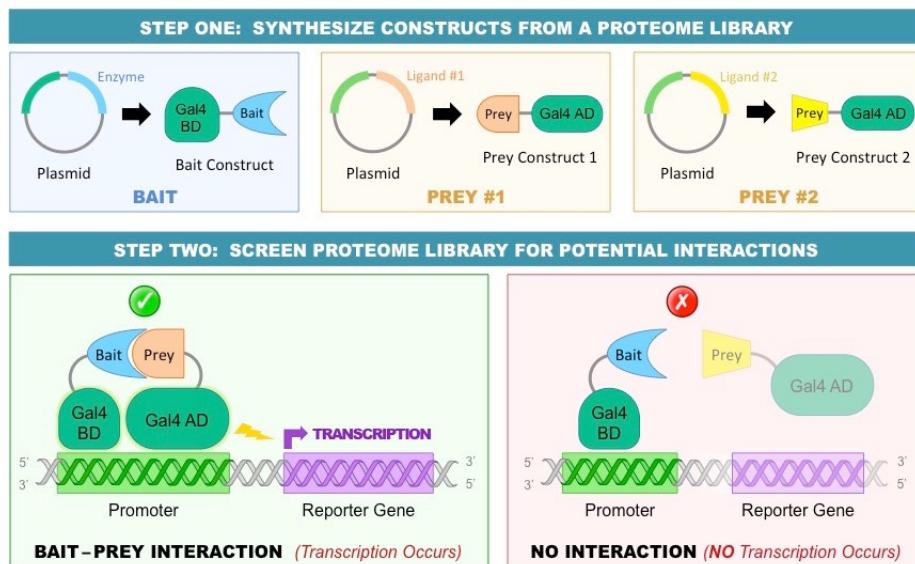
### 1.3.1.2 Μέθοδοι *in vivo*

Στις μεθόδους *in vivo*, μια συγκεκριμένη διαδικασία εκτελείται σε ελεγχόμενο περιβάλλον εντός του ζωντανού οργανισμού. Οι μέθοδοι *in vivo* για τον εντοπισμό PPIs είναι: yeast two-hybrid (Y2H, Y3H) και synthetic lethality.

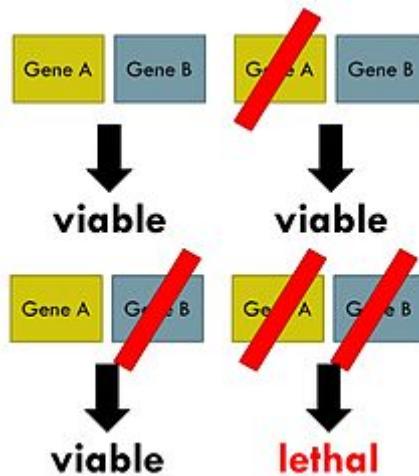
*Yeast two-hybrid* ή two-hybrid screening είναι μια τεχνική μοριακής βιολογίας που χρησιμοποιείται για τον εντοπισμό PPIs ερευνώντας για πιθανές φυσικές αλληλεπιδράσεις μεταξύ είτε δύο πρωτεΐνων που αναμένουμε να αλληλεπιδρούν είτε βρίσκοντας πρωτεΐνες (prey) που αλληλεπιδρούν με μια συγκεκριμένη πρωτεΐνη (bait) [23]. Αποτελεί μια *in vivo* τεχνική protein-fragment complementation, οικογένεια πειραμάτων που αναφέρθηκε στις *in vitro* τεχνικές. Ειδικότερα, το πείραμα εκτελείται εισάγοντας δύο πρωτεΐνες ενδιαφέροντος σε μαγιά, με κάθε πρωτεΐνη να ενώνεται με έναν transcription factor - TF που έχει χωριστεί σε δύο πανομοιότυπα κομμάτια. Πρωτεΐνες που αλληλεπιδρούν μεταξύ τους οδηγούν στη δημιουργία ενός λειτουργικού TF όταν έρθουν σε κοντινή απόσταση. Παρ' ότι ιδιαίτερα χρήσιμη τεχνική, οδηγεί στη δημιουργία πολλών false positive αλληλεπιδράσεων, ενώ το τελικό σύνολο PPIs είναι μικρό σε αριθμό, καθώς εισάγουν έναν μεγάλο παράγοντα false negative αλληλεπιδράσεων. Τέλος, επειδή χρησιμοποιεί τη μαγιά ως host, δημιουργεί προβλήματα κατά τη μελέτη πρωτεΐνων που περιέχουν συγκεκριμένες για τα θηλαστικά τροποποιήσεις [24].

**Synthetic lethality:** Πρόκειται για μια *in vivo* γενετική τεχνική που προσπαθεί να κατανοήσει τους μηχανισμούς που επιτρέπουν φαινοτυπική ισορροπία σε ένα κύταρρο/ οργανισμό παρά τις αλλαγές σε γενετικό και περιβαλλοντικό επίπεδο. Η έννοια synthetic lethality προκύπτει όταν η ταυτόχρονη διαταραχή δυο γονιδίων οδηγεί σε κυτταρικό θάνατο ή στον θάνατο του οργανισμού, ενώ η διαταραχή μονάχα ενός εκ των δυο γονιδίων δεν έχει τέτοια αποτελέσματα. Συνεπώς, η τεχνική αυτή προσπαθεί να δημιουργήσει μεταλλάξεις σε γονίδια που είναι βιώσιμα μόνα τους αλλά

προκαλούν τοξικότητα σε συνδυασμό μεταξύ τους ύπο συγκεκριμένες συνθήκες [25]. Ωστόσο, αξίζει να αναφερθεί ότι οι σχέσεις που εντοπίζονται από την τεχνική synthetic lethality δεν απαιτούν φυσικές αλληλεπιδράσεις μεταξύ των πρωτεϊνών ενδιαφέροντος, με αποτέλεσμα να αποκαλούμε αυτού του είδους τις αλληλεπιδράσεις ως λειτουργικές αλληλεπιδράσεις (*functional interactions*).



Σχήμα 1.7: Μέθοδος Yeast Two-Hybrid



Σχήμα 1.8: Παράδειγμα Synthetic Lethality

Τεχνική	Περιγραφή
Tandem affinity purification (TAP)	Βασίζεται στον διπλό χαρακτηρισμό μιας πρωτεΐνης σε επίπεδο χρωμοσωμάτων.
Affinity Chromatography	Τεχνική επιλεκτικής κάθαρσης ενός μορίου ή μιας ομάδας μορίων από πολύπλοκες δομές βασισμένη σε εξαιρετικά συγκεκριμένες αλληλεπιδράσεις μεταξύ δυο μορίων.
Co-immunoprecipitation	Επιβεβαιώνει αλληλεπιδράσεις με τη χρήση με τη χρήση εξειδικευμένων για την πρωτεΐνη ενδιαφέροντος αντισωμάτων.
Protein Microarrays	Ανάλυση βασισμένη σε μικροσυστοιχίες που επιτρέπει την ταυτόχρονη επεξεργασία χιλιάδων παραμέτρων μέσα σε ένα μόνο πείραμα.
Protein-fragment complementation assays	Χρηση για τον εντοπισμό PPIs μεταξύ πρωτεϊνών ανεξαρτήτως μεγέθους και εκφρασμένες σε ενδογενές επίπεδο.
Phage Display	Στηρίζεται στην ενσωμάτωση της δομικής και γενετικής πληροφορίας μιας πρωτεΐνης σε έναν βακτηριοφάγο.
X-ray crystallography	Οπτικοποίηση υψηλής ευχρίνειας μέσω κρυσταλλογραφίας με ακτίνες X.
NMR spectroscopy	Εντοπισμός PPIs μέσω απορρόφησης ηλεκτρομαγνητικής ακτινοβολίας των πυρήνων των πρωτεϊνών.

Πίνακας 1.3: Σύνοψη *in vitro* τεχνικών εντοπισμού PPIs

Τεχνική	Περιγραφή
Yeast two-hybrid (Y2H)	Τεχνική που εκτελείται αξιολογώντας μια πρωτεΐνη ενδιαφέροντος πάνω σε μια τυχαία βιβλιοθήκη από πιθανές πρωτεΐνες αλληλεπιδρασης.
Synthetic lethality	Γενετική τεχνική που αξιολογεί τις λειτουργικές αλληλεπιδράσεις (και όχι τις φυσικές αλληλεπιδράσεις).

Πίνακας 1.4: Σύνοψη *in vivo* τεχνικών εντοπισμού PPIs

### 1.3.2 Υπολογιστικές μέθοδοι εντοπισμού

Η μέθοδος yeast two-hybrid, καθώς και πολλές άλλες *in vitro* και *in vivo* μέθοδοι, οδήγησαν στην ανάπτυξη μιας μεγάλης γκάμας από χρήσιμα εργαλεία για τον εντοπισμό PPIs μεταξύ ορισμένων πρωτεΐνων που μπορούν να αλληλεπιδράσουν με πολλαπλούς συνδυασμούς. Παρ' όλα αυτά, τα αποτελέσματα των παραπάνω μεθόδων μπορεί να μην είναι αξιόπιστα λόγω της περιορισμένης γνώσης γύρω από τις πιθανές PPIs. Για την πλήρη κατανόηση των πιθανών αλληλεπιδράσεων, κρίθηκε σκόπιμη η δημιουργία προσεγγίσεων που προβλέπουν όλες τις πιθανές αλληλεπιδράσεις μεταξύ πρωτεΐνων [26].

Μια πληθώρα από μεθόδους *in silico*, δηλαδή εκτελεσμένες σε υπολογιστή ή μέσω υπολογιστικής προσωμοίωσης, αναπτύχθηκαν για να στηρίξουν τις αλληλεπιδράσεις που προέκυψαν μέσω των παραπάνω πειραματικών μεθόδων. Ορισμένες από τις υπολογιστικές μεθόδους που έχουν αναπτυχθεί είναι: structure-based approaches, sequence-based approaches, εγγύτητα χρωμοσωμάτων (chromosome proximity), gene fusion, *in silico* 2 hybrid, mirror tree, phylogenetic profiling και gene expression-based approaches.

*Structure-based approaches:* Η βασική ιδέα που διέπει τις μεθόδους εντοπισμού PPIs μέσω δομικών προσεγγίσεων είναι η πρόβλεψη αλληλεπιδράσεων εαν δύο πρωτεΐνες έχουν παρόμοια δομή. Για παράδειγμα, ένας πρόσφατος αλγόριθμος για structure-based prediction of PPIs, ο **Coev2Net**, ακολουθεί μια διαδικασία τριών βημάτων, που χωρίζεται σε:

- πρόβλεψη της επιφάνειας αλληλεπιδρασης,
- αξιολόγηση της συμβατότητας της επιφάνειας για αλληλεπιδραση μέσω ενός βασικού μοντέλου *coevolution* και
- αξιολόγηση του confidence score της αλληλεπιδρασης [27].

*Sequence-based approaches:* Οι προβλέψεις για PPIs βασίζονται στην ενσωμάτωση γνώσης γνωστών PPIs με τη γνώση που αφορά την ακολουθιακή ομοιογία. Με άλλα λόγια, μια αλληλεπιδραση που εντοπίζεται σε ένα είδος μπορεί να χρησιμοποιηθεί ως βάση για τον εντοπισμό της ίδιας αλληλεπιδρασης σε άλλα είδη. Οι sequence-based approaches χωρίζονται σε δύο διακριτές κατηγορίες:

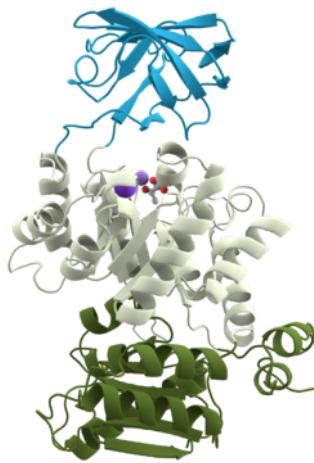
1. Ortholog-Based προσεγγίσεις και
2. Domain-Pairs-Based προσεγγίσεις.

1. *Ortholog-Based προσέγγιση:* Στηρίζεται στην μεταφορά παρατηρήσεων από μια λειτουργικά ορισμένη πρωτεΐνηκή ακολουθία στην πρωτεΐνηκή ακολουθία ενδιαφέροντος, βασισμένοι στην ομοιότητά τους. Η παρατήρηση μέσω ομοιότητας προκύπτει με την χρήση αλγορίθμων τοπικής αναζήτησης ανα ζεύγη και στηρίζεται στο βαθμό ομοιότητας της πρωτεΐνης ενδιαφέροντος με πρωτεΐνες σε μια βάση δεδομένων πρωτεϊνών [28]. Η προσέγγιση ονομάζεται ortholog-based καθώς στηρίζεται στην έννοια της ορθολογικότητας, δηλαδή της ανάπτυξης γονιδίων σε διαφορετικά είδη που προέρχονται από ένα κοινό προγονικό γονίδιο μέσω συσχέτισης.

2. *Domain-Pairs-Based προσέγγιση:* Η προσέγγιση αυτή αφορά τα domains, που αποτελούν ένα διακριτό, συμπαγές και ευσταθές μέρος της πρωτεΐνικής ακολουθίας (συγκεκριμένα της τεταρτοταγούς δομής των πρωτεϊνών) που μπορεί να εξελιχθεί, να λειτουργήσει και να υπάρξει ανεξάρτητα από την υπόλοιπη πρωτεΐνηκή αλυσίδα. Ως ατομικά δομικά και λειτουργικά μέρη μιας πρωτεΐνης, τα protein domains παίζουν σημαντικό ρόλο στην ανάπτηξη της πρόβλεψης πρωτεϊνών δομικών κλάσεων, στην προβλεψη τύπων πρωτεΐνικών μεμβρανών, στην πρόβλεψη ενζυμικής κλάσης κ.α. Θεωρούνται θεμελιώδη για την αλληλεπιδραση πρωτεϊνών καθώς είναι άμεσα συνδεδεμένα με την διαμοριακή αλληλεπιδραση, επομένως είναι αρκετά αξιόπιστη η χρήση των domains και των αλληλεπιδράσεων τους για την πρόβλεψη PPIs και αντίστροφα [29].

*Chromosome Proximity:* Είναι γνωστό ότι λειτουργικά παρόμοιες πρωτεΐνες τείνουν να οργανώνονται πολύ κοντά σε περιοχές στα γονιδιώματα των προκαρυωτικών οργανισμών. Συνεπώς, δημιουργήθηκε η ιδέα πως αν αυτή η σχέση γειτνίασης των πρωτεϊνών παρατηρηθεί σε πολλαπλά γονιδιώματα, τότε μπορεί να υπάρχει πιθανή λειτουργική σύνδεση μεταξύ των πρωτεϊνών στα αντίστοιχα γονιδιώματα. Η ιδέα αυτή επιβεβαιώθηκε πειραματικά [30] και χρησιμοποιήθηκε για τη μελέτη της λειτουργικής σχέσης των πρωτεϊνών, καθώς και την πρόβλεψη αλληλεπιδράσεων.

*Gene Fusion:* Γνωστή και ως *Rosetta stone method*, βασίζεται στην τάση ορισμένων πρωτεϊνών που περιέχουν ένα μόνο domain σε έναν οργανισμό να συγχωνεύονται για να δημιουργήσουν multidomain πρωτεΐνες σε άλλους οργανισμούς [31]. Το άνοθεν φαινόμενο πρωτεΐνικής συγχώνευσης υπονοεί την πιθανή λειτουργική αλληλεπιδραση των ξεχωριστών πρωτεϊνών. Με τη χρήση πληροφορίας σχετικά με τη διάταξη των domains σε



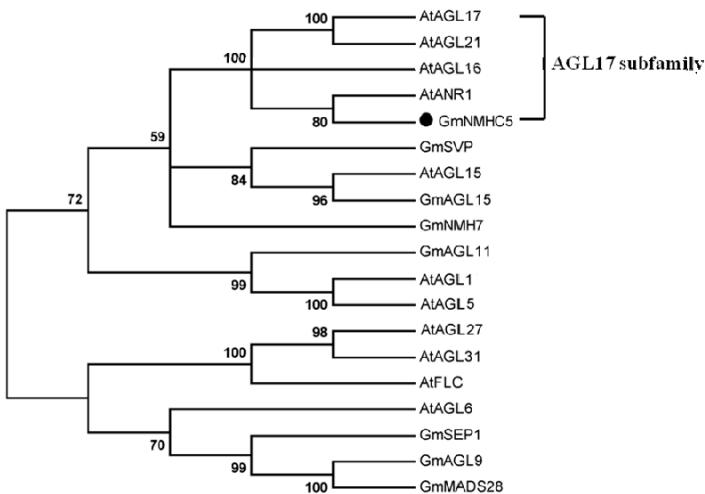
Σχήμα 1.9: Παράδειγμα πρωτεΐνη με 3 domains (*Pyruvate kinase*)

διαφορετικά γονιδιώματα, γίνεται πρόβλεψη πρωτεΐνικών αλληλεπιδράσεων [32]. Ωστόσο, μπορεί να εφαρμοστεί μόνο σε πρωτεΐνες όπου έχουν παρατηρηθεί οι συγκεκριμένες διατάξεις των domains.

*In Silico Two-Hybrid:* Θεωρία που βασίζεται στον ισχυρισμό ότι οι πρωτεΐνες που αλληλεπιδρούν θα πρέπει να υφίστανται ταυτόχρονη εξέλιξη (*co-evolution*) προκειμένου να διατηρήσουν τις λειτουργίες τους αξιόπιστες. Με άλλα λόγια, αν ορισμένα αμινοξέα της μιας πρωτεΐνης υποστούν κάποια αλλαγή, τότε τα αντίστοιχα αμινοξέα της άλλης πρωτεΐνης θα πρέπει να υποστούν τις αναγκαίες μεταλλάξεις, προκειμένου η σύνδεση να παραμείνει λειτουργική. Καθώς η συγκεκριμένη ανάλυση στηρίζεται στην πρόβλεψη της φυσικής εγκύτητας μεταξύ ζευγαριών residues ανάμεσα σε δυο πρωτεΐνες, αυτόματα τα αποτελέσματα υποδεικνύουν και πιθανές φυσικές αλληλεπιδράσεις [33].

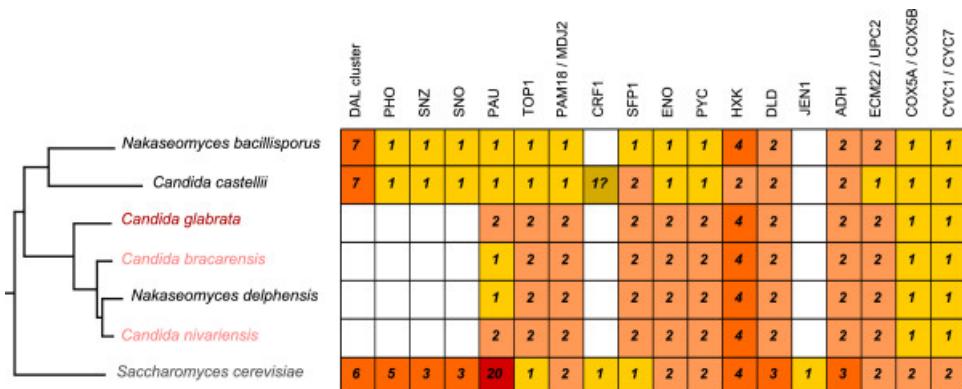
*Phylogenetic Tree - Mirror Tree Method:* Μια ακόμη υπολογιστική μέθοδος που χρησιμοποείται για τον εντοπισμό PPIs είναι το φυλογενετικό δέντρο. Το φυλογενετικό δέντρο παρέχει την πληροφορία της εξέλιξης μιας πρωτεΐνης. Σε αυτό βασίζεται η μέθοδος *mirror tree*, όπου υποστηρίζεται ότι οι πρωτεΐνες που αλληλεπιδρούν μεταξύ τους θα παρουσιάζουν παρόμοια φυλογενετικά δέντρα, λόγω του παράγοντα *co-evolution* που υπάρχει καθ' όλη την αλληλεπίδραση [34]. Ειδικότερα, η βασική αρχή πίσω από την θεωρία είναι ότι ο παράγοντας *co-evolution* αντανακλάται στο βαθμό ομοιότητας που παρουσιάζουν οι πίνακες απόστασης των φυλογενετικών δέντρων των αλληλεπιδρόντων πρωτεΐνών [35]. Υψηλοί βαθμοί συσχέτισης υποδεικνύουν την σχέση *co-evolution* των πρωτεΐνών και με τη χρήση αυτής της σχέσης εξάγονται συμπεράσματα για την πιθανότητα φυσικής αλληλεπιδρασης μεταξύ αυτών.

*Phylogenetic Profile:* Στηρίζεται στην θεωρία ότι λειτουργικά συνδεδεμένες πρωτεΐνες τείνουν να συνυπάρχουν κατά την εξέλιξη ενος οργανι-



Σχήμα 1.10: Παράδειγμα φυλογενετικού δέντρου

σμού. Με άλλα λόγια, αν δύο πρωτεΐνες έχουν μια λειτουργική σύνδεση σε ένα γονιδίωμα, θα υπάρχει ισχυρή "πίεση" ώστε να υιοθετηθούν μαζί κατά τη διαδικασία της εξέλιξης [36]. Επομένως, τα ορθόλογα τους σε άλλα γονιδιώματα είτε θα έχουν διατηρηθεί είτε θα έχουν απορριφθεί (η έννοια της ορθολογικότητας παρουσιάστηκε στη σελίδα 15 - Ortholog-Based προσέγγιση). Για τον εντοπισμό της ύπαρξης ή όχι των πρωτεΐνων αυτών χρησιμοποιείται το φυλογενετικό προφίλ, που περιγράφει την εμφάνιση μιας πρωτεΐνης σε ένα σετ γονιδιωμάτων. Συνεπώς, αν δύο πρωτεΐνες μοιράζονται το ίδιο φυλογενετικό προφίλ, τότε είναι αρκετά πιθανή η λειτουργική τους σύνδεση. Όμως, παρ' ότι η μέθοδος έχει υποσχόμενα αποτελέσματα στον εντοπισμό PPIs, βασίζεται σε ολοκληρωμένες ακολουθίες γονιδιωμάτων (γεγονός που περιορίζει την γενίκευση της μεθόδου) ενώ παράλληλα διάφορα γεγονότα κατά την διαδικασία του co-evolution, οπως η αντιγραφή γονιδιωμάτων ή η απώλεια λειτουργιών, μπορούν να διαφθείρουν το φυλογενετικό προφίλ, επηρεάζοντας την απόδοση της μεθόδου [37].



Σχήμα 1.11: Παράδειγμα φυλογενετικού προφίλ

*Gene Expression:* Η έννοια gene expression σημαίνει την ποσοτικο-ποιότηση του επιπέδου στο οποίο ένα συγκεκριμένο γονίδιο εκφράζεται μέσα σε ένα κύτταρο/ιστό/οργανισμό κάτω από διαφορετικές περιβαλλοντικές και χρονικές συνθήκες. Μέσω της εφαρμογής αλγορίθμων συσταδοποίησης, γίνεται ομαδοποίηση γονιδίων ανάλογα με τον βαθμό έκφρασής τους, επομένως το αποτέλεσμα μπορεί να βοηθήσει στη διατύπωση των λειτουργικών σχέσεων των ομαδοποιημένων γονιδίων, καθώς πρωτεΐνες από γονίδια που ανήκουν στην ίδια συστάδα έχουν μεγαλύτερη πιθανότητα αλληλεπίδρασης μεταξύ τους απότι με πρωτεΐνες από άλλες συστάδες [38].

Τεχνική	Περιγραφή
Structure-based approach	Πρόβλεψη PPIs εαν δύο πρωτεΐνες έχουν παρόμοια δομή (πρωτοταγή, δευτεροταγή ή τεταρτοταγή).
Ortholog-based approach	Sequence-based τεχνική βασισμένη στην ομόλογη φύση της πρωτεΐνης ενδιαφέροντος σε σχέση με πρωτεΐνες μιας βάσης δεδομένων με χρήση αλγορίθμων τοπικής αναζήτησης ανα ζεύγη.
Domain-pairs-based approach	Sequence-based τεχνική που προβλέπει PPIs με βάση την αλληλεπίδραση των domains των πρωτεΐνων.
Chromosome Proximity	Βασίζεται στην ιδέα ότι υπάρχει πιθανή λειτουργική σύνδεση μεταξύ πρωτεΐνων που οργανώνονται κοντά σε περιοχές.
Gene Fusion	Γνωστή και ως Rosetta Stone method, στηρίζεται στην θεωρία όπου πρωτεΐνες με μόνο ένα domain σε έναν οργανισμό συνδυάζονται για να δημιουργήσουν multidomain πρωτεΐνες σε άλλους οργανισμούς.

Πίνακας 1.5: Σύνοψη in silico τεχνικών εντοπισμού PPIs (1)

Τεχνική	Περιγραφή
In silico 2 hybrid	Τεχνική που βασίζεται στη θεωρία ότι οι πρωτεΐνες που αλληλεπιδρούν μεταξύ τους θα πρέπει να υποστούν co-evolution για να διατηρήσουν τις λειτουργίες τους αξιόπιστες.
Mirror tree	Τεχνική φυλογενετικού δέντρου που προβλέπει PPIs με βάση την εξελικτική πορεία μιας πρωτεΐνης.
Phylogenetic profile	Στηρίζεται στον εντοπισμό PPIs εαν δύο πρωτεΐνες μοιράζονται το ίδιο φυλογενετικό προφίλ.
Gene expression	Τεχνική που βασίζεται στη θεωρία ότι πρωτεΐνες που ανήκουν σε συστάδες με παρόμοιο φυλογενετικό προφίλ είναι πιθανότερο να αλληλεπιδρούν μεταξύ τους απ' ότι με πρωτεΐνες που ανήκουν σε άλλες συστάδες.

Πίνακας 1.6: Σύνοψη in silico τεχνικών εντοπισμού PPIs (2)

Συμπερασματικά, παρ' ότι οι παραπάνω μέθοδοι δεν μπορούν να προβλέψουν τις αλληλεπιδράσεις μεταξύ πρωτεΐνών με 100% ακρίβεια, οι υπολογιστικές μέθοδοι περιορίζουν το σύνολο των πιθανών αλληλεπιδράσεων στις πιο πιθανές αλληλεπιδράσεις. Το υποσύνολο αυτό στη συνέχεια χρησιμεύει ως το έναυσμα για περαιτέρω εργαστηριακά πειράματα προκειμένου να επιβεβαιωθούν και πειραματικά, μειώνοντας έτσι σημαντικά το κόστος και το χρόνο εκτέλεσης των πειραμάτων.

### 1.3.3 Βάσεις δεδομένων αλληλεπιδράσεων πρωτεΐνης με πρωτεΐνη

Παλαιότερα, το βασικό μέσο παροχής δεδομένων για αλληλεπιδράσεις πρωτεΐνών ήταν ατομικές επιστημονικές δημοσιεύσεις. Τα τελευταία χρόνια, η δημιουργία τεράστιου όγκου πειραματικών δεδομένων για PPIs οδήγησε στη δημιουργία βιολογικών βάσεων δεδομένων προκειμένου να είναι δυνατή η οργάνωση και η επεξεργασία τους. Στη συνέχεια

παρουσιάζονται μερικές από τις πιο δημοφιλείς βάσεις δεδομένων για αλληλεπιδράσεις πρωτεΐνών.

**DIP:** Γνωστή ως *Database of Interacting Proteins*, είναι μια βάση δεδομένων που περιέχει πειραματικά εξακριβωμένες αλληλεπιδράσεις μεταξύ ζεύγων πρωτεΐνών. Ταυτόχρονα, παρέχει ολοκληρωμένα εργαλεία για την αναζήτηση και την εξαγωγή πληροφοριών σχετικά με τις αλληλεπιδράσεις, ενώ παράλληλα επιτρέπει την οπτικοποίηση καθώς και την πλοήγηση σε δίκτυα αλληλεπιδράσεων [39].

**BIND:** *Biomolecular Interaction Network Database*. Διαθέτει μια δομή δεδομένων που αποθηκεύει μια πληθώρα αλληλεπιδράσεων, με λεπτομερή περιγραφή του τρόπου πειραματικής εξαγωγής των αλληλεπιδράσεων (συχνά παρατίθενται και σύνδεσμοι που παραπέμπουν στη σχετική βιβλιογραφία).

**BioGRID:** Γνωστή με το πλήρες όνομα *Biological General Repository for Interaction Datasets*, αποτελεί μια βάση δεδομένων που περιέχει πληροφορίες για γενετικές και πρωτεΐνικές αλληλεπιδράσεις σε 13 διαφορετικά είδη [40]. Τα δεδομένα εξάγωνται από τη βιβλιογραφία, καθώς και από άλλες βάσεις δεδομένων (π.χ. BIND) και αυτή τη στιγμή περιλαμβάνει πάνω από 1,740,000 αλληλεπιδράσεις από 70,000 ερευνητικές δημοσιεύσεις.

**MINT:** *Molecular INTeraction database*. Αναπτύχθηκε από το πανεπιστήμιο Tor Vergata της Ρώμης και περιέχει πειραματικά επιβεβαιωμένες αλληλεπιδράσεις πρωτεΐνών που έχουν εξαχθεί από την επιστημονική βιβλιογραφία και ελεγχθεί από ειδικούς, με την πρόσθετη πληροφορία της αξιολόγησης του βαθμού σημαντικότητας της δημοσίευσης που περιέχει την εκάστοτε αλληλεπίδραση [41]. Περιέχει περισσότερες από 130,000 αλληλεπιδράσεις σε 600+ οργανισμούς που εντοπίσθηκαν σε πάνω από 6,000 ερευνητικά άρθρα.

**HPID:** *the Human Protein Interaction Database*. Βάση δεδομένων που σχεδιάστηκε στο πανεπιστήμιο Inha της πόλης Incheon της Κορέας για να παρέχει πληροφορίες όσον αφορά αλληλεπιδράσεις μεταξύ πρωτεΐνών στους ανθρώπινους οργανισμούς, που προέρχονται από στατιστικά μοντέλα σε πειραματικά ή και δομικά δεδομένα. Παράλληλα, ενσωματώνει πληροφορίες που βρίσκονται σε άλλες βάσεις δεδομένων όπως BIND, DIP και HPRD (Human Protein Reference Database) και παράλληλα, επιτρέπει την δυνατότητα αναζήτησης πρωτεΐνών που πιθανώς αλληλεπιδρούν με πρωτεΐνες που παραθέτουν οι χρήστες μέσω πληθώρας υπολογιστικών εργαλείων [42].

**IntAct:** Βάση δεδομένων ανοιχτού κώδικα για την αποθήκευση, παρουσίαση και ανάλυση πρωτεΐνικών αλληλεπιδράσεων. Οι αλληλεπιδράσεις προέρχονται είτε από ερευνητική βιβλιογραφία είτε από καταθέσεις χρηστών. Επιτρέπει αναπαράσταση των πρωτεΐνικών αλληλεπιδράσεων τόσο

σε επίπεδο κειμένου όσο και γραφικά, με περίπου 1,000,000 αλληλεπιδράσεις από 20,000+ δημοσιεύσεις [43].

**HitPredict:** Μέσο συγκέντρωσης PPIs από γνωστές βάσεις δεδομένων (π.χ. BioGRID, IntAct, HPRD, MINT, DIP etc). Τα αποτελέσματα συνδυάζονται, σχολιάζονται και βαθμολογούνται. Το σκορ υπολογίζεται με βάση τις πειραματικές λεπτομέριες κάθε αλληλεπίδρασης καθώς και τα δομικά και λειτουργικά γνωρίσματα των πρωτεΐνων που συμμετέχουν. Επομένως, δίνει στο χρήστη μια εικόνα σχετικά με την αξιοπιστία των αλληλεπιδράσεων. Ως και τον Αύγουστο του 2019, περιλάμβανε 750,000 αλληλεπιδράσεις περίπου 88,000 διαφορετικών πρωτεΐνων σε 124 διαφορετικά είδη [44].

**APID: Agile Protein Interaction Data Analyzer.** Αποτελεί ένα διαδραστικό εργαλείο βιοπληροφορικής που δημιουργήθηκε για να επιτρέπει εξερεύνηση και ανάλυση των γνωστών μέχρι τώρα PPIs και της οπτικοποίησής τους μέσω της ενσωμάτωσης και ενοποίησης των μεγαλύτερων βάσεων δεδομένων (BIND, BioGRID, DIP, HPRD, IntAct, MINT) σε μια κοινή πλατφόρμα [45].

**PDB: Protein Data Bank.** Αποτελεί μια βάση δεδομένων για την τρισδιάστατη δομική απεικόνιση μεγάλων μακρομορίων όπως οι πρωτεΐνες. Τα δεδομένα που περιέχει εξάγονται πειραματικά μέσω X-ray crystallography ή NMR spectroscopy και καταθέτονται από επιστήμονες από όλο τον κόσμο, με αποτέλεσμα τη δημιουργία μιας από τις μεγαλύτερες βάσεις δεδομένων στον τομέα της δομικής βιολογίας [46].

Παρ' όλο που υπάρχει ένας σημαντικός αριθμός από βάσεις δεδομένων και εργαλεία για τις αλληλεπιδράσεις πρωτεΐνων, η διασταύρωση και η επικάλυψη των δεδομένων είναι σχετικά μικρή, ενώ ακόμη και τα αποτελέσματα για συγκεκριμένες αλληλεπιδράσεις από δημοσιεύσεις διαφέρουν. Συνήθως, οι διαφορές αυτές οφείλονται σε διαφορετικά κατώφλια "εμπιστευτικότητας" σχετικά με την αλληλεπίδραση, ωστόσο αυτό αποτελεί σημαντικό εμπόδιο κατά την εξαγωγή αξιόπιστων δεδομένων και την αξιολόγησή τους.

---

---

## ΚΕΦΑΛΑΙΟ 2

---

### ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

#### 2.1 Εισαγωγή

##### 2.1.1 Ορισμός

Η έννοια της μηχανικής μάθησης (*machine learning*), ως κλάδος της επιστήμης των υπολογιστών, βρίσκεται στο επίκεντρο της σύγχρονης βιομηχανίας και έρευνας. Πρόκειται για μια μέθοδο ανάλυσης δεδομένων που αυτοματοποιεί την δημιουργία μοντέλων σε μηχανές και υπολογιστικά συστήματα, βοηθώντας στη σταδιακή βελτίωση της απόδοσής τους, χωρίς την παροχή συγκεκριμένων οδηγιών προκειμένου να το πετύχουν. Ουσιαστικά, αποτελεί μια εφαρμογή τεχνητής νοημοσύνης σε συστήματα, δίνοντάς τους τη δυνατότητα της μάθησης και της βελτίωσης της απόδοσής τους χωρίς να έχουν προγραμματισθεί για αυτό.

Η μηχανική μάθηση, σαν όρος, αποδίδεται στον Αμερικανό Arthur Samuel, που τον χρησιμοποίησε στη δημοσίευσή του το 1959 που αφορούσε τη δημιουργία έξυπνων αλγορίθμων μάθησης στο παιχνίδι της ντάμας [47]. Ένας πιο αυστηρός ορισμός των αλγορίθμων που ανήκουν στο πεδίο της μηχανικής μάθησης δόθηκε το 1997 από τον Αμερικανό επιστήμονα Tom M. Mitchell και είναι ο εξής:

Ενα πρόγραμμα υπολογιστή μαθαίνει από εμπειρία  $E$  συναρτήσει κάποιων διεργασιών  $T$  και απόδοσης  $P$  αν η απόδοσή του  $P$  σε όλες τις διεργασίες  $T$  βελτιώνεται με την εμπειρία  $E$ .

### 2.1.2 Ιστορική Αναδρομή

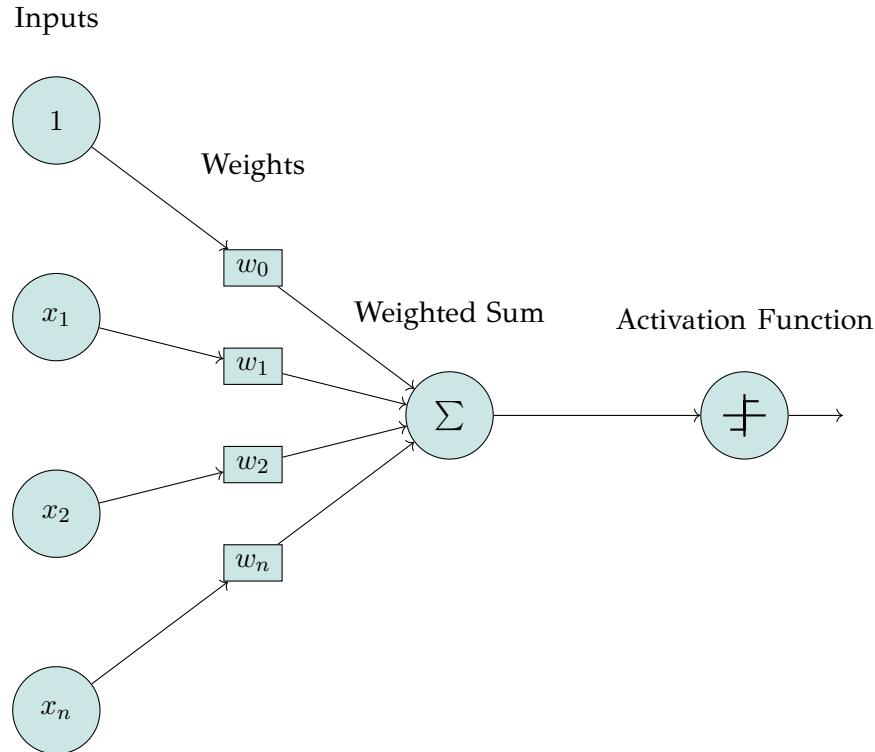
Προτού ωστόσο αποδοθεί ο όρος μηχανική μάθηση, είχαν ήδη δημιουργηθεί οι βάσεις για την ανάπτυξη του εν λόγω τομέα. Στατιστικές μέθοδοι είχαν ήδη αναπτυχθεί και βελτιστοποιηθεί. Το 1947 έχουμε την πρώτη αναφορά στον όρο νευρωνικά δίκτυα από τους *Warren McCulloch* και *Walter Pitts* σε μια δημοσίευση αναφορικά με τους νευρώνες του εγκεφάλου και το πως λειτουργούν [48]. Την ίδια περίοδο, ο *Donald Hebb* στο βιβλίο του *The Organization of Behaviour* που δημοσιεύτηκε το 1949 μοντελοποιεί τις αλγηλεπιδράσεις μεταξύ των κυττάρων του εγκεφάλου και παράλληλα αναφέρει θεωρίες σχετικά με την ενεργοποίηση και την επικοινωνία των νευρώνων [49]. Σε αυτό το βιβλίο παρουσιάστηκε και για πρώτη φορά η έννοια του *Hebbian learning*, που αποτελεί μια από τις βασικότερες θεωρίες στον χώρο της μη επιβλεπόμενης μάθησης (στην οποία θα αναφερθούμε παρακάτω).

Η δεκαετία του 1950 αποτέλεσε την απαρχή της ανάπτυξης αλγορίθμων μηχανικής μάθησης, με την εφαρμογή τους να πραγματοποιείται σε πολύ απλά προβλήματα. Ειδικότερα, το 1957 ο Αμερικανός φυσιολόγος και πρωτοπόρος στον χώρο της τεχνητής νοημοσύνης *Frank Rosenblatt* συνδύασε τη θεωρία μοντελοποίησης των κυτταρικών αλγηλεπιδράσεων του εγκεφάλου του *D. Hebb* με τη θεωρία περί μηχανικής μάθησης του *A. Samuel* και δημιούργησε τον αλγόριθμο *Perceptron*, που αποτελεί έναν αλγόριθμο επιβλεπόμενης μάθησης για δυαδική ταξινόμηση, με σκοπό τη χρήση του σε προβλήματα αναγνώρισης εικόνας (συγκεκριμένα αναγνώριση μοτίβων και σχημάτων).

Η δεκαετία του 1960 είδε την εισαγωγή μπεϋζιανών μεθόδων στην μηχανική μάθηση καθώς και την εισαγωγή πολλαπλών επιπέδων στους ήδη υπάρχοντες αλγορίθμους μηχανικής μάθησης. Παράλληλα, το 1967 αναπτύχθηκε ο αλγόριθμος *Nearest Neighbor* [50]. Ο αλγόριθμος χρησιμοποιήθηκε αρχικά σε προβλήματα χαρτογράφησης διαδρομών και αποτέλεσε την απαρχή της βασικής αναγνώρισης προτύπων.

Ακολούθησε μια περίοδος στασιμότητας για την έρευνα και την ανάπτυξη της μηχανικής μάθησης, γεγονός που οφείλεται κυρίως στον σκεπτικισμό γύρω από τους περιορισμούς που είχαν όλες οι μέθοδοι, περιορισμοί άρρηκτα συνδεδεμένοι με τους περιορισμούς επεξεργαστικής δύναμης της εποχής. Μέσα σε αυτό το κλίμα αναπτύχθηκε το 1970 από τον Φινλανδό μαθηματικό *Seppo Linnainmaa* η ιδέα του *reverse automatic differentiation*, που αντιστοιχεί στην σύγχρονη έννοια του *backpropagation* (αν και δεν ονομάστηκε έτσι πριν το 1986) [51]. Ειδικότερα, η ιδέα περιγράφει την οπισθοδρομική διάδοση του σφάλματος, με το σφάλμα να υπολογίζεται στην έξοδο και να επανατροφοδοτείται σε επίπεδα του δικτύου για λόγους μάθησης.

Η δεκαετία του 1980 είδε την εισαγωγή πολλών νέων και βασικών αλγορίθμων και θεωριών στον τομέα της μηχανικής μάθησης. Το 1982 ο



Σχήμα 2.1: Δομή του Perceptron

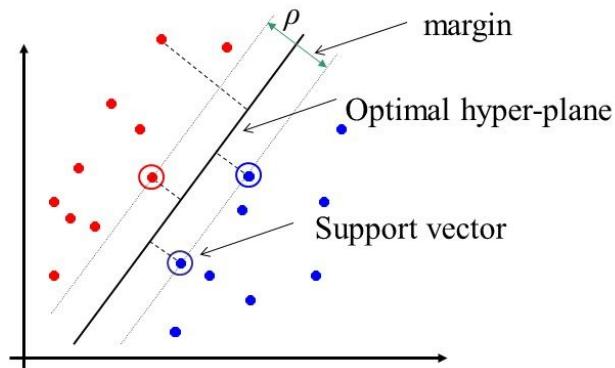
Αμερικανός επιστήμονας *John Hopfield* αναπτύσσει τη θεωρία των *associative neural networks*, που αποτελούν την πρώτη μορφή *recurrent neural networks*, μιας αρχιτεκτονικής νευρωνικών δικτύων με μεγάλη διάδοση τις επόμενες δεκαετίες.

Ένας άλλος λόγος που η ανάπτυξη της μηχανικής μάθησης αντιμετώπισε μια στασιμότητα τις τελευταίες δεκαετίες ήταν η επικέντρωση της έρευνας σε γνωσιοκεντρικές (knowledge-based) προσεγγίσεις. Αυτό άλλαξε τη δεκαετία του 1990 όπου η έρευνα μετατοπίστηκε σε προσεγγίσεις με βάση τα δεδομένα (data-driven). Το 1995 δημοσιεύεται η εργασία της *Tin Kam Ho* όπου παρουσιάζονται οι αλγόριθμοι κατηγοριοποίησης *random forest* [52]. Το 1995 γίνεται η σύγχρονη διατύπωση άλλης μιας σημαντικής θεωρίας, αυτής των *support vector machines*, που αποτελούν μοντέλα επιβλεπόμενης μάθησης με χρήση σε προβλήματα κατηγοριοποίησης ή *regression* [53]. Παράλληλα, μια σημαντική τεχνική που χρησιμοποείται κατα κόρον μέχρι και σήμερα για την αναγνώριση ομιλίας αναπτύχθηκε το 1997 από τους *Jürgen Schmidhuber* και *Sepp Hochreiter*. Πρόκειται για τα μοντέλα νευρωνικών δικτών *Long Short-Term Memory*, που αποτελούν μια ειδική περίπτωση των προαναφερθέντων *recurrent neural networks*, με την ικανότητα να μαθαίνουν διεργασίες που απαιτούν μνήμη χιλιιάδων διαχριτών βημάτων [54].

Κάπως έτσι, φτάνουμε στον 21ο αιώνα, και ειδικότερα μέτα το 2010, όπου η ραγδαία αύξηση της υπολογιστικής ισχύος επιτρέπει τη δημιουργία *deep neural network* αρχιτεκτονικών. Πλέον, η μηχανική μάθηση είναι διαδεδομένη σε πολλαπλούς τομείς και βρίσκεται χρήση σε πληθώρα εφαρμογών και υπηρεσιών λογισμικού, ενώ η έρευνα πάνω στο αντικείμενο έχει αυξηθεί κατα κόρον, με τις εταιρίες να επενδύουν σημαντικά ποσά προκειμένου να μείνουν μπροστά από τον ανταγωνισμό.

## SVM: separable classes

Support vectors uniquely characterize optimal hyper-plane



Σχήμα 2.2: Support Vector Machines

## 2.2 Είδη μηχανικής μάθησης

Η μηχανική μάθηση χωρίζεται σε πολλά και ξεχωριστά είδη, με τα συστήματα που την χρησιμοποιούν ως εργαλείο να χωρίζονται ανάλογα με:

- την προσέγγισή τους ως προς τη "μάθηση"
- τον τύπο δεδομένων που δέχονται στην είσοδο
- τον τύπο δεδομένων που δίνουν ως έξοδο
- τον τύπο προβλημάτων/διεργασιών που καλούνται να επιλύσουν κ.α.

Διακρίνονται τρεις κύριες κατηγορίες μηχανικής μάθησης ως προς τον παράγοντα "μάθηση": επιβλεπόμενη μάθηση (supervised learning), μη επιβλεπόμενη μάθηση (unsupervised learning) και ενισχυτική μάθηση (reinforcement learning).

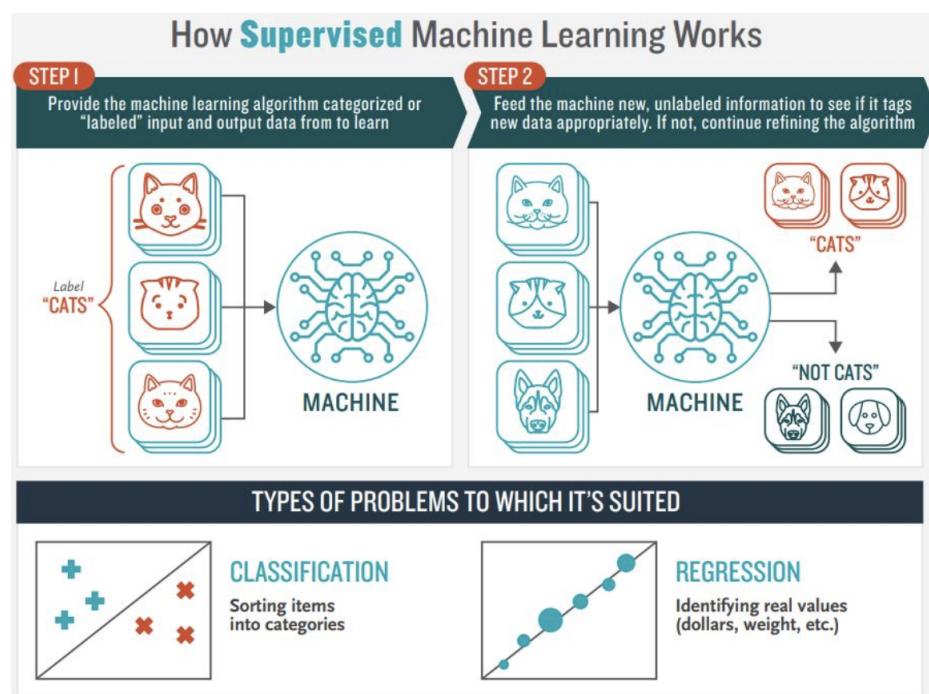
### 2.2.1 Επιβλεπόμενη Μάθηση

Η επιβλεπόμενη μάθηση περιγράφει μια κατηγορία προβλημάτων, που περιλαμβάνουν τη δημιουργία μοντέλων με σκοπό την χαρτογράφηση δεδομένων εισόδου σε μια επιθυμητή έξοδο. Ένας πιο αυστηρός ορισμός δίνεται στο βιβλίο *Pattern Recognition and Machine Learning* [55] και είναι ο ακόλουθος:

Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems.

Τα μοντέλα εκπαιδεύονται σε δεδομένα "εκπαίδευσης", που αποτελούνται από τα αντικείμενα εισόδου και τις επιθυμητές τιμές εξόδου τους, και χρησιμοποιούνται για να κάνουν προβλέψεις σε "τεστ" δεδομένα, όπου παρέχεται μόνο το κομμάτι της εισόδου και το αποτέλεσμα της πρόβλεψης συγκρίνεται με τις πραγματικές τιμές εξόδου προκειμένου να υπολογιστεί η ακρίβεια του μοντέλου.

Υπάρχουν δύο βασικοί τύποι προβλημάτων επιβλεπόμενης μάθησης και διαφοροποιούνται με βάση το είδος της εξόδου που καλούνται να προβλέψουν. Συγκεκριμένα, ονομάζονται *classification* τα προβλήματα που καλούνται να προβλέψουν μια κατηγορική έξοδο, ενώ ονομάζονται *regression* όταν καλούνται να προβλέψουν μια αριθμητική τιμή.



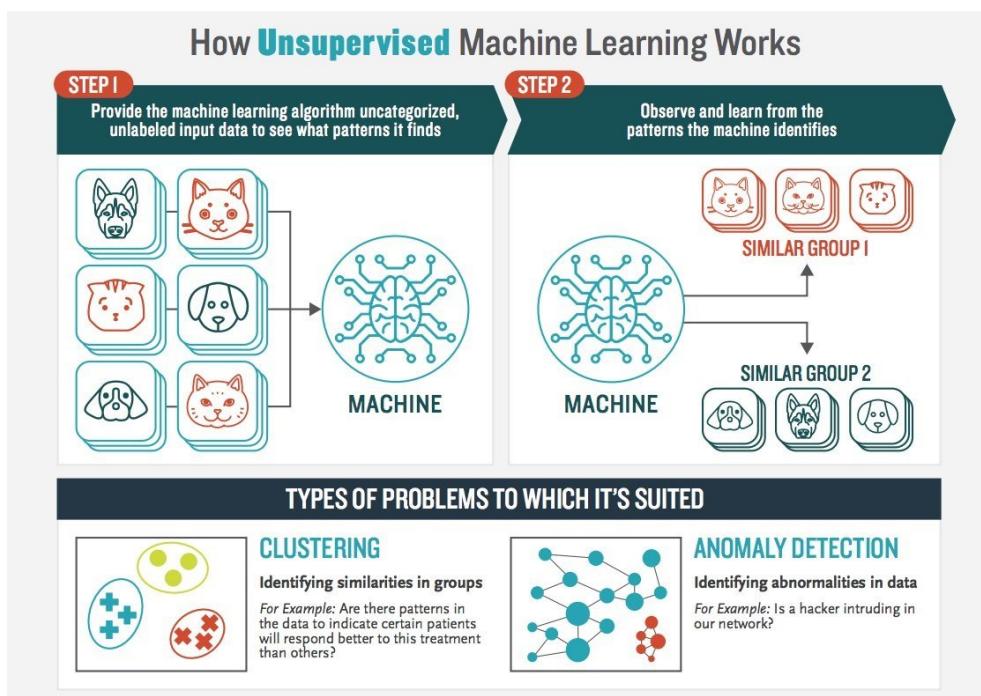
Σχήμα 2.3: Επιβλεπόμενη Μάθηση

### 2.2.2 Μη Επιβλεπόμενη Μάθηση

Η μη επιβλεπόμενη μάθηση αναφέρεται σε προβλήματα που περιλαμβάνουν την ανάπτυξη μοντέλων για την περιγραφή ή την εξαγωγή σχέσεων από δεδομένα. Αποτελεί το αντίθετο της επιβλεπόμενης μάθησης, με την έννοια ότι δεν περιλαμβάνονται οι επιθυμητές τιμές στην εκπαίδευση του μοντέλου. Αντίθετα, τροφοδοτούμε το μοντέλο με μια πληθώρα δεδομένων και παρέχονται τα εργαλεία προκειμένου αυτό να κατανοήσει τις ιδιότητες των δεδομένων. Έστερα, το μοντέλο μαθαίνει να ομαδοποιεί/οργανώνει τα δεδομένα με τέτοιον τρόπο ώστε κάποιος άνθρωπος να μπορεί να εξάγει συμπεράσματα για αυτά με βάση τη νέα οργάνωσή τους.

Αποτελεί πολύ σημαντικό αντικείμενο στον τομέα της μηχανικής μάθησης, καθώς τα περισσότερα δεδομένα στον κόσμο είναι ανοργάνωτα και ασυσχέτιστα μεταξύ τους, επομένως η δυνατότητα υπολογιστικών συστημάτων να εξάγουν συμπεράσματα για αυτά αποτελεί πλεονέκτημα σε πολλές βιομηχανίες.

Παρ' ότι υπάρχουν πολλές κατηγορίες μη επιβλεπόμενης μάθησης, οι κυριότερες είναι το *clustering*, που περιλαμβάνει την εύρεση ομαδοποιήσεων σε ανοργάνωτα δεδομένα και το *density estimation*, δηλαδή την εξαγωγή συμπερασμάτων σχετικά με την κατανομή των δεδομένων.



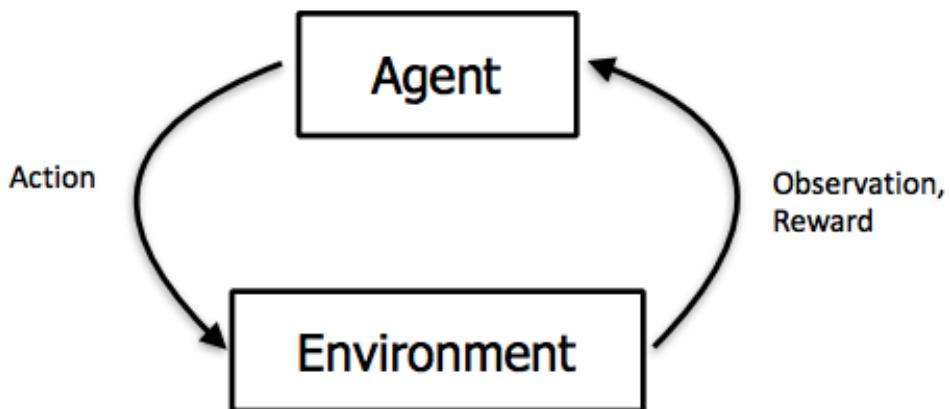
Σχήμα 2.4: Μη Επιβλεπόμενη Μάθηση

### 2.2.3 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση περιγράφει μια κατηγορία προβλημάτων όπου ένας πράκτορας ενεργεί σε ένα περιβάλλον και πρέπει να μάθει να λειτουργεί προκειμένου να μεγιστοποιήσει κάποιο κέρδος. Πιο συγκεκριμένα, η ενισχυτική μάθηση έχει τον ακόλουθο ορισμό:

Reinforcement learning is learning what to do - how to map situations to actions - so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them [56].

Η διαδικασία μάθησης στη συγκεκριμένη μέθοδο παρομοιάζεται με τον τρόπο που οι άνθρωποι μαθαίνουν εμπειρικά από τα λάθη τους. Ο πράκτορας τοποθετείται σε ένα περιβάλλον και αρχικά πραγματοποιεί πολλά λάθη. Ωστόσο, εφόσον υπάρχει κάποιο σήμα προς τον πράκτορα που συσχετίζει τις σωστές συμπεριφορές με κάποια ανταμοιβή και τις λάθος με καποια ποινή, ενισχύεται η προτίμηση στις καλές συμπεριφορές. Ως αποτέλεσμα, με το πέρασμα του χρόνου ο αλγόριθμος μάθησης πραγματοποιεί λιγότερα λάθη απ' ότι προηγουμένως. Συνήθως υλοποιείται μέσω μοντέλων αποφάσεων Markov, που αποτελούν στοχαστικές διαδικασίες διακριτού χρόνου για τη μοντέλοποίηση λήψης αποφάσεων.



Σχήμα 2.5: Ενισχυτική Μάθηση

## 2.3 Νευρωνικά Δίκτυα

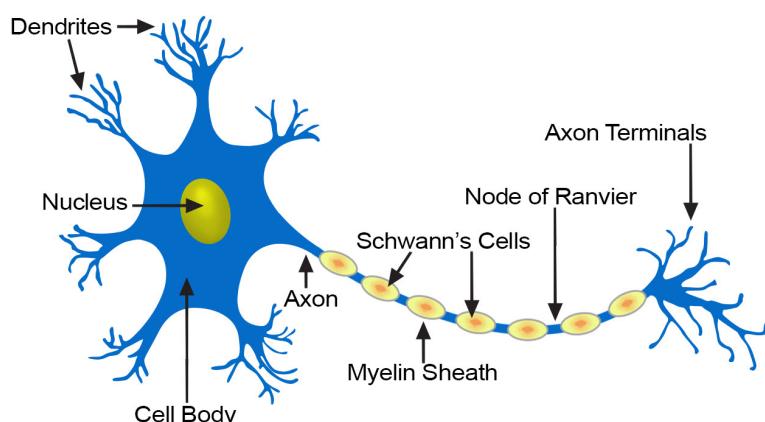
Τα νευρωνικά δίκτυα (*Artificial Neural Networks*) βρίσκονται στο επίκεντρο της μηχανικής μάθησης και αποτελούν την πιο διαδεδομένη μορφή αλγορίθμων που υλοποιούν μοντέλα μάθησης. Πρόκειται για υπολογιστικά συστήματα που είναι σχεδιασμένα ώστε να "μιμούνται" τη λειτουργία των νευρικών δικτύων που υπάρχουν στους εγκεφάλους των ζωϊκών οργανισμών. Μαθαίνουν να υλοποιούν διεργασίες εξάγωντας χαρακτηριστικά από τα παραδείγματα που τους τροφοδοτούνται, δίχως να γνωρίζουν λεπτομέριες για τις διεργασίες και χωρίς να τους ορίζονται σαφείς κανόνες για τον τρόπο επίλυσης [57].

Η απαρχή της θεωρίας των νευρωνικών δικτύων ήρθε, όπως αναφέρθηκε προηγουμένως, το 1947 από τους *Warren McCulloch* και *Walter Pitts*, που δημιούργησαν το πρώτο υπολογιστικό μοντέλο με βάση τα νευρωνικά δίκτυα. Το 1949 ο *Donald Hebb* δημιούργει ένα μοντέλο μάθησης (*Hebbian Learning*) που βασίζεται στην ικανότητα των νευρώνων να τροποποιούν τη δομή και τη λειτουργία τους για να προσαρμοστούν σε περιβαλλοντικές αλλαγές (*neural plasticity*), με τους *W. Clark* και *B. Farley* να υλοποιούν το παραπάνω μοντέλο μάθησης του Hebb σε υπολογιστικά συστήματα, γνωστά ως *calculators* [58].

### 2.3.1 Νευρώνες

Τα νευρωνικά δίκτυα καθιστώνται από τεχνητούς νευρώνες, που αποτελούν το ανάλογο των βιολογικών νευρώνων. Ένας νευρώνας αποτελεί ένα από τα βασικότερα στοιχεία του νευρικού συστήματος και πρόκειται για ένα κύτταρο που λαμβάνει, επεξεργάζεται και μεταδίδει πληροφορία μέσω ηλεκτρικών και χημικών σημάτων.

**Structure of a Typical Neuron**



Σχήμα 2.6: Βιολογική Δομή Νευρώνα

Σε δομικό επίπεδο, διακρίνονται τρια βασικά μέρη που αποτελούν έναν νευρώνα:

- το κυτταρικό σώμα,
- τους δενδρίτες και
- τον άξονα.

Το κυτταρικό σώμα αποτελεί το κεντρικό μέρος του νευρώνα, περιέχει τον πυρήνα του κυττάρου καθώς και άλλα οργανίδια που επιτρέπουν πρωτεΐνική σύνθεση (π.χ. endoplasmic reticulum) και παραγωγή ενέργειας (μιτοχόνδρια). Μέσω των δενδριτών και του άξονα, ένας νευρώνας επεκτείνει τις διεργασίες του σε άλλα κύτταρα. Οι δενδρίτες συνήθως διακλαδώνονται έντονα με σταδιακή μείωση στο πάχος τους έπειτα από κάθε διακλαδωση και περιέχουν το κύριο μέρος της πληροφορίας εισόδου του νευρώνα. Ο άξονας είναι πιο λεπτός και μπορεί να επεκταθεί εως και χιλιάδες φορές όσο η διάμετρος του κυτταρικού σώματος σε μήκος, ενώ μεταφέρει νευρικά σήματα μακριά από το σώμα και προς άλλα κύτταρα.

Οσον αφορά τη λειτουργία τους, η πληροφορία εισέρχεται ως ηλεκτρική διέγερση μέσω των δενδριτών στο σώμα του νευρώνα, επεξεργάζεται και στη συνέχεια μεταδίδεται μέσω του άξονα (που λειτουργεί πρακτικά ως καλώδιο) και καταλήγει στα άκρα του, όπου και μεταδίδει την πληροφορία σε άλλα κύτταρα μέσω δομών που ονομάζονται συνάψεις.

### 2.3.2 Λειτουργία και Οργάνωση

Στα ANNs, οι νευρώνες αποτελούν ένα είδος συνάρτησης που εκτελεί ένα weighted sum πάνω στις εισόδους τους και στη συνέχεια εφαρμόζουν έναν μη γραμμικό μετασχηματισμό στο αποτέλεσμα, το οποίο και μεταδίδουν στους νευρώνες με τους οποίους είναι συνδεδεμένοι.

Αν θεωρήσουμε ως  $y$  την έξοδο ενός νευρώνα, η παραπάνω περιγραφή μοντελοποιείται μαθηματικά ως:

$$y = f(\sum(x_1 * w_1 + x_2 * w_2 + \dots + x_n * w_n)) \quad (2.1)$$

where:

$$\begin{aligned} x_n &= \text{είσοδοι του νευρώνα} \\ w_n &= \text{βάρη του νευρώνα} \\ f(x) &= \text{συνάρτηση ενεργοποίησης του νευρώνα} \end{aligned}$$

Η συνάρτηση ενεργοποίησης είναι η συνάρτηση που περιγράφει την έξοδο του νευρώνα. Αποτελεί τον παράγοντα που εισάγει τη μη γραμμικότητα στη λειτουργία των νευρωνικών δικτύων και επομένως επιτρέπει την περιγραφή πολύπλοκων συστημάτων, διατηρώντας ωστόσο την επεξεργασία των εισόδων με απλό τρόπο. Οι συνηθέστερες συναρτήσεις ενεργοποίησης είναι οι βηματικές συναρτήσεις (step), οι σιγμοειδής (sigmoid) και οι συναρτήσεις γραμμικής ανόρθωσης (rectified linear units - ReLU ).

- **Βηματικές συναρτήσεις:** Η έξοδος είναι δυαδική και παίρνει τιμή ανάλογα με το αν το weighted sum είναι πάνω ή κάτω από ένα ορισμένο κατώφλι .

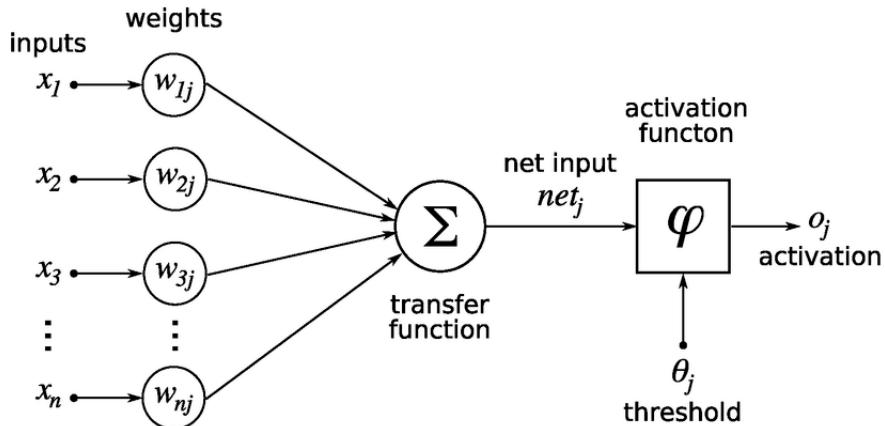
$$y = \begin{cases} 0 & \text{if } \sum(x_n * w_n) < \theta \\ 1 & \text{if } \sum(x_n * w_n) > \theta \end{cases} \quad (2.2)$$

- **Σιγμοειδείς συναρτήσεις:** Πρόκειται για φραγμένες, διαφορίσιμες και πραγματικές συναρτήσεις με χαρακτηριστική S-μορφή που χρησιμοποιούνται κυρίως για την υπολογιστική τους απλότητα. Ένα απλό παράδειγμα είναι η logistic function που έχει την ακόλουθη μορφή:

$$y = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1} \quad (2.3)$$

- **Συναρτήσεις γραμμικής ανόρθωσης:** Αποτελεί πλέον την πιο διαδομένη συνάρτηση ενεργοποίησης για πολλούς τύπους νευρωνικών δικτύων καθώς αντιμετωπίζει πολλά προβλήματα παραγώγισης που εμφανίζουν οι σιγμοειδής (π.χ. vanishing gradients κ.α.). Ο μαθηματικός τους ορισμός είναι ο εξής:

$$y = \max(0, x) \quad (2.4)$$



Σχήμα 2.7: Μαθηματική δομή τεχνητού νευρώνα

Οι νευρώνες συνήθως οργανώνονται σε πολλαπλά επίπεδα, με τους νευρώνες ενός επιπέδου να συνδέονται μόνο με τους νευρώνες του αμέσως προηγούμενου και του αμέσως επόμενου επιπέδου. Το πρώτο επίπεδο των νευρωνικών δικτύων, δηλαδή το επίπεδο των νευρώνων που δέχεται την είσοδο, ονομάζεται επίπεδο εισόδου, ενώ το επίπεδο που παρέχει τις εξόδους του νευρωνικού δικτύου ονομάζεται επίπεδο εξόδου. Ενδιάμεσα μπορεί να μεσολαβούν κανένα, ένα ή και περισσότερα επίπεδα αναλόγως

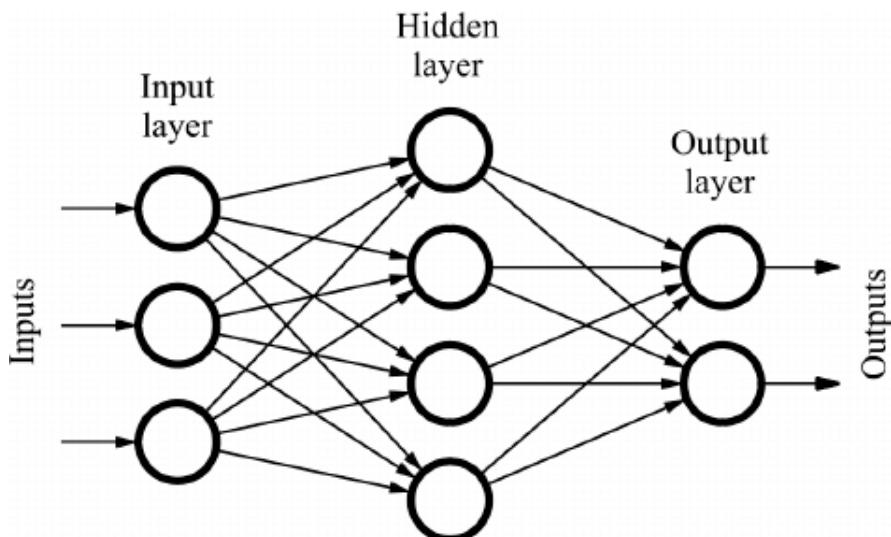
με την πολυπλοκότητα του δικτύου. Παράλληλα, ο τρόπος σύνδεσης των νευρώνων μεταξύ διαφορετικών επιπέδων μπορεί να διαφέρει. Μπορεί να είναι πλήρως διασυνδεδεμένα, δηλαδή κάθε νευρώνας ενός επιπέδου να συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου είτε να υπάρχουν για π.χ. επίπεδα *pooling*, οπου μια ομάδα νευρώνων από ένα επίπεδο να συνδέεται σε έναν νευρώνα του επόμενου επιπέδου, μειώνοντας τον αριθμό των νευρώνων σε αυτο το επίπεδο.

Ένας άλλος παράγοντας για την ορθή λειτουργία των νευρωνικών δικτύων είναι οι λεγόμενες υπερπαράμετροι. Οι υπερπαράμετροι αποτελούν παραμέτρους που ορίζονται πριν την διαδικασία μάθησης των μοντέλων. Αν και δεν επηρεάζουν την απόδοση του μοντέλου, έχουν επίδραση πάνω στην ταχύτητα σύγκλισης και στην ποιότητα της μάθησης. Μερικά παραδείγματα υπερπαραμέτρων είναι ο ρυθμός μάθησης, ο αριθμός κρυφών επιπέδων και το *batch size*.

### 2.3.3 Αρχιτεκτονικές νευρωνικών δικτύων

Η εξέλιξη της τεχνολογίας έχει επιτρέψει την άνοδο πολλών αρχιτεκτονικών νευρωνικών δικτύων, οι οποίες με τη σειρά τους έχουν πετύχει state-of-the-art αποτελέσματα σε πολλαπλούς τομείς.

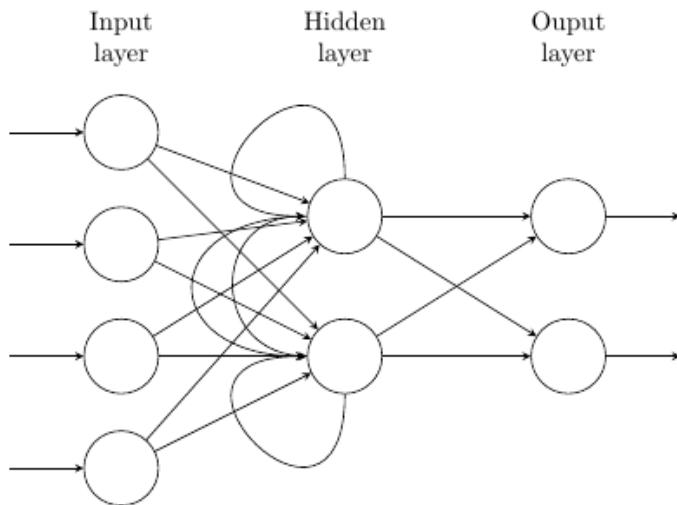
Η πρώτη και απλούστερη δομή νευρωνικών δικτύων είναι η *feedforward* αρχιτεκτονική. Στη συγκεκριμένη δομή, η πληροφορία κινείται μονάχα προς μια κατεύθυνση, από το επίπεδο εισόδου προς το επίπεδο εξόδου μέσω των κρυφών επιπέδων (αν υπάρχουν).



Σχήμα 2.8: Feedforward αρχιτεκτονική

Αυτή η μονοδιάστατη ροή πληροφορίας άλλαξε με την είσοδο στον χώρο των *recurrent neural networks*. Αποτελούν μια δομή νευρωνικών δικτύων

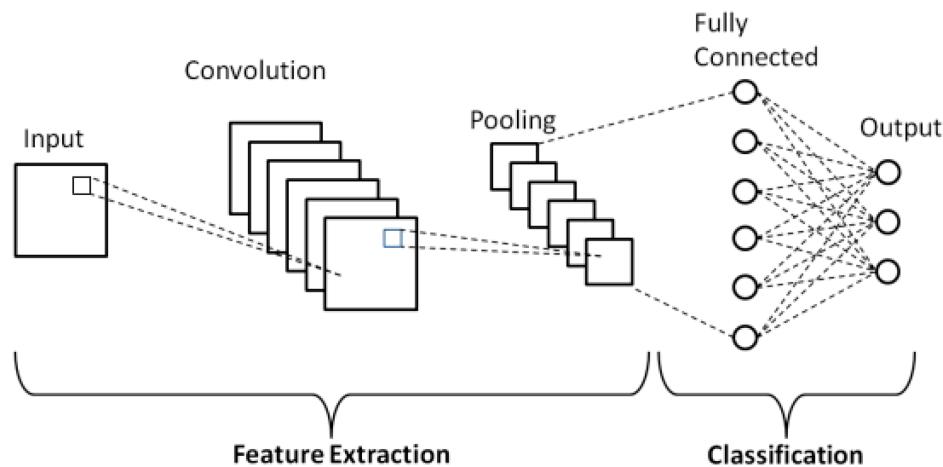
όπου εκτός από την εμπρόσθια ροή της πληροφορίας, δίνεται η δυνατότητα στους νευρώνες να χρησιμοποιήσουν την εσωτερική τους κατάσταση για να αξιολογήσουν δεδομένα εισόδου, με αποτέλεσμα την ροή πληροφορίας και προς τα πίσω. Τα recurrent neural networks αποτελούν πλέον την πιο διαδεδομένη μορφή νευρωνικών δικτύων για χρήση σε προβλήματα επεξεργασίας φυσικής γλώσσας, επεξεργασίας ομιλίας καθώς και αναγνώρισης γραφικού χαρακτήρα.



Σχήμα 2.9: Recurrent αρχιτεκτονική

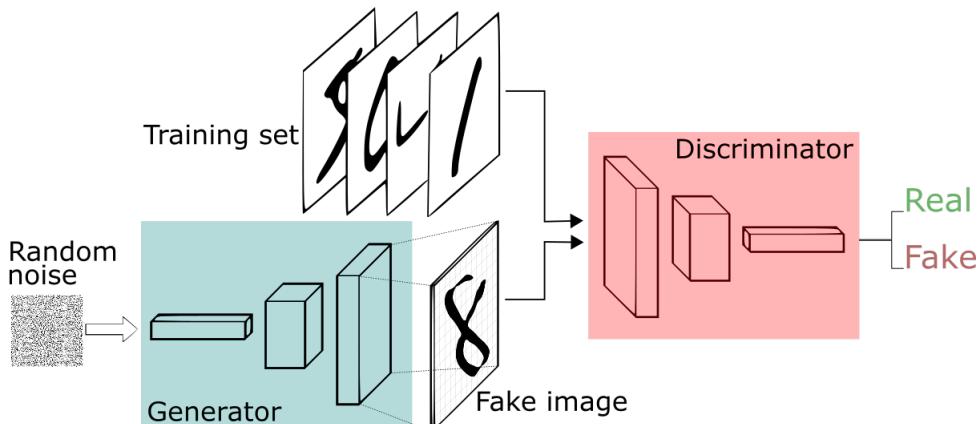
Μια ακόμη διαδεδομένη μορφή νευρωνικών δικτύων, που χρησιμοποιείται κατά κόρον σε προβλήματα επεξεργασίας εικόνων, είναι η αρχιτεκτονική συνελικτικών νευρωνικών δικτύων (*convolutional neural networks*). Ως συνελικτικά νευρωνικά δίκτυα θεωρούνται τα νευρωνικά δίκτυα που συμπεριλαμβάνουν επίπεδα όπου πραγματοποιείται συνέλιξη. Πρακτικά, από μαθηματική άποψη σε ένα συνελικτικό επίπεδο πραγματοποιείται ένα εσωτερικό γινόμενο μιας περιοχής των εξόδων των νευρώνων του προηγούμενου επιπέδου με κάποια φίλτρα (που ορίζονται πριν την εκπαίδευση των μοντέλων), ακολουθεί ένα στάδιο *pooling* οπου μειώνεται η διαστασιμότητα του μοντέλου και τα αποτελέσματα τροφοδοτούνται συνήθως σε πλήρως διασυνδεδεμένα επίπεδα.

Τέλος, αξίζει να αναφερθεί μια ακόμη αρχιτεκτονική που αναπτύχθηκε το 2014 από τον Ian Goodfellow και την ομάδα του [59]. Πρόκειται για τα *General Adversarial Networks* (GANs), τα οποία δοσμένου ενός σετ δεδομένων εκπαίδευσης μαθαίνουν να δημιουργούν νέα δεδομένα με τα ίδια στατιστικά χαρακτηριστικά με τα δεδομένα εκπαίδευσης. Όσον αφορά τη λειτουργία τους, δύο νευρωνικά δίκτυα "ανταγωνίζονται" το ένα το άλλο σε ένα "παιχνίδι", με το ένα δίκτυο (*generative*) να παράγει



Σχήμα 2.10: Convolutional αρχιτεκτονική

υποφήφια νέα δεδομένα και το άλλο (*discriminative*) να τα αξιολογεί. Τα GANs έχουν εξαιρετικά αποτελέσματα στην παραγωγή ρεαλιστικών φωτογραφιών, στο βαθμό που ένας άνθρωπος δεν μπορεί να ξεχωρίσει αν η φωτογραφία είναι πραγματική ή κατασκευασμένη.



Σχήμα 2.11: General Adversarial Network αρχιτεκτονική

---

---

## ΚΕΦΑΛΑΙΟ 3

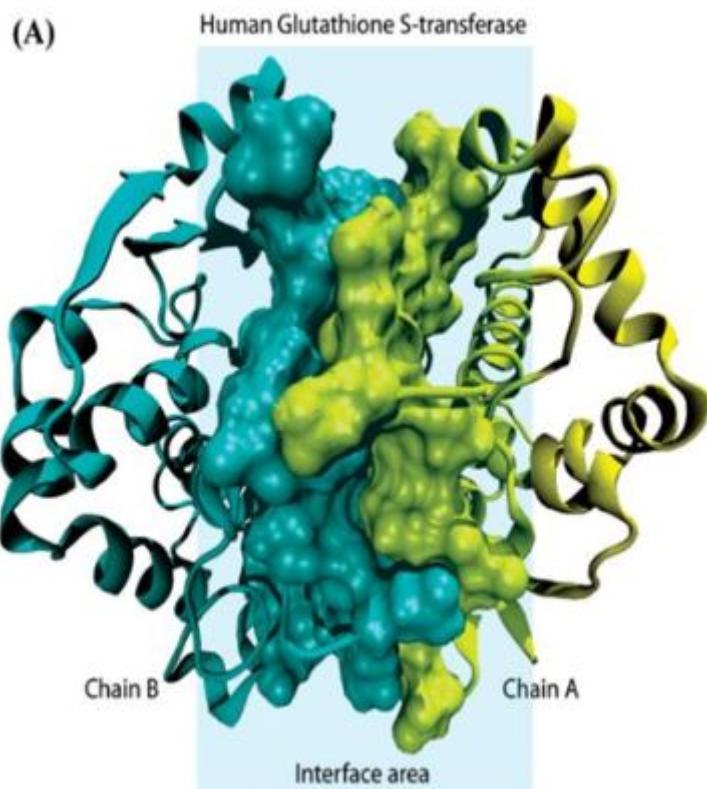
---

### ΥΛΟΠΟΙΗΣΗ

Στο κεφάλαιο 1 έγινε μια σύντομη εισαγωγή στις βιολογικές έννοιες των πρωτεΐνων και των πρωτεΐνικών αλληλεπιδράσεων, που αποτελούν το αντικείμενο της συγκεκριμένης εργασίας. Στη συνέχεια, στο κεφάλαιο 2 παρουσιάστηκαν οι έννοιες της μηχανικής μάθησης και των νευρωνικών δικτύων, τα οποία αποτέλεσαν τα εργαλεία που χρησιμοποιήθηκαν για την εκπόνηση της παρούσας εργασίας. Στο συγκεκριμένο κεφάλαιο παρουσιάζεται η προσέγγιση που επιλέχθηκε για την επίλυση του προβλήματος του εντοπισμού πρωτεΐνικών αλληλεπιδράσεων με τη χρήση μοντέλων μηχανικής μάθησης και συγκεκριμένα νευρωνικών δικτύων. Αρχικά, γίνεται μια ανάλυση του συγκεκριμένου προβλήματος το οποίο πρέπει να επιλυθεί και αναφέρεται αναλυτικά η διαδικασία εξαγωγής βιολογικών δεδομένων για τις ανάγκες του μοντέλου μας. Γίνεται λεπτομερής περιγραφή των διαδικασιών προεπεξεργασίας προκειμένου τα δεδομένα να μετασχηματιστούν στην κατάλληλη μορφή για την εκπαίδευση των διαφόρων μοντέλων. Ταυτόχρονα, παρουσιάζονται τα εργαλεία που χρησιμοποιήθηκαν για την ανάπτυξη των μοντέλων. Παράλληλα, παρουσιάζεται υλοποίηση μεθόδου παραγοντοποίησης μητρώων (*matrix factorization*) καθώς και τεχνική αποδόμησης τανυστή (*tensor decomposition*) προκειμένου να συμπληρωθούν τυχών μη έγκυρες τιμές. Τέλος, αναλύονται οι αρχιτεκτονικές νευρωνικών δικτύων που υλοποιήθηκαν, οι λόγοι για τους οποίους επιλέχθηκαν καθώς και περαιτέρω παρατηρήσεις πάνω στο συγκεκριμένο πρόβλημα.

### 3.1 Ορισμός προβλήματος

Οπως αναφέρθηκε προηγουμένως, οι αλληλεπιδράσεις μεταξύ πρωτεϊνών παίζουν καθοριστικό ρόλο στις λειτουργίες των πρωτεϊνών, επιτρέποντας τους να εκτελέσουν τις βασικές βιολογικές διαδικασίες της ζωής των οργανισμών. Οι πρωτεΐνες αλληλεπιδρούν μεταξύ τους μέσω των *interfaces*, που αποτελούν περιοχές της επιφάνειάς τους με συγκεκριμένες γεωμετρικές και φυσικοχημικές ιδιότητες. Επομένως, το πρώτο βήμα για τον εντοπισμό πρωτεϊνικών αλληλεπιδράσεων είναι η εύρεση πληροφοριών σχετικά με πρωτεϊνικά *interfaces*.



Σχήμα 3.1: Παράδειγμα επιφάνειας αλληλεπίδρασης σε PPI

Η βασικότερη πηγή δεδομένων σχετικά με πρωτεϊνικά *interfaces* προέρχεται από τις πρωτεϊνικές δομές που είναι αποθηκευμένες στην *Protein Data Bank* και συγκεκριμένα εξάγονται μέσω X-ray crystallography. Ωστόσο, ο προσδιορισμός *interfaces* μέσω της παραπάνω βάσης δεδομένων αποτελεί μια χρονοβόρα και κοστοβόρα διαδικασία, ενώ παράλληλα μόνο το 50% των δομών που περιέχονται στην PDB είναι αλληλεπιδράσεις πρωτεϊνών, με το άλλο 50% να είναι μονομερή, νουκλεοτιδικές αλυσίδες κ.α. Ακόμη, μονάχα ένα μικρό κομμάτι από τις πρωτεϊνικές αλληλεπιδράσεις που περιέχονται στην PDB αποτελούν πραγματικές βιολογικές δομές, επο-

μένως η επιβεβαίωση των πρωτεΐνικών αλληλεπιδράσεων με έναν τρόπο υψηλής ακρίβειας αποτελεί δύσκολο πρόβλημα. Παράλληλα, η φύση της X-ray κρυσταλλογραφίας οδηγεί στην αποτύπωση κρυσταλλικών δομών που περιέχουν βιολογικά ασήμαντες κρυσταλλικές επαφές ή δεν περιέχουν σχετικές επαφές που υπάρχουν, με αποτέλεσμα την ανάγκη επαναπροσδιορισμού και διαχωρισμού των πραγματικών βιολογικών επαφών από τις κρυσταλλικές επαφές μεταξύ των πρωτεΐνων.

Για τους παραπάνω (και όχι μόνο) λόγους, δημιουργήθηκε η ανάγκη πρόβλεψης πρωτεΐνικών αλληλεπιδράσεων *in silico* προκειμένου να κατανοήσουμε περαιτέρω τις βιολογικές διαδικασίες αλλά και να διευρύνουμε την γνώση πάνω στον τομέα της κατασκευής φαρμάκων [60].

Υπάρχει ήδη ένας μεγάλος αριθμός μεθόδων για τον προσδιορισμό PPIs, με τις περισσότερες να εφαρμόζουν αλγορίθμους μηχανικής μάθησης πάνω σε σετ χαρακτηριστικών που εξάγονται από την ακολουθιακή ομολογία ή/και τη δομή των πρωτεΐνών με γνωστά interfaces (δηλαδή είναι χαρακτηριστικά sequence-based ή/και structure-based, που αναλύθηκαν στο κεφάλαιο 1). Οι μέθοδοι διαφέρουν στα δεδομένα που χρησιμοποιούν για την εκπαίδευση και αξιολόγηση των αλγορίθμων, στην φύση των interfaces (αν δηλαδή είναι transient ή/και obligate), στην φύση των προβλέψεων, στην επιλογή των residues για αξιολόγηση και άλλα. Στη συγκεκριμένη διπλωματική, παρουσιάζεται μια τεχνική μηχανικής μάθησης που εκπαιδεύεται σε sequence-based και structure-based χαρακτηριστικά και προβλέπει αλληλεπιδράσεις πρωτεΐνών τόσο σε επίπεδο patch οσο και σε επίπεδο residue.

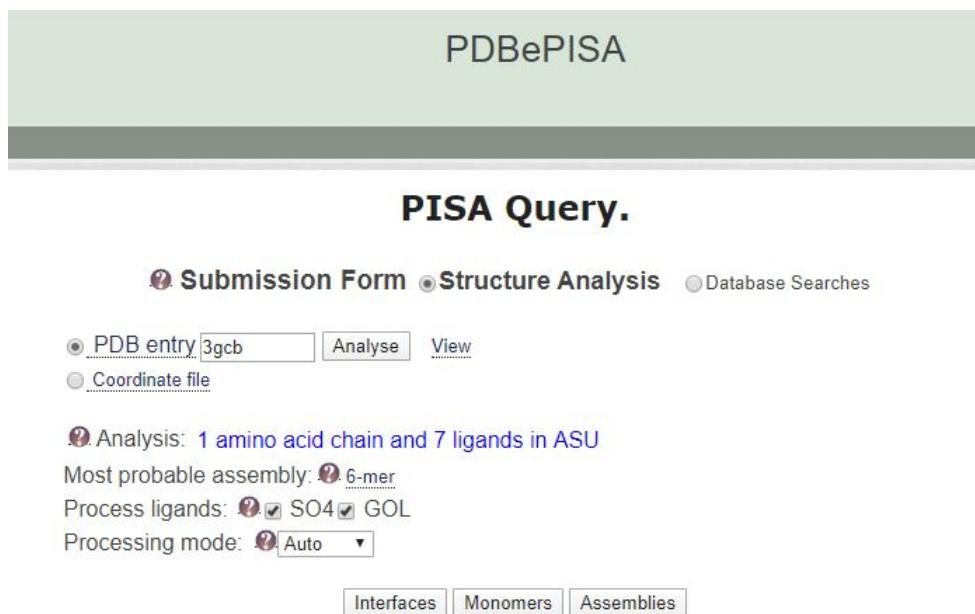
## 3.2 Δεδομένα

### 3.2.1 Ανάκτηση δεδομένων πρωτεΐνικών αλληλεπιδράσεων

Η εξαγωγή του σετ δεδομένων βασίζεται σε επιστημονική εργασία που δημοσιεύτηκε το 2017 με τίτλο *IntPred: a structure-based predictor of protein-protein interaction sites* [11] και ακολουθήθηκε η ίδια διαδικασία προκειμένου τα αποτελέσματα να είναι συγκρίσιμα.

Όπως αναφέρθηκε προηγουμένως, η φύση της X-ray κρυσταλλογραφίας οδηγεί στην αποτύπωση επαφών που δεν έχουν βιολογική σημασία. Για τον λόγο αυτό, χρησιμοποιείται το εργαλείο PISA (*Protein, Interfaces, Structures and Assemblies*), που εξάγει δεδομένα από την PDB και χρησιμοποιώντας μεθόδους χημικής θερμοδυναμικής διαχωρίζει τις συγκεντρώσεις μακρομορίων από τις κρυσταλλικές επαφές [61]. Μέσω του εργαλείου PISA, ανακτήθηκαν 58,397 βιολογικές μονάδες.

Στη συνέχεια πραγματοποιήθηκε ένας "καθαρισμός" των αποτελεσμάτων που ανακτήθηκαν. Αρχικά, διαγράφηκαν αποτελέσματα που πε-



Σχήμα 3.2: Εργαλείο PISA

ριείχαν πρωτεΐνικά περιβλήματα ιών (*viral capsids*) καθώς και αποτελέσματα που προέκυψαν μέσω NMR spectroscopy (η διαδικασία περιγράφηκε στο κεφάλαιο 1). Ακόμη, για λόγους ποιότητας των αποτελεσμάτων διαγράφηκαν αποτελέσματα με ανάλυση χειρότερη των 3Å R-factor μεγαλύτερο του 30%. R-factor είναι μια μετρική ομοιότητας μεταξύ ενός κρυσταλλογραφικού μοντέλου και των πειραματικών X-ray αποτελεσμάτων διάθλασης, όπου ουσιαστικά υποδειχνύει την ποιότητα του μοντέλου κρυσταλλογραφίας. Για ανάλυση μεγαλύτερη από 2Å, το R-factor δεν πρέπει να υπερβαίνει σημαντικά το 0.25. Τέλος, αφαιρέθηκαν τυχών πεπτιδικές αλυσίδες που προέκυψαν μέσω της ανάκτησης των αποτελεσμάτων (δηλαδή δομές με λιγότερα από 30 αμινοξέα). Με αυτό τον τρόπο, από τα αρχικά 58,397 αποτελέσματα προέκυψαν 25,876 δομές κατασκευασμένες από 87,738 αλυσίδες.

Επιπλέον, για την αποφυγή πανομοιότυπων ή παρόμοιων αποτελεσμάτων σε διαφορετικά entries, χρησιμοποιήθηκε το εργαλείο PISCES. Το PISCES (A Protein Sequence Culling Server) είναι ένα εργαλείο που δημιουργεί υποσύνολα υψηλής ποιότητας από μεγάλα σετ πρωτεΐνων, με βάση μια πληθώρα παραγόντων όπως η δομική ποιότητα, η μέγιστη κοινή ακολουθία ταυτοτήτων κ.α. Μέσω του εργαλείου PISCES, ομαδοποιήθηκαν οι αλυσίδες κατά 25% ακολουθιακή ομοιότητα (sequence similarity) και στη συνέχεια από κάθε ομάδα επιλέχθηκε ο καλύτερος "αντιπρόσωπος" με βάση την ποιότητα ανάλυσης ή (σε περίπτωση ισοβαθμίας) του καλύτερου R-factor. Έτσι, καταλήξαμε σε 4,345 αλυσίδες.

>>PISCES --server: Taking input parameters for culling whole PDB



### Choose your desired thresholds:

Maximum percentage identity:	<input type="text" value="25"/>
Minimum resolution:	<input type="text" value="0.0"/>
Maximum resolution:	<input type="text" value="3.0"/>
Maximum R-value:	<input type="text" value="0.3"/>
Minimum chain length:	<input type="text" value="40"/>
Maximum chain length:	<input type="text" value="10000"/>
Skip non-X-ray entries?	<input checked="" type="radio"/> Yes <input type="radio"/> No
Skip CA-only entries?	<input checked="" type="radio"/> Yes <input type="radio"/> No
How do you want to cull PDB? <small>(Help?)</small>	<input checked="" type="radio"/> By chains <input type="radio"/> By entries <small>(Help?)</small>

---

Following parameters are only for culling PDB by entries (Help?)

Cull chains within entries?	<input checked="" type="radio"/> Yes <input type="radio"/> No
Chain culling seq. id. threshold:	<input type="text" value="100"/>

---

Σχήμα 3.3: Εργαλείο PISCES

Η παραπάνω διαδικασία που περιγράφηκε αφορά την εξαγωγή των δεδομένων εκπαίδευσης εκτελέσθηκε ξεχωριστά και για τη δημιουργία του σετ αξιολόγησης, το οποίο περιείχε 4204 αλυσίδες.

### 3.2.2 Δημιουργία κομματιών επιφάνειας

Προκειμένου να υπολογιστούν οι ιδιότητες της πρωτεΐνικής επιφάνειας, η επιφάνεια πρέπει να διαιρεθεί σε μικρότερα μέρη (fragments). Για τον σκοπό αυτό χρησιμοποιήθηκε το πρόγραμμα *pdbmakepatch*, που ανήκει στο σύνολο προγραμμάτων υπολογιστικής βιολογίας *BiopTools* και αναπτύχθηκε το 2015 από το πανεπιστήμιο UCL [62], το οποίο εξάγει κομμάτια επιφάνειας από δοθείσες πρωτεΐνες.

Ειδικότερα, προτού αναλυθεί η λειτουργία του προγράμματος, πρέπει να γίνει μια εισαγωγή στους ακόλουθους όρους:

- **Patch centre atom :** Είναι το άτομο που δίνεται ως είσοδος στο πρόγραμμα και με βάση το οποίο υπολογίζονται τα κομμάτια (patches)

της επιφάνειας.

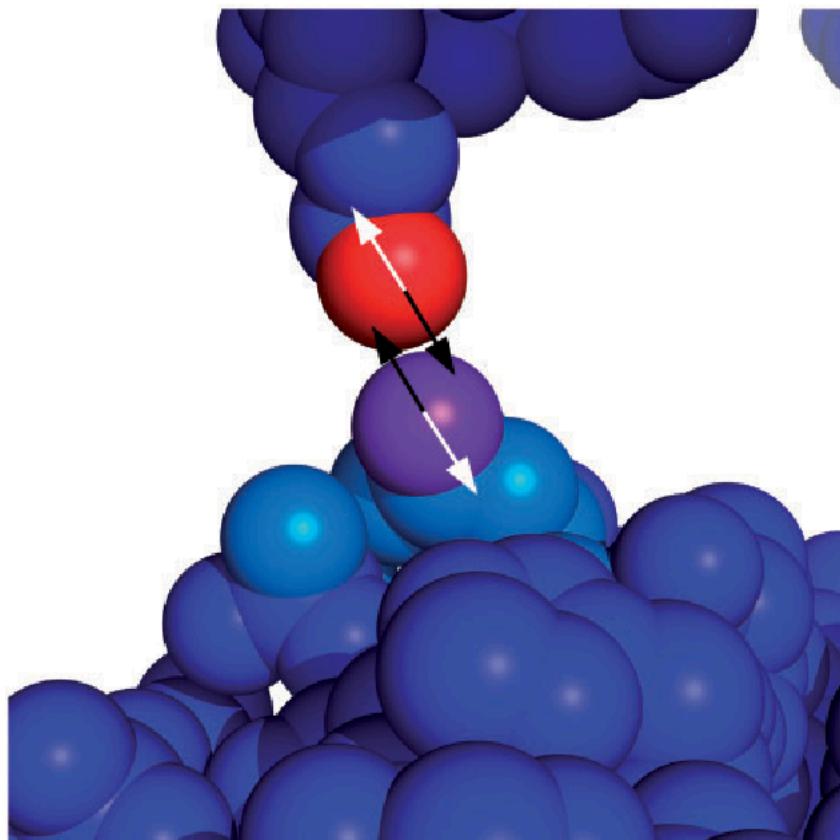
- **Patch radius** : Είναι η ελάχιστη απόσταση με βάση την οποία επιλέγονται τα υποψήφια residues για το τελικό κομμάτι.
- **Contact radius** : Ορίζεται για ένα ζεύγος ατόμων ώς το άθροισμα των ακτίνων Van der Waals τους, συν έναν παράγοντα σφάλματος (στο συγκεκριμένο πρόβλημα ορίζεται στα 2 Å). Δύο άτομα θεωρούμε ότι βρίσκονται σε επαφή αν η απόσταση των κέντρων τους είναι μικρότερη από το contact radius.
- **Residue geometry vector** : Πρόκειται για έναν πίνακα που ορίζεται για δοσμένα residue ξεκινώντας από το κέντρο του ατόμου ( Ca ) και τερματίζοντας στο γεωμετρικό κέντρο των 10 γειτονικά πλησιέστερων residues. Το γεωμετρικό κέντρο ορίζεται ως ο μέσος όρος των συντεταγμένων των κέντρων τους.
- **Residue solvent vector** : Πίνακας με παρόμοια αρχή όπως ο residue geometry vector, ωστόσο κινείται στην αντίθετη κατεύθυνση.
- **Solvent angle** : Ορίζεται για ένα ζεύγος residues ως η γωνία μεταξύ των residue solvent vectors τους.

Για ένα δοσμένο αρχείο pdb και ένα patch centre atom, το πρόγραμμα `pdbmakepatch` δημιουργεί ένα κομμάτι ακολουθώντας την εξής επαναληπτική διαδικασία:

1. Ορίζεται  $P$  ενα αρχικά άδειο σύνολο ατόμων του κομματιού και προσθέτουμε το patch centre atom στο  $P$ .
2. Βρίσκουμε όλα τα residues με τουλάχιστον ένα κέντρο ατόμου μέσα στην ακτίνα (patch radius) του κεντρικού ατόμου (patch centre atom). Τα συγκεκριμένα residues αποτελούν το σύνολο  $C$  των υποψήφιων για εισαγωγή στο κομμάτι.
3. Για κάθε μέλος του συνόλου  $P$ , ελέγχουμε αν υπάρχει επαφή με κάποιο από τα μέλη του συνόλου  $C$ . Αν κάποιο μέλος του συνόλου  $C$  βρίσκεται σε επαφή με μέλος του  $P$  και το solvent angle που σχηματίζουν είναι μικρότερο από  $120^\circ$ , τότε το μεταφέρουμε στο  $P$ .
4. Επαναλαμβάνουμε το βήμα 3 μέχρις ότου να μην υπάρχει μεταφορά μελών από το σύνολο  $C$  στο σύνολο  $P$ .

5. Κάθε residue με άτομο στο σύνολο  $C$  ορίζεται ως *patch residue*.

Η χρήση του solvent angle γίνεται για να μην συμπεριληφθούν residues από αντίθετες πλευρές στο ίδιο κομμάτι, με αποτέλεσμα τη δημιουργία διακεκομμένων κομματιών, οπως φαίνεται στην παρακάτω εικόνα. Το άτομο υποψήφιος (χόκκινο) βρίσκεται εντός της απόστασης επαφής (contact radius) από ένα άτομο που ανήκει στο κομμάτι (μωβ). Οι residue geometry vectors (άσπρο) χρησιμοποιούνται για να υπολογιστούν οι solvent angle vectors (μαύρο), με βάση τους οποίους υπολογίζεται το solvent angle. Στη συγκεκριμένη περίπτωση, καθώς η γωνία ξεπερνά τις  $120^\circ$ , το άτομο υποψήφιος δεν συμπεριλαμβάνεται στο κομμάτι.



Σχήμα 3.4: Παράδειγμα χρήσης solvent angle

Για όλες τις δομές που προέκυψαν κατά την δημιουργία των δεδομένων εκπαίδευσης, δημιουργούνται σετ επικαλυπτόμενων κομματιών (*overlapping patches*) για να απεικονισθεί η επιφάνειά τους. Για να δημιουργηθούν τα παραπάνω σετ, επιλέχθηκαν residues με *relative solvent accessibility* (RSA)  $> 25\%$ . Η συγκεκριμένη μετρική, που ονομάζεται και σχετικά προσβάσιμη επιφάνεια (relative solvent accessible area - RASA),

αποτελεί μια μετρική έκθεσης ενός residue και υπολογίζεται με τον ακόλουθο τύπο:

$$RASA = \frac{ASA}{\max(ASA)} \quad (3.1)$$

where:

$$\begin{aligned} ASA &= \text{προσβάσιμη επιφάνεια από διαλυτικό μέσο} \\ \max(ASA) &= \text{μέγιστη δυνατή προσβάσιμη επιφάνεια για το residue.} \end{aligned}$$

Τα παραπάνω επιλεγμένα residues αποτελούν το σύνολο των κέντρων κομματιού. Για κάθε residue που ανήκει σε αυτό το σύνολο, το άτομο με την υψηλότερη ASA βρίσκεται και επιλέγεται ως patch centre atom για είσοδο στο πρόγραμμα *pdbmakepatch*.

### 3.2.3 Ανάθεση κατηγορίας

Για την ανάθεση της κατηγορίας του κάθε κομματιού (class label), υπολογίζεται το ποσοστό της RASA (Relative Solvent Accessible Area) που οφείλεται σε residues τα οποία έχουν χαρακτηριστεί ως residues επιφάνειας. Ένα residue χαρακτηρίζεται ως residue επιφάνειας εαν ικανοποιεί την παραχάτω σχέση:

$$RASA_i^n - RASA_i^c \geq 10\% \quad (3.2)$$

where:

$$\begin{aligned} RASA_i^n &= \text{non complex RASA τιμή του residue i.} \\ RASA_i^c &= \text{complex RASA τιμή του residue i.} \end{aligned}$$

Το ποσοστό της RASA που οφείλεται σε residues επιφάνειας, γνωστό και ως interface fraction  $fASA_p$ , για ένα κομμάτι επιφάνειας  $p$  που περιέχει ένα σύνολο από residues  $r_p$ , και ένα υποσύνολο από residues επιφάνειας  $r_{intf}$  υπολογίζεται από τον ακόλουθο τύπο:

$$fASA_p = \frac{\sum_{j \in r_{intf}} RASA_j^n}{\sum_{i \in r_p} RASA_i^n} \quad (3.3)$$

Τελικά, η κατηγορία που δίνεται σε ένα κομμάτι βασίζεται στον παραχάτω κανόνα:

$$C_p = \begin{cases} I & \text{if } fASA_p \geq 0.5 \\ S & \text{if } fASA_p = 0 \\ U & \text{otherwise} \end{cases} \quad (3.4)$$

where:

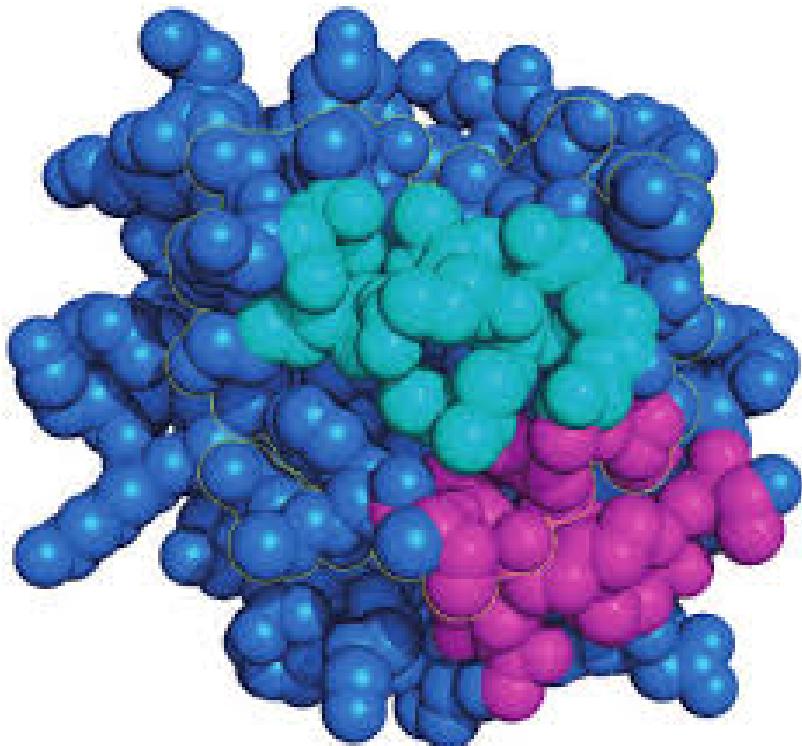
$I$  = Interaction

$S$  = Surface

$U$  = Unlabelled

Ο χαρακτηρισμός ενός κομματιού ως *unlabelled*, δηλαδή δίχως κατηγορία, γίνεται σε περιπτώσεις κομματιών που βρίσκονται στο άκρο της αλληλεπίδρασης ( γίνεται περαιτέρω επεξήγηση σε φωτογραφία στη συνέχεια ). Τα κομμάτια που χαρακτηρίζονται έτσι παραλείπονται από την εκπαίδευση του μοντέλου, προκειμένου να διατηρηθεί η δυαδική φύση του προβλήματος κατηγοριοποίησης, ωστόσο περιλαμβάνονται στα δεδομένα αξιολόγησης όταν οι προβλέψεις κομματιού μεταφέρονται σε επίπεδο residues.

Η παρακάτω φωτογραφία περιέχει μια τοποθεσία αλληλεπίδρασης (το περίγραμμα της οποίας τονίζεται με κίτρινο χρώμα). Το κομμάτι με το κυανό χρώμα χαρακτηρίζεται ως κομμάτι αλληλεπίδρασης (interface patch) ενώ το κομμάτι με το κόκκινο χρώμα ως *unlabelled* καθώς το ποσοστό των residues του κομματιού που εντάσσονται στην αλληλεπίδραση δεν είναι επαρκώς υψηλό.



Σχήμα 3.5: Παράδειγμα *unlabelled* κομματιού

### 3.2.4 Χαρακτηριστικά δεδομένων

Τα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν για τον εντοπισμό των αλληλεπιδράσεων χρησιμοποίησαν 8 χαρακτηριστικά (*features*) για την εκπαίδευση και την αξιολόγηση τους, τα οποία μπορούν να χωρισθούν σε ακολουθιακά (*sequential*) και δομικά (*structural*) χαρακτηριστικά

#### 3.2.4.1 Ακολουθιακά χαρακτηριστικά

Τα παρακάτω χαρακτηριστικά λαμβάνουν υπόψιν τους μονάχα ακολουθιακές ιδιότητες. Με τον όρο ακολουθιακό χαρακτηριστικό (*sequential feature*) εννοούμε μια ομάδα αμινοξέων μέσα σε μια πρωτεΐνη που παρέχει ορισμένες ιδιότητες. Ορισμένα παραδείγματα ακολουθιακών χαρακτηριστικών είναι τα binding sites καθώς και τα post-translational modification (PTM) sites. Καθώς τα ακολουθιακά χαρακτηριστικά υπολογίζονται σε επίπεδο residues, η τιμή των χαρακτηριστικών για κάθε κομμάτι αποτελεί απλώς τον μέσο όρο των τιμών των residues τους.

- **Hydrophobicity:** Η υδροφοβικότητα αποτελεί τη φυσική ιδιότητα ενός μορίου να απωθείται φαινομενικά από μια μάζα νερού. Σε έρευνες που έχουν διεξαχθεί τα τελευταία χρόνια παρατηρείται ότι οι περιοχές αλληλεπίδρασης σε μια πρωτεΐνη είναι πιο υδροφοβικές από την υπόλοιπη πρωτεΐνική επιφάνεια [63] [64]. Για το συγκεκριμένο πρόβλημα, η υδροφοβικότητα ενός residue αντιπροσωπεύεται από την τιμή της στην κλίμακα που ορίστηκε από τους Kyte και Doolittle, η οποία υποδεικνύει υδροφοβικότητα σε αμινοξέα [65] και συμφωνα με την οποία περιοχές με θετικές τιμές ορίζονται ως υδροφοβικές.
- **Propensity:** Ένας άλλος τρόπος διαχωρισμού των συνθέσεων αμινοξέων μεταξύ περιοχών αλληλεπίδρασης και περιοχών επιφάνειας είναι μέσω της ροπής/τάσης των residues (*propensity*). Στο συγκεκριμένο πρόβλημα χρησιμοποιείται η τάση των residues ως προς την ολική προσβάσιμη επιφάνεια (ASA), καθώς αποτελεί ένα από τα πιο διαδεδομένα χαρακτηριστικά για την μελέτη της λειτουργίας των αλληλεπιδράσεων [66]. Η ροπή ενός residue  $i$  τύπου  $X$  υπολογίζεται ως:

$$Pr(i, X) = \left( \ln \frac{F_{intf}(X)}{F_{surf}(X)} \right) \times \left( \frac{ASA(i)}{\overline{ASA}_{surf}(X)} \right) \quad (3.5)$$

where:

- |                            |   |
|----------------------------|---|
| $F_{intf}(X)$              | = ποσοστό αλληλεπίδρασης του residue τύπου X.             |
| $F_{surf}(X)$              | = ποσοστό επιφάνειας του residue τύπου X.                 |
| $ASA(i)$                   | = Absolute Solvant-accessible surface Area του residue i. |
| $\overline{ASA}_{surf}(X)$ | = μέση ASA για όλα τα residues επιφάνειας τύπου X.        |

Συμπεριλαμβάνεται ο όρος της απόλυτης προσβάσιμης επιφάνειας από διαλυτικό μέσο (ASA) για να συμπεριλαμβάνεται η προσφορά της ASA κάθε residue i ανεξαρτήτως τύπου, αντί να αντιμετωπίζονται οι κατανομές κάθε residue ενός τύπου X ως ίδιες. Ταυτόχρονα, η εισαγωγή του όρου  $\overline{ASA}_{surf}(X)$  γίνεται ως ρυθμιστικός παράγοντας στις διαφορές μεγέθους στις αλυσίδες αμινοξέων, προκειμένου να αποφευχθεί η υπερ-αντιπροσώπευση (over-representation) μεγάλων αλυσίδων.

Για ένα residue τύπου X, το ποσοστό αλληλεπίδρασης  $F_{intf}$  υπολογίζεται ως εξής:

$$F_{intf}(X) = \frac{\sum ASA_{intf}^n(X)}{\sum ASA_{intf}^n} \quad (3.6)$$

οπου ο αριθμητής υποδηλώνει την ολική απόλυτη προσβάσιμη επιφάνεια από διαλυτικό μέσο στα δεδομένα εκπαίδευσης για τα residues τύπου X, ενώ ο παρονομαστής είναι η ολική απόλυτη προσβάσιμη επιφάνεια από διαλυτικό μέσο για όλα τα residues αλληλεπίδρασης.

Αντίστοιχα, το ποσοστό επιφάνειας  $F_{surf}(X)$  υπολογίζεται ως:

$$F_{surf}(X) = \frac{\sum ASA_{surf}^n(X)}{\sum ASA_{surf}^n} \quad (3.7)$$

με τις τιμές του αριθμητή και παρονομαστή να αντιστοιχούν με αυτές της 3.6 για τα residues επιφάνειας των δεδομένων εκπαίδευσης και τα συνολικά residues επιφάνειας αντίστοιχα.

Θετικές τιμές στη ροπή υποδεικνύουν υπερ-αντιπροσώπευση ενός residue τύπου X στο σύνολο των αλληλεπιδράσεων, ενώ αρνητικές τιμές υποδεικνύουν το αντίθετο. Με αυτό τον τρόπο, residues τα οποία έχουν υψηλές θετικές τιμές είναι πιθανότερο να συμμετέχουν στις αλληλεπιδράσεις απ' ότι residues με χαμηλές τιμές, γεγονός που υποδηλώνει μια συσχέτιση που μπορεί να αξιοποιηθεί από τα μοντέλα μηχανικής μάθησης.

- **Conservation Scores:** Conservation score είναι ένα σχετικό μέτρο της εξελικτικής διατήρησης (evolutionary conservation) μιας αλυσίδας αμινοξέων σε μια πρωτεΐνη και βασίζεται στις φυλογενετικές σχέσεις μεταξύ ομόλογων ακολουθιών αμινοξέων (οι όροι "φυλογενετικός" και "ομόλογος" παρουσιάστηκαν συνοπτικά στο κεφάλαιο 1). Ειδικότερα, ο βαθμός στον οποίο μια αλυσίδα αμινοξέων είναι εξελικτικά διατηρημένη (για παράδειγμα ο ρυθμός εξέλιξης της αλυσίδας) συνδέεται με την δομική και λειτουργική σημασία της πρωτεΐνης. Στη συγκεκριμένη υλοποίηση χρησιμοποιήθηκαν 2 conservation scores για

κάθε residue, ενα *functionally equivalent protein (FEP) score* και ένα *homologue score*.

Για τον υπολογισμό του *FEP score*, χρησιμοποιήθηκε αρχικά το εργαλείο *PDBWS*, που αντιστοιχίζει αλυσίδες που προέρχονται από τη βάση δεδομένων PDB με εγγραφές στην βάση δεδομένων *UnitProtKB/SwissProt* [67]. Η βάση δεδομένων *UnitProtKB* περιέχει ακολουθιακές και λειτουργικές πληροφορίες για πρωτεΐνες. Η *SwissProt* αποτελεί μια επιμελημένη βάση δεδομένων σχετικά με ακολουθίες πρωτεΐνων, μέρος του knowledgebase (KB) της βάσης δεδομένων *UnitProtKB* που περιέχει πληροφορίες από επιστημονικές εργασίες και επιβλεπόμενες υπολογιστικές αναλύσεις, με τις πληροφορίες να ελέγχονται και να σχολιάζονται χειροκίνητα από ειδικούς με σκοπό την παραγωγή αποτελεσμάτων με υψηλή ανάλυση και εύκολη σύνδεση με άλλες βάσεις δεδομένων.

Στη συνέχεια, χρησιμοποιείται η βάση δεδομένων *FOSTA* για την εύρεση της οικογένειας των λειτουργικά ισοδύναμων ορθόλογων όπου κάθε *UnitProtKB/SwissProt* εγγραφή είναι μέλος. Η βάση δεδομένων *FOSTA* είναι μια βάση δεδομένων FEPs που παρέχει πληροφορίες και αυτοματοποιημένη ανάλυση με σκοπό την εξαγωγή ομάδων πρωτεΐνων που έχουν χαρακτηρισθεί ως λειτουργικά ισοδύναμες (*functionally equivalent*) [68]. Από τα αποτελέσματα διατηρούνται και προχωρούν για επεξεργασία μόνο οι οικογένειες που περιέχουν τουλάχιστον 9 ακόμη μέλη.

Για τον υπολογισμό του *homologue score*, πραγματοποιείται μια αναζήτηση *BLAST* στην βάση δεδομένων *UnitProtKB/SwissProt*. *BLAST* (Basic Local Alignment Search Tool) ονομάζεται ένας αλγόριθμος βιοπληροφορικής που συγκρίνει βασικές πληροφορίες μεταξύ πρωτεΐνων σε ακολουθιακό επίπεδο και εντοπίζει περιοχές τοπικής ομοιότητας ενώ παράλληλα υπολογίζει την στατιστική σημαντικότητα των αποτελεσμάτων [69]. Στη συνέχεια πραγματοποιείται ένας καθαρισμός των αποτελεσμάτων, όπου εγγραφές με τους χαρακτηρισμούς *putative,predicted* ή *hypothetical* απορρίπτονται, όπως και εγγραφές με  $E-value > 0.01$ . *E-value* ονομάζεται ο αριθμός των αναμενόμενων αποτελεσμάτων με παρόμοιο σκορ που μπορεί να σημειωθούν κατά τύχη. Επομένως, όσο πιο κοντά στο 0 είναι η συγκεκριμένη μετρική, τόσο πιο "σημαντικό" είναι το αποτέλεσμα. Αν έχουν διατηρηθεί το ελάχιστο 10 αποτελέσματα, τότε μπορούν να προχωρήσουν για επεξεργασία εως και 200 αποτελέσματα, κατεταγμένα με βάση το μικρότερο *E-value*.

Για κάθε σύνολο αποτελεσμάτων που πέρασε για επεξεργασία, χρησιμοποιείται το εργαλείο *MUSCLE* για την παραγωγή ευθυγραμμίσεων (alignments). Το εργαλείο *MUSCLE* (*MUltiple Sequence Comparison by Log Expectation*) παράγει ευθυγραμμίσεις μεταξύ τριών ή και

περισσοτέρων βιολογικών αλυσίδων παρόμοιου μήκους, με εξαιρετικά αποτελέσματα για την ευθυγράμμιση πρωτεϊνών. Από το αποτέλεσμα εξάγονται συμπεράσματα σχετικά με την ομολογία (homology), καθώς και τις εξελικτικές σχέσεις μεταξύ των αλυσίδων που μελετήθηκαν [70].

Τέλος, κάθε ευθυγράμμιση που εντοπίζεται χρησιμοποιείται για τον υπολογισμό conservation scores με τη μέθοδο Valdar01. Η μέθοδος Valdar01 παρουσιάστηκε το 2001 από τους William S. J Valdar και Janet M. Thornton για την ποσοτική αξιολόγηση της διατήρησης των residues κάτα την εξελικτική πορεία ενός βιολογικού οργανισμού [71]. Η βαθμολόγηση των δύο συνόλων έγινε με τη χρήση του προγράμματος scorecons, άλλο ένα πρόγραμμα που αποτελεί κομμάτι του συνόλου εργαλείων υπολογιστικής βιολογίας BiopTools. Κατά την βαθμολόγηση, το σκορ ενός κομματιού προκύπτει από τον μέσο όρο των σκορ των residues του.

### 3.2.4.2 Δομικά χαρακτηριστικά

Τα παρακάτω χαρακτηριστικά χρειάζονται δομική πληροφορία προκειμένου να υπολογιστούν. Ως δομική πληροφορία εννοείται η πληροφορία σχετικά με την 3-D διευθέτηση των ατόμων σε μια αλυσίδα αμινοξέων. Όπως και στην περίπτωση των ακολουθιακών χαρακτηριστικών, έτσι και τα δομικά χαρακτηριστικά στη συγκεκριμένη υλοποίηση υπολογίζονται για κάθε κομμάτι ως ο μέσος όρος των τιμών των residues τους.

- **Intra-Chain disulphide bonds :** Γνωστοί και ως S-S bonds, πρόκειται για ομοιοπολικούς δεσμούς που αναπτύσσονται μεταξύ δύο ομάδων θειόλης (*thiol*), δηλαδή δύο θειικών αναλόγων αλκοόλης (όπου ένα άτομο θείου αντικαθιστά ένα άτομο οξυγόνου στο υδροξύλικό κομμάτι μιας αλκοόλης). Οι δεσμοί υπολογίζονται από το πρόγραμμα *pdblistss*, που εντοπίζει S-S bonds, σύμφωνα με τον αλγόριθμο των Hazes και Dijkstra [72], φάχνοντας για ζεύγη *S* με αποστάσεις μεταξύ τους μικρότερες από 2.5 Å, συν έναν παράγοντα αβεβαιότητας 10 %. Ένα residue παίρνει την τιμή 1 αν σχηματίζει δεσμό S-S και 0 αν δεν σχηματίζει.
- **Intra-Chain hydrogen bonds:** Οι δεσμοί υδρογόνου υπολογίζονται από το πρόγραμμα *rdhhbond*, που ακολουθεί τους κανόνες για την εύρεση δεσμών υδρογόνου που τέθηκαν από τους Baker και Hubbard το 1984 [73] και διατυπώνονται ως εξής:

Δοθέντος ενός ατόμου "δότη" *D* (με το οποίο δένεται το άτομο υδρογόνου) και ενός ατόμου αποδέκτη *A*, σχηματίζεται δεσμός υδρογόνου εαν η απόσταση  $H \cdots A \leq 2.5 \text{ Å}$  και η γωνία του υδρογόνου είναι μεταξύ 90-180 °. Σε περίπτωση που η θέση του υδρογόνου δεν μπορεί να υπολογιστεί, τότε σχηματίζεται δεσμός υδρογόνου εαν η απόσταση

$D \cdots A \leq 3.35 \text{ \AA}$  και η γωνία μεταξύ  $D - A$  είναι μεταξύ 90-180 °.

Με βάση τον παραπάνω κανόνα, ένα residue παίρνει την τιμή 1 εαν σχηματίζει δεσμό υδρογόνου ενώ παίρνει την τιμή 0 στην αντίθετη περίπτωση.

- **Secondary structure:** Η δευτεροταγής δομή μιας πρωτεΐνης αναφέρεται στην τρισδιάστατη μορφή των τοπικών μερών της. Στο συγκεκριμένο πρόβλημα, χρησιμοποιήθηκε το πρόγραμμα *pdbsecstr* της πλατφόρμας εργαλείων *Bio3D Tools*, το οποίο αναθέτει δευτεροταγή δομή σε ένα residue με βάση τον κανόνα που θέτουν οι *W. Kabsch* και *C. Sander* [74]. Σύμφωνα με τον παραπάνω κανόνα, ανατίθεται σε ένα κομμάτι  $SS_p$  δευτεροταγή δομή με βάση τις παρακάτω προϋποθέσεις:

$$SS_p = \begin{cases} H & \text{if } \alpha > 20\% \text{ and } \beta \leq 20\% \\ E & \text{if } \alpha \leq 20\% \text{ and } \beta > 20\% \\ EH & \text{if } \alpha > 20\% \text{ and } \beta > 20\% \\ C & \text{if } \alpha \leq 20\% \text{ and } \beta \leq 20\% \end{cases} \quad (3.8)$$

where:

$H$	=	$\alpha$ -helix secondary structure
$E$	=	$\beta$ -sheet secondary structure
$EH$	=	mixed secondary structure
$C$	=	coil secondary structure
$\alpha$	=	% of residues assigned as $\alpha$ -helix
$\beta$	=	% of residues assigned as $\beta$ -sheet

- **Planarity:** Ύπολογίζεται μέσω της ρίζας της μέσης τετραγωνικής απόστασης (root mean squared distance) όλων των ατόμων από το επίπεδο καλύτερης επαφής. Το επίπεδο της καλύτερης επαφής βρίσκεται κεντράροντας τις συντεταγμένες ( $x, y, z$ ) των ατόμων ενός κομματιού και εκτελώντας PCA (Principal Component Analysis), όπου τα πρώτα δύο primary components της μεθόδου ορίζουν το επίπεδο.

### 3.2.5 Σύνοψη δεδομένων εισόδου

Στις προηγούμενες υποενότητες, έγινε μια περιγραφή του τρόπου εύρεσης δεδομένων που περιέχουν πληροφορίες σχετικά με αλληλεπιδράσεις μεταξύ πρωτεΐνων. Παρουσιάστηκαν εργαλεία επεξεργασίας καθώς και μέθοδοι καθαρισμού των δεδομένων ενώ στη συνέχεια αναλύθηκε η διαδικασία εξαγωγής πληροφορίας κομματιών (*patches*) καθώς και το τελικό *label* κάθε κομματιού. Τέλος, υπήρξε μια σύντομη παρουσίαση των χαρακτηριστικών που υπολογίστηκαν για κάθε κομμάτι, χαρακτηριστικά τα οποία θα χρησιμοποιηθούν κατά την εκπαίδευση των μοντέλων μηχανικής μάθησης.

Στον παρακάτω πίνακα παρουσιάζεται συνοπτικά η μορφή των δεδομένων τα οποία θα χρησιμοποιηθούν για την εκπαίδευση μοντέλων μηχανικής μάθησης, καθώς και για την αξιολόγησή τους, για προβλήματα εντοπισμού PPIs:

Χαρακτηριστικό	Περιγραφή	Τύπος
<b>Ακολουθιακά Χαρακτηριστικά</b>		
<b>prop</b>	propensity score	Συνεχής αριθμητική τιμή
<b>hpho</b>	hydrophobicity	Συνεχής αριθμητική τιμή
<b>homology</b>	homology conservation score	Συνεχής αριθμητική τιμή
<b>FEP</b>	FEP conservation score	Συνεχής αριθμητική τιμή
<b>Δομικά Χαρακτηριστικά</b>		
<b>SS</b>	disulphide bonds	Συνεχής αριθμητική τιμή
<b>Hb</b>	hydrogen bonds	Συνεχής αριθμητική τιμή
<b>SecStr</b>	secondary structure	Κατηγορική τιμη ( $H, E, EH, C$ )
<b>Κατηγορία</b>		
<b>intf</b>	Κατηγορία δεδομένων εισόδου	Δυαδική κατηγορική τιμή ( $I, S$ )

Πίνακας 3.1: Σύνοψη δεδομένων εισόδου

### 3.3 Προεπεξεργασία

Οι προηγούμενες ενότητες παρουσίασαν τις διαδικασίες δημιουργίας ενός συνόλου από δεδομένα αλληλεπιδράσεων πρωτεϊνών. Στη συγκεχριμένη ενότητα, παρουσιάζεται η επεξεργασία των συγκεκριμένων δεδομένων προτού αυτά τοποθετηθούν στα μοντέλα μηχανικής μάθησης.

#### 3.3.1 Εργαλεία

Αρχικά, ας αναφέρουμε τα εργαλεία που χρησιμοποιήθηκαν για την προεπεξεργασία των δεδομένων. Ειδικότερα, χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python (**Python 3.6.9**) στο περιβάλλον προγραμματισμού του **Google Colab** καθώς και η γλώσσα προγραμματισμού **Matlab**. Το **Google Colab**, ένα προϊόν της Google Research, είναι μια μορφή υπηρεσίας **Jupyter Notebook**, δηλαδή εργαλείο που επιτρέπει τόσο την συγγραφή κώδικα όσο και την υποστήριξη πλούσιων στοιχείων κειμένου (π.χ. εξισώσεις, εικόνες κ.α.). Δίνει τη δυνατότητα συγγραφής και εκτέλεσης κώδικα Python μέσω browser, χωρίς την ανάγκη περαιτέρω εγκατάστασης, ενώ παρέχει δωρεάν πρόσβαση σε υπολογιστικές μονάδες όπως GPU και TPU για την επιτάχυνση του χρόνου εκτέλεσης των προγραμμάτων των χρηστών. Επιπλέον, επιτρέπει την απευθείας σύνδεση του κώδικα με τον προσωπικό αποθηκευτικό χώρο του Google Drive, καθιστώντας την όλη διαδικασία κατανεμημένη καθώς πραγματοποιείται στο cloud. Από την άλλη, η **Matlab** αποτελεί ένα περιβάλλον αριθμητικών υπολογισμών, το οποίο αναπτύχθηκε από την Mathworks και επιτρέπει μεταξύ άλλων την επεξεργασία και τροποποίηση πινάκων, την οπτικοποίηση γραφικών παραστάσεων, την υλοποίηση αλγορίθμων, τη δημιουργία διεπαφών χρήστη κ.α., ενώ έχει και τη δυνατότητα αλληλεπίδρασης με προγράμματα γραμμένα σε άλλες γλώσσες προγραμματισμού.

#### 3.3.2 Διαδικασία

Το αποτέλεσμα της διαδικασίας που παρουσιάστηκε στην παράγραφο 3.2 είναι ένα αρχείο σε μορφή **.csv**, το οποίο φορτώνεται στο περιβάλλον του **Google Colab** και αποθηκεύεται σε μια δομή δεδομένων **Dataframe** για περαιτέρω επεξεργασία.

Μελετώντας τα δεδομένα, παρατηρούνται ορισμένα ”προβλήματα”, τα οποία πρέπει να μελετηθούν. Συγκεκριμένα:

1. Τα δεδομένα περιέχουν ταυτόχρονα αριθμητικά και κατηγορικά δεδομένα.
2. Μικρή διακύμανση στις τιμές ορισμένων πεδίων (π.χ στο πεδίο **SS-bonds**).
3. Παρουσία μη έγκυρων (**NaN** τιμών).

Το πρώτο ζήτημα χρειάζεται αντιμετώπιση καθώς τα μοντέλα μηχανικής μάθησης απαιτούν την ύπαρξη αριθμητικών τιμών προκειμένου να

πραγματοποιήσουν τους υπολογισμούς τους και επιλύεται πραγματοποιώντας μια απλή αντιστοίχιση των κατηγορικών τιμών σε αριθμητικές.

Το δεύτερο ζήτημα είναι περισσότερο παρατήρηση για τη μορφή των δεδομένων και μπορεί να οπτικοποιηθεί καλύτερα βλέπωντας την τυπική απόκλιση των αριθμητικών χαρακτηριστικών των δεδομένων.

Τυπική απόκλιση χαρακτηριστικών	
Χαρακτηριστικό	Τιμή
propensity	0.114462
hydrophobicity	0.886340
planarity	0.767370
SSbonds	0.017156
Hbonds	0.164863
fosta_scorecons	0.193172
blast_scorecons	0.170614

Πίνακας 3.2: Τυπική απόκλιση προ-επεξεργασμένων δεδομένων

Εύκολα παρατηρείται ότι η τυπική απόκλιση των τιμών του χαρακτηριστικού SSbonds είναι ποσοτικά μια τάξη μεγέθους κάτω από τις υπόλοιπες, γεγονός που μπορεί να επηρεάσει την απόδοση των μοντέλων μηχανικής μάθησης. Αυτό, σε συνδυασμό με τον μικρό αριθμό διακριτών τιμών στο συγκεκριμένο χαρακτηριστικό (μόλις 141 διακριτές τιμές από 317,531 εγγραφές) μας προϊδεάζουν για την μάθηση του μοντέλου από το συγκεκριμένο χαρακτηριστικό.

Number of NaN values per column in pre-processed data:	
patchID	0
propensity	43
hydrophobicity	43
planarity	0
secondary_str	26
SSbonds	0
Hbonds	0
fosta_scorecons	249078
blast_scorecons	83601
intf_class	0
dtype: int64	

Σχήμα 3.6: Αριθμός ελλειπών τιμών

Οσον αφορά το τρίτο ζήτημα, αυτό αποτελεί αρκετά σημαντικό πρόβλημα και πρέπει να αντιμετωπιστεί. Κατά την εκτέλεση των προγραμμάτων που αναφέρθηκαν στην παράγραφο 3.2, παρατηρήθηκαν ορισμένες μη έγκυρες τιμές, ο αριθμός των οποίων φαίνεται στην εικόνα. Παρατηρούμε έναν πολύ μεγάλο αριθμό μη έγκυρων τιμών για τα conservation scores χαρακτηριστικά, όπου στο FEP conservation score έχουμε 83,601 μη έγκυρες τιμές αλλά ειδικότερα στην περίπτωση του homologue conservation score, όπου από τις 317,531 εγγραφές, οι 249,078 είναι

μη-έγκυρες. Η ύπαρξη τους, παρ' ότι κατανοητή λόγω του τρόπου υπολογισμού των συγκεκριμένων χαρακτηριστικών, δημιουργεί πρόβλημα κατά την εκπαίδευση των μοντέλων, καθώς ένα νευρωνικό δίκτυο δεν μπορεί να τρέξει με μη έγκυρες τιμές ως είσοδο. Για την αντιμετώπιση του συγκεκριμένου προβλήματος δοκιμάστηκαν διάφορες μέθοδοι που θα παρουσιαστούν στη συνέχεια.

Στη συνέχεια, πραγματοποιείται μια μορφή κανονικοποίησης των δεδομένων που ονομάζεται *standardization*. Με τον όρο *standardization* ή *Z-score normalization* εννοούμε την επεξεργασία των δεδομένων, προκειμένου κάθε χαρακτηριστικό να έχει τις ιδιότητες μιας κανονικής κατανομής, δηλαδή μέση τιμή 0 και τυπική απόκλιση 1. Η κανονικοποίηση αυτή είναι αρκετά σημαντική κατά την επεξεργασία τιμών με διαφορετικές κλίμακες, αλλά αποτελεί και προϋπόθεση για πολλούς αλγορίθμους μηχανικής μάθησης. Η κανονικοποίηση αυτή γίνεται με τον ακόλουθο τρόπο, αφαιρώντας από κάθε δείγμα την μέση τιμή του και διαιρώντας με την τυπική του απόκλιση. Αξίζει να δοθεί προσοχή για την κανονικοποίηση ανά χαρακτηριστικό και όχι για όλο το σύνολο των δεδομένων, προκειμένου να έχουμε θετικά αποτελέσματα.

$$z = \frac{x - \mu_{col}}{\sigma_{col}} \quad (3.9)$$

where:

$$\begin{aligned} \mu_{col} &= \text{μέση τιμή κάθε χαρακτηριστικού} \\ \sigma_{col} &= \text{τυπική απόκλιση κάθε χαρακτηριστικού} \end{aligned}$$

Mean of features after Standardization:	
propensity	-0.000122
hydrophobicity	-0.000056
planarity	-0.000024
secondary_str	0.000002
SSbonds	-0.000066
Hbonds	0.000022
fosta_scorecons	0.000023
blast_scorecons	0.000028
dtype:	float32

(α') Μέση Τιμή

Standard Deviation of features after Standardization:	
propensity	0.999990
hydrophobicity	0.999972
planarity	0.999999
secondary_str	1.000711
SSbonds	0.999995
Hbonds	0.999969
fosta_scorecons	1.000010
blast_scorecons	1.000005
dtype:	float32

(β') Τυπική απόκλιση

Σχήμα 3.7: Standardization

### 3.4 Συμπλήρωση μητρώου

Για την αντιμετώπιση των πολλαπλών *NaN* τιμών που υπήρχαν στα παραδείγματα εκπαίδευσης, χρησιμοποιήθηκε μια τεχνική συμπλήρωσης των τιμών που ονομάζεται *σημπλήρωση μητρώων* (*matrix completion*). Συμπλήρωση μητρώου ονομάζεται η ανάκτηση ενός πίνακα από υποδειγματολειπτημένες ή ελλειπείς/μη έγκυρες εγγραφές. Ειδικότερα, αν θεωρήσουμε τον επιθυμητό πίνακα ως  $X$  και τον πίνακα που περιέχει τις ελλειπείς τιμές ως  $X_\Omega$ , τότε ισχύει η παρακάτω σχέση:

$$X_\Omega = H_\Omega \odot X + N \quad (3.10)$$

where:

- $\Omega$  = υποσύνολο που περιέχει τις συντεταγμένες των έγκυρων εγγραφών
- $\odot$  = τελεστής πολυμού στοιχείο προς στοιχείο
- $\Omega$  = πίνακας δειγματοληψίας
- $N$  = πίνακας θορύβου

Στην δημοσίευση των *E. J. Candes* και *B. Recht* [75], προτείνεται η ελαχιστοποίηση της τάξης του πίνακα προκειμένου να ανακτηθεί ο αρχικός πίνακας  $X$ , δηλαδή:

$$\min_M \text{rank}(M), \quad s.t. \quad M_\Omega = X_\Omega \quad (3.11)$$

where:

- $M$  = πίνακας ανάκτησης,  $M \in R^{m \times n}$
- $X_\Omega$  =  $H_\Omega \odot X$
- $M_\Omega$  = προβολή του  $M$  στο  $\Omega$

Στην περίπτωση που ο αρχικός πίνακας έχει υποβαθμιστεί από θόρυβο, υπάρχει ανάγκη περιορισμού του επιπέδου θορύβου μέσα σε ένα επιθυμητό εύρος, κάτι που περιγράφεται από την ακόλουθη σχέση:

$$\min_M \text{rank}(M), \quad s.t. \quad \|M_\Omega - X_\Omega\|_F < \delta \quad (3.12)$$

where:

- $\|\cdot\|_F$  = Frobenius norm ενός πίνακα
- $\delta$  = παράμετρος περιορισμού σφάλματος,  $\delta > 0$

Ωστόσο, το συγκεκριμένο πρόβλημα είναι NP-hard οσον αφορά την επίλυση του, καθώς όλες οι μέθοδοι ακριβούς επίλυσης είναι διπλά

εκθετικοί της μέγιστης διάστασης του μητρώου  $X$   $\max(m, n)$  όσον αφορά τη διαστασιμότητα. Αυτός είναι και ο λόγος που οι "βέλτιστοι" αλγόριθμοι επιχειρούν να λύσουν προσεγγιστικά το πρόβλημα ελαχιστοποίησης της τάξης (*rank minimization*). Στη συγκεκριμένη εργασία χρησιμοποιήθηκαν δύο τεχνικές συμπλήρωσης μητρώων, με σκοπό την ανάκτηση της πληροφορίας των *NAN* τιμών. Στη συνέχεια αναλύεται η κάθε τεχνική, γίνεται η μαθηματική τους διατύπωση και στο τέλος συγκρίνονται οι δύο τεχνικές σχετικά με την ακρίβειά τους.

### 3.4.1 Συμπλήρωση μητρώου μέσω SGD

Η πρώτη τεχνική που χρησιμοποιήθηκε για την συμπλήρωση του dataset βασίζεται στην εργασία των *R. Gemulla, P. J. Haas, E. Nijkamp και Y. Sismanis* με τίτλο *Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent* [76]. Στη συγκεκριμένη δημοσίευση παρουσιάζεται μια τεχνική παραγοντοποίησης μητρώων μέσω *Stochastic Gradient Descent*.

Η μέθοδος gradient descent αποτελεί έναν από τους βασικότερους αλγορίθμους βελτιστοποίησης καθώς και τον πιο διαδεδομένο αλγόριθμο βελτιστοποίησης νευρωνικών δικτύων. Ελαχιστοποιεί μια αντικειμενική συνάρτηση  $J(\theta)$  με παραμέτρους  $\theta \in R^\theta$  τροποποιώντας τις παραμέτρους στην αντίθετη κατεύθυνση της παραγώγου της αντικειμενικής συνάρτησης  $\nabla_\theta J(\theta)$  συναρτήσει των παραμέτρων. Μια απλή μαθηματική έκφραση των μεθόδων gradient descent είναι η ακόλουθη:

$$\theta_t = \theta_{t-1} - n * \nabla_\theta J(\theta) \quad (3.13)$$

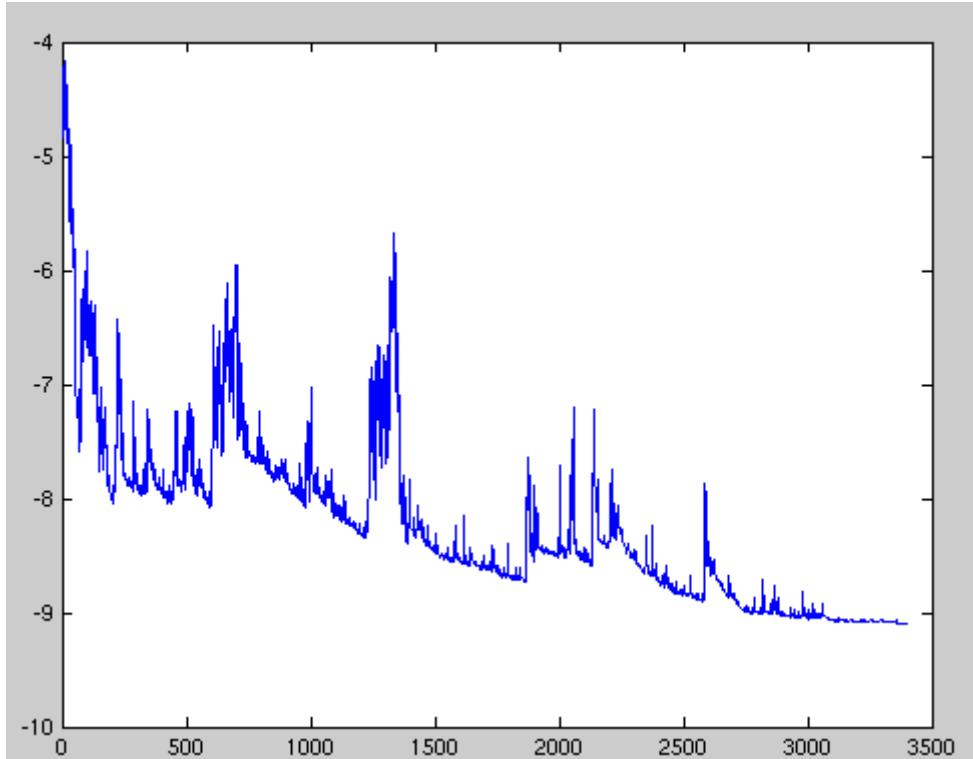
where:

- $\theta$  = παράμετροι του μοντέλου.
- $J(\theta)$  = αντικειμενική συνάρτηση.
- $n$  = ρυθμός μάθησης (learning rate).

Οστόσο, καθώς για τον υπολογισμό ενός μόνο βήματος τροποποίησης των παραμέτρων απαιτείται η εύρεση της παραγώγου ολόκληρου του σετ δεδομένων, οι αλγόριθμοι gradient descent τείνουν να συγκλίνουν πολύ αργά και να δημιουργούν προβλήματα για την βελτιστοποίηση δεδομένων που δεν χωρούν στη μνήμη. Για το λόγο αυτό προτάθηκε η μέθοδος *Stochastic Gradient Descent*, που αποτελεί τροποποίηση της μεθόδου gradient descent. Αντικαθιστά τον υπολογισμό της παραγώγου ολόκληρου του σετ των δεδομένων με μια προσέγγιση αυτής, υπολογισμένη από ένα τυχαίο υποσύνολο των δεδομένων.

Ενώ η μέθοδος gradient descent συγκλίνει σε ένα τοπικό ελάχιστο, η μεγάλη διακύμανση με την οποία συγκλίνει η μέθοδος Stochastic Gradient Descent επιτρέπει τη σύγκλιση σε νέα και πιθανώς βελτιωμένα τοπικά

ελάχιστα. Από την άλλη, η έντονη αυτή διακύμανση καθιστά τη διαδικασία σύγκλισης (λόγω άλματων στις τιμές της αντικειμενικής συνάρτησης). Για την αντιμετώπιση αυτού του προβλήματος, χρησιμοποιείται μειούμενος ρυθμός μάθησης και παρατηρείται παρόμοιος ρυθμός σύγκλισης με αυτόν της gradient descent.



Σχήμα 3.8: Διακύμανση SGD

Η μαθηματική διατύπωση της Stochastic Gradient Descent, όπου πραγματοποιείται τροποποίηση των παραμέτρων με βάση ένα τυχαίο υποσύνολο των δεδομένων  $x^{(i)}, y^{(i)}$  είναι η εξής:

$$\theta_t = \theta_{t-1} - n * \nabla_{\theta} J(\theta; x^{(i)}; y^{(i)}) \quad (3.14)$$

where:

- $\theta$  = παράμετροι του μοντέλου.
- $J(\theta)$  = αντικειμενική συνάρτηση.
- $n$  = ρυθμός μάθησης (learning rate).

Για την εφαρμογή της μεθόδου Stochastic Gradient Descent στην παραγοντοποίηση του συνόλου δεδομένων μας, θεωρούμε ως παραμέτρους τα μητρώα  $W$  και  $H$  (τέτοια ώστε  $V = WH$  οπου  $V$  το αρχικό σύνολο

των δεδομένων μας). Ακόμη, θεωρούμε ως αντικειμενική συνάρτηση τη συνάρτηση απωλειών  $L_{NZSL}$ , που ορίζεται ως εξής:

$$L_{NZSL} = \sum_{i,j: V_{ij} \neq 0} (V_{ij} - [WH]_{ij})^2 \quad (3.15)$$

Παράλληλα, αναλύουμε την  $L_{NZSL}$  ως άθροισμα συναρτήσεων τοπικών απωλειών  $l$  σε ένα υποσύνολο των δεδομένων  $V$ , δηλαδή:

$$L = \sum_{i,j \in \mathbb{Z}} l(V_{ij}, W_{i*}, H_{*j}) \quad (3.16)$$

Ο αλγόριθμος που χρησιμοποιήθηκε με χρήση Stochastic Gradient Descent για την παραγοντοποίηση του μητρώου δεδομένων μας είναι ο εξής:

---

#### Αλγόριθμος 1: SGD for Matrix Factorization

---

**Data:**

$Z$ : σύνολο δεδομένων,

$W, H$ : αρχικοποιημένοι με τυχαίες τιμές.

**while** δεν συγκλίνει **do**

  Επιλογή τυχαίου παραδείγματος  $(i, j) \in Z$ .

$W'_{i*} \leftarrow W_{i*} - \epsilon_n \frac{\partial}{\partial W_{i*}} l(V_{ij}, W_{i*}, H_{*j})$

$H'_{*j} \leftarrow H_{*j} - \epsilon_n \frac{\partial}{\partial H_{*j}} l(V_{ij}, W_{i*}, H_{*j})$

$W_{i*} \leftarrow W'_{i*}$

$H_{*j} \leftarrow H'_{*j}$

**end**

---

Μετά την επιλογή του τυχαίου παραδείγματος  $(i, j) \in Z$ , χρειάζεται μόνο η τροποποίηση των  $W_{i*}$  και  $H_{*j}$ , μειώνοντας σημαντικά τους αναγκαίους υπολογισμούς. Η μείωση αυτή οφείλεται και στην έκφραση των συνολικών απωλειών ως άθροισμα τοπικών απωλειών. Όσον αφορά την αντικατάσταση της ακριβούς παραγώγου με την προσέγγισή της, εκτός από την μείωση της υπολογιστικής πολυπλοκότητας, διευκολύνεται η αποφυγή τοπικών ελαχίστων, εντοπίζεται επαναληφθιμότητα σε δεδομένα ενώ παράλληλα οι τροποποιήσεις βασισμένες σε δεδομένα συγκεκριμένων γραμμών/στηλών μειώνει τις απώλειες σε αντίστοιχες γραμμές/στήλες. Επομένως, όσο μεγαλύτερη ομοιότητα εμφανίζεται στα δεδομένα, τόσο ταχύτερη είναι η σύγκλιση του αλγορίθμου [77]. Στην περίπτωση του dataset μας, καθώς κάθε αλληλεπίδραση χωρίζεται σε patches που διαθέτουν παρόμοιες

(αν όχι πανομοιότυπες) ιδιότητες, παρατηρούνται πολλαπλές όμοιες εγγραφές, με αποτέλεσμα την γρήγορη σύγκλιση του αλγορίθμου. Με απλές αλγεβρικές τροποποιήσεις προκύπτουν και οι παράγωγοι της αντικειμενικής συνάρτησης  $L_{NZSL}$  ως:

<b>Συνάρτηση Απώλειας - Ορισμός και Παράγωγοι</b>
---

$$L_{NZSL} = \sum_{(i,j) \in Z} (V_{ij} - [WH]_{ij})^2$$

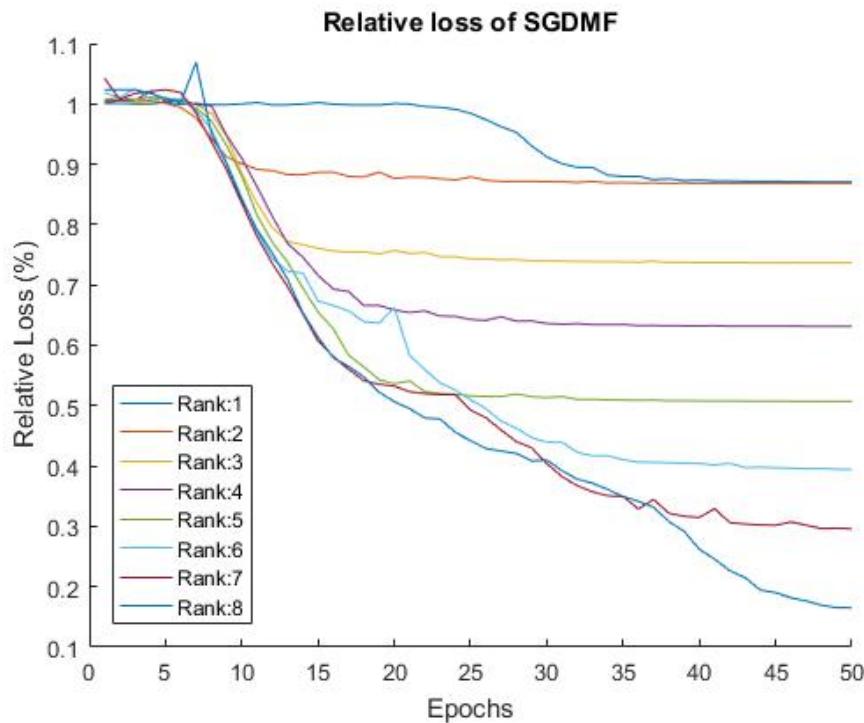
$$\frac{\partial}{\partial W_{ik}} L_{ij} = -2 * (V_{ij} - [WH]_{ij}) * H_{kj}$$

$$\frac{\partial}{\partial H_{kj}} L_{ij} = -2 * (V_{ij} - [WH]_{ij}) * W_{ik}$$

Πίνακας 3.3: Συνάρτηση Απώλειας - Ορισμός, Παράγωγοι

Για την υλοποίηση του αλγορίθμου 1, χρησιμοποιήθηκε το υπολογιστικό περιβάλλον της Matlab και για ρυθμό εκπαίδευσης χρησιμοποιήθηκε ως αρχική τιμή το  $2^{-27}$  (που αποτελούσε τη μέγιστη τιμή θετική τιμή για την οποία δεν απέκλεινε ο αλγόριθμος στα πρώτα βήματα) ενώ στη συνέχεια χρησιμοποιήθηκε μια ευρετική μέθοδος που ονομάζεται **bold driver** και χρησιμοποιείται συχνά για την τροποποίηση του βήματος μεθόδων gradient descent. Ειδικότερα, σε κάθε εποχή αυξάνεται ο ρυθμός εκπαίδευσης κατά ένα μικρό ποσοστό (στην περίπτωσή μας 5%) εφόσον παρατηρήθει μείωση στις απώλειες, ενώ σε περίπτωση αύξησης των απώλειών μειώνεται δραστικά ο ρυθμός εκπαίδευσης (π.χ. 50%). Τέλος, κατά τη διάρκεια κάθε εποχής ο ρυθμός εκπαίδευσης παραμένει σταθερός. Όσον αφορά τη σύγκλιση του αλγορίθμου, χρησιμοποιήθηκε ως αντικειμενική συνάρτηση η σχετική απόσταση μεταξύ των γνωστών (μη-μηδενικών) στοιχείων του αρχικού συνόλου δεδομένων και των αντίστοιχων στοιχείων του μητρώου που υπολογίστηκε ( $WH$ ).

Στη συνέχεια, πραγματοποιήθηκαν πειράματα προκειμένου να βρεθεί η τάξη της βέλτιστης αναπαράστασης χαμηλότερης τάξης του αρχικού μας συνόλου δεδομένων και παρακάτω παρουσιάζονται τα αποτελέσματα. Παρατηρούμε ότι η μέθοδος συγκλίνει αρκετά γρήγορα, ενώ όσο αυξάνεται η τάξη του ανακατασκετασμένου πίνακα μειώνονται σημαντικά οι σχετικές απώλειες. Για τάξη  $r = 8$  παρατηρήθηκαν οι ελάχιστες σχετικές απώλειες  $Loss = 0.16255$ . Πειράματα που διεξήχθησαν και για τάξεις με  $r > 8$  δεν έδειξαν σημαντική βελτίωση, οπότε χρησιμοποιήθηκε η συγκεκριμένη ανακατασκευή για την εκπαίδευση των νευρωνικών μοντέλων.



Σχήμα 3.9: Σχετικές απώλειες SGD

### 3.4.2 Παραγοντοποίηση μητρώου μέσω τανυστικής αποδόμησης

Στο χώρο της μηχανικής μάθησης, ένας τανυστής (*tensor*) ονομάζεται μια πολυδιάστατη δομή ή, πιο αυστηρά, ένας τανυστής  $N$ -οστής τάξης είναι το αποτέλεσμα του τανυστικού γινομένου  $N$  διανυσματικών χώρων, ο καθένας εκ των οποίων έχει το δικό του σύστημα συντεταγμένων [78]. Προτού περάσουμε στην παρουσίαση της μεθόδου που χρησιμοποιήθηκε για την παραγοντοποίηση μητρώου μέσω τανυστικής αποδόμησης, θα οριστούν κάποιες βασικές έννοιες σχετικά με τους τανυστές. Ως τάξη (*order*) ενός τανυστή ορίζουμε τη διαστασιμότητά του. Στη συγκεκριμένη διπλωματική εργασία, οι τανυστές με τους οποίους εργαζόμαστε είναι δεύτερης τάξης και αποτελούν μητρώα, επομένως μελλοντικές αναφορές στον όρο "τανυστής" θα απασχολούν τανυστές 2 τάξης για λόγους απλότητας, ενώ ο συμβολισμός των τανυστών δεύτερης τάξης θα γίνεται με κεφαλαία γράμματα με έντονη γραφή (π.χ.  $A$ ). Τα στοιχεία  $(i, j)$  ενός τανυστή  $A$  συμβολίζονται ως  $a_{ij}$  και έχουν δείκτες  $i = 1, \dots, I$ . Ως νόρμα ενός τανυστή  $A \in \mathbb{R}^{I \times J}$  ορίζεται η τετραγωνική ρίζα του αθροίσματος των τετραγώνων όλων των στοιχείων του, δηλαδή:

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^I \sum_{j=1}^J a_{ij}^2} \quad (3.17)$$

Ιδιαίτερη σημασία στην ανάλυση τανυστών στη μηχανική μάθηση εμφανίζουν και τα γινόμενα μητρώαν, προκειμένου να εκφραστεί ένας τανυστής ως το προϊόν αυτών. Μερικά από αυτά τα γινόμενα είναι το γινόμενο Kronecker, το γινόμενο Khatri-Rao, το γινόμενο Hadamard, με εκτενή ανάλυση αυτών στη σχετική βιβλιογραφία [79].

Ως εισαγωγή στην αποδόμηση τανυστών, πρέπει να ορίσουμε την έννοια του βαθμού ενός τανυστή. Ένας τανυστής  $N$ -οστής τάξης θεωρείται πρώτου βαθμού (rank one) όταν μπορεί να γραφεί ως το εξωτερικό γινόμενο  $N$  διανυσμάτων, π.χ.:  $\mathbf{A} = a^{(1)} \circ a^{(2)}$ , οπου το  $\circ$  συμβολίζει το εξωτερικό γινόμενο διανυσμάτων. Ο βαθμός (rank) ενός τανυστή  $\mathbf{A}$  ορίζεται ως ο ελάχιστος αριθμός τανυστών πρώτου βαθμού που το άθροισμα αυτών μας δίνει τον  $\mathbf{A}$ . Η ιδέα της έκφρασης ενός τανυστή ως το πεπερασμένο άθροισμα τανυστών πρώτου βαθμού παρουσιάστηκε το 1927 από τον F. Hitchcock ως πολυαδική μορφή ενός τανυστή [80], ενώ το 1944 ο R. Cattell παρουσίασε ιδέες για παράλληλη ανάλυση πολυαδικών μορφών και χρήση πολλαπλών αξόνων για ανάλυση [81]. Οι έννοιες αυτές έμειναν σχετικά άγνωστες μέχρι το 1970 και την εισαγωγή τους στο χώρο της ψυχοακουστικής, με τη μορφή των CANDECOMP (Canonical Decomposition) από τους J. D. Carroll και J. J. Chang [82] και PARAFAC (Parallel Factors) από τον R. Harshman [83]. Θα αναφερόμαστε στην αποδόμηση τανυστών CANDECOMP/PARAFAC ως αποδόμηση CP (CP decomposition). Η αποδόμηση CP παραγοντοποιεί έναν τανυστή ως ένα άθροισμα  $R$  τανυστών πρώτου βαθμού [84] [85]. Για π.χ., δοσμένου ενός τανυστή τρίτης τάξης  $\mathbf{X} \in \mathbb{R}^{I \times J \times K}$ , επιθυμούμε να τον γράψουμε ως:

$$\mathbf{X} \approx \sum_{r=1}^R a_r \circ b_r \circ c_r \quad (3.18)$$

οπου:  $R$  θετικός ακέραιος,  $a_r \in \mathbb{R}^I$ ,  $b_r \in \mathbb{R}^J$  και  $c_r \in \mathbb{R}^K$  για  $r = 1, \dots, R$ . Στην εικόνα 3.10 παρουσιάζεται η CP αποδόμηση ενός τανυστή τρίτης τάξης.

Το ζητούμενο στην τανυστική αποδόμηση είναι η εύρεση του αριθμού  $R$  των διανυσμάτων που προσεγγίζουν τον τανυστή. Συνεπώς, σύμφωνα με τους παραπάνω ορισμούς, το  $R$  είναι ο βαθμός του τανυστή και η διαδικασία εύρεσής του είναι μια αρκετά πολύπλοκη διαδικασία ( εχει αποδειχθεί ότι είναι NP-hard πρόβλημα [86] ). Πρακτικά, η εύρεση της τάξης  $R$  ενός τανυστή γίνεται εφαρμόζοντας διαφορετικές τιμές στο  $R$  και

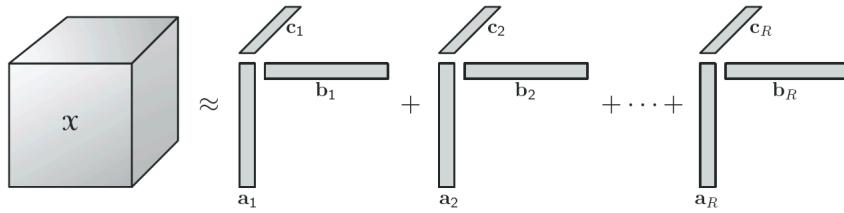


Fig. 3.1 CP decomposition of a three-way array.

## Σχήμα 3.10: Αποδόμηση τανυστή τρίτης τάξης

βλέποντας πόσο καλά προσεγγίζεται ο αρχικός τανυστής από τον ανακατασκευασμένο. Η μαθηματική έκφραση της παραπάνω πρότασης παρουσιάζεται στην εξίσωση 3.19, όπου με λ συμβολίζουμε ένα διάνυσμα στο οποίο αποθηκεύονται τα βάρη και το οποίο πολλαπλασιάζουμε με το εξωτερικό γινόμενο των τριών διανυσμάτων.

$$\min_{\widehat{X}} \|X - \widehat{X}\| \text{ where } \widehat{X} = \sum_{r=1}^R \lambda_r a_r \circ b_r \circ c_r \quad (3.19)$$

Στη διπλωματική εργασία χρησιμοποιήθηκε μια τροποποιημένη εκδοχή της αποδόμησης CP που ονομάζεται βεβαρημένη βελτιστοποίηση μεσω CP αποδόμησης (CP Weighted Optimization η πίο απλά CP-WOPT) [87] [88]. Ειδικότερα, στην περίπτωση τανυστών με ελλιπή δεδομένα, η αποδόμηση CP μπορεί να μοντελοποιηθεί ως ένα βεβαρημένο πρόβλημα ελαχίστων τετραγώνων με έμφαση μόνο στις γνωστές τιμές. Η μέθοδος CP-WOPT επιλύει το βεβαρημένο πρόβλημα ελαχίστων τετραγώνων με τη χρήση βελτιστοποίησης πρώτης τάξης. Τροποποιεί την συνάρτηση σφάλματος ωστε να αγνοούνται οι μη υπάρχουσες τιμές και να μοντελοποιούνται μονάχα οι γνωστές τιμές, επομένως μετά μπορούν να χρησιμοποιηθούν μέθοδοι μη γραμμικής βελτιστοποίησης για την απευθείας επίλυση του βεβαρημένου προβλήματος ελαχίστων τετραγώνων με CP αποδόμηση. Η τροποποιημένη της συνάρτησης σφάλματος φαίνεται στην εξίσωση 3.20:

$$f_w(A, B) = \frac{1}{2} \sum_{i=1}^I \sum_{j=1}^J \left\{ w_{ij} (x_{ij} - \sum_{r=1}^R a_{ir} b_{jr}) \right\}^2 \quad (3.20)$$

οπου  $\mathbf{W}$  ενας μη-αρνητικός τανυστής βαρών ίδιου μεγέθους με τον  $\mathbf{X}$  που τα στοιχεία του ορίζονται ως:

$$w_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_{ij} \text{ is known.} \\ 0 & \text{if } \mathbf{x}_{ij} \text{ is missing.} \end{cases} \quad (3.21)$$

Πρακτικά ο τανυστής  $\mathbf{W}$  λειτουργεί ως μάσκα προκειμένου να ληφθούν υπόψη μόνο οι γνωστές τιμές. Χρησιμοποιώντας τους ορισμούς που παρουσιάστηκαν προηγουμένως, μπορούμε να ορίσουμε τη συνάρτηση απωλειών ως:

$$f_w(\mathbf{A}, \mathbf{B}) = \frac{1}{2} \|\mathbf{W} \cdot (\mathbf{X} - \mathbf{AB}^T)\|^2 \quad (3.22)$$

όπου \* ο τελεστής του γινομένου Hadamard (ή αλλιώς το γινόμενό τους στοιχείο προς στοιχείο). Σκοπός η εύρεση των  $\mathbf{A}$  και  $\mathbf{B}$  που ελαχιστοποιούν την αντικειμενική συνάρτηση 3.22.

Για την ελαχιστοποίηση της αντικειμενικής συνάρτησης χρησιμοποιήθηκε ως αλγόριθμος βελτιστοποίησης η μέθοδος μη γραμμικής συζύγους παραγώγου (*Nonlinear conjugate gradient method - ncg*) [89]. Πρόκειται για μια μέθοδο που επεκτείνεται ικανοποιητικά για προβλήματα μεγάλου μεγέθους και έχει την ακόλουθη έκφραση:

Δεδομένου ενός προβλήματος ελαχιστοποίησης χωρίς περιορισμούς  $\min f(x), x \in \mathbb{R}^n$ , η μέθοδος *ncg* έχει τη μορφή:

$$x_{k+1} = x_k + a_k d_k, \quad k = 0, 1, 2, \dots \quad (3.23)$$

όπου

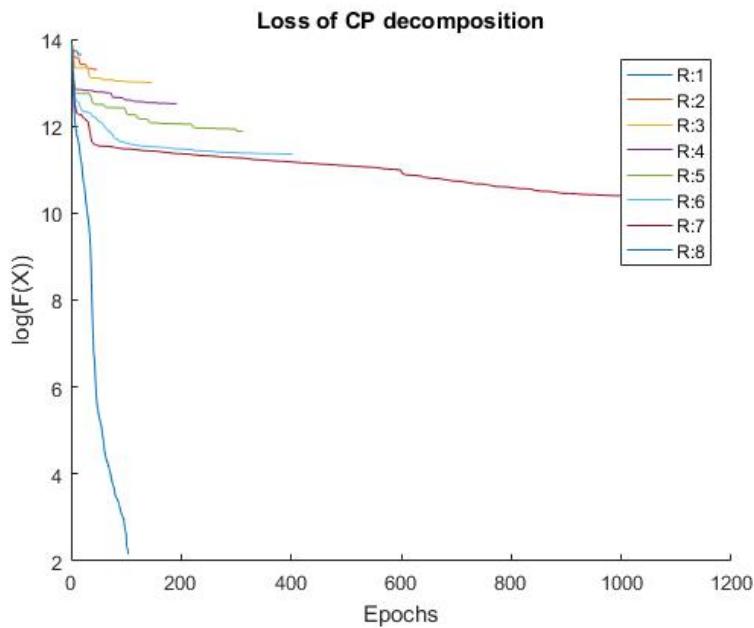
$x_0$  = αρχικό σημείο εκκίνησης.

$a_k$  = ρυθμός εκπαίδευσης (learning rate).

$$d_k = \begin{cases} -g_k & k = 0; \\ -g_k + \beta_k d_{k-1} & k \geq 1, \end{cases} \quad \text{where } g_k = \nabla f(x_k)$$

Διαφορετικά  $k$  ορίζουν διαφορετικές μεθόδους *ncg*. Μερικοί δημοφιλείς τύποι για το  $k$  παρουσιάζονται στον πίνακα 3.4

Τύποι $\beta_k$ για μεθόδους ncg	
Μέθοδος	Τύπος
Fletcher-Reeves	$\beta^{FR} = \frac{\ g_k\ ^2}{\ g_{k-1}\ ^2}$
Polak-Ribiere-Polyak	$\beta^{PRP} = \frac{g_k^T(g_k - g_{k-1})}{\ g_{k-1}\ ^2}$
Hestenes-Stiefel	$\beta^{HS} = \frac{g_k^T(g_k - g_{k-1})}{d_{k-1}^T(g_k - g_{k-1})}$
Conjugate Descent	$\beta^{CD} = -\frac{\ g_k\ ^2}{d_{k-1}^T g_{k-1}}$
Dai-Yuan	$\beta^{DY} = \frac{\ g_k\ ^2}{d_{k-1}^T(g_k - g_{k-1})}$
Liu-Storey	$\beta^{LS} = -\frac{g_k^T(g_k - g_{k-1})}{d_{k-1}^T g_{k-1}}$

Πίνακας 3.4: Δημοφιλείς τύποι  $\beta_k$  [2]

Σχήμα 3.11: Απώλειες CP αποδόμησης

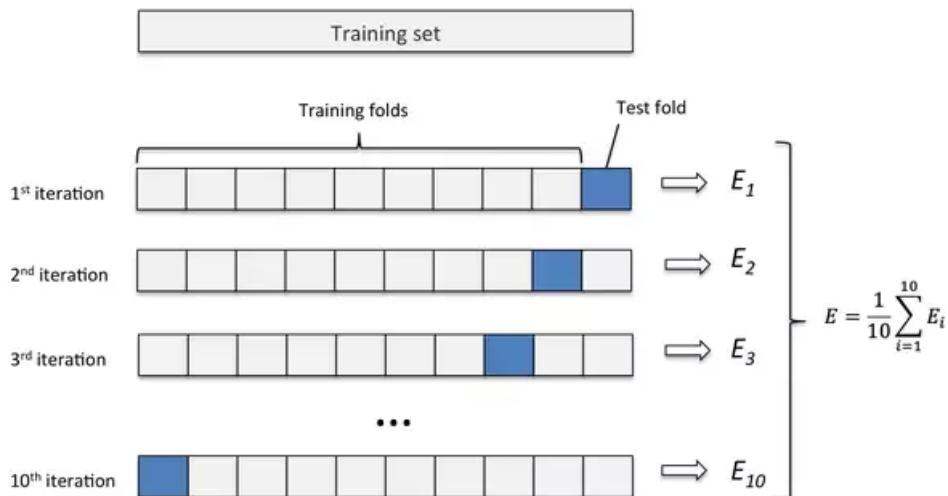
Στα πλαίσια της διπλωματικής χρησιμοποιήθηκε ο τύπος των *Polak-Ribiere-Polyak* για την τροποποίηση του  $\beta_k$ . Στην εικόνα 3.11 παρουσιάζονται οπτικά τα αποτελέσματα της CP αποδόμησης για διαφορετικούς βαθμούς  $r$ , με τις απώλειες εμφανίζονται λογαριθμικά λόγω της τεράστιας διαφοράς των τιμών. Παρατηρούμε ότι για  $r = 8$  ο ανακατασκευασμένος τανυστής προσεγγίζει σημαντικά τον αρχικό τανυστή ( $F(x) = 8.75$ ), επομένως λαμβάνουμε τη συγκεκριμένη ανακατασκευή ως σύνολο δεδομένων προς εκπαίδευση.

### 3.5 Εκπαίδευση

Έπειτα από την ανακατασκευή των δεδομένων με τις μεθόδους της παραγοντοποίησης μητρώου μέσω SGD και της τανυστικής αποδόμησης, τα δεδομένα είναι σε κατάλληλη μορφή για να τροφοδοτηθούν σε νευρωνικά δίκτυα. Για τη σωστή αντιμετώπιση του προβλήματος, πρέπει να επιλεγεί τόσο η κατάλληλη δομή/αρχιτεκτονική νευρωνικών δικτύων, αλλά και να γίνει σωστή ρύθμιση των υπερπαραμέτρων, στις οποίες έγινε αναφορά στο κεφάλαιο 2. Η κατασκευή των νευρωνικών δικτύων καθορίζει τον τρόπο μη γραμμικής συσχέτισης μεταξύ εισόδων και εξόδου. Δοκιμάστηκαν πολλές διαφορετικές αρχιτεκτονικές και ρυθμίστηκε πειραματικά μια πληθώρα παραμέτρων, ωστόσο στα πλαίσια της εργασίας θα παρουσιαστούν οι δύο καλύτερες αρχιτεκτονικές, μια αρχιτεκτονική πλήρως διασυνδεδεμένου δικτύου (*fully connected neural network*) και μια αρχιτεκτονική συνελικτικού δικτύου (*convolutional neural network*). Αρχικά, η εκπαίδευση καθώς και η αξιολόγηση των δεδομένων πραγματοποιήθηκε με τη μέθοδο  $k$ -απλή επικυρωμένη διαστάυρωση *k-fold cross-validation*. Επικυρωμένη διαστάυρωση (*Cross-validation*) ονομάζεται μια στατιστική μέθοδος που χρησιμοποιείται για την εκτίμηση της ικανότητας των νευρωνικών δικτύων. Στην περίπτωση του *k*-fold cross validation, τα δεδομένα χωρίζονται σε  $k$  ομάδες, με τις  $k - 1$  να αποτελούν το σετ δεδομένων εκπαίδευσης (training set) και την 1 να αποτελεί το σετ δεδομένων αξιολόγησης (test set). Το μοντέλο εκπαιδεύεται με τα δεδομένα εκπαίδευσης, αξιολογείται η απόδοση του στα δεδομένα αξιολόγησης και αποθηκεύονται οι μετρικές αξιολόγησης. Η διαδικασία επαναλαμβάνεται για κάθε μια εκ των  $k$  ομάδων και η απόδοση του μοντέλου ορίζεται ως ο μέσος όρος των αποδόσεων για κάθε εκπαίδευση του μοντέλου. Στην περίπτωσή μας, επιλέχθηκε η αξιολόγηση των μοντέλων μέσω 10-fold cross validation, προκειμένου τα αποτελέσματα να είναι συγκρίσιμα με τα αποτελέσματα της δημοσίευσης [11].

#### 3.5.1 Πλήρως διασυνδεδεμένη αρχιτεκτονική

Η πρώτη αρχιτεκτονική που χρησιμοποιήθηκε, καθώς εμφανίζεται γενικότερα στην σχετική βιβλιογραφία και έχει χρησιμοποιηθεί με επιτυχία στο παρελθόν σε αντίστοιχα προβλήματα υπολογιστικής βιολογίας, είναι

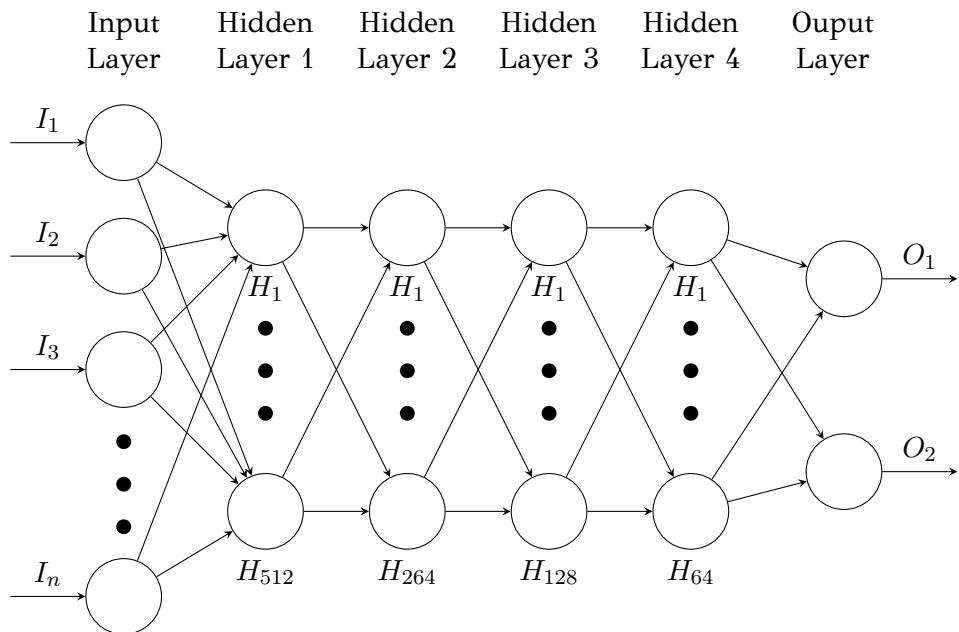


Σχήμα 3.12: 10-fold cross validation

αυτή του πλήρως διασυνδεδεμένου νευρωνικού δικτύου. Υστερα από πολλαπλά πειράματα, η "βέλτιστη" αρχιτεκτονική αποτελείται:

- ένα επίπεδο εισόδου με 1024 νευρώνες,
- τέσσερα κρυφά επίπεδα με 512, 264, 128, 64 νευρώνες αντίστοιχα και
- ένα επίπεδο εξόδου με 2 νευρώνες.

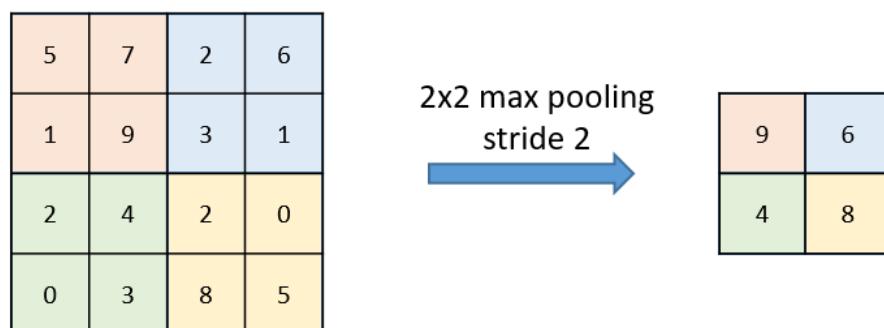
Ως συνάρτηση ενεργοποίησης, για το επίπεδο εισόδου και τα τέσσερα κρυφά επίπεδα χρησιμοποιήθηκε η *ReLU*, ενώ για το επίπεδο εξόδου καθώς το πρόβλημα μας είναι κατηγοριοποίηση δυαδικής μορφής (*binary classification*), χρησιμοποιήθηκε η σιγμοειδής συνάρτηση (*sigmoid*). Κατά την εκπαίδευση χρησιμοποιήθηκε ως βελτιστοποιητής ο *Adam (Adaptive Moment Estimation)* με ρυθμό εκπαίδευσης  $L_r = 0.0001$ . Ακόμη, δοκιμάστηκε η προσθήκη επιπέδων *απόρριψης* (*dropout layers*). Η λειτουργία των επιπέδων απόρριψης συνοφίζεται στην απενεργοποίηση νευρώνων του προηγούμενου επιπέδου, με τον αριθμό των απενεργοποιημένων νευρώνων σε κάθε εποχή να ορίζεται από μια υπερπαράμετρο που ονομάζεται ρυθμός *απόρριψης* (*Dropout rate*). Το επίπεδο απόρριψης εφαρμόζεται για να αποφευχθεί το φαινόμενο του *υπερταιριάσματος* (*overfitting*), κατά το οποίο το νευρωνικό "απομνημονεύει" τα δεδομένα εκπαίδευσης έχοντας εξαιρετική απόδοση σε αυτά αλλά έχει χαμηλή απόδοση σε νέα, "άγνωστα" δεδομένα. Ωστόσο, στην περίπτωση μας η προσθήκη επιπέδων απόρριψης δεν βοήθησε, γεγονός που αποδίδεται στον μικρό αριθμός χαρακτηριστικών (8 χαρακτηριστικά) των δεδομένων εκπαίδευσης.



Σχήμα 3.13: Fully connected αρχιτεκτονική

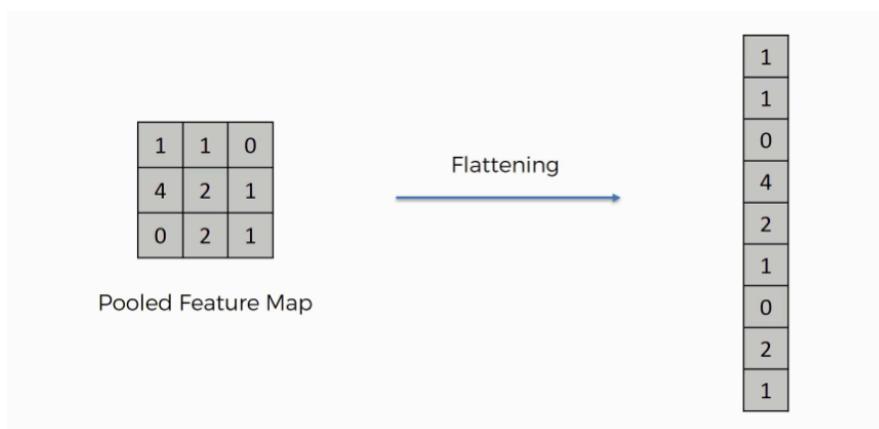
### 3.5.2 Συνελικτική αρχιτεκτονική

Γιατέρα από τα πειράματα που πραγματοποιήθηκαν σε πλήρως διασυνδεδεμένες αρχιτεκτονικές, ακολούθησε η δημιουργία και ο πειραματισμός με συνελικτικές δομές νευρωνικών δικτύων. Τα συνελικτικά νευρωνικά δίκτυα έχουν αναπτυχθεί σημαντικά τα τελευταία χρόνια λόγω της ικανότητάς τους να αναλύουν χωρική πληροφορία. Όσον αφορά την αρχιτεκτονική του συνελικτικού δικτύου, αποτελείται από 6 1 –  $D$  συνελικτικά επίπεδα με 32, 32, 64, 64, 128, 128 φίλτρα αντίστοιχα, με συνάρτηση ενεργοποίησης ReLu, όπου κάθε φίλτρο αποτελείται από ένα παράθυρο  $3 \times 3$ , με κάθε συνελικτικό φίλτρο να ακολουθείται από ένα επίπεδο υποδειγματοληψίας που ονομάζεται Max Pooling, με μέγεθος παραθύρου  $3 \times 3$ .



Σχήμα 3.14: Max Pooling

Κατά τη διαδικασία max pooling, από κάθε παράθυρο επιλέγεται η μέγιστη τιμή ως τιμή ως τιμή αντιπρόσωπος για το επόμενο επίπεδο. Με αυτό τον τρόπο επιτυγχάνεται μείωση της διαστασιμότητας του δικτύου ενώ παράλληλα προγματοποιείται περιορισμός του θορύβου, καθώς οι θορυβώδεις τιμές απορρίπτονται. Στο τέλος του βου συνελικτικού επιπέδου εντοπίζεται ένα επίπεδο εξομάλυνσης (*flatten layer*), που μετατρέπει την έξοδο σε ένα διάνυσμα στήλης. Στη συνέχεια, η έξοδος αυτή τροφοδοτείται σε ένα πλήρως διασυνδεδεμένο νευρωνικό δίκτυο με δύο επίπεδα, με 128 νευρώνες το πρώτο και με 2 νευρώνες το δεύτερο, που αποτελεί και το επίπεδο εξόδου.



Σχήμα 3.15: Flattening

---

---

## ΚΕΦΑΛΑΙΟ 4

---

### ΑΠΟΤΕΛΕΣΜΑΤΑ

Στο κεφάλαιο 3 παρουσιάστηκε αναλυτικά η υλοποίηση της διπλωματικής εργασίας, από την προεπεξεργασία και τροποποίηση των δεδομένων εως την παρουσίαση των αρχιτεκτονικών νευρωνικών δικτύων που χρησιμοποιήθηκαν. Σε αυτό το κεφάλαιο θα παρουσιαστούν τα αποτελέσματα και οι προβλέψεις των νευρωνικών δικτύων και θα πραγματοποιηθεί μια σύγκριση με άλλα υπάρχοντα μοντέλα πρόβλεψης αλληλεπιδράσεων πρωτεΐνων, αφού πρώτα ορισθούν όλες οι μετρικές αξιολόγησης των μοντέλων.

#### 4.1 Αποτελέσματα Εκπαίδευσης

Προκειμένου να αξιολογήσουμε τα αποτελέσματα εκπαίδευσης, πρέπει να καθοριστεί το σύνολο των δεδομένων εκπαίδευσης με βάση το οποίο θα πραγματοποιηθεί η εκπαίδευση. Καθώς τα δεδομένα μας τροποποιήθηκαν από δύο αλγορίθμους, η λογική προσέγγιση θα ήταν η σύγκριση των αποτελεσμάτων εκπαίδευσης για τα δύο σύνολα και η επιλογή του συνόλου που αποδίδει καλύτερα. Ωστόσο, παρατηρώντας το περιεχόμενο των τιμών των δύο συνόλων ύστερα από την επεξεργασία, συμπεραίνουμε ότι έχουν παρόμοιες (σχεδόν πανομοιότυπες) εγγραφές, γεγονός που επιβεβαιώθηκε υπολογίζοντας την απόσταση *Frobenius* των δύο μητρώων μέσω του τύπου:

$$dist(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^I \sum_{j=1}^J (a_{ij} - b_{ij})^2} \approx 0$$

Επομένως, θα μπορούσαμε να επιλέξουμε ένα από τα δύο σύνολα εκπαίδευσης τυχαία για την εκπαίδευση/αξιολόγηση των μοντέλων μας. Ακόμη, παρατηρώντας τους χρόνους σύγκλισης των δύο αλγορίθμων, καταλήγουμε στη χρήση του συνόλου εκπαίδευσης που προέκυψε μέσω

τανυστικής αποδόμησης για την εκπαίδευση των μοντέλων μας, καθώς η διαφορά στην ταχύτητα σύγκλισης καθιστά απαγορευτική την εκτέλεση της παραγοντοποίησης μητρώου μέσω SGD σε σχέση με την τανηστική αποδόμηση.

Χρόνοι Σύγκλισης Μεθόδων	
Μέθοδος	Χρόνος (σε s)
Παραγοντοποίηση μητρώου μέσω SGD	1895.38
CP αποδόμηση	26.04

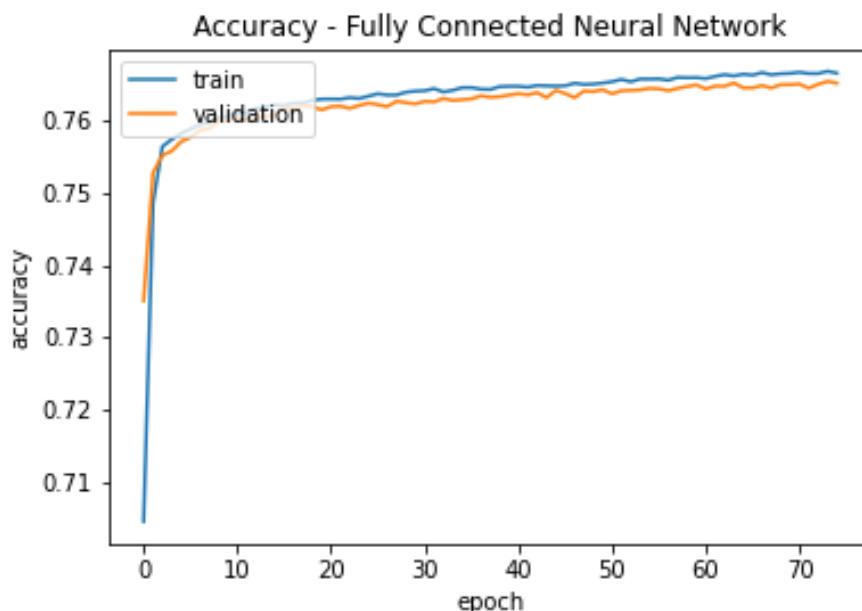
Πίνακας 4.1: Χρόνοι σύγκλισης μεθόδων συμπλήρωσης μητρώων

Όσον αφορά τα χρονικά αποτελέσματα της εκπαίδευσης, εμπειρικά παρατηρείται ότι η εκπαίδευση των πλήρως διασυνδεδεμένων επιπέδων ενός νευρωνικού δικτύου πραγματοποιείται πολύ πιο γρήγορα απ' ότι στην περίπτωση των συνελικτικών επιπέδων, ενώ τα συνελικτικά επίπεδα απαιτούν περισσότερη μνήμη για την αποθήκευση των ενδιάμεσων παραγώγων. Στην περίπτωσή μας, καθώς τα αποτελέσματα που παρουσιάζονται στη συνέχεια προέκυψαν από επικυρωμένη διασταύρωση (cross-validation), παρουσιάζονται οι χρόνοι για την εκπαίδευση μας πτυχής (1 fold), με τα αποτελέσματα να εμφανίζονται στον πίνακα 4.2. Όπως παρατηρούμε, η εκπαίδευση του πλήρως διασυνδεδεμένου δικτύου χρειάστηκε 30 λιγότερες εποχές από αυτήν του συνελικτικού δικτύου, ωστόσο λόγω της γρηγορότερης εκτέλεσης των εποχών του συνελικτικού η εκπαίδευση του χρειάστηκε σχεδόν τον ίδιο χρόνο ( $\approx 20$  λεπτά). Αξίζει να παρατηρηθεί επίσης ότι κάθε εποχή του συνελικτικού μοντέλου ήταν 1.33 φορές πιο γρήγορη από την αντίστοιχη του πλήρως διασυνδεδεμένου. Η συνθήκη σύγκλισης ήταν ίδια και για τα δύο μοντέλα και θεωρούσαμε σύγκλιση όταν δεν παρατηρούταν μεταβολή μικρότερη από  $\delta = 2 \times 10^{-4}$  στο σφάλμα επικύρωσης (validation loss) για ένα διάστημα 10 εποχών.

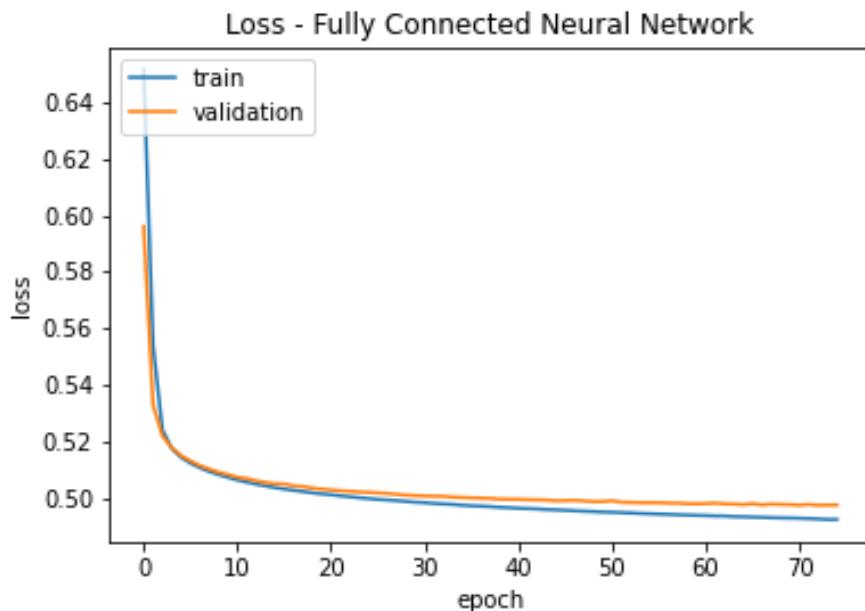
Χρόνοι Σύγκλισης Μοντέλων		
Μοντέλο	Εποχές (ανά εποχή)	Χρόνος (σε s)
Πλήρως Διασυνδεδεμένο	75 (16.82s)	1220.61 (20.3 λεπτά)
Συνελικτικό	105 (12.61s)	1225.28 (20.4 λεπτά)

Πίνακας 4.2: Χρόνοι σύγκλισης νευρωνικών δικτύων

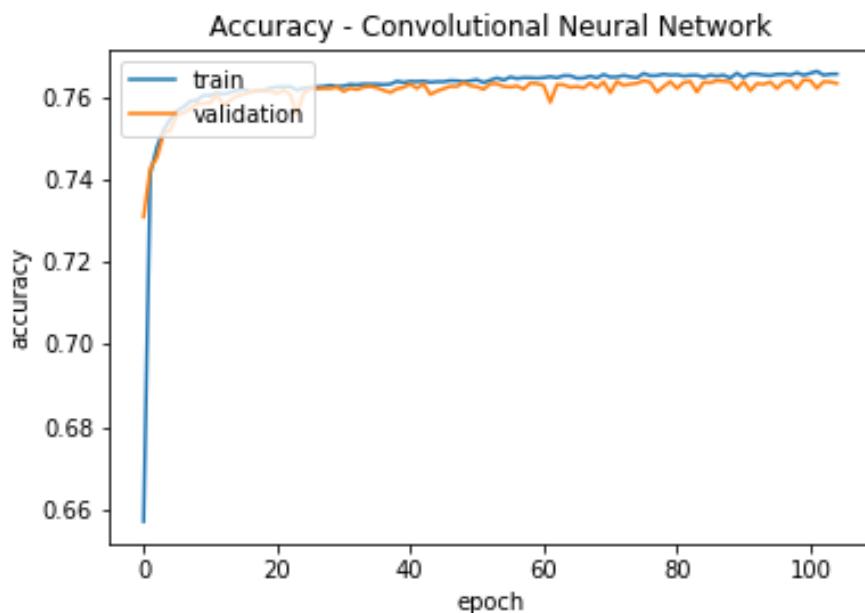
Παρακάτω, παρουσιάζονται οι γραφικές ακρίβειας και απωλειών των νευρωνικών δικτύων. Καθώς τα αποτελέσματα που παρουσιάζονται στη συνέχεια προέκυψαν από επικυρωμένη διασταύρωση (cross-validation), επιλέξαμε να αποτυπώσουμε τις απώλειες και την ακρίβεια ενδεικτικά για μια από τις 10 πτυχές (folds). Φαίνεται από τις γραφικές παραστάσεις ότι το πλήρως διασυνδεδεμένο δίκτυο συγκλίνει αρκετά γρήγορα (75 εποχές) και μαθαίνει αρκετά γρήγορα το σύνολο εκπαίδευσης, κάτι που φαίνεται από την γρήγορη αύξηση της ακρίβειας και την πτώση των απωλειών. Ωστόσο, οι καμπύλες για τα δεδομένα επικύρωσης συγκλίνουν πολύ γρήγορα σε μια τιμή και στη συνέχεια παρουσιάζουν μια ταλαντωτική συμπεριφορά γύρω από αυτή, γεγονός που για να αποφευχθεί μας οδήγησε στην εκπαίδευση του πλήρως διασυνδεδεμένου μοντέλου με πολύ μικρό ρυθμό μάθησης ( $= 0.00001$ ). Αυτό δείχνει την μεγάλη ευαισθησία του πλήρως διασυνδεδεμένου δικτύου, ενώ ταυτόχρονα δεν γενικεύει σε μεγάλο βαθμό. Από την άλλη, το συνελικτικό νευρωνικό δίκτυο απαιτεί περισσότερες εποχές για να συγκλινεί (105 εποχές), ωστόσο συγκλίνει με πιο ομαλό τρόπο ενώ δεν δείχνει να παρουσιάζει την έντονη ταλαντωτική συμπεριφορά του πλήρως διασυνδεδεμένου δικτύου, γι' αυτό και χρησιμοποιήθηκε μεγαλύτερη τιμή για τον ρυθμό μάθησης ( $= 0.0001$ ).



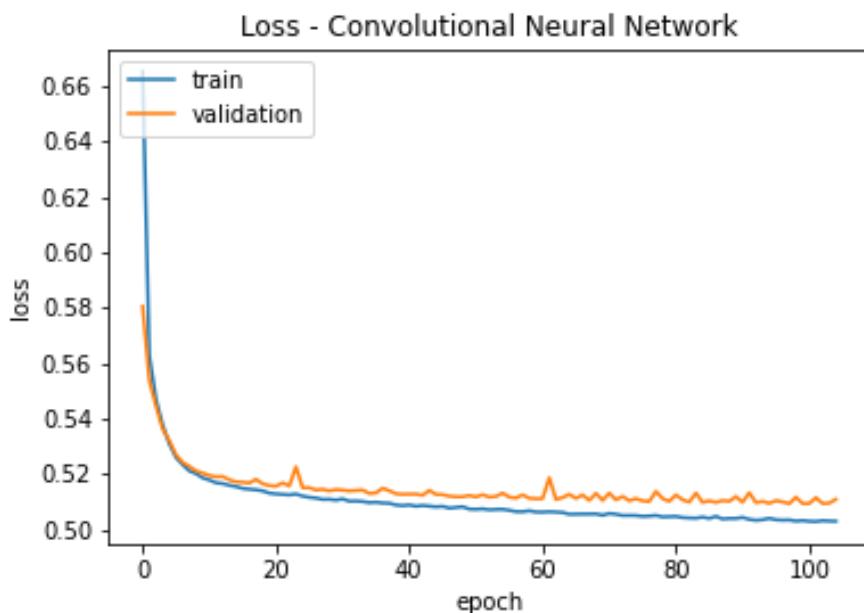
Σχήμα 4.1: Ακρίβεια πλήρως διασυνδεδεμένου νευρωνικού δικτύου



Σχήμα 4.2: Απώλειες πλήρως διασυνδεδεμένου νευρωνικού δικτύου



Σχήμα 4.3: Ακρίβεια συνελικτικού νευρωνικού δικτύου



Σχήμα 4.4: Απώλειες συνελικτικού νευρωνικού δικτύου

Στη συνέχεια θα αναφερθούμε στις μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων μας. Αρχικά, η αντικειμενική συνάρτηση με βάση την οποία βελτιστοποιήθηκαν τα μοντέλα μας ήταν η απώλεια δυαδικής διασταυρωμένης εντροπίας (*binary crossentropy loss*).

$$J(w) = \frac{1}{m} \sum_{i=1}^m [y_i * \log \sigma(w^T x_i) + (1 - y_i) * \log (1 - \sigma(w^T x_i))] \quad (4.1)$$

Για την αξιολόγηση της απόδοσης των μοντέλων και τη σύγκριση των αποτελεσμάτων χρησιμοποιήθηκαν οι ακόλουθες μετρικές:

*Sensitivity*: Μετρική που εκφράζει την ποσότητα των θετικών περιπτώσεων που κατηγοριοποιούνται σωστά. Αποτελεί την σημαντικότερη μετρική όταν η αποφυγή των *false negatives* έχει την ύψιστη σημασία.

*Specificity*: Αντίστοιχη μετρική με το sensitivity, εκφράζει την ποσότητα των αρνητικών περιπτώσεων που κατηγοριοποιούνται σωστά, σημαντικό όταν μας απασχολούν τα *false positives*.

*Precision*: Μετρική που εκφράζει την πιθανότητα μιας θετικής κατηγοριοποίησης να είναι σωστή σε σχέση με όλες τις θετικές περιπτώσεις.

*Accuracy*: Η πιο διαδεδομένη μετρική, εκφράζει την πιθανότητα μια κατηγοριοποίηση να είναι σωστή, ωστόσο δεν είναι πάντοτε αξιόπιστη στην περίπτωση δεδομένων που παρουσιάζουν ανισσοροπία (*imbalanced data sets*).

*F1-score*: Αποτελεί τον αριμονικό μέσο των sensitivity και recall και αποτελεί ένα μέτρο αξιολόγησης της αποδοτικότητας ενός κατηγοριοποιητή, ωστόσο δεν λαμβάνει υπόψιν τις περιπτώσεις των *true negatives*, εστιάζοντας κυρίως στις θετικές κατηγοριοποιήσεις.

*MCC*: Γνωστή ως *Matthew's Correlation Coefficient*, είναι ένα μέτρο συσχέτισης μεταξύ της πραγματικής τιμής και της πρόβλεψης. Λαμβάνει τιμές από  $-1 - 1$ , με το 0 να δηλώνει ότι οι προβλέψεις γίνονται τυχαία.

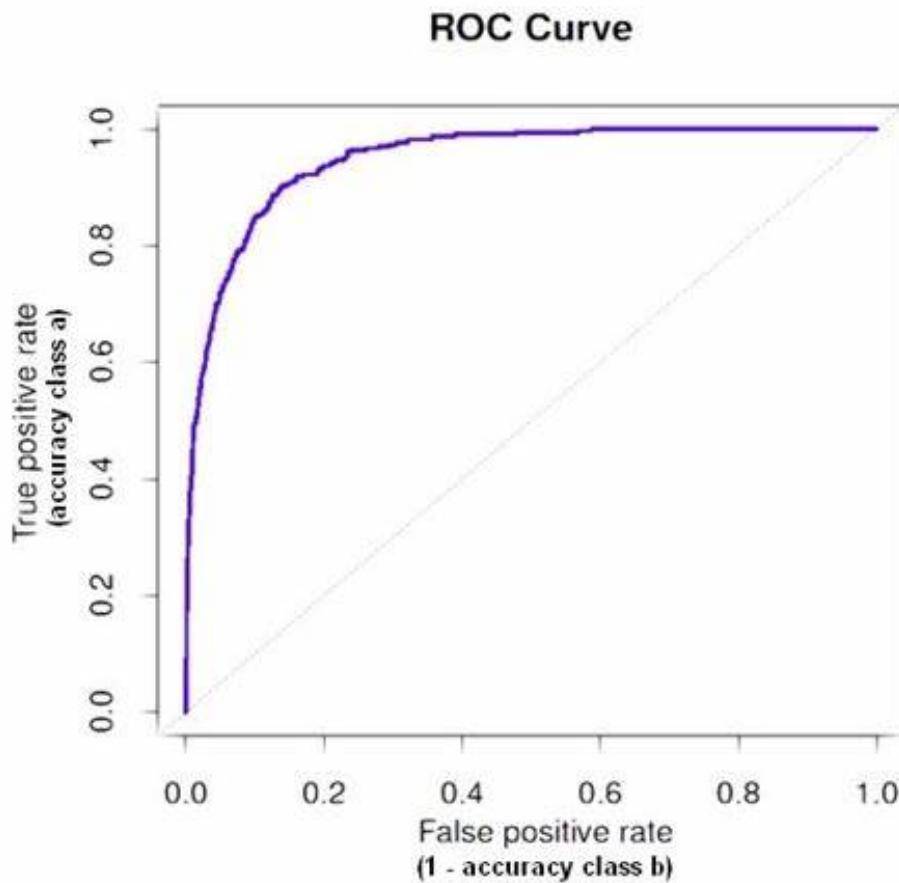
Μετρικές απόδοσης δυαδικής κατηγοριοποίησης		
Μετρική	Τύπος	Εύρος τιμών
Sensitivity	$\frac{TP}{TP+FN}$	[0, 1]
Specificity	$\frac{TN}{TN+FP}$	[0, 1]
Precision	$\frac{TP}{TP+FP}$	[0, 1]
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	[0, 1]
F1	$\frac{2TP}{2TP+FP+FN}$	[0, 1]
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	[-1, 1]

Πίνακας 4.3: Μετρικές απόδοσης δυαδικής κατηγοριοποίησης

Στον πίνακα 4.4 παρουσιάζονται οι μαθηματικοί τύποι που ορίζουν τις παραπάνω μετρικές. Ως *true positive* (*TP*) ορίζεται μια θετική περίπτωση που κατηγοριοποιείται ως θετική, ως *false positive* (*FP*) μια αρνητική περίπτωση που κατηγοριοποιείται ως θετική, ως *true negative* (*TN*) μια αρνητική περίπτωση που κατηγοριοποιείται ως αρνητική ενώ ως *false negative* (*FN*) μια αρνητική περίπτωση που κατηγοριοποιείται ως θετική.

Εχοντας ορίσει τις μετρικές στον πίνακα 4.4, προχωράμε στην εισαγωγή μερικών ακόμη εννοιών για την περαιτέρω κατανόηση των αποτελεσμάτων μας. Η πρώτη έννοια στην οποία θα αναφερθούμε είναι η καμπύλη *AUC-ROC* (*Area Under Curve - Reciever Operation Characteristics*). Η καμπύλη ROC αποτελεί μια μετρική που χρησιμοποιείται σε προβλήματα δυαδικής κατηγοριοποίησης προκειμένου να αξιολογηθεί η ποιότητα της εξόδου του κατηγοριοποιητή. Πρόκειται για μια καμπύλη πιθανότητας για διαφορετικές κατηγορίες, που εκφράζει την ικανότητα του μοντέλου να ξεχωρίσει τις δεδομένες κατηγορίες ως προς τις πιθανότητες πρόβλεψης. Στην εικόνα 4.5 φαίνεται μια τυπική καμπύλη ROC. Στον άξονα X υπάρχει ο ρυθμός *False Positive*, δηλαδή την πιθανότητα μια αρνητική πρόβλεψη να είναι θετική, ενώ στον άξονα Y εχουμε τον ρυθμό *True Positive*, δηλαδή την

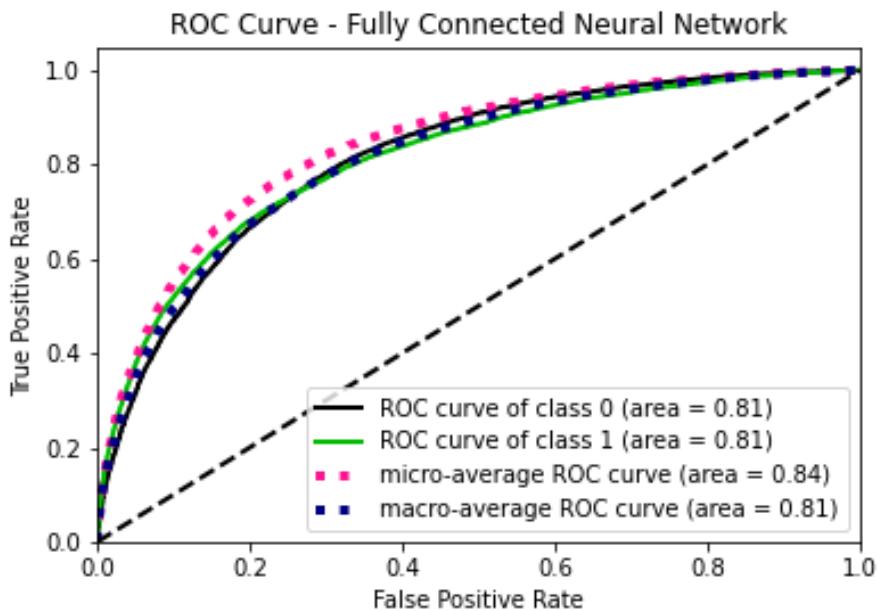
πιθανότητα μια θετική πρόβλεψη να είναι θετική. Εναλλακτικά, ο ρυθμός *False Positive* αντιστοιχεί στην τιμή 1-Specificity ενώ ο ρυθμός *True Positive* αντιστοιχεί στην τιμή του Sensitivity.



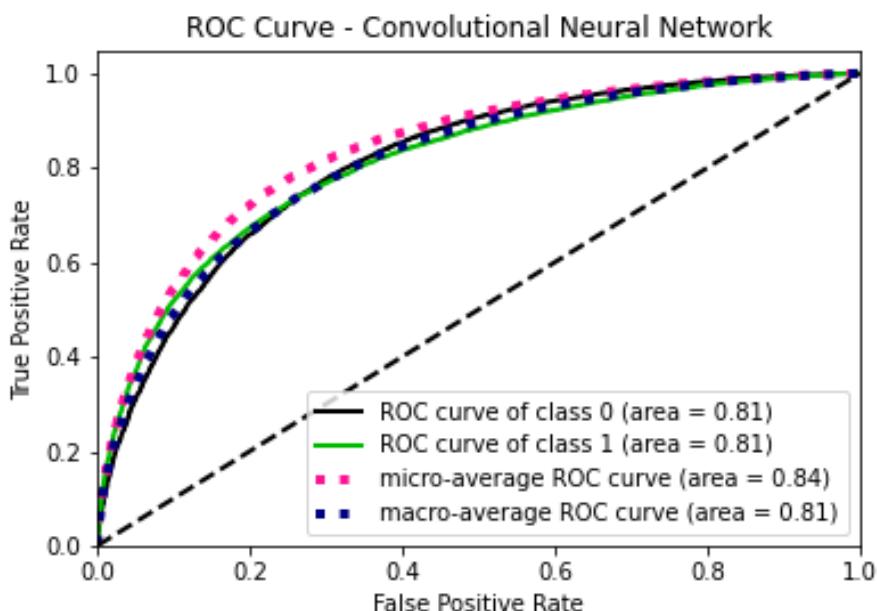
Σχήμα 4.5: Τυπική Καμπύλη ROC

Σχετικά με το περιεχόμενο της εικόνας 4.5, ενα μοντέλο που προβλέπει τυχαία παρουσιάζει καμπύλη ROC που θα μοιάζει με την διαγώνια γκρι γραμμή και δεν έχει ικανότητα διάκρισης. Αντίθετα, όσο απομακρύνεται η καμπύλη από την διαγώνια γραμμή, τόσο καλύτερο είναι το μοντέλο στο να ξεχωρίζει τις θετικές περιπτώσεις από τις αρνητικές. Με βάση την παραπάνω καμπύλη, μπορούμε να εξάγουμε ορισμένες χρήσιμες μετρικές, όπως το σκορ AUC. Οπως αναφέρει και το όνομα του, το AUC υπολογίζεται ως το εμβαδόν που βρίσκεται κάτω από την καμπύλη ROC και έχει αρκετές εφιληγαντές. Συνοπτικά, είναι "η πιθανότητα ο κατηγοριοποιητής να κατατάξει ένα τυχαίο θετικό δείγμα υψηλότερα από ένα τυχαίο αρνητικό δείγμα" [90]. Το σκόρο AUC για έναν κατηγοριοποιητή που κάνει τυχαίες προβλέψεις (εμβαδόν κάτω από την γκρι διαγώνιο) θα είναι 0.5, ενώ όσο αυξάνεται πλησιάζοντας την μονάδα δηλώνει και καλύτερη ικανότητα πρόβλεψης. Παρακάτω παρατίθενται οι καμπύλες AUC-ROC για

κάθε μοντέλο που αναπτύξαμε, με την επιπλέον προσθήκη όπου θεωρούμε διαδοχικά κάθε κατηγορία ως την "θετική" κατηγορία.



Σχήμα 4.6: Καμπύλη ROC πλήρως διασυνδεδεμένου νευρωνικού δικτύου



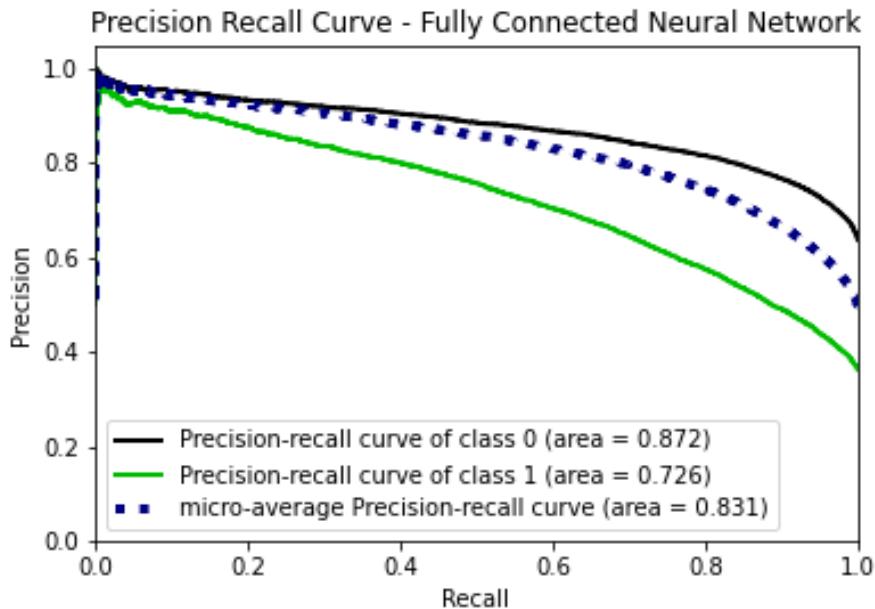
Σχήμα 4.7: Καμπύλη ROC συνελικτικού νευρωνικού δικτύου

Από τα αποτελέσματα παρατηρούμε ότι και οι δύο αρχιτεκτονικές

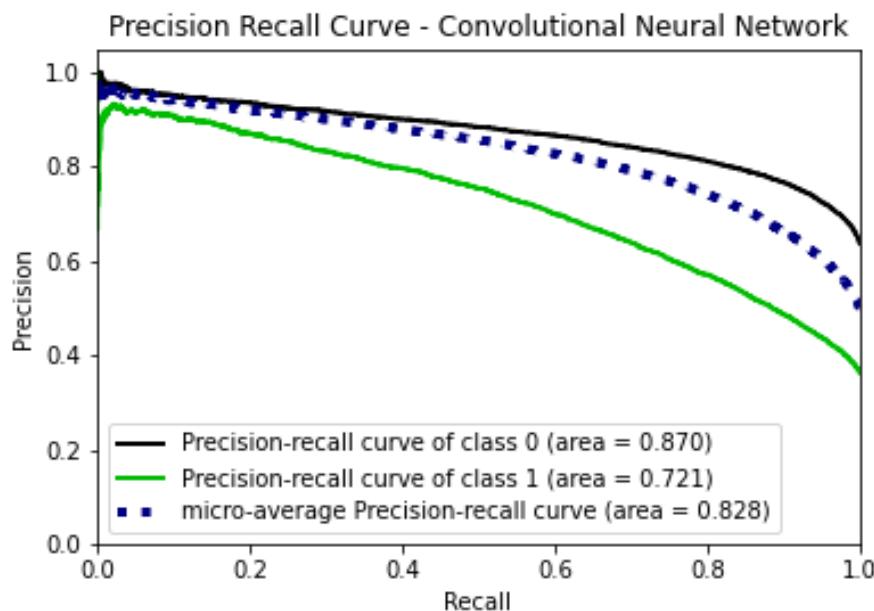
παρέχουν αρκετά ικανοποιητικά αποτελέσματα όσον αφορά τις προβλέψεις. Συγκεκριμένα, και για τις δύο αρχιτεκτονικές υπολογίζουμε:

$AUC_{DNN} = 0.8148$  και  $AUC_{CNN} = 0.8121$ . Οι τιμές βρίσκονται αρκετά κοντά μεταξύ τους για να μπορούμε να ξεχωρίσουμε κάποια αρχιτεκτονική απλώς παρατηρούμε ότι η πλήρως διασυνδεδεμένη αρχιτεκτονική αποδίδει ελαφρώς καλύτερα από την συνελικτική όσον αφορά τη συγκεκριμένη μετρική.

Άλλη μια γραφική που μας βοηθάει στην κατανόηση της αποδοτικότητας των μοντέλων μας είναι η καμπύλη ακρίβειας ανάκλησης (*Precision Recall curve*). Η διαφορά με την καμπύλη ROC έγκειται στο γεγονός ότι η καμπύλη ακρίβειας ανάκλησης εστιάζει στην κατηγορία που ορίζεται ως κατηγορία μειονότητας, δεν λαμβάνει υπόψιν τις αρνητικές κατηγοριοποιήσεις και μας παρέχει χρήσιμη πληροφορία στην περίπτωση που δουλεύουμε με ανισόρροπα (*imbalanced*) σύνολα εκπαίδευσης. Στον άξονα Y βρίσκεται το precision, που εκφράζει την πιθανότητα μιας θετικής σωστής κατηγοριοποίησης ως προς το σύνολο των θετικών περιπτώσεων, ενώ στον άξονα X το recall, που εκφράζει την πιθανότητα μιας θετικής κατηγοριοποίησης ως προς τον αριθμό των πιθανών θετικών κατηγοριοποιήσεων που θα μπορούσαν να συμβούν. Είναι επιθυμητό τα μοντέλα μας να έχουν ταυτόχρονα υψηλή ακρίβεια και υψηλή ανάκληση, ωστόσο στην πραγματικότητα υπάρχει ενα trade-off μεταξύ των δυο.



Σχήμα 4.8: Καμπύλη Precision Recall πλήρως διασυνδεδεμένου νευρωνικού δικτύου



Σχήμα 4.9: Καμπύλη Precision Recall συνελικτικού νευρωνικού δικτύου

Παρατηρούμε ότι και στην περίπτωση της καμπύλης ακρίβειας-ανάκλησης, το πλήρως διασυνδεδεμένο νευρωνικό δίκτυο έχει οριακά καλύτερα αποτελέσματα από το συνελικτικό νευρωνικό δίκτυο. Στην περίπτωσή μας, η καμπύλες που μας απασχολούν είναι κυρίως οι καμπύλες που αφορούν την κατηγορίας 1, δηλαδή αυτές που ορίζουν ως θετική περίπτωση την ύπαρξη αλληλεπίδρασης.

Στα αποτελέσματα της δημοσίευσης [11] με την οποία συγχρίνουμε τα αποτελέσματά μας, λόγω της ύπαρξης *Nan* τιμών στα χαρακτηριστικά του αρχικού συνόλου δεδομένων (ιδίως στα χαρακτηριστικά *Conservation Scores*), παρατίθενται από τους συγγραφείς της δημοσίευσης τα αποτελέσματα για διαφορετικές περιπτώσεις εγγραφών. Στον παρακάτω πίνακα για τα αποτελέσματα των νευρωνικών δικτύων και των random forests επιλέχθηκαν τα αποτελέσματα όπου λαμβάνονται υπόψιν όλα τα χαρακτηριστικά. Στην περίπτωση που οι τιμές των χαρακτηριστικών είναι μη-έγκυρες, για τα νευρωνικά δίκτυα επιλέχθηκε μια τιμή που βασίζεται στον μέσο όρο της κατανομής των δεδομένων, ενώ για τα random forests χρησιμοποιήθηκε η μέθοδος *fractional instances*, κατά την οποία εγγραφές που περιέχουν ελλιπείς τιμές διαιρούνται σε πολλαπλές τιμές με διαφορετικά βάρη.

Σύγκριση μεθόδων						
Method	ACC	PREC	SPEC	SENS	F	MCC
Neural Network [11]	0.735	0.653	0.892	0.415	0.507	0.355
Random Forest [11]	<b>0.760</b>	0.679	<b>0.906</b>	0.439	0.533	0.398
CP + Fully Connected NN	0.756	0.756	0.756	0.756	0.756	0.512
CP + Convolutional NN	0.758	<b>0.758</b>	0.758	<b>0.758</b>	<b>0.758</b>	<b>0.516</b>

Πίνακας 4.4: Αποτελέσματα - Σύγκριση Μεθόδων

Τα παραπάνω αποτελέσματα αποδεικνύουν την βελτίωση της απόδοσης του κατηγοριοποιητή μέσω της συμπλήρωσης των ελλειπών τιμών και της χρήσης νευρωνικών δικτύων. Ειδικότερα, λαμβάνοντας υπόψιν τα αποτελέσματα του καλύτερου μοντέλου μας (απόδομηση ταυνυστών και συνελικτική δομή νευρωνικών δικτύων), ενώ η απόδοση των νευρωνικών δικτύων που αναπτύξαμε είναι παρόμοια με αυτή της δημοσίευσης [11], η σωστή κατηγοριοποίηση των κομματιών επιφανείας (αύξηση του sensitivity κατά 72.66% σε σχέση με τα random forests και κατά 82.65% σε σχέση με την αρχιτεκτονική νευρωνικών δικτύων) καθώς και η συνολική αποδοτικότητα του κατηγοριοποιητή μας είναι σημαντικά βελτιωμένη. Χρησιμοποιώντας τον MCC ως κύρια μετρική αξιολόγησης, όπως πραγματοποιείται και στην δημοσίευση των Northe et al. και καθώς θεωρείται η καλύτερη μετρική αξιολόγησης ενός δυαδικού κατηγοριοποιητή σε θέματα υπολογιστικής βιολογίας, πετύχαμε αύξηση κατά 29.65% σε σχέση με την αρχιτεκτονική των random forests και κατά 45.35% σε σχέση με την αρχιτεκτονική νευρωνικών δικτύων [11].

## 4.2 Μελλοντικές Εξελίξεις

Στην ενότητα αυτή παρουσιάζονται διάφορες ιδέες σχετικά με την επέκταση της παρούσας διπλωματικής. Αρχικά, η πιο απλή διαδικασία για την βελτίωση των αποτελεσμάτων προέρχεται από την εκπαίδευση των μοντέλων μηχανικής μάθησης με ολοένα και περισσότερα δεδομένα και η αξιολόγηση τους σε ανεξάρτητα σύνολα αξιολόγησης, επομένως ως πρώτο βήμα θα μπορούσαν να εξαχθούν νέα δεδομένα από την πρωτεϊνική βάση δεδομένων PDB και να κατασκευαστούν νέα σύνολα δεδομένων εκπαίδευσης και αξιολόγησης με πιο σύγχρονα δεδομένα.

Παράλληλα, όσον αφορά το σύνολο των δεδομένων εκπαίδευσης, θα μπορούσαν να προστεθούν περισσότερα χαρακτηριστικά, τόσο δομικά όσο και ακολουθιακά, για την εξαγωγή επιπλέον πληροφορίας σχετικά με τις αλληλεπιδράσεις μεταξύ πρωτεϊνών. Μερικά χαρακτηριστικά που παρουσιάζουν θετικά αποτελέσματα είναι οι φυσικοχημικές ιδιότητες των αμινοξέων (*physicochemical properties of amino acids*) [5] και το σκορ αταξίας (*disorder score*) [91].

Όσον αφορά την προεπεξεργασία των δεδομένων, οι μέθοδοι συμπλήρωσης μητρώου λειτουργούν ιδανικά με μεγάλες δομές δεδομένων, γεγονός που δεν συμβαίνει στην περίπτωσή μας όπου έχουμε μόνο 8 χαρακτηριστικά. Ωστόσο, η ήδη υπάρχουσα απόδοση των αλγορίθμων θα μπορούσε να συγκριθεί με διαφορετικές μεθόδους. Μια σκέψη περίλαμβάνει την αξιολόγηση της τανηστικής αποδόμησης με πειραματισμούς όσον αφορά την τιμή των  $\beta_k$  3.4 κατά την εφαρμογή των μεθόδων μη γραμμικής συζηγγούς παραγώγου (ncg). Αντίστοιχα, θα μπορούσε να ελεγχθεί η αύξηση της ταχύτητας σύγκλισης του αλγορίθμου παραγοντοποίησης μητρώου μέσω SGD εφόσον υλοποιηθεί με κατανεμημένο τρόπο μέσω της τροποποιημένης μορφής του SGD που ονομάζεται *Stratified Stochastic Gradient Descent (SSGD)* [76].

---

## ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] V. S. Rao, K. Srinivas, G. N. Sujini, and G. N. S. Kumar, “Protein-protein interaction detection: Methods and analysis,” *International Journal of Proteomics*, vol. 2014, pp. 1–12, 2014.
- [2] Z.-J. Shi and J. Guo, “A new algorithm of nonlinear conjugate gradient method with strong convergence,” *Computational Applied Mathematics*, vol. 27, pp. 93 – 106, 00 2008.
- [3] J. Bradford and D. Westhead, “Improved prediction of protein-protein binding sites using a support vector machines approach.,” *Bioinformatics (Oxford, England)*, vol. 21, pp. 1487–94, 05 2005.
- [4] M. Zhang, Q. Su, Y. Lu, M. Zhao, and B. Niu, “Application of machine learning approaches for protein-protein interactions prediction,” *Medicinal chemistry (Shariqah (United Arab Emirates))*, vol. 13, 05 2017.
- [5] Z. Xie, X. Deng, and K. Shu, “Prediction of protein–protein interaction sites using convolutional neural network and improved data sets,” *International Journal of Molecular Sciences*, vol. 21, p. 467, 01 2020.
- [6] H. Neuvirth, R. Raz, and G. Schreiber, “Promate: A structure based prediction program to identify the location of protein–protein binding sites,” *Journal of Molecular Biology*, vol. 338, no. 1, pp. 181 – 199, 2004.
- [7] L. Terveen and W. Hill, “Beyond recommender systems: Helping people help each other,” 02 2001.
- [8] F. H. Stephenson, “Protein,” in *Calculations for Molecular Biology and Biotechnology*, pp. 375–429, Elsevier, 2016.
- [9] A. Blanco and G. Blanco, “Proteins,” in *Medical Biochemistry*, pp. 21–71, Elsevier, 2017.
- [10] M. Yanagida, “Functional proteomics current achievements,” *Journal of Chromatography B*, vol. 771, pp. 89–106, may 2002.

- [11] T. C. Northe, A. Barešić, and A. C. R. Martin, “IntPred: a structure-based predictor of protein–protein interaction sites,” *Bioinformatics*, vol. 34, pp. 223–229, sep 2017.
- [12] J. D. L. Rivas and C. Fontanillo, “Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks,” *PLoS Computational Biology*, vol. 6, p. e1000807, jun 2010.
- [13] I. M. Nooren, “NEW EMBO MEMBER’s REVIEW: Diversity of protein–protein interactions,” *The EMBO Journal*, vol. 22, pp. 3486–3492, jul 2003.
- [14] K. Miura, “An overview of current methods to confirm protein–protein interactions,” *Protein & Peptide Letters*, vol. 25, pp. 728–733, oct 2018.
- [15] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, “Functional organization of the yeast proteome by systematic analysis of protein complexes,” *Nature*, vol. 415, pp. 141–147, jan 2002.
- [16] M. Urh, D. Simpson, and K. Zhao, “Chapter 26 affinity chromatography,” in *Methods in Enzymology*, pp. 417–438, Elsevier, 2009.
- [17] E. Golemis, *Protein-protein interactions : a molecular cloning manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2002.
- [18] G. MacBeath and S. Schreiber, “Printing proteins as microarrays for high-throughput function determination,” *Science (New York, N.Y.)*, vol. 289, p. 1760—1763, September 2000.
- [19] S. W. Michnick, P. H. Ear, C. Landry, M. K. Malleshaiah, and V. Messier, “Protein-fragment complementation assays for large-scale analysis, functional dissection and dynamic studies of protein–protein interactions in living cells,” in *Methods in Molecular Biology*, pp. 395–425, Humana Press, 2011.
- [20] G. Smith, “Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface,” *Science*, vol. 228, pp. 1315–1317, jun 1985.
- [21] A. H. Y. Tong, “Systematic genetic analysis with ordered arrays of yeast deletion mutants,” *Science*, vol. 294, pp. 2364–2368, dec 2001.
- [22] M. R. O’Connell, R. Gamsjaeger, and J. P. Mackay, “The structural analysis of protein–protein interactions by NMR spectroscopy,” *PROTEOMICS*, vol. 9, pp. 5224–5232, dec 2009.
- [23] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-

- Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, “A comprehensive analysis of protein–protein interactions in *saccharomyces cerevisiae*,” *Nature*, vol. 403, pp. 623–627, feb 2000.
- [24] M. Schneider, J. R. Johnson, N. J. Krogan, and S. K. Chanda, “The virus–host interactome,” in *Viral Pathogenesis*, pp. 157–167, Elsevier, 2016.
- [25] S. L. Ooi, X. Pan, B. D. Peyser, P. Ye, P. B. Meluh, D. S. Yuan, R. A. Irizarry, J. S. Bader, F. A. Spencer, and J. D. Boeke, “Global synthetic-lethality analysis and yeast functional profiling,” *Trends in Genetics*, vol. 22, pp. 56–63, jan 2006.
- [26] A. Zhang, *Protein interaction networks : computational analysis*. Cambridge New York: Cambridge University Press, 2009.
- [27] R. Hosur, J. Xu, J. Bienkowska, and B. Berger, “iWRAP: An interface threading approach with application to prediction of cancer-related protein–protein interactions,” *Journal of Molecular Biology*, vol. 405, pp. 1295–1310, feb 2011.
- [28] S.-A. Lee, C. hsiung Chan, C.-H. Tsai, J.-M. Lai, F.-S. Wang, C.-Y. Kao, and C.-Y. F. Huang, “Ortholog-based protein-protein interaction prediction and its application to inter-species interactions,” *BMC Bioinformatics*, vol. 9, no. Suppl 12, p. S11, 2008.
- [29] J. Wojcik and V. Schachter, “Protein-protein interaction map inference using interacting domain profile pairs,” *Bioinformatics*, vol. 17, pp. S296–S305, jun 2001.
- [30] M. Yamada, M. S. Kabir, and R. Tsunedomi, “Divergent promoter organization may be a preferred structure for gene control in *Escherichia coli*,” *Journal of Molecular Microbiology and Biotechnology*, vol. 6, no. 3-4, pp. 206–210, 2003.
- [31] A. J. Enright, I. Iliopoulos, N. C. Kyriides, and C. A. Ouzounis, “Protein interaction maps for complete genomes based on gene fusion events,” *Nature*, vol. 402, pp. 86–90, nov 1999.
- [32] E. M. Marcotte, “Detecting protein function and protein-protein interactions from genome sequences,” *Science*, vol. 285, pp. 751–753, jul 1999.
- [33] F. Pazos and A. Valencia, “In silico two-hybrid system for the selection of physically interacting protein pairs,” *Proteins: Structure, Function, and Genetics*, vol. 47, pp. 219–227, mar 2002.
- [34] T. Sato, Y. Yamanishi, M. Kanehisa, and H. Toh, “The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships,” *Bioinformatics*, vol. 21, pp. 3482–3489, jun 2005.
- [35] R. A. Craig and L. Liao, “Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance

- matrices,” *BMC Bioinformatics*, vol. 8, jan 2007.
- [36] K. Srinivas, “Methodology for phylogenetic tree construction,” *Journal of Proteomics & Bioinformatics*, vol. s1, 2008.
  - [37] T.-W. Lin, J.-W. Wu, and D. T.-H. Chang, “Combining phylogenetic profiling-based and machine learning-based techniques to predict functional related proteins,” *PLoS ONE*, vol. 8, p. e75940, sep 2013.
  - [38] A. Grigoriev, “A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast *saccharomyces cerevisiae*,” *Nucleic Acids Research*, vol. 29, pp. 3513–3519, sep 2001.
  - [39] I. Xenarios and D. Eisenberg, “Protein interaction databases,” *Current Opinion in Biotechnology*, vol. 12, pp. 334–339, aug 2001.
  - [40] C. Stark, “BioGRID: a general repository for interaction datasets,” *Nucleic Acids Research*, vol. 34, pp. D535–D539, jan 2006.
  - [41] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, “MINT: a molecular INTeraction database,” *FEBS Letters*, vol. 513, pp. 135–140, dec 2001.
  - [42] K. Han, B. Park, H. Kim, J. Hong, and J. Park, “HPID: The human protein interaction database,” *Bioinformatics*, vol. 20, pp. 2466–2470, apr 2004.
  - [43] H. Hermjakob, “IntAct: an open source molecular interaction database,” *Nucleic Acids Research*, vol. 32, pp. 452D–455, jan 2004.
  - [44] A. Patil, K. Nakai, and H. Nakamura, “HitPredict: a database of quality assessed protein–protein interactions in nine species,” *Nucleic Acids Research*, vol. 39, pp. D744–D749, oct 2010.
  - [45] C. Prieto and J. D. L. Rivas, “APID: Agile protein interaction DataAnalyzer,” *Nucleic Acids Research*, vol. 34, pp. W298–W302, jul 2006.
  - [46] and Stephen K Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. D. Costanzo, C. Christie, J. M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D. S. Goodsell, R. K. Green, V. Gurjanovic, D. Guzenko, B. P. Hudson, Y. Liang, R. Lowe, E. Peisach, I. Periskova, C. Randle, A. Rose, M. Sekharan, C. Shao, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Young, C. Zardecki, M. Zhuravleva, G. Kurisu, H. Nakamura, Y. Kengaku, H. Cho, J. Sato, J. Y. Kim, Y. Ikegawa, A. Nakagawa, R. Yamashita, T. Kudou, G.-J. Bekker, H. Suzuki, T. Iwata, M. Yokochi, N. Kobayashi, T. Fujiwara, S. Velankar, G. J. Kleywegt, S. Anyango, D. R. Armstrong, J. M. Berrisford, M. J. Conroy, J. M. Dana, M. Deshpande, P. Gane, R. Gáborová, D. Gupta, A. Gutmanas, J. Koča, L. Mak, S. Mir, A. Mukhopadhyay, N. Nadzirin, S. Nair, A. Patwardhan, T. Paysan-Lafosse, L. Pravda, O. Salih, D. Sehnal, M. Varadi, R. Vařeková, J. L. Markley, J. C. Hoch, P. R. Romero, K. Baskaran, D. Maziuk, E. L. Ulrich, J. R. Wedell, H. Yao,

- M. Livny, and Y. E. Ioannidis, “Protein data bank: the single global archive for 3d macromolecular structure data,” *Nucleic Acids Research*, vol. 47, pp. D520–D528, oct 2018.
- [47] A. L. Samuel, “Some studies in machine learning using the game of checkers,” *IBM Journal of Research and Development*, vol. 3, pp. 210–229, jul 1959.
- [48] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, dec 1943.
- [49] D. O. Hebb, *The organization of behavior; a neuropsychological theory, (by) D.O. Hebb. Science Editions*. New York: John Wiley and Sons, 1967.
- [50] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, jan 1967.
- [51] S. Linnainmaa, “Taylor expansion of the accumulated rounding error,” *BIT*, vol. 16, pp. 146–160, jun 1976.
- [52] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, IEEE Comput. Soc. Press.
- [53] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, sep 1995.
- [54] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, pp. 1735–1780, nov 1997.
- [55] C. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.
- [56] A. G. B. Richard S. Sutton, *Reinforcement Learning*. The MIT Press, 2018.
- [57] Y.-Y. Chen, Y.-H. Lin, C.-C. Kung, M.-H. Chung, and I.-H. Yen, “Design and implementation of cloud analytics-assisted smart power meters considering advanced artificial intelligence as edge analytics in demand-side management for smart homes,” *Sensors*, vol. 19, p. 2047, may 2019.
- [58] B. Farley and W. Clark, “Simulation of self-organizing systems by digital computer,” *Transactions of the IRE Professional Group on Information Theory*, vol. 4, pp. 76–84, sep 1954.
- [59] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, (Cambridge, MA, USA), p. 2672–2680, MIT Press, 2014.
- [60] S. Fletcher and A. D. Hamilton, “Targeting protein–protein interactions by rational design: mimicry of protein surfaces,” *Journal of The Royal Society Interface*, vol. 3, pp. 215–233, mar 2006.

- [61] E. Krissinel and K. Henrick, “Inference of macromolecular assemblies from crystalline state,” *Journal of Molecular Biology*, vol. 372, pp. 774–797, sep 2007.
- [62] C. T. Porter and A. C. Martin, “BiopLib and BiopTools—a c programming library and toolset for manipulating protein structure,” *Bioinformatics*, p. btv482, aug 2015.
- [63] S. Jones and J. M. Thornton, “Analysis of protein-protein interaction sites using surface patches 1 1edited by g.von heijne,” *Journal of Molecular Biology*, vol. 272, pp. 121–132, sep 1997.
- [64] T. J. Magliery and L. Regan *BMC Bioinformatics*, vol. 6, no. 1, p. 240, 2005.
- [65] J. Kyte and R. F. Doolittle, “A simple method for displaying the hydrophobic character of a protein,” *Journal of Molecular Biology*, vol. 157, pp. 105–132, may 1982.
- [66] H.-X. Zhou and Y. Shan, “Prediction of protein interaction sites from sequence profile and residue neighbor list,” *Proteins: Structure, Function, and Genetics*, vol. 44, no. 3, pp. 336–343, 2001.
- [67] A. C. R. Martin, “Mapping PDB chains to UniProtKB entries,” *Bioinformatics*, vol. 21, pp. 4297–4301, sep 2005.
- [68] L. E. McMillan and A. C. Martin, “Automatically extracting functionally equivalent proteins from SwissProt,” *BMC Bioinformatics*, vol. 9, oct 2008.
- [69] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of Molecular Biology*, vol. 215, pp. 403–410, oct 1990.
- [70] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Research*, vol. 32, pp. 1792–1797, mar 2004.
- [71] W. Valdar and J. Thornton, “Protein–protein interfaces: Analysis of amino acid conservation in homodimers,” 02 2001.
- [72] B. Hazes and B. W. Dijkstra, “Model building of disulfide bonds in proteins with known three-dimensional structure,” *”Protein Engineering, Design and Selection”*, vol. 2, no. 2, pp. 119–125, 1988.
- [73] E. Baker and R. Hubbard, “Hydrogen bonding in globular proteins,” *Progress in Biophysics and Molecular Biology*, vol. 44, pp. 97–179, jan 1984.
- [74] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, pp. 2577–2637, dec 1983.
- [75] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,”
- [76] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis, “Large-scale

- matrix factorization with distributed stochastic gradient descent,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’11*, ACM Press, 2011.
- [77] L. Bottou and O. Bousquet, “The tradeoffs of large scale learning.,” vol. 20, 01 2007.
  - [78] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” *SIAM Review*, vol. 51, pp. 455–500, 2009.
  - [79] S. Liu and O. TRENKLER, “Hadamard, khatri-rao, kronecker and other matrix products,” *International Journal of Information Systems Sciences*, vol. 4, 01 2008.
  - [80] F. L. Hitchcock, “The expression of a tensor or a polyadic as a sum of products,” *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927.
  - [81] R. Cattell, ““Parallel proportional profiles” and other principles for determining the choice of factors by rotation,” *Psychometrika*, vol. 9, pp. 267–283, December 1944.
  - [82] J. D. Carroll and J.-J. Chang, “Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition,” *Psychometrika*, vol. 35, pp. 283–319, 1970.
  - [83] R. A. Harshman, “Determination and proof of minimum uniqueness conditions for parafac1,” *UCLA working papers in phonetics*, vol. 22, no. 111-117, p. 3, 1972.
  - [84] T. Papastergiou, E. I. Zacharaki, and V. Megalooikonomou, “Tensmil2: Improved multiple instance classification through tensor decomposition and instance selection,” in *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2019.
  - [85] T. Papastergiou, E. I. Zacharaki, and V. Megalooikonomou, “Tensor decomposition for multiple-instance classification of high-order medical data,” *Complexity*, vol. 2018, p. 8651930, Dec 2018.
  - [86] J. Håstad, *Tensor rank is NP-complete*, vol. 11, pp. 451–460. 04 2006.
  - [87] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, “Scalable tensor factorizations for incomplete data,” *Chemometrics and Intelligent Laboratory Systems*, vol. 106, p. 41–56, Mar 2011.
  - [88] T. Papastergiou and V. Megalooikonomou, “A distributed proximal gradient descent method for tensor completion,” in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 2056–2065, 2017.
  - [89] S. Sanmatías and E. Vercher, “A generalized conjugate gradient algorithm,” *Journal of Optimization Theory and Applications*, vol. 98, no. 2, pp. 489–502, 1998.
  - [90] D. J. Hand, “Measuring classifier performance: a coherent alternative to the area under the roc curve,” *Machine Learning*, vol. 77, pp. 103–123, 2009.

- [91] B.-Q. Li, K.-Y. Feng, L. Chen, T. Huang, and Y.-D. Cai, “Prediction of protein-protein interaction sites by random forest algorithm with mrmr and ifs,” *PLOS ONE*, vol. 7, pp. 1–10, 08 2012.