



ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΙΩΑΝΝΙΝΩΝ

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ**  
**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**  
**ΠΜΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΔΙΚΤΥΩΝ**

**ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

**ΠΡΟΒΛΕΨΗ ΑΛΛΗΛΕΠΙΔΡΑΣΕΩΝ ΠΡΩΤΕΪΝΩΝ ΑΠΟ**  
**ΑΛΛΗΛΟΥΧΙΕΣ ΑΜΙΝΟΞΕΩΝ ΜΕ ΧΡΗΣΗ ΒΑΘΙΑΣ**  
**ΜΑΘΗΣΗΣ**

Δημήτριος Μητρογιώργος

Επιβλέπων: Πέτρος Καρβέλης  
Αναπληρωτής Καθηγητής

Άρτα, Ιούλιος, 2026

**ΠΡΟΒΛΕΨΗ ΑΛΛΗΛΕΠΙΔΡΑΣΕΩΝ ΠΡΩΤΕΪΝΩΝ ΑΠΟ  
ΑΛΛΗΛΟΥΧΙΕΣ ΑΜΙΝΟΞΕΩΝ ΜΕ ΧΡΗΣΗ ΒΑΘΙΑΣ  
ΜΑΘΗΣΗΣ**

# **Εγκρίθηκε από τριμελή εξεταστική επιτροπή**

Τόπος, Ημερομηνία

## **ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ**

- Επιβλέπων καθηγητής  
Πέτρος Καρβέλης,  
Αναπληρωτής καθηγητης
- Μέλος επιτροπής  
Χρυσόστομος Στύλιος,  
Καθηγητης
- Μέλος επιτροπής  
Ιωάννης Τσούλος,  
Καθηγητης

## Δήλωση μη λογοκλοπής

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

Μητρογιώργος, Δημήτριος

Υπογραφή \_\_\_\_\_

## ΠΕΡΙΛΗΨΗ

Η πρόβλεψη αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης (ΑΠΠ) αποτελεί κρίσιμο πρόβλημα της βιοπληροφορικής, με άμεσες εφαρμογές στην κατανόηση κυτταρικών μηχανισμών και στην ανακάλυψη νέων θεραπευτικών στόχων. Στην παρούσα εργασία διερευνάται η δυνατότητα πρόβλεψης ΑΠΠ αποκλειστικά από αλληλουχίες αμινοξέων, αξιοποιώντας αναπαραστάσεις που παράγονται από το προ-εκπαιδευμένο γλωσσικό μοντέλο πρωτεϊνών ProtT5. Συγκεντρώθηκε σύνολο δεδομένων άνω των 330.000 ζευγών πρωτεϊνών από επτά διαφορετικά είδη, με βαθμολογίες αλληλεπίδρασης προερχόμενες από τη βάση δεδομένων STRING. Αναπτύχθηκαν και συγκρίθηκαν δύο αρχιτεκτονικές βαθιάς μάθησης: ένα Αντίληπτρο Βαθιάς Μάθησης, το οποίο συνδυάζει τις αναπαραστάσεις των δύο πρωτεϊνών μέσω του γινομένου Hadamard και της απόλυτης διαφοράς τους, και ένας Σιαμέζικος Ταξινομητής με κοινόχρηστο κωδικοποιητή, ο οποίος προβάλλει τις δύο πρωτεΐνες σε έναν κοινό λανθάνοντα χώρο πριν από τη σύγκρισή τους. Η αξιολόγηση των μοντέλων πραγματοποιήθηκε σε τρία πειραματικά σενάρια: ανά είδος ξεχωριστά, στο σύνολο του ενοποιημένου συνόλου δεδομένων, και σε είδη που απουσίαζαν πλήρως από την εκπαίδευση. Ο Σιαμέζικος Ταξινομητής υπερτερούσε σταθερά στα δύο πρώτα σενάρια, πετυχαίνοντας ακρίβεια 96% και F1 87,22% στο ενοποιημένο σύνολο δεδομένων, ενώ στο σενάριο γενίκευσης σε άγνωστα είδη παρουσίασε υψηλότερο AUC και ανάκληση, υποδεικνύοντας ισχυρή και μεταφέρσιμη διακριτική ικανότητα. Τα ευρήματα επιβεβαιώνουν ότι οι αναπαραστάσεις ProtT5 φέρουν επαρκή βιολογική πληροφορία για αξιόπιστη πρόβλεψη ΑΠΠ, ακόμη και χωρίς ρητή μοντελοποίηση της τρισδιάστατης δομής των πρωτεϊνών.

**Λέξεις-κλειδιά:** πρόβλεψη αλληλεπιδράσεων πρωτεϊνών, βαθιά μάθηση, σιαμέζικο νευρωνικό δίκτυο, ProtT5, αναπαραστάσεις αλληλουχιών, βάση STRING, βιοπληροφορική

# ABSTRACT

Protein-protein interaction (PPI) prediction is a fundamental problem in bioinformatics, with direct implications for understanding cellular mechanisms and identifying novel therapeutic targets. This thesis investigates the feasibility of predicting PPIs directly from amino acid sequences, using representations generated by the pre-trained protein language model ProtT5. A dataset of over 330,000 protein pairs was compiled from seven different species, with interaction confidence scores obtained from the STRING database. Two deep learning architectures were developed and compared: a Deep Learning Perceptron, which combines the two protein representations through a Hadamard product and their absolute difference, and a Siamese Classifier with a shared encoder, which projects both proteins into a common latent space prior to their comparison. Model evaluation was carried out under three experimental scenarios: per-species cross-validation, cross-validation on the unified dataset, and generalization to species entirely absent from training. The Siamese Classifier consistently outperformed the perceptron in the first two scenarios, achieving 96% accuracy and an F1-score of 87.22% on the unified dataset, while in the cross-species generalization scenario it exhibited higher AUC and recall, indicating strong and transferable discriminative ability. These findings confirm that ProtT5 embeddings encode sufficient biological information to enable reliable PPI prediction, even without explicit modeling of protein three-dimensional structure.

**Keywords:** protein-protein interaction prediction, deep learning, Siamese neural network, ProtT5, sequence embeddings, cancer, STRING database, bioinformatics

# ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

|                                                              |           |
|--------------------------------------------------------------|-----------|
| <b>Περίληψη</b>                                              | <b>4</b>  |
| <b>Abstract</b>                                              | <b>5</b>  |
| <b>Πίνακας Περιεχομένων</b>                                  | <b>7</b>  |
| <b>Κατάλογος Πινάκων</b>                                     | <b>8</b>  |
| <b>Κατάλογος Διαγραμμάτων/Εικόνων</b>                        | <b>9</b>  |
| <b>Πίνακας Συντομογραφιών</b>                                | <b>10</b> |
| <b>1 Εισαγωγή</b>                                            | <b>11</b> |
| 1.1 Πληροφορικό Υπόβαθρο . . . . .                           | 12        |
| 1.1.1 Τεχνητά Νευρωνικά Δίκτυα και Οπισθοδιάδοση . . . . .   | 12        |
| 1.1.2 Σιαμέζικα Νευρωνικά Δίκτυα . . . . .                   | 13        |
| 1.1.3 Ανάλυση Κύριων Συνιστωσών . . . . .                    | 14        |
| 1.1.4 Συνημιτονική Ομοιότητα . . . . .                       | 14        |
| 1.1.5 Αλγόριθμοι Βελτιστοποίησης: Adam και AdamW . . . . .   | 15        |
| 1.1.6 Τεχνικές Κανονικοποίησης και Αποφυγής Υπερπροσαρμογής  | 16        |
| 1.1.7 Μετασχηματιστές . . . . .                              | 17        |
| 1.2 Βιολογικό Υπόβαθρο . . . . .                             | 19        |
| 1.2.1 Αναδίπλωση Πρωτεϊνών . . . . .                         | 19        |
| 1.2.2 Πρωτεϊνικές αλληλεπιδράσεις . . . . .                  | 21        |
| 1.3 Στόχοι και Κίνητρο της Εργασίας . . . . .                | 23        |
| <b>2 Βιβλιογραφική Ανασκόπηση</b>                            | <b>24</b> |
| <b>3 Συλλογή Δεδομένων</b>                                   | <b>29</b> |
| 3.1 Πρωτεϊνικές αναπαραστάσεις . . . . .                     | 31        |
| 3.2 Πιθανότητες αλληλεπίδρασης . . . . .                     | 35        |
| <b>4 Τα Μοντέλα</b>                                          | <b>36</b> |
| 4.1 Αντίληπτο Βαθιάς Μάθησης . . . . .                       | 37        |
| 4.2 Σιαμέζικος Ταξινομητής . . . . .                         | 40        |
| <b>5 Αποτελέσματα</b>                                        | <b>42</b> |
| 5.1 Δοκιμή μοντέλων σε κάθε είδος ξεχωριστά . . . . .        | 43        |
| 5.2 Δοκιμή μοντέλων σε όλο το σύνολο των δεδομένων . . . . . | 47        |
| 5.3 Δοκιμή μοντέλων σε ξένα είδη οργανισμών . . . . .        | 52        |

|          |                   |           |
|----------|-------------------|-----------|
| <b>6</b> | <b>Συζήτηση</b>   | <b>59</b> |
| <b>7</b> | <b>Συμπέρασμα</b> | <b>62</b> |

## ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

|   |                                                                                                                                                          |    |
|---|----------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1 | Επιδόσεις ανά είδος για το Αντίληπτρο Βαθιάς Μάθησης (ABM) και τον Σιαμέζικο Ταξινομητή (ΣΤ) (μέσος όρος πενταπλής διασταυρωμένης επικύρωσης) . . . . .  | 44 |
| 2 | Ποσοστό ορθής ταξινόμησης ανά είδος στο ενοποιημένο σύνολο δεδομένων, για το Αντίληπτρο Βαθιάς Μάθησης (ABM) και τον Σιαμέζικο Ταξινομητή (ΣΤ) . . . . . | 50 |
| 3 | Οι επιδόσεις των μοντέλων σε ολόκληρο το σύνολο δεδομένων . . . .                                                                                        | 52 |
| 4 | Ποσοστό ορθής ταξινόμησης ανά ξένο είδος, για το Αντίληπτρο Βαθιάς Μάθησης (ABM) και τον Σιαμέζικο Ταξινομητή (ΣΤ) . . . . .                             | 57 |
| 5 | Οι επιδόσεις των μοντέλων σε ξένα είδη οργανισμών . . . . .                                                                                              | 59 |

# ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ/ΕΙΚΟΝΩΝ

|      |                                                                                                            |    |
|------|------------------------------------------------------------------------------------------------------------|----|
| 1.1  | Η αρχιτεκτονική ενός μετασχηματιστή με κωδικοποιητή και αποκωδικοποιητή (πηγή: Vaswani et al.) . . . . .   | 18 |
| 1.2  | Η αιτιολογική αλληλουχία από την αλυσίδα αμινοξέων μέχρι την αλληλεπίδραση της πρωτεΐνης . . . . .         | 20 |
| 1.3  | Ένα δίκτυο αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης (πηγή: db-string.org) . . . . .                             | 22 |
| 2.1  | Η αρχιτεκτονική του μοντέλου των Yao et al. (2019) . . . . .                                               | 26 |
| 2.2  | Η αρχιτεκτονική του μοντέλου ISPRED-SEQ των Manfredi et al. (2023) . . . . .                               | 28 |
| 3.1  | Λεπτομερής σύνθεση του σετ δεδομένων . . . . .                                                             | 31 |
| 3.2  | Η διαδικασία υπολογισμού των αναπαραστάσεων για τις πρωτεΐνες (πηγή: Elnaggar et al. 2025) . . . . .       | 33 |
| 3.3  | Διάγραμμα διασποράς με διαφορετικό χρώμα για κάθε είδος . . . . .                                          | 35 |
| 4.1  | Η βασική αρχιτεκτονική του Αντίληπτρου Βαθιάς Μάθησης . . . . .                                            | 39 |
| 4.2  | Η βασική αρχιτεκτονική του Σιαμέζικου Ταξινομητή . . . . .                                                 | 40 |
| 4.3  | Η αρχιτεκτονική του κοινόχρηστου κωδικοποιητή . . . . .                                                    | 41 |
| 5.1  | Απόδοση του Αντίληπτρου Βαθιάς Μάθησης ανά είδος, ανά μετρική (πενταπλή διασταυρωμένη επικύρωση) . . . . . | 45 |
| 5.2  | Απόδοση του Σιαμέζικου Ταξινομητή ανά είδος, ανά μετρική (πενταπλή διασταυρωμένη επικύρωση) . . . . .      | 45 |
| 5.3  | Πίνακας σύγκρισης του Αντίληπτρου Βαθιάς Μάθησης στο ενοποιημένο σύνολο δεδομένων . . . . .                | 48 |
| 5.4  | Πίνακας σύγκρισης του Σιαμέζικου Ταξινομητή στο ενοποιημένο σύνολο δεδομένων . . . . .                     | 49 |
| 5.5  | Καμπύλη ROC του Αντίληπτρου Βαθιάς Μάθησης στο ενοποιημένο σύνολο δεδομένων . . . . .                      | 50 |
| 5.6  | Καμπύλη ROC του Σιαμέζικου Ταξινομητή στο ενοποιημένο σύνολο δεδομένων . . . . .                           | 51 |
| 5.7  | Πίνακας σύγκρισης του Αντίληπτρου Βαθιάς Μάθησης στα ξένα είδη οργανισμών . . . . .                        | 53 |
| 5.8  | Πίνακας σύγκρισης του Σιαμέζικου Ταξινομητή στα ξένα είδη οργανισμών . . . . .                             | 54 |
| 5.9  | Ποσοστό ορθής ταξινόμησης ανά ξένο είδος για το Αντίληπτρο Βαθιάς Μάθησης . . . . .                        | 55 |
| 5.10 | Ποσοστό ορθής ταξινόμησης ανά ξένο είδος για τον Σιαμέζικο Ταξινομητή . . . . .                            | 56 |
| 5.11 | Καμπύλη ROC του Αντίληπτρου Βαθιάς Μάθησης στα ξένα είδη οργανισμών . . . . .                              | 57 |
| 5.12 | Καμπύλη ROC του Σιαμέζικου Ταξινομητή στα ξένα είδη οργανισμών . . . . .                                   | 58 |

## ΠΙΝΑΚΑΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

|      |       |                                   |
|------|-------|-----------------------------------|
| ΑΠΠ  | ..... | Αλληλεπίδραση Πρωτεΐνης-Πρωτεΐνης |
| ΘΠΑ  | ..... | Θετική Προγνωστική Αξία           |
| ΜΑΑ  | ..... | Μέση Απόλυτη Απόκλιση             |
| ΡΜΤΑ | ..... | Ρίζα Μέσης Τετραγωνικής Απόκλισης |
| ΤΝΔ  | ..... | Τεχνητό Νευρωνικό Δίκτυο          |

# 1 Εισαγωγή

Η μηχανική μάθηση αποτελεί έναν από τους πιο δυναμικά αναπτυσσόμενους κλάδους της επιστήμης υπολογιστών, με εφαρμογές που εκτείνονται σε ένα ευρύ φάσμα επιστημονικών και τεχνολογικών πεδίων. Στον πυρήνα της βρίσκεται η ιδέα ότι ένα υπολογιστικό σύστημα μπορεί να «μαθαίνει» από δεδομένα, εντοπίζοντας μοτίβα και κανονικότητες χωρίς να απαιτείται η ρητή προγραμματιστική κωδικοποίηση κάθε κανόνα [1]. Η προσέγγιση αυτή αντιπροσωπεύει μια θεμελιώδη αλλαγή στον τρόπο με τον οποίο επιλύονται πολύπλοκα προβλήματα: αντί να σχεδιάζεται ένας εξαντλητικός αλγόριθμος που καλύπτει κάθε πιθανή περίπτωση, το μοντέλο εκπαιδεύεται σε παραδείγματα και γενικεύει τη γνώση που αποκτά σε νέες, αόρατες καταστάσεις.

Η βαθιά μάθηση, ως η πιο εξελιγμένη και ευρέως χρησιμοποιούμενη υποκατηγορία της μηχανικής μάθησης, βασίζεται στη χρήση πολυεπίπεδων νευρωνικών δικτύων που είναι ικανά να μαθαίνουν ιεραρχικές αναπαραστάσεις από ακατέργαστα δεδομένα [2]. Η ικανότητα αυτή αποδείχθηκε ιδιαίτερα σημαντική σε τομείς όπου η χειροκίνητη εξαγωγή χαρακτηριστικών είναι δύσκολη ή αδύνατη, όπως η αναγνώριση εικόνων, η επεξεργασία φυσικής γλώσσας και, πιο πρόσφατα, η βιολογική έρευνα. Η επιτυχία της βαθιάς μάθησης οφείλεται σε τρεις αλληλένδετους παράγοντες: τη διαθεσιμότητα τεραστίων ποσοτήτων δεδομένων, την εκθετική αύξηση της υπολογιστικής ισχύος μέσω εξειδικευμένων επεξεργαστών γραφικών, και την ανάπτυξη νέων αλγορίθμων εκπαίδευσης και αρχιτεκτονικών που επιτρέπουν την αποτελεσματική βελτιστοποίηση δικτύων με δισεκατομμύρια παραμέτρους.

Η βιοπληροφορική αποτελεί έναν από τους πιο γόνιμους τομείς εφαρμογής της μηχανικής μάθησης, καθώς η σύγχρονη βιολογική έρευνα παράγει ογκώδη σύνολα δεδομένων που υπερβαίνουν κατά πολύ τις δυνατότητες παραδοσιακής ανάλυσης. Αλληλουχίες γονιδιώματος, δεδομένα γονιδιακής έκφρασης, τρισδιάστατες δομές πρωτεϊνών και δίκτυα βιολογικών αλληλεπιδράσεων αποτελούν παραδείγματα δεδομένων υψηλής διαστατικότητας και πολυπλοκότητας, για τα οποία η μηχανική μάθηση έχει αναδειχθεί ως απαραίτητο εργαλείο [3]. Η σύγκλιση μηχανικής μάθησης και βιοπληροφορικής έχει οδηγήσει σε σειρά επιστημονικών επιτευγμάτων, από την πρόβλεψη της τρισδιάστατης δομής πρωτεϊνών με το AlphaFold έως την ανακάλυψη νέων φαρμακευτικών μορίων και την ανάλυση δεδομένων μονοκύτταρης γονιδιωματικής [4].

Στο πλαίσιο της βιοπληροφορικής, η πρόβλεψη αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης αντιπροσωπεύει ένα πρόβλημα ιδιαίτερης δυσκολίας και σημασίας. Ο χώρος των πιθανών ζευγών πρωτεϊνών είναι τεράστιος — μόνο στον ανθρώπινο οργανισμό εκτιμάται ότι υπάρχουν δεκάδες χιλιάδες λειτουργικές αλληλεπιδράσεις — ενώ η πειραματική τεκμηρίωσή τους είναι εξαιρετικά χρονοβόρα και δαπανηρή. Η μηχανική μάθηση, και ειδικότερα η βαθιά μάθηση, προσφέρει μια ελκυστική

εναλλακτική: αξιοποιώντας τις αλληλουχίες αμινοξέων ως είσοδο και τα υπάρχοντα πειραματικά δεδομένα ως εκπαιδευτικό σήμα, μπορεί να εκμάθει τις υποκείμενες βιολογικές αρχές που διέπουν τις αλληλεπιδράσεις και να τις εφαρμόσει σε κλίμακα που υπερβαίνει κατά πολύ τις δυνατότητες της πειραματικής έρευνας. Η παρούσα εργασία εντάσσεται ακριβώς σε αυτό το πλαίσιο, διερευνώντας την αποτελεσματικότητα διαφορετικών αρχιτεκτονικών βαθιάς μάθησης για την πρόβλεψη αλληλεπιδράσεων πρωτεϊνών από αλληλουχίες αμινοξέων.

## 1.1 Πληροφορικό Υπόβαθρο

Για την ανάπτυξη και κατανόηση των προτεινόμενων μοντέλων μηχανικής μάθησης, απαιτείται η εξοικείωση με τις μαθηματικές και υπολογιστικές δομές που τα θεμελιώνουν. Στην ενότητα αυτή παρουσιάζονται τα βασικά εργαλεία που αξιοποιούνται στην εργασία: η αρχιτεκτονική και ο αλγόριθμος εκπαίδευσης των τεχνητών νευρωνικών δικτύων, η δομή και τα πλεονεκτήματα των σιαμέζικων δικτύων για την επεξεργασία ζευγών εισόδου, οι αλγόριθμοι βελτιστοποίησης που χρησιμοποιούνται κατά την εκπαίδευση, οι τεχνικές κανονικοποίησης για την αποφυγή υπερπροσαρμογής, καθώς και η αρχιτεκτονική των μετασχηματιστών που αποτελεί τη βάση του γλωσσικού μοντέλου ProtT5.

### 1.1.1 Τεχνητά Νευρωνικά Δίκτυα και Οπισθοδιάδοση

Τα τεχνητά νευρωνικά δίκτυα (ΤΝΔ) αποτελούν υπολογιστικά μοντέλα εμπνευσμένα από τη δομή του βιολογικού νευρικού συστήματος [5]. Αποτελούνται από επίπεδα νευρώνων — το επίπεδο εισόδου, ένα ή περισσότερα κρυφά επίπεδα, και το επίπεδο εξόδου — που συνδέονται μέσω παραμετροποιημένων ακμών με βάρη  $w$  και μεροληψίες  $b$ . Κάθε νευρώνας  $j$  στο επίπεδο  $\ell$  υπολογίζει μια γραμμική συνδυαστική τιμή και στη συνέχεια εφαρμόζει μια μη γραμμική συνάρτηση ενεργοποίησης  $\sigma$ :

$$z_j^{(\ell)} = \sum_i w_{ij}^{(\ell)} \cdot a_i^{(\ell-1)} + b_j^{(\ell)}$$

$$a_j^{(\ell)} = \sigma(z_j^{(\ell)})$$

όπου  $a_i^{(\ell-1)}$  είναι η ενεργοποίηση του νευρώνα  $i$  στο προηγούμενο επίπεδο. Σε μορφή πινάκων, για ολόκληρο το επίπεδο  $\ell$  η σχέση γράφεται:

$$a^{(\ell)} = \sigma(W^{(\ell)} a^{(\ell-1)} + b^{(\ell)})$$

Η εκπαίδευση ενός ΤΝΔ στοχεύει στην ελαχιστοποίηση μιας συνάρτησης κόστους  $L(\theta)$  ως προς τις παραμέτρους  $\theta = W, b$  μέσω κατηγορηματικής κλίσης. Ο αλγόριθμος οπισθοδιάδοσης αποτελεί την πρακτική εφαρμογή του κανόνα αλυσίδας για τον υπολογισμό των μερικών παραγώγων της συνάρτησης κόστους ως προς κάθε παράμετρο [5]. Ορίζεται το σφάλμα  $\delta^{(\ell)}$  για κάθε επίπεδο  $\ell$  ως:

$$\begin{aligned}\delta^{(L)} &= \nabla_a L \odot \sigma' \left( z^{(L)} \right) \\ \delta^{(\ell)} &= \left( W^{(\ell+1)\top} \delta^{(\ell+1)} \right) \odot \sigma' \left( z^{(\ell)} \right)\end{aligned}$$

όπου  $\odot$  συμβολίζει το στοιχειομετρικό γινόμενο Hadamard και  $\sigma'$  την παράγωγο της συνάρτησης ενεργοποίησης. Οι κλίσεις για τα βάρη και τις μεροληψίες υπολογίζονται τότε ως:

$$\begin{aligned}\delta^{(\ell)} &= \left( W^{(\ell+1)\top} \delta^{(\ell+1)} \right) \odot \sigma' \left( z^{(\ell)} \right) \\ \theta &\leftarrow \theta - \eta \nabla_{\theta} L(\theta)\end{aligned}$$

Η ενημέρωση των παραμέτρων με ρυθμό μάθησης  $\eta$  γίνεται ως:

$$\theta \leftarrow \theta - \eta \nabla_{\theta} L(\theta)$$

### 1.1.2 Σιαμέζικα Νευρωνικά Δίκτυα

Τα σιαμέζικα νευρωνικά δίκτυα αποτελούν μια ειδική κατηγορία αρχιτεκτονικής που σχεδιάστηκε για προβλήματα σύγκρισης ζευγών εισόδων [6]. Αποτελούνται από δύο ή περισσότερα υποδίκτυα με πανομοιότυπες αρχιτεκτονικές που μοιράζονται ακριβώς τα ίδια βάρη  $\theta$ , επεξεργαζόμενα παράλληλα δύο εισόδους  $x_1$  και  $x_2$ . Τυπικά, κάθε υποδίκτυο  $f$  παράγει μια αναπαράσταση για την αντίστοιχη είσοδο:

$$h_1 = f_{\theta}(x_1), \quad h_2 = f_{\theta}(x_2)$$

Η κοινή χρήση βαρών εξασφαλίζει δύο θεμελιώδεις ιδιότητες: (α) η μετατροπή των εισόδων σε αναπαραστάσεις ακολουθεί τους ίδιους κανόνες, επιτρέποντας άμεση σύγκριση στον ίδιο λανθάνοντα χώρο, και (β) μειώνεται ο αριθμός των εκπαιδευσιμων παραμέτρων περίπου στο μισό σε σχέση με δύο ανεξάρτητα δίκτυα. Κρίσιμη παράμετρος είναι η επιλογή της συνάρτησης συνδυασμού  $g$  που παράγει μια ενιαία αναπαράσταση από τις δύο αναπαραστάσεις. Για να διατηρηθεί η συμμετρία ως προς τη σειρά εισόδου — δηλαδή  $g(h_1, h_2) = g(h_2, h_1)$  — επιλέγονται συμμετρικές πράξεις. Στην παρούσα εργασία χρησιμοποιούνται δύο τέτοιες πράξεις σε συνδυασμό:

$$\begin{aligned}c_{\text{mul}} &= h_1 \odot h_2 \\ c_{\text{diff}} &= |h_1 - h_2| \\ c &= [c_{\text{mul}}; c_{\text{diff}}] \in \mathbb{R}^{2d}\end{aligned}$$

Το γινόμενο Hadamard εντοπίζει τα κοινά ισχυρά χαρακτηριστικά μεταξύ των δύο αναπαραστάσεων, ενώ η απόλυτη διαφορά αποτυπώνει τις αποκλίσεις τους. Η

ταυτόχρονη χρήση τους μειώνει την αμφισημία που θα προέκυπτε από μία μόνο πράξη, καθώς για παράδειγμα  $h_1 \odot h_2 = h'_1 \odot h'_2$  δεν συνεπάγεται  $|h_1 - h_2| = |h'_1 - h'_2|$ .

### 1.1.3 Ανάλυση Κύριων Συνιστωσών

Η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis, PCA) είναι μία γραμμική μέθοδος μείωσης της διάστασης, η οποία μετασχηματίζει τις αρχικές μεταβλητές σε ένα νέο σύνολο ασυσχέτιστων μεταβλητών, τις κύριες συνιστώσες[48]. Έστω  $X \in \mathbb{R}^{n \times p}$  ο πίνακας δεδομένων και  $X_c$  ο κεντραρισμένος πίνακας, ο οποίος προκύπτει αφαιρώντας από κάθε μεταβλητή τη μέση τιμή της. Ο πίνακας συνδιακύμανσης των δεδομένων ορίζεται ως

$$S = \frac{1}{n-1} X_c^T X_c.$$

Οι κύριες συνιστώσες προκύπτουν από την ιδιοανάλυση του  $S$ ,

$$S \mathbf{v}_k = \lambda_k \mathbf{v}_k,$$

όπου  $\mathbf{v}_k$  είναι το  $k$ -οστό ιδιοδιάνυσμα και  $\lambda_k$  η αντίστοιχη ιδιοτιμή. Οι ιδιοτιμές ταξινομούνται κατά φθίνουσα σειρά,  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ , και εκφράζουν τη διακύμανση που εξηγείται από την αντίστοιχη κύρια συνιστώσα. Η πρώτη κύρια συνιστώσα είναι η κατεύθυνση  $\mathbf{v}_1$  που μεγιστοποιεί τη διακύμανση της προβολής των δεδομένων υπό τον περιορισμό  $\|\mathbf{v}_1\| = 1$ ,

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|=1} \mathbf{v}^T S \mathbf{v}.$$

Οι προβολές των δεδομένων στον νέο χώρο υπολογίζονται ως

$$Z = X_c V,$$

όπου  $V = [\mathbf{v}_1, \dots, \mathbf{v}_p]$ . Η αναλογία της συνολικής διακύμανσης που εξηγείται από τις πρώτες  $m$  κύριες συνιστώσες δίνεται από

$$\text{EVR}_m = \frac{\sum_{k=1}^m \lambda_k}{\sum_{k=1}^p \lambda_k},$$

και χρησιμοποιείται για την επιλογή του αριθμού των συνιστωσών που διατηρούνται κατά τη μείωση της διάστασης.

### 1.1.4 Συνημιτονική Ομοιότητα

Η συνημιτονική ομοιότητα αποτελεί ένα ευρέως χρησιμοποιούμενο μέτρο ομοιότητας μεταξύ διανυσμάτων σε υψηλοδιάστατους χώρους, και αξιοποιείται εκτενώς στην επεξεργασία φυσικής γλώσσας, στην ανάκτηση πληροφορίας και στη βιοπληροφορική [7]. Σε αντίθεση με άλλα μέτρα απόστασης, όπως η ευκλείδεια απόσταση, η συνημιτονική ομοιότητα δεν λαμβάνει υπόψη το μέγεθος των διανυσμάτων αλλά αποκλειστικά τη γωνία που σχηματίζουν μεταξύ τους στον διανυσματικό χώρο. Για δύο διανύσματα  $h_1, h_2 \in \mathbb{R}^d$ , η συνημιτονική ομοιότητα ορίζεται ως:

$$\cos\_sim(h_1, h_2) = \frac{h_1 \cdot h_2}{\|h_1\| \cdot \|h_2\|} = \frac{\sum_{i=1}^d h_{1i} \cdot h_{2i}}{\sqrt{\sum_{i=1}^d h_{1i}^2} \cdot \sqrt{\sum_{i=1}^d h_{2i}^2}} \quad (1)$$

όπου  $h_1 \cdot h_2$  είναι το εσωτερικό γινόμενο των δύο διανυσμάτων και  $\|h_1\|, \|h_2\|$  οι αντίστοιχες ευκλείδειες νόρμες τους. Η τιμή της συνημιτονικής ομοιότητας ανήκει στο διάστημα  $[-1, 1]$ : τιμή 1 υποδηλώνει ότι τα δύο διανύσματα είναι παράλληλα και έχουν την ίδια κατεύθυνση, τιμή 0 ότι είναι κάθετα μεταξύ τους, και τιμή  $-1$  ότι είναι αντιπαράλληλα. Η ιδιότητα αυτή την καθιστά ιδιαίτερα κατάλληλη για σύγκριση διανυσμάτων σε υψηλοδιάστατους χώρους, όπου οι ευκλείδειες αποστάσεις τείνουν να συγκλίνουν και να χάνουν την εκφραστικότητά τους λόγω της λεγόμενης κατάρας της διαστατικότητας. Μια σημαντική ιδιότητα της συνημιτονικής ομοιότητας είναι η αναλλοίωσή της ως προς την κλίμακα των διανυσμάτων: η πολλαπλασιασμός ενός διανύσματος με οποιαδήποτε θετική σταθερά δεν μεταβάλλει την τιμή της. Τυπικά, για  $\lambda > 0$ :

$$\cos\_sim(\lambda h_1, h_2) = \cos\_sim(h_1, h_2) \quad (2)$$

Επιπλέον, η συνημιτονική ομοιότητα είναι εκ κατασκευής συμμετρική, δηλαδή  $\cos\_sim(h_1, h_2) = \cos\_sim(h_2, h_1)$ , ιδιότητα που την καθιστά ιδανική για την αξιολόγηση συμμετρικών σχέσεων όπως η ομοιότητα δύο πρωτεϊνικών αναπαραστάσεων. Προκειμένου οι τιμές της συνημιτονικής ομοιότητας να είναι σε μορφή ποσοστού, εφαρμόζεται ο γραμμικός μετασχηματισμός:

$$\frac{\cos\_sim(h_1, h_2) + 1}{2} \quad (3)$$

### 1.1.5 Αλγόριθμοι Βελτιστοποίησης: Adam και AdamW

Ο αλγόριθμος Adam αποτελεί επέκταση της στοχαστικής κατηγορηματικής κλίσης με προσαρμοστικούς ρυθμούς μάθησης ανά παράμετρο [8]. Διατηρεί εκτιμήσεις των δύο πρώτων ροπών των κλίσεων: τον εκθετικά σταθμισμένο μέσο όρο  $m_t$  (πρώτη ροπή) και τον εκθετικά σταθμισμένο μέσο τετράγωνο  $v_t$  (δεύτερη ροπή). Σε κάθε βήμα  $t$ , με κλίση  $g_t = \nabla_{\theta} L_{t(\theta)}$ :

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \end{aligned}$$

Λόγω της αρχικοποίησης  $m_0 = v_0 = 0$ , εφαρμόζεται διόρθωση μεροληψίας:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

Η ενημέρωση των παραμέτρων γίνεται τότε:

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

Τυπικές τιμές υπερπαραμέτρων:  $\beta_2 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\varepsilon = 10^{-8}$ . Ο διαιρέτης  $\sqrt{\hat{v}_t} + \varepsilon$  ουσιαστικά κανονικοποιεί τον ρυθμό μάθησης ανάλογα με το μέγεθος των κλίσεων για κάθε παράμετρο χωριστά. Ο αλγόριθμος AdamW διαφοροποιείται από τον Adam στον τρόπο εφαρμογής της απομείωσης βαρών. Στον κλασικό Adam, η L2 κανονικοποίηση συγχωνεύεται στην κλίση  $g_t$ , με αποτέλεσμα να αλληλεπιδρά με τον προσαρμοστικό μηχανισμό και να παράγει υποβέλτιστη ρύθμιση [9]. Ο AdamW αποσυνδέει την απομείωση βαρών από την ενημέρωση κλίσεων:

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} - \eta \lambda \theta_{t-1}$$

όπου  $\lambda$  ο συντελεστής απομείωσης βαρών. Με αυτό τον τρόπο κάθε παράμετρος συρρικνώνεται ανεξάρτητα από τον προσαρμοστικό ρυθμό μάθησης, οδηγώντας σε καλύτερη γενίκευση.

### 1.1.6 Τεχνικές Κανονικοποίησης και Αποφυγής Υπερπροσαρμογής

Η πρώτη μέθοδος κατά της υπερπροσαρμογής που χρησιμοποιήθηκε ήταν η κανονικοποίηση παρτίδας, η οποία εισήχθη για να αντιμετωπίσει το φαινόμενο της εσωτερικής μεταβολής της συνδιακύμανσης, δηλαδή τη συνεχή μεταβολή της κατανομής των ενεργοποιήσεων κάθε επιπέδου κατά την εκπαίδευση [10]. Για μια παρτίδα  $B = \{x_1, \dots, x_m\}$ , η διαδικασία για κάθε χαρακτηριστικό  $k$  είναι:

$$\begin{aligned} \mu_B^{(k)} &= \frac{1}{m} \sum_{i=1}^m x_i^{(k)}, & \sigma_B^{2(k)} &= \frac{1}{m} \sum_{i=1}^m \left( x_i^{(k)} - \mu_B^{(k)} \right)^2 \\ \hat{x}_i^{(k)} &= \frac{x_i^{(k)} - \mu_B^{(k)}}{\sqrt{\sigma_B^{2(k)} + \varepsilon}}, & y_i^{(k)} &= \gamma^{(k)} \hat{x}_i^{(k)} + \beta^{(k)} \end{aligned}$$

όπου  $\gamma^{(k)}$  και  $\beta^{(k)}$  είναι εκπαιδύσιμες παράμετροι κλίμακας και μετατόπισης αντίστοιχα. Η κανονικοποίηση παρτίδας επιτρέπει τη χρήση υψηλότερων ρυθμών μάθησης, μειώνει την εξάρτηση από την αρχικοποίηση βαρών και λειτουργεί ως ρυθμιστικός παράγοντας.

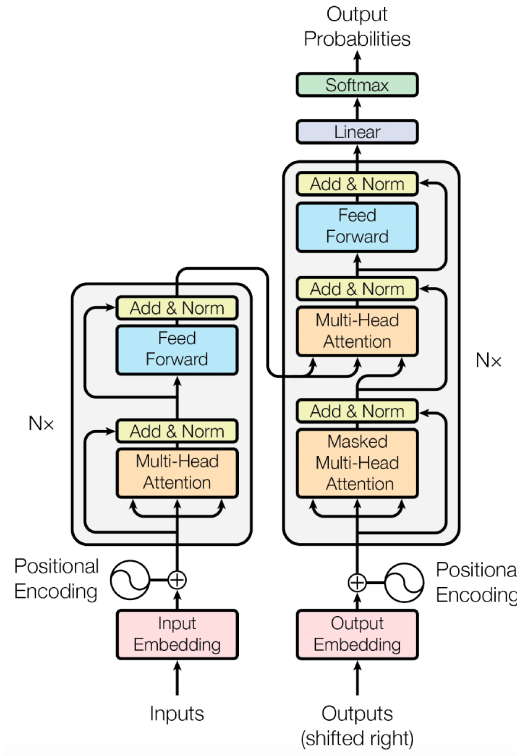
Η δεύτερη μέθοδος που χρησιμοποιήθηκε ήταν η dropout που αποτελεί ισχυρή τεχνική αποφυγής υπερπροσαρμογής [11]. Κατά την εκπαίδευση, κάθε νευρώνας απενεργοποιείται ανεξάρτητα με πιθανότητα  $p$  σε κάθε βήμα εκπαίδευσης. Τυπικά, για ένα επίπεδο με ενεργοποιήσεις  $a$ :

$$r_i \sim \text{Bernoulli}(1 - p), \quad \tilde{a}_i = \frac{r_i \cdot a_i}{1 - p}$$

όπου η κλιμάκωση με  $1/(1-p)$  διασφαλίζει ότι η αναμενόμενη τιμή των ενεργοποιήσεων παραμένει σταθερή μεταξύ εκπαίδευσης και αξιολόγησης. Κατά την αξιολόγηση, το dropout απενεργοποιείται και χρησιμοποιούνται όλοι οι νευρώνες. Η τεχνική αναγκάζει το δίκτυο να αναπτύξει πλεονάζουσες αναπαραστάσεις που δεν εξαρτώνται υπερβολικά από μεμονωμένους νευρώνες, βελτιώνοντας τη γενίκευση. Στο προτεινόμενο μοντέλο, κανονικοποίηση παρτίδας και dropout εφαρμόζονται συνδυαστικά μετά από τα γραμμικά επίπεδα του κωδικοποιητή, διαμορφώνοντας ένα ισορροπημένο πλαίσιο βελτιστοποίησης για την εξαγωγή γενικεύσιμων πρωτεϊνικών αναπαραστάσεων.

### 1.1.7 Μετασχηματιστές

Ο μετασχηματιστής εισήχθη από τους Vaswani et al. (2017) και αποτελεί σήμερα τη θεμελιώδη αρχιτεκτονική για την πλειονότητα των σύγχρονων μοντέλων γλώσσας. Ο βασικός του στόχος είναι να μοντελοποιεί εξαρτήσεις μεταξύ στοιχείων μιας ακολουθίας ανεξάρτητα από την απόστασή τους, κάτι που οι προηγούμενες αρχιτεκτονικές, όπως τα αναδρομικά νευρωνικά δίκτυα, αντιμετώπιζαν με δυσκολία. Αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή, καθένας από τους οποίους είναι στοίβα επιπέδων. Κάθε επίπεδο περιέχει δύο υποστρώματα: έναν μηχανισμό αυτο-προσοχής και ένα δίκτυο πρόσθιας τροφοδότησης, με υπολειπόμενες συνδέσεις και κανονικοποίηση στρώματος γύρω από κάθε ένα [12].



**Εικόνα 1.1.** Η αρχιτεκτονική ενός μετασχηματιστή με κωδικοποιητή και αποκωδικοποιητή (πηγή: Vaswani et al.)

Το κεντρικό μαθηματικό εργαλείο είναι ο μηχανισμός κλιμακωτής εσωτερικής προσοχής. Η είσοδος μετασχηματίζεται σε τρεις πίνακες — ερωτήματα  $Q$ , κλειδιά  $K$  και τιμές  $V$  — μέσω εκπαιδευσιμων γραμμικών προβολών, και η έξοδος υπολογίζεται ως:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

Το εσωτερικό γινόμενο  $QK^\top$  μετράει την ομοιότητα μεταξύ κάθε ζεύγους θέσεων, ενώ το softmax το μετατρέπει σε κατανομή βαρών προσοχής που χρησιμοποιούνται για σταθμισμένο άθροισμα των τιμών  $V$ . Η διαίρεση με  $\sqrt{d_k}$ , όπου  $d_k$  είναι η διάσταση των διανυσμάτων κλειδιών, αποτρέπει τον αριθμητικό κορεσμό του softmax για υψηλές διαστάσεις, διατηρώντας έτσι την ομαλή ροή των κλίσεων κατά την εκπαίδευση. Στην πράξη, ο μηχανισμός εκτελείται παράλληλα σε  $h$  κεφαλές, κάθε μία με διαφορετικούς πίνακες προβολής, και τα αποτελέσματα συνενώνονται και προβάλλονται μέσω ενός εκπαιδευσιμου πίνακα  $W^O$ :

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O$$

Επειδή ο μηχανισμός αυτο-προσοχής δεν έχει από μόνος του αίσθηση της σειράς των στοιχείων, προστίθεται στην είσοδο μία κωδικοποίηση θέσης. Στην αρχική εργασία χρησιμοποιήθηκαν ημιτονοειδείς συναρτήσεις σταθερής μορφής, ενώ νεότερα μοντέλα όπως ο `ProT5` υιοθετούν εκπαιδευσιμες σχετικές κωδικοποιήσεις θέσης. Μετά τον

μηχανισμό αυτο-προσοχής, κάθε θέση της αλληλουχίας διέρχεται ανεξάρτητα από ένα δίκτυο πρόσθιας τροφοδότησης δύο στρωμάτων με μη γραμμική ενεργοποίηση, το οποίο εμπλουτίζει περαιτέρω την αναπαράσταση. Το συνολικό αποτέλεσμα είναι μία αρχιτεκτονική που επεξεργάζεται ολόκληρη την ακολουθία παράλληλα, καθιστώντας τον μετασχηματιστή τόσο υπολογιστικά αποδοτικό όσο και ιδιαίτερα κατάλληλο για εκπαίδευση σε μεγάλης κλίμακας δεδομένα [12].

Στο πλαίσιο της παρούσας εργασίας, οι μετασχηματιστές αξιοποιήθηκαν για την παραγωγή πλούσιων αναπαραστάσεων πρωτεϊνικών αλληλουχιών. Η θεμελιώδης παραδοχή της προσέγγισης αυτής είναι ότι οι πρωτεϊνικές αλληλουχίες μπορούν να αντιμετωπιστούν ως φυσική γλώσσα: κάθε αμινοξύ αντιστοιχεί σε μία λέξη και κάθε αλληλουχία σε μία πρόταση. Με αυτόν τον τρόπο, προ-εκπαιδευμένα μοντέλα γλώσσας πρωτεϊνών μπορούν να εφαρμοστούν απευθείας στις ακατέργαστες αλληλουχίες, χωρίς να απαιτείται καμία πρόσθετη βιολογική επισήμανση. Μέσω του μηχανισμού αυτο-προσοχής, ο μετασχηματιστής παράγει για κάθε αμινοξύ μία πλαίσιο-εξαρτώμενη αναπαράσταση — δηλαδή ένα διάνυσμα του οποίου η τιμή εξαρτάται από το σύνολο των γειτονικών αμινοξέων στην αλληλουχία και όχι μόνο από την ταυτότητα του ίδιου του αμινοξέος. Τα διανύσματα αυτά σταθερής διαστατικότητας — ανεξάρτητα από το μήκος της αλληλουχίας εισόδου — αποτέλεσαν τις εισόδους για την εκπαίδευση των μοντέλων πρόβλεψης που αναπτύχθηκαν στα κατάντη στάδια της εργασίας, επιτρέποντας την αποτελεσματική μεταφορά της βιολογικής πληροφορίας που έχει ενσωματωθεί κατά την προ-εκπαίδευση.

## **1.2 Βιολογικό Υπόβαθρο**

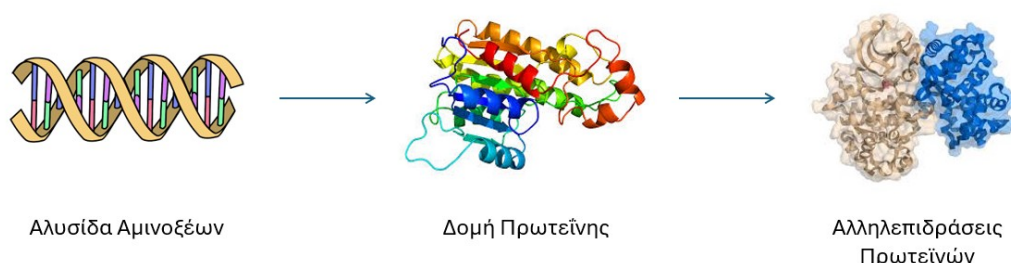
Η κατανόηση των αλληλεπιδράσεων πρωτεϊνών προϋποθέτει εξοικείωση με τις θεμελιώδεις αρχές της μοριακής βιολογίας που τις διέπουν. Στην ενότητα αυτή παρουσιάζεται το βιολογικό υπόβαθρο που είναι απαραίτητο για την κατανόηση της προσέγγισης που ακολουθείται στην εργασία: αρχικά εξετάζεται η διαδικασία αναδίπλωσης των πρωτεϊνών και ο τρόπος με τον οποίο η αλληλουχία αμινοξέων καθορίζει την τελική τρισδιάστατη δομή τους, και στη συνέχεια αναλύεται η φύση και η σημασία των πρωτεϊνικών αλληλεπιδράσεων ως βασικού μηχανισμού οργάνωσης των κυτταρικών λειτουργιών.

### **1.2.1 Αναδίπλωση Πρωτεϊνών**

Οι πρωτεΐνες αρχικά σχηματίζονται ως γραμμικές αλυσίδες αμινοξέων, σύμφωνα με τις γενετικές οδηγίες του κυττάρου. Αυτή η σειρά των αμινοξέων — γνωστή ως πρωτοταγής δομή — καθορίζει τον τρόπο με τον οποίο η αλυσίδα θα αρχίσει να διπλώνει [13]. Η σύνθεση της πολυπεπτιδικής αλυσίδας πραγματοποιείται στα ριβοσώματα μέσω της διαδικασίας της μετάφρασης, κατά την οποία το mRNA αποκωδικοποιείται και τα αμινοξέα ενώνονται μεταξύ τους με πεπτιδικούς δεσμούς

[14]. Η πρωτοταγής δομή αποτελεί ουσιαστικά το «αλφάβητο» της πρωτεΐνης: κάθε αμινοξύ διαθέτει διαφορετική πλευρική ομάδα με μοναδικές χημικές ιδιότητες, και η διαδοχή τους κωδικοποιεί πλήρως την τελική τρισδιάστατη μορφή που θα αποκτήσει το μόριο. Ακόμη και μια μεμονωμένη αλλαγή σε ένα αμινοξύ μπορεί να μεταβάλλει δραστικά τη λειτουργία ή τη σταθερότητα μιας πρωτεΐνης, όπως παρατηρείται σε παθολογικές μεταλλάξεις [13].

Κατά την αναδίπλωση, δημιουργούνται τοπικές δομές όπως  $\alpha$ -έλικες και  $\beta$ -πτυχωτές επιφάνειες — στοιχεία που αποτελούν τη δευτεροταγή δομή της πρωτεΐνης. Αυτές οι δομές σταθεροποιούνται κυρίως μέσω δεσμών υδρογόνου μεταξύ των ατόμων του κύριου σκελετού της πολυπεπτιδικής αλυσίδας [15]. Οι  $\alpha$ -έλικες σχηματίζονται όταν διαδοχικά αμινοξέα στρέφονται γύρω από έναν κεντρικό άξονα, δημιουργώντας μια σπειροειδή διαμόρφωση συγκρατούμενη από ενδοαλυσιδικούς δεσμούς υδρογόνου, ενώ οι  $\beta$ -πτυχωτές επιφάνειες προκύπτουν από την παράλληλη ή αντιπαράλληλη διάταξη τμημάτων της αλυσίδας. Στη συνέχεια, η αλυσίδα αποκτά μια πιο σύνθετη τρισδιάστατη μορφή — την τριτοταγή δομή — η οποία αποτελεί το αποτέλεσμα πολύπλοκων αλληλεπιδράσεων μεταξύ των πλευρικών ομάδων των αμινοξέων, συμπεριλαμβανομένων υδρόφοβων αλληλεπιδράσεων, ηλεκτροστατικών δυνάμεων και δεσμών van der Waals. Ορισμένες πρωτεΐνες, τέλος, συναρμολογούνται σε πολυμερικές δομές αποτελούμενες από δύο ή περισσότερες πολυπεπτιδικές αλυσίδες, σχηματίζοντας την τεταρτοταγή δομή, η οποία συχνά είναι απαραίτητη για τη βιολογική τους δραστηριότητα [16].



**Εικόνα 1.2.** Η αιτιολογική αλληλουχία από την αλυσίδα αμινοξέων μέχρι την αλληλεπίδραση της πρωτεΐνης

Οι χημικές ιδιότητες των πλευρικών ομάδων των αμινοξέων — αν είναι υδρόφοβες, φορτισμένες ή πολικές — επηρεάζουν καθοριστικά τις αλληλεπιδράσεις μεταξύ τους και οδηγούν στη σταθεροποίηση της τελικής δομής. Συγκεκριμένα, το λεγόμενο υδρόφοβο αποτέλεσμα αποτελεί έναν από τους κυριότερους κινητήριους παράγοντες της αναδίπλωσης: τα υδρόφοβα κατάλοιπα τείνουν να συγκεντρώνονται στο εσωτερικό της πρωτεΐνης, μακριά από το υδατικό περιβάλλον, ενώ τα υδρόφιλα κατάλοιπα εκτίθενται στην επιφάνεια [17]. Αυτή η διαδικασία ελαχιστοποιεί την ενεργειακά δυσμενή επαφή μεταξύ μη πολικών ομάδων και νερού, οδηγώντας το σύστημα σε θερμοδυναμικά σταθερότερη κατάσταση. Επιπλέον, δισουλφιδικοί δεσμοί μεταξύ

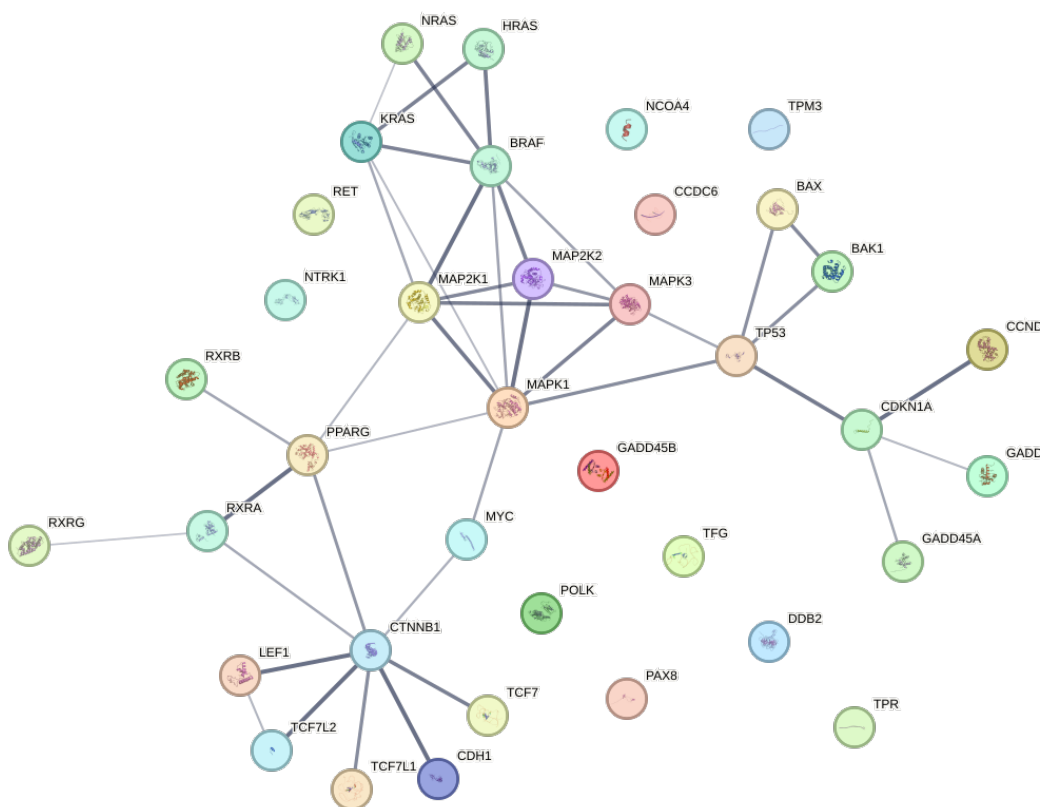
καταλοίπων κυστεΐνης μπορούν να συμβάλλουν στη μηχανική σταθερότητα της αναδιπλωμένης δομής, ιδιαίτερα σε πρωτεΐνες που εκκρίνονται εκτός κυττάρου και εκτίθενται σε πιο ακραίες συνθήκες [18]. Ιόντα μετάλλων και συμπαραγοντες μπορούν επίσης να συμμετέχουν στη σταθεροποίηση της δομής, λειτουργώντας ως «γέφυρες» μεταξύ διαφορετικών τμημάτων της αλυσίδας.

Η τελική αναδιπλωμένη μορφή της πρωτεΐνης καθορίζει όλες τις χημικές και βιολογικές της ιδιότητες και, ως εκ τούτου, τον ρόλο που θα επιτελέσει μέσα στο κύτταρο [13]. Η σύνδεση αυτή μεταξύ δομής και λειτουργίας αποτελεί θεμελιώδη αρχή της μοριακής βιολογίας: πρωτεΐνες με παρόμοια τρισδιάστατη μορφή συχνά επιτελούν ανάλογες λειτουργίες, ακόμη και αν η αλληλουχία των αμινοξέων τους διαφέρει σημαντικά. Ιδιαίτερης σημασίας είναι η έννοια του ενεργού κέντρου, δηλαδή της περιοχής της πρωτεΐνης στην οποία συνδέεται το υπόστρωμα ή ο προσδέτης — μια λειτουργία που εξαρτάται εξ ολοκλήρου από τη χωρική διάταξη συγκεκριμένων αμινοξέων [19]. Ακόμη, πολλές πρωτεΐνες υπόκεινται σε δομικές αλλαγές κατά τη σύνδεσή τους με άλλα μόρια — ένα φαινόμενο γνωστό ως επαγόμενη εφαρμογή — το οποίο επιτρέπει τη ρύθμιση της βιολογικής τους δραστηριότητας.

Η αλληλουχία των αμινοξέων καθορίζει την τρισδιάστατη αναδίπλωση μιας πρωτεΐνης, και η τελική αυτή δομή είναι εκείνη που προσδιορίζει τις βιολογικές της ιδιότητες και λειτουργίες. Η πρόβλεψη της τρισδιάστατης μορφής μιας πρωτεΐνης αποτελεί από μόνη της ένα ιδιαίτερα απαιτητικό πρόβλημα στον χώρο της μηχανικής μάθησης και της υπολογιστικής βιολογίας [20], λόγω της πολυπλοκότητας των αλληλεπιδράσεων που λαμβάνουν χώρα κατά την αναδίπλωση. Το λεγόμενο «πρόβλημα αναδίπλωσης πρωτεϊνών» αναγνωρίστηκε ως ένα από τα μεγαλύτερα ανοικτά ζητήματα της βιολογίας και της χημείας για δεκαετίες [21]. Η ανάπτυξη εργαλείων όπως το AlphaFold από την DeepMind σηματοδότησε τομή στο πεδίο αυτό, επιτυγχάνοντας ακρίβεια πρόβλεψης συγκρίσιμη με πειραματικές μεθόδους και ανοίγοντας νέους ορίζοντες στη βιοϊατρική έρευνα και στον σχεδιασμό φαρμάκων.

### 1.2.2 Πρωτεϊνικές αλληλεπιδράσεις

Οι αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης (ΑΠΠ) αποτελούν τον θεμελιώδη μηχανισμό μέσω του οποίου η κυτταρική βιολογία οργανώνεται σε λειτουργικά δίκτυα. Κάθε βιοχημική διεργασία — από τη μεταγραφή του DNA ως την απόκριση σε εξωκυτταρικά σήματα — εξαρτάται από την ικανότητα μιας πρωτεΐνης να αναγνωρίζει με ακρίβεια μια άλλη και να σχηματίζει μαζί της ένα λειτουργικό σύμπλεγμα. Η ειδικότητα αυτή δεν προκύπτει από έναν μεμονωμένο ισχυρό δεσμό, αλλά από τη συνεργιστική δράση πολλών ασθενών αλληλεπιδράσεων στη διεπιφάνεια επαφής: δεσμοί υδρογόνου, υδρόφοβες δυνάμεις, ηλεκτροστατικές έλξεις και αλληλεπιδράσεις van der Waals συνδυάζονται ώστε να δημιουργήσουν σύνδεση υψηλής συγγένειας και ειδικότητας [22][23].



**Εικόνα 1.3.** Ένα δίκτυο αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης (πηγή: db-string.org)

Η γεωμετρία της διεπιφάνειας είναι κρίσιμη. Το κλασικό μοντέλο «κλειδιού-κλειδαριάς» του Fischer, αν και ιστορικά σημαντικό, έχει σε μεγάλο βαθμό αντικατασταθεί από το μοντέλο της επαγόμενης εφαρμογής και, πιο πρόσφατα, από την έννοια της επιλογής διαμόρφωσης. Σύμφωνα με αυτή την τελευταία προσέγγιση, οι πρωτεΐνες δεν είναι άκαμπτες δομές αλλά συνολα ισορροπίας μεταξύ πολλαπλών διαμορφώσεων· ο προσδέτης δεν «πλάθει» την πρωτεΐνη, αλλά σταθεροποιεί μια προϋπάρχουσα ενεργή διαμόρφωση που εντοπίζεται σε μικρό ποσοστό στη φυσιολογική κατάσταση. Ιδιαίτερη κατηγορία αποτελεί η αλλοστερική ρύθμιση, στην οποία η δέσμευση ενός μορίου σε θέση απομακρυσμένη από το ενεργό κέντρο μεταδίδει δομικές αλλαγές σε όλη την πρωτεΐνη. Το φαινόμενο αυτό — μεταφορά πληροφορίας μέσω της τρισδιάστατης πτύχωσης — αποτελεί τη βάση της ανατροφοδότησης σε μεταβολικά μονοπάτια: το τελικό προϊόν μιας αλυσίδας αντιδράσεων αναστέλλει αλλοστερικά το πρώτο ένζυμο της αλυσίδας, ρυθμίζοντας έτσι την παραγωγή χωρίς να χρειαστεί μεταγραφική απόκριση [24].

Στον αντίποδα βρίσκονται οι υψηλής τάξης συμπλεγματοισμοί — σύμπλοκα πολλαπλών υπομονάδων όπου η βιολογική λειτουργία εμφανίζεται αποκλειστικά στο επίπεδο του συγκροτήματος και όχι στις επιμέρους πρωτεΐνες. Χαρακτηριστικό παράδειγμα είναι το πρωτεάσωμα, ένα βαρέλι 2.5 MDa από εικοσιοκτώ υπομονάδες που αποικοδομεί πρωτεΐνες σημασμένες με ουβικουτίνη· καμία μεμονωμένη υπομονάδα δεν φέρει πρωτεολυτική δραστηριότητα. Παρόμοια λογική ισχύει για

το ριβόσωμα, τις DNA πολυμεράσες και τους κινητήριους χαρτογραφημένους στα κυτταροσκελετικά νημάτια [25].

Οι διεπιφάνειες των πρωτεϊνών είναι συχνά επίπεδες και εκτεταμένες, στερούμενες της βαθιάς δεσμευτικής κοιλότητας που απαιτείται για τους κλασικούς φαρμακολογικούς αγωνιστές. Τα τελευταία χρόνια, ωστόσο, μελέτες υψηλής ανάλυσης έχουν αναδείξει ότι η ενέργεια δέσμευσης συχνά συγκεντρώνεται σε μικρές υποπεριοχές της διεπιφάνειας — τα αποκαλούμενα hot spots — οι οποίες μπορούν να αποτελέσουν στόχο μικρών μορίων ή πεπτιδικών αναλόγων, ανοίγοντας δρόμο για έναν νέο τύπο φαρμακολογικής παρέμβασης [26][27].

### 1.3 Στόχοι και Κίνητρο της Εργασίας

Οι αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης (ΑΠΠ) αποτελούν τη ραχοκοκαλιά των κυτταρικών λειτουργιών. Σχεδόν κάθε βιολογική διεργασία — από τη μεταγωγή σήματος και τη ρύθμιση της γονιδιακής έκφρασης έως την κυτταρική διαίρεση και την ανοσολογική απόκριση — βασίζεται στην ικανότητα των πρωτεϊνών να αναγνωρίζουν και να δεσμεύουν η μία την άλλη με εξαιρετική εξειδίκευση [28][29]. Η εξειδίκευση αυτή δεν είναι τυχαία, αλλά αποτέλεσμα εκατομμυρίων ετών εξελικτικής πίεσης που διαμόρφωσε συγκεκριμένες δομικές και φυσικοχημικές συμπληρωματικότητες μεταξύ των πρωτεϊνικών διεπιφανειών. Πέρα από τις ζευγαρωτές αλληλεπιδράσεις, οι πρωτεΐνες συχνά συγκροτούν μεγάλα πολυπρωτεϊνικά σύμπλοκα, των οποίων η σύνθεση και η δυναμική ρύθμιση καθορίζουν την έκβαση κρίσιμων κυτταρικών αποφάσεων. Η χαρτογράφηση αυτών των αλληλεπιδράσεων σε γενωμική κλίμακα, γνωστή ως ανάλυση του πρωτεϊνικού διαδράματος, αποτελεί έναν από τους κεντρικούς στόχους της σύγχρονης συστημικής βιολογίας [30]. Η κατανόηση του interactome δεν εξαντλείται στην απαρίθμηση των αλληλεπιδράσεων, αλλά επεκτείνεται στην ανάλυση της χωροχρονικής τους οργάνωσης, καθώς οι ίδιες πρωτεΐνες μπορεί να συμμετέχουν σε διαφορετικά δίκτυα ανάλογα με τον τύπο του κυττάρου, τη φάση του κυτταρικού κύκλου ή τα εξωτερικά ερεθίσματα.

Στο πλαίσιο της βιοϊατρικής έρευνας, η πρόβλεψη ΑΠΠ έχει άμεση πρακτική αξία. Πολλές ασθένειες όπως ο καρκίνος, οι νευροεκφυλιστικές παθήσεις και οι μολυσματικές νόσοι, οφείλονται εν μέρει σε διαταραχές του φυσιολογικού δικτύου αλληλεπιδράσεων — είτε μέσω παθολογικής ενεργοποίησης είτε μέσω απώλειας λειτουργικών συνδέσεων [31]. Χαρακτηριστικό παράδειγμα αποτελούν ογκογόνες μεταλλάξεις που οδηγούν στην εμφάνιση νέων, παθολογικών αλληλεπιδράσεων ή στην αποδιοργάνωση ρυθμιστικών συμπλόκων που φυσιολογικά ελέγχουν τον κυτταρικό πολλαπλασιασμό. Η ικανότητα εντοπισμού αυτών των διαταραγμένων αλληλεπιδράσεων ανοίγει τον δρόμο για την ανακάλυψη νέων θεραπευτικών στόχων, ενώ παράλληλα τροφοδοτεί την ανάπτυξη φαρμάκων που στοχεύουν συγκεκριμένα στις διεπιφάνειες πρωτεϊνικών αλληλεπιδράσεων (αναστολείς

πρωτεϊνικών αλληλεπιδράσεων) [32]. Τα φάρμακα αυτά αντιπροσωπεύουν μια ιδιαίτερα ελπιδοφόρα κατηγορία θεραπευτικών μορίων, καθώς επιτρέπουν τη στόχευση αλληλεπιδράσεων που παραδοσιακά θεωρούνταν «μη φαρμακεύσιμες» λόγω της έλλειψης ευδιάκριτων ενζυμικών θέσεων πρόσδεσης. Η συστηματική χαρτογράφηση των παθολογικά τροποποιημένων ΑΠΠ σε διάφορες νόσους αναμένεται να επεκτείνει σημαντικά τον θεραπευτικό ορίζοντα της σύγχρονης φαρμακολογίας.

Παραδοσιακά, ο πειραματικός εντοπισμός ΑΠΠ — μέσω τεχνικών όπως η αντοχή δύο-υβριδίων ζύμης, η συν-ανοσοκατακρήμνιση ή η φασματομετρία μάζας — είναι εξαιρετικά χρονοβόρος, δαπανηρός και δύσκολα επεκτάσιμος σε γενωμική κλίμακα [33]. Επιπλέον, κάθε πειραματική μέθοδος εμφανίζει συγκεκριμένες τεχνικές αδυναμίες και τάσεις σφάλματος, με αποτέλεσμα τα δεδομένα από διαφορετικές πηγές να παρουσιάζουν συχνά χαμηλή επικάλυψη και προβλήματα αναπαραγωγιμότητας. Ο αριθμός των πιθανών ζευγών πρωτεϊνών στον ανθρώπινο οργανισμό εκτιμάται στις δεκάδες χιλιάδες, εκ των οποίων μόνο ένα μικρό κλάσμα έχει πειραματικά τεκμηριωθεί [30], γεγονός που αναδεικνύει το τεράστιο χάσμα μεταξύ του γνωστού και του άγνωστου τμήματος του ανθρώπινου interactome. Η ανάπτυξη αξιόπιστων υπολογιστικών μοντέλων πρόβλεψης ΑΠΠ αποτελεί επομένως όχι απλώς μια ακαδημαϊκή πρόκληση, αλλά μια πρακτική αναγκαιότητα για την επιτάχυνση της βιοϊατρικής έρευνας. Τα μοντέλα αυτά, αξιοποιώντας πληροφορίες από αλληλουχίες, τρισδιάστατες δομές, εξελικτική συντήρηση και δεδομένα γονιδιακής έκφρασης, έχουν τη δυνατότητα να διερευνούν τον χώρο των πιθανών αλληλεπιδράσεων με ταχύτητα και κλίμακα που υπερβαίνουν κατά πολύ τις δυνατότητες της πειραματικής προσέγγισης.

## 2 Βιβλιογραφική Ανασκόπηση

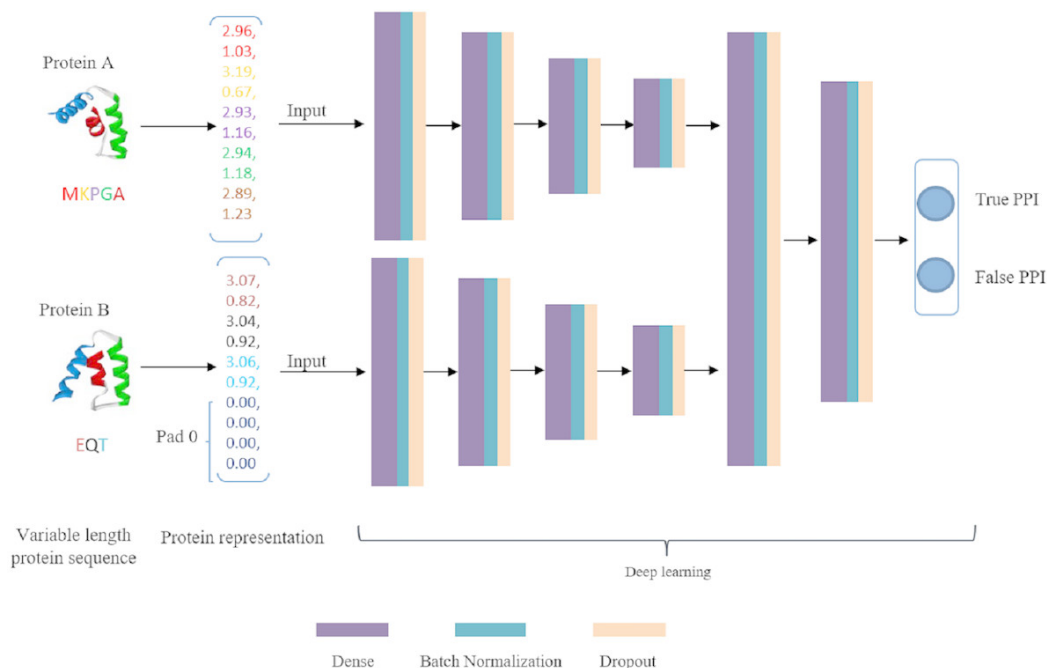
Στην προτεινόμενη προσέγγιση, παρακάμπτεται το ενδιάμεσο αυτό στάδιο της ρητής πρόβλεψης της δομής. Αντί να υπολογίζεται πρώτα η τρισδιάστατη διαμόρφωση και στη συνέχεια η πιθανή αλληλεπίδραση, το μοντέλο επιχειρεί να εκτιμήσει απευθείας την πιθανότητα αλληλεπίδρασης μεταξύ δύο πρωτεϊνών χρησιμοποιώντας αποκλειστικά τις αλληλουχίες αμινοξέων τους. Συγκεκριμένα, ως είσοδο λαμβάνονται οι αναπαραστάσεις των δύο αλληλουχιών, τα οποία ενσωματώνουν σημασιολογική και δομική πληροφορία σε συμπαγή διανύσματα χαρακτηριστικών. Με βάση αυτά, το μοντέλο παράγει ως έξοδο μια εκτίμηση της πιθανότητας αλληλεπίδρασης των δύο πρωτεϊνών.

Με τον τρόπο αυτό, η πληροφορία που σχετίζεται με την αναδίπλωση και τη δομή δεν υπολογίζεται ρητά, αλλά μαθαίνεται έμμεσα κατά την εκπαίδευση του μοντέλου, καθώς ενσωματώνεται στις εσωτερικές αναπαραστάσεις του. Αυτό επιτρέπει την αποδοτικότερη και ταχύτερη πρόβλεψη αλληλεπιδράσεων, αποφεύγοντας το

υπολογιστικά δαπανηρό βήμα της πλήρους δομικής πρόβλεψης. Οι πρωτεϊνικές αναπαραστάσεις έχουν αξιοποιηθεί εκτενώς στη βιοπληροφορική για την πρόβλεψη λειτουργικών ιδιοτήτων των πρωτεϊνών, συμπεριλαμβανομένης της ικανότητάς τους να αλληλεπιδρούν με διαφορετικές κατηγορίες μορίων [34]. Ιδιαίτερα αποτελεσματική έχει αναδειχθεί η συνδυαστική χρήση πρωτεϊνικών αναπαραστάσεων με τεχνικές βαθιάς μάθησης, η οποία έχει καταφέρει να αποτυπώσει πολύπλοκες σχέσεις μεταξύ των πρωτεϊνών και να επιτύχει υψηλή ακρίβεια στην πρόβλεψη αλληλεπιδράσεων μεταξύ τους.

Για παράδειγμα, οι Wan & Zeng (2016) υλοποίησαν ένα μοντέλο βαθιάς μάθησης που αντιμετωπίζει τις αλληλουχίες πρωτεϊνών σαν φράσεις και τα επιμέρους k-mer τμήματα σαν λέξεις. Στη συνέχεια, κάνει χρήση ενός Skip-Gram μοντέλου με αρνητική δειγματοληψία για την μετατροπή αυτών των «φράσεων» σε διανύσματα σταθερού μεγέθους μέσω της τεχνικής word2vec. Η δημιουργία των αναπαραστάσεων γίνεται με αυτόματο, μη επιβλεπόμενο τρόπο από μεγάλες βάσεις δεδομένων χωρίς ετικέτες, τόσο για τις πρωτεΐνες όσο και για τις χημικές ενώσεις. Οι βέκτορες αυτοί στη συνέχεια ενώνονται και δίνονται ως είσοδος σε ένα βαθύ νευρωνικό δίκτυο (BNN) εκπαιδευμένο με στοχαστική καταβατική κλίση, dropout και κανονικοποίηση παρτίδας για να προβλέψει αν μία χημική ένωση και μία πρωτεΐνη αλληλεπιδρούν ή όχι [35].

Μια παρόμοια μέθοδος προτείνεται από τους Yao et al. (2019) οι οποίοι επίσης συνδύασαν αναπαραστάσεις πρωτεϊνών με βαθιά μάθηση, αλλά με μερικές σημαντικές διαφορές. Συγκεκριμένα, η μέθοδος για τη δημιουργία των αναπαραστάσεων, που ονομάζεται Res2vec, αντιμετωπίζει κάθε μεμονωμένο αμινοξύ ως μία «λέξη» (και όχι k-mer υποακολουθίες) μειώνοντας έτσι το μέγεθος του λεξιλογίου και απλοποιώντας την αναπαράσταση. Και αυτή η μέθοδος χρησιμοποιεί Skip-Gram μέσω της βιβλιοθήκης Gensim, ενώ ο βέλτιστος αριθμός διαστάσεων των διανυσμάτων ορίστηκε στις 20. Ως προς την αρχιτεκτονική, το μοντέλο τους (γνωστό ως DeepFE-PPI) υιοθετεί μία διπλή αρχιτεκτονική: δύο ξεχωριστά βαθιά νευρωνικά δίκτυα επεξεργάζονται κάθε πρωτεΐνη ανεξάρτητα, και τα αποτελέσματά τους ενώνονται στη συνέχεια σε ένα κοινό επίπεδο ολοκλήρωσης, το οποίο καταλήγει σε ταξινόμηση μέσω softmax. Το μοντέλο αξιολογήθηκε για πρόβλεψη ΑΠΠ και πέτυχε ακρίβεια έως 98,71% σε ανθρώπινα δεδομένα [36].



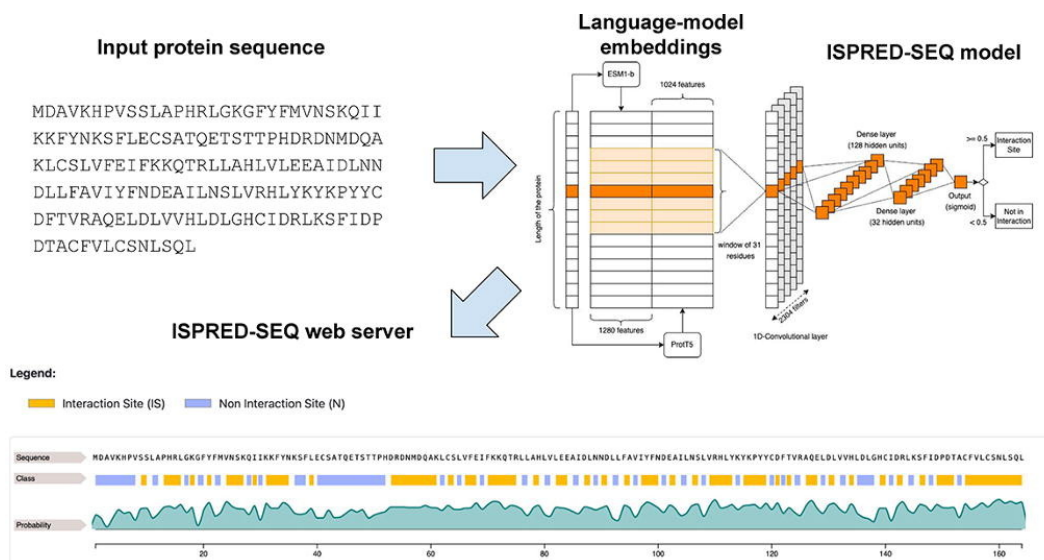
**Εικόνα 2.1.** Η αρχιτεκτονική του μοντέλου των Yao et al. (2019)

Σε ένα διαφορετικό πλαίσιο εφαρμογής, οι Madan et al. (2022) παρουσιάζουν μία μέθοδο για την πρόβλεψη αλληλεπιδράσεων πρωτεΐνης-ιού, γνωστή ως STEP (Siamese Tailored deep sequence Embedding of Proteins). Η μέθοδος συνδυάζει συμφραστικές ενσωματώσεις από το προ-εκπαιδευμένο μοντέλο Μετασχηματιστή ProtBERT, εκπαιδευμένο σε πάνω από 2 δισεκατομμύρια αμινοξικές αλληλουχίες. Για το μοντέλο έγινε χρήση Σιαμέζικης αρχιτεκτονικής η οποία επεξεργάζεται κάθε πρωτεΐνη ξεχωριστά μέσω του ίδιου κοινόχρηστου δικτύου, ενώ ο συνδυασμός των δύο αναπαραστάσεων γίνεται μέσω του γινομένου Hadamard) μιας τεχνικής κατάλληλης για τη μοντελοποίηση συμμετρικών αλληλεπιδράσεων. Η εκπαίδευση ακολουθεί ένα σχήμα Positive Unlabeled (PU) learning, αντιμετωπίζοντας έτσι τη χαρακτηριστική ανισορροπία κλάσεων που προκύπτει από το γεγονός ότι οι βάσεις δεδομένων ΑΠΠ περιλαμβάνουν αποκλειστικά γνωστές θετικές αλληλεπιδράσεις, ενώ οι «αρνητικές» είναι στην πραγματικότητα αδύνατο να επαληθευτούν. Το μοντέλο αξιολογήθηκε σε δύο βιολογικά σχετικά σενάρια (τις αλληλεπιδράσεις του ιού JCV με ανθρώπινους υποδοχείς στον εγκέφαλο και των πρωτεϊνών του SARS-CoV-2 με ανθρώπινες πρωτεΐνες-ξενιστές) επιτυγχάνοντας επίδοση αιχμής σε σύγκριση με μεθόδους που βασίζονται σε παραδοσιακές word2vec αναπαραστάσεις [37].

Μια ουσιαστικά διαφορετική προσέγγιση ακολουθούν οι Dallago et al. (2021), οι οποίοι προτείνουν τη χρήση αναπαραστάσεων από προ-εκπαιδευμένα Γλωσσικά Μοντέλα Πρωτεϊνών μέσω του εργαλείου bio\_embeddings. Αντί να εκπαιδεύουν από την αρχή ένα μοντέλο word2vec σε εξειδικευμένα δεδομένα, εκμεταλλεύονται μοντέλα βασισμένα στην αρχιτεκτονική Μετασχηματιστή (όπως το ProtBERT, εκπαιδευμένο στη βάση BFD με πάνω από 2,3 δισεκατομμύρια αλληλουχίες) για την παραγωγή

πλούσιων, πολυδιάστατων (1024 διαστάσεων) αναπαραστάσεων. Σε αντίθεση με τις στατικές αναπαραστάσεις των προηγούμενων μεθόδων, οι συμφραστικές αναπαραστάσεις των Μετασχηματιστών αποδίδουν διαφορετική αναπαράσταση σε κάθε αμινοξύ ανάλογα με το πλήρες αλληλουχιακό του πλαίσιο. Τα διανύσματα ανά πρωτεΐνη παράγονται μέσω global average pooling των per-residue αναπαραστάσεων, και στη συνέχεια χρησιμοποιούνται ως χαρακτηριστικά εισόδου σε ελαφριούς ταξινομητές μέσω transfer learning. Επιπλέον, η μέθοδος ενσωματώνει βήματα οπτικοποίησης μέσω UMAP ή t-SNE για τη διερεύνηση του χώρου των πρωτεϊνών σε δύο ή τρεις διαστάσεις [38].

Μια σημαντική συνεισφορά στο πεδίο αποτελεί η εργασία των Manfredi et al. (2023), οι οποίοι παρουσίασαν το ISPRED-SEQ, ένα μοντέλο βαθιάς μάθησης για την πρόβλεψη θέσεων αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης αποκλειστικά από αλληλουχίες, χωρίς να απαιτείται η γνώση της τρισδιάστατης δομής [39]. Ως είσοδος χρησιμοποιείται η αλληλουχία της πρωτεΐνης, η οποία αρχικά κωδικοποιείται μέσω των γλωσσικών μοντέλων ProtT5 και ESM-1v, παράγοντας ένα διανυσματικό χαρακτηριστικό 2304 διαστάσεων για κάθε κατάλοιπο. Στη συνέχεια, η αρχιτεκτονική αξιοποιεί ένα παράθυρο 31 καταλοίπων γύρω από κάθε υπό εξέταση κατάλοιπο, που επεξεργάζεται από ένα μονοδιάστατο συνελκτικό επίπεδο ακολουθούμενο από τρία διαδοχικά πυκνά επίπεδα. Αξιοσημείωτο είναι ότι το ISPRED-SEQ υπερτερεί έναντι των υπάρχοντων μεθόδων αιχμής σε πολλαπλά ανεξάρτητα σύνολα δεδομένων, επιβεβαιώνοντας ότι η χρήση εμπλουτισμένων αναπαραστάσεων από προ-εκπαιδευμένα γλωσσικά μοντέλα πρωτεϊνών μπορεί να υποκαταστήσει αποτελεσματικά τα παραδοσιακά χειροκίνητα χαρακτηριστικά, όπως φυσικοχημικές ιδιότητες και προφίλ αλληλουχιών από πολλαπλές στοιχίσεις. Η εργασία αυτή είναι ιδιαίτερα συναφής με την παρούσα, καθώς και οι δύο προσεγγίσεις αξιοποιούν αναπαραστάσεις ProtT5 για την κωδικοποίηση πρωτεϊνικών αλληλουχιών, με τη διαφορά ότι το ISPRED-SEQ στοχεύει στον εντοπισμό συγκεκριμένων καταλοίπων αλληλεπίδρασης σε επίπεδο υπολοίπου, ενώ η παρούσα εργασία επικεντρώνεται στην πρόβλεψη της πιθανότητας αλληλεπίδρασης μεταξύ ολόκληρων ζευγών πρωτεϊνών



**Εικόνα 2.2.** Η αρχιτεκτονική του μοντέλου ISPRED-SEQ των Manfredi et al. (2023)

Οι Dang & Vu (2024) παρουσίασαν το xCAPT5, ένα υβριδικό μοντέλο πρόβλεψης ΑΠΠ που συνδυάζει γλωσσικά μοντέλα πρωτεϊνών με συνελκτικά νευρωνικά δίκτυα [40]. Το μοντέλο αξιοποιεί το T5-XL-UniRef50 για την παραγωγή πλούσιων αναπαραστάσεων αμινοξέων από αλληλουχίες, ενώ ο πυρήνας του αποτελείται από ένα σιαμέζικο συνελκτικό δίκτυο πολλαπλών πυρήνων που εντοπίζει χαρακτηριστικά αλληλεπίδρασης τόσο σε μικρο- όσο και σε μακρο-επίπεδο, ενισχυμένο με τον αλγόριθμο XGBoost. Η εργασία αυτή είναι ιδιαίτερα συναφής με την παρούσα, καθώς και οι δύο προσεγγίσεις χρησιμοποιούν αναπαραστάσεις από το ίδιο γλωσσικό μοντέλο (ProtT5/T5-XL-UniRef50) σε συνδυασμό με σιαμέζικη αρχιτεκτονική. Η κύρια διαφορά έγκειται στο ότι το xCAPT5 αξιοποιεί συνελκτικά επίπεδα πολλαπλών πυρήνων για την εξαγωγή τοπικών χαρακτηριστικών από τις αναπαραστάσεις σε επίπεδο καταλοίπου, ενώ η παρούσα εργασία χρησιμοποιεί μέσες αναπαραστάσεις ολόκληρης της πρωτεΐνης και επικεντρώνεται στη σύγκριση διαφορετικών αρχιτεκτονικών συνδυασμού τους. Τα πειραματικά αποτελέσματα δείχνουν ότι το xCAPT5 υπερτερεί έναντι των υπάρχουσών μεθόδων αιχμής στη δυαδική πρόβλεψη ΑΠΠ, επιδεικνύοντας ισχυρή γενίκευση σε ενδο-είδους, διαείδους και διαφορετικά πλαίσια ομοιότητας αλληλουχιών.

Μια διαφορετική προσέγγιση ακολουθούν οι Wu et al. (2024) με το MAPE-PPI, οι οποίοι επιχειρούν να γεφυρώσουν το χάσμα μεταξύ αλληλουχιακών και δομικών μεθόδων πρόβλεψης ΑΠΠ [41]. Η μέθοδος ορίζει το μικροπεριβάλλον κάθε καταλοίπου βάσει του αλληλουχιακού και δομικού του πλαισίου, κωδικοποιώντας αυτά τα μικροπεριβάλλοντα σε χημικά ουσιαστικούς διακριτούς κώδικες μέσω ενός εκτεταμένου «λεξιλογίου», με στόχο να αποτυπωθεί τόσο η αλληλουχιακή όσο και η δομική πληροφορία κάθε αμινοξέος. Προς αυτό τον σκοπό προτείνεται μια νέα στρατηγική προ-εκπαίδευσης, γνωστή ως Μοντελοποίηση κωδικοβιβλίου με μάσκα, που μαθαίνει τις εξαρτήσεις μεταξύ διαφορετικών μικροπεριβαλλόντων. Το

μοντέλο επιδεικνύει εξαιρετική κλιμακωσιμότητα, καταφέροντας να επεξεργαστεί εκατομμύρια ΑΠΠ με ανώτερους συμβιβασμούς μεταξύ αποτελεσματικότητας και υπολογιστικής αποδοτικότητας σε σχέση με τις υπάρχουσες μεθόδους αιχμής. Η εργασία αυτή αναδεικνύει ότι η ενσωμάτωση δομικής πληροφορίας — έστω και έμμεσα μέσω μικροπεριβαλλόντων — μπορεί να βελτιώσει σημαντικά την προβλεπτική ικανότητα, και αποτελεί φυσική επέκταση της παρούσας εργασίας, η οποία επικεντρώνεται αποκλειστικά στις αλληλουχιακές αναπαραστάσεις.

Οι προσεγγίσεις αυτές αξιοποιούν την πλούσια αναπαράσταση των αλληλουχιών και τις δομές των πρωτεϊνών που ενσωματώνεται στις αναπαραστάσεις, επιτρέποντας στα μοντέλα να γενικεύουν αποτελεσματικά ακόμα και για πρωτεϊνικά ζεύγη που δεν έχουν παρατηρηθεί πειραματικά. Κοινός παρονομαστής και των τριών προσεγγίσεων είναι η εκμάθηση αναπαραστάσεων από αλληλουχίες χωρίς επιβλεπόμενη εκπαίδευση, αξιοποιώντας την εγγενή «γλώσσα» των πρωτεϊνών. Η εξελικτική τάση που παρατηρείται (από τις αναπαραστάσεις k-mer του word2vec, στις μονο-αμινοξικές Res2vec αναπαραστάσεις, έως τις βαθιές ενσωματώσεις των Μετασχηματιστών) αντικατοπτρίζει τη συνεχή επιδίωξη για πιο εκφραστικές και γενικεύσιμες αναπαραστάσεις πρωτεϊνικών αλληλουχιών.

### 3 Συλλογή Δεδομένων

Στην υλοποίηση χρησιμοποιήθηκαν τα πρωτεώματα επτά διαφορετικών ειδών, τα οποία ανήκουν σε διαφορετικές ταξινομικές τάξεις: μηλιά (*Malus domestica*), σολομός (*Salmo salar*), χρυσαετός (*Aquila chrysaetos*), μύγα των φρούτων (*Drosophila melanogaster*), μανιτάρι πλευρώτους (*Pleurotus ostreatus*), βάτραχος (*Xenopus laevis*) και άνθρωπος (*Homo sapiens*). Η επιλογή αυτή αποσκοπούσε στη διασφάλιση της ποικιλομορφίας του συνόλου δεδομένων, προκειμένου να διευκολυνθεί ο εντοπισμός των δομικών και λειτουργικών μοτίβων που διέπουν τις αλληλεπιδράσεις μεταξύ πρωτεϊνών σε όλα τα είδη. Στο πλαίσιο της ογκολογίας, η μελέτη των αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης (ΑΠΠ) κατέχει ιδιαίτερη σημασία. Διαταραχές στα δίκτυα αυτά συνδέονται άμεσα με τη δυσλειτουργία καρκινικών πρωτεϊνών, ενώ παράλληλα ενδέχεται να αποκαλύψουν νέους θεραπευτικούς στόχους [42]. Η ίδια μεθοδολογία δύναται να επεκταθεί και σε σημαντικά μεγαλύτερα και πιο ετερογενή σύνολα δεδομένων, καλύπτοντας ευρύτερο φάσμα οργανισμών και τύπων νόσου. Η επιλογή ειδών που καλύπτουν διαφορετικές ταξινομικές ομάδες — από φυτά και μύκητες έως σπονδυλωτά — δεν αποσκοπούσε μόνο στην ποικιλομορφία των αλληλουχιών, αλλά και στη δυνατότητα εξέτασης του κατά πόσο οι αναπαραστάσεις ProtT5 συλλαμβάνουν χαρακτηριστικά αλληλεπίδρασης που είναι κοινά ανάμεσα σε εξελικτικά απομακρυσμένους οργανισμούς. Η ετερογένεια αυτή αποδείχθηκε ιδιαίτερα χρήσιμη τόσο για την αξιολόγηση της απόδοσης ανά είδος όσο και για

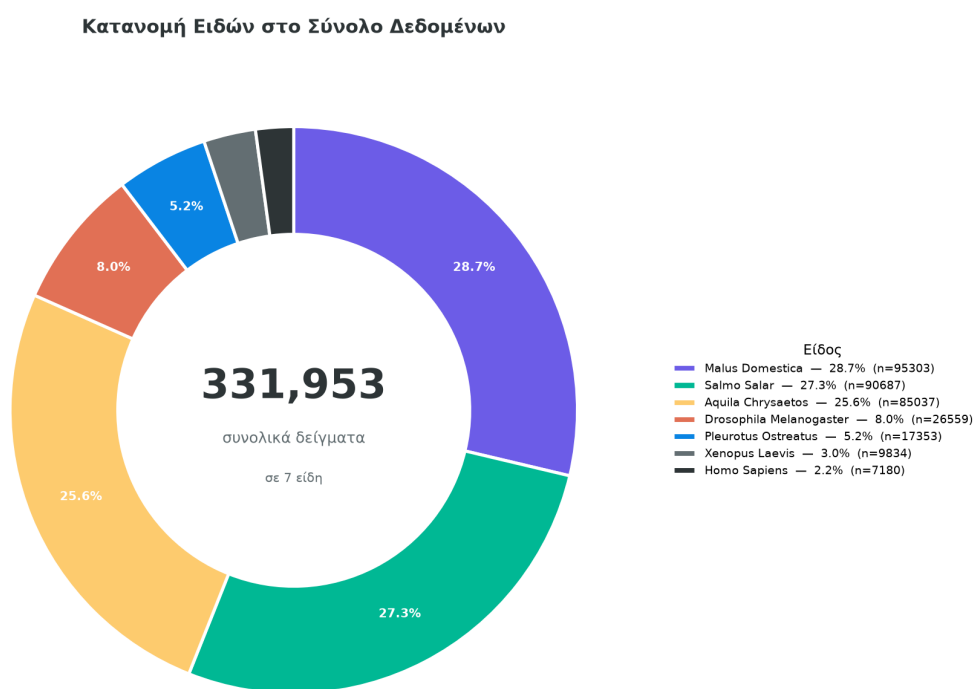
τον σχεδιασμό του πειράματος γενίκευσης σε άγνωστα είδη που παρουσιάζεται στην ενότητα 5.

Από τη βάση δεδομένων STRING αντλήθηκαν τα σχετικά πρωτεϊνικά δίκτυα αλληλεπίδρασης μαζί με τις αντίστοιχες πιθανότητες αλληλεπίδρασης μεταξύ των πρωτεϊνών, καθώς και οι διανυσματικές αναπαραστάσεις για κάθε πρωτεΐνη ξεχωριστά. Οι αναπαραστάσεις αυτές αποτελούν πυκνές διανυσματικές αναπαραστάσεις που κωδικοποιούν τις βιολογικές και λειτουργικές ιδιότητες κάθε πρωτεΐνης σε έναν συνεχή χώρο χαρακτηριστικών, επιτρέποντας υπολογιστικές μεθόδους να αξιοποιήσουν τη δομική και εξελικτική πληροφορία που εμπεριέχεται στις αμινοξικές αλληλουχίες [43]. Η χρήση τέτοιων προ-υπολογισμένων αναπαραστάσεων, αντί της απευθείας επεξεργασίας των ακατέργαστων αλληλουχιών κατά τη διάρκεια της εκπαίδευσης, μειώνει σημαντικά το υπολογιστικό κόστος και επιτρέπει την ταχύτερη πειραματική διερεύνηση διαφορετικών αρχιτεκτονικών σύγκρισης, χωρίς να απαιτείται επανάληψη της χρονοβόρας διαδικασίας κωδικοποίησης σε κάθε νέα δοκιμή.

Από τα δίκτυα αλληλεπιδράσεων που αποκτήθηκαν από τη βάση δεδομένων STRING ([string-db.org](http://string-db.org)), εξαγάγαμε πάνω από 2 εκατομύρια μοναδικά ζεύγη πρωτεϊνών. Η STRING αποτελεί σήμερα μία από τις πληρέστερες βάσεις δεδομένων πρωτεϊνικών αλληλεπιδράσεων, συγκεντρώνοντας και ενοποιώντας στοιχεία από πολλαπλές πηγές: πειραματικά δεδομένα, εξόρυξη κειμένου από επιστημονική βιβλιογραφία, υπολογιστικές προβλέψεις βασισμένες σε συν-έκφραση γονιδίων και γονιδιωματικό πλαίσιο, καθώς και βάσεις δεδομένων γνωστών μοριακών μονοπατιών [44]. Αφαιρέθηκαν όλα τα ζεύγη πρωτεϊνών στα οποία η πιθανότητα αλληλεπίδρασης κυμαινόταν μεταξύ 0.2 και 0.8, καθώς η αβεβαιότητα ως προς την ύπαρξη ή μη αλληλεπίδρασης τα καθιστούσε ακατάλληλα για εκπαίδευση και αξιολόγηση. Στα εναπομείναντα ζεύγη αποδόθηκαν οι κατηγορίες «μη αλληλεπίδραση» για πιθανότητα μικρότερη ή ίση του 0.2 και «αλληλεπίδραση» για πιθανότητα μεγαλύτερη ή ίση του 0.8. Η επιλογή αυτών των αυστηρών ορίων αποσκοπούσε στη διασφάλιση ότι τα δείγματα εκπαίδευσης και αξιολόγησης αντιπροσωπεύουν με σαφήνεια τις δύο κλάσεις, αποφεύγοντας την εισαγωγή θορύβου από ζεύγη πρωτεϊνών των οποίων η κατάσταση αλληλεπίδρασης παραμένει αμφίβολη βάσει των διαθέσιμων στοιχείων. Η στρατηγική αυτή, αν και μειώνει το μέγεθος του τελικού συνόλου δεδομένων σε σχέση με τη χρήση όλων των διαθέσιμων ζευγών, αναμένεται να βελτιώσει την ποιότητα του εκπαιδευτικού σήματος και, κατ' επέκταση, την αξιοπιστία των συμπερασμάτων που εξάγονται από τα μοντέλα.

Στην συνέχεια εφαρμόστηκε μια διαδικασία αφαίρεσης διπλότυπων όπου υπολογίστηκε η ομοιότητα μεταξύ των διανυσματικών αναπαραστάσεων των πρωτεϊνών, σε δύο διαδοχικά στάδια: αρχικά πραγματοποιήθηκε αφαίρεση διπλότυπων εντός κάθε είδους, κατά την οποία για κάθε πρωτεΐνη αναζητούνταν οι 100 πλησιέστεροι γείτονές της βάσει συνημιτονικής ομοιότητας και όσες εμφάνιζαν

ομοιότητα μεγαλύτερη ή ίση του ορίου 0.90 χαρακτηρίζονταν ως διπλότυπα και αφαιρούνταν, διατηρώντας μόνο μία αντιπροσωπευτική εμφάνιση ανά ομάδα παρόμοιων ακολουθιών. Στη συνέχεια, οι πρωτεΐνες που επιβίωναν από το πρώτο στάδιο υποβάλλονταν σε αφαίρεση διπλότυπων μεταξύ ειδών, συγκρινόμενες με το σύνολο των μοναδικών πρωτεϊνών που είχαν ήδη διατηρηθεί από όλα τα προηγούμενως επεξεργασμένα αρχεία, ώστε να αποφευχθεί η επανεμφάνιση της ίδιας πρωτεΐνης σε διαφορετικά αρχεία, με αποτέλεσμα η τελική συλλογή να περιλαμβάνει αποκλειστικά μοναδικές πρωτεΐνες τόσο σε επίπεδο εντός όσο και μεταξύ των αρχείων. Το βήμα αυτό ήταν αναγκαίο για την αφαίρεση παρόμοιων πρωτεϊνών, ώστε να αποφευχθεί η πιθανή διαρροή πληροφορίας. Στο τελικό σύνολο δεδομένων έμειναν 331,953 ζευγάρια πρωτεϊνών. Η διαδικασία αυτή, αν και υπολογιστικά απαιτητική λόγω του μεγάλου αριθμού συγκρίσεων ομοιότητας που έπρεπε να πραγματοποιηθούν, κρίθηκε απαραίτητη προκειμένου να αποφευχθεί η τεχνητή διόγκωση της απόδοσης των μοντέλων που θα προέκυπτε αν σχεδόν πανομοιότυπες πρωτεΐνες εμφανίζονταν ταυτόχρονα στα σύνολα εκπαίδευσης και δοκιμής. Με τον τρόπο αυτό, οποιαδήποτε αναφερόμενη επίδοση των μοντέλων αντικατοπτρίζει πραγματική ικανότητα γενίκευσης και όχι απλή απομνημόνευση παρόμοιων αλληλουχιών.

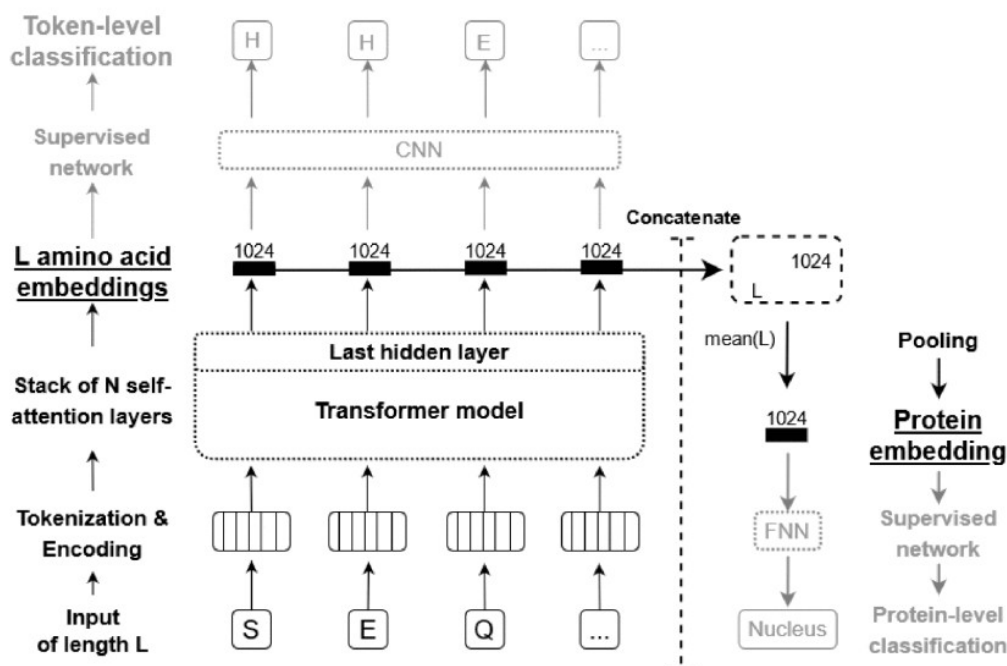


**Εικόνα 3.1.** Λεπτομερής σύνθεση του σετ δεδομένων

### 3.1 Πρωτεϊνικές αναπαραστάσεις

Για την παραγωγή των αναπαραστάσεων αυτών χρησιμοποιήθηκε το μοντέλο ProtT5, ένα γλωσσικό μοντέλο πρωτεϊνών που εδράζεται στην αρχιτεκτονική T5

(Text-to-Text Transfer Transformer) και εκπαιδεύτηκε αρχικά στη βάση δεδομένων BFD (Big Fantastic Database), που περιλαμβάνει πάνω από 2,1 δισεκατομμύρια μετα-γονιδιωματικές πρωτεϊνικές αλληλουχίες, και στη συνέχεια βελτιστοποιήθηκε περαιτέρω στο σύνολο δεδομένων UniRef50 [43]. Το UniRef50 αντιπροσωπεύει ένα σύνολο πρωτεϊνικών αλληλουχιών από ευρύ φάσμα οργανισμών, ομαδοποιημένων σε 50% ομοιότητα αλληλουχίας, με στόχο την αντιπροσωπευτική και μη πλεονάζουσα κάλυψη του πρωτεϊνικού χώρου [45]. Σε αντίθεση με τα παραδοσιακά χαρακτηριστικά αλληλουχίας, τα γλωσσικά μοντέλα πρωτεϊνών μαθαίνουν πλαίσιο-εξαρτώμενες αναπαραστάσεις, λαμβάνοντας υπόψη τη συνολική δομή της αλληλουχίας και όχι μόνο μεμονωμένα αμινοξέα [46]. Για κάθε πρωτεϊνική αλληλουχία, το μοντέλο ProtT5 παρήγαγε αρχικά μια αναπαράσταση για κάθε αμινοξύ ξεχωριστά — ένα διάνυσμα 1024 διαστάσεων ανά κατάλοιπο — και στη συνέχεια υπολόγισε τον μέσο όρο όλων αυτών των επιμέρους αναπαραστάσεων [46]. Με αυτή την προσέγγιση, κάθε πρωτεΐνη αντιστοιχίστηκε σε ένα διάνυσμα σταθερής διάστασης 1024 στοιχείων, ανεξάρτητα από το μήκος της αλληλουχίας της, επιτρέποντας έτσι την ομοιόμορφη αναπαράσταση και επεξεργασία πρωτεϊνών με διαφορετικό αριθμό αμινοξέων [45]. Αυτή η τεχνική μέσης συγκέντρωσης διατηρεί τη συνολική πληροφορία της αλληλουχίας σε έναν συμπαγή φορέα, που στη συνέχεια μπορεί να χρησιμοποιηθεί ως είσοδος σε μοντέλα μηχανικής μάθησης για την πρόβλεψη αλληλεπιδράσεων [47]. Το βασικό πλεονέκτημα αυτής της επιλογής είναι η υπολογιστική απλότητα και η δυνατότητα άμεσης σύγκρισης πρωτεϊνών διαφορετικού μήκους μέσω κοινών τεχνικών όπως το γινόμενο Hadamard ή η ευκλείδεια/συνημιτονική απόσταση, χωρίς να απαιτούνται μηχανισμοί ευθυγράμμισης ή περικοπής αλληλουχιών. Το ενδεχόμενο κόστος αυτής της απλοποίησης, δηλαδή η πιθανή απώλεια πληροφορίας που σχετίζεται με συγκεκριμένες θέσεις καταλοίπων, συζητείται εκτενέστερα στην ενότητα της Συζήτησης.



**Εικόνα 3.2.** Η διαδικασία υπολογισμού των αναπαραστάσεων για τις πρωτεΐνες (πηγή: Elnaggar et al. 2025)

Η αρχιτεκτονική του T5 βασίζεται στο κλασικό μοντέλο Μετασχηματιστή που εισήγαγαν οι Vaswani et al. (2017), με τη διαφορά ότι διατηρεί και τα δύο βασικά συστατικά του: τον κωδικοποιητή και τον αποκωδικοποιητή. Ο κωδικοποιητής δέχεται ως είσοδο την αλληλουχία αμινοξέων με τα αποκρυμμένα τμήματα και την επεξεργάζεται μέσω στοίβας επιπέδων αυτο-προσοχής, παράγοντας μία πυκνή διανυσματική αναπαράσταση για κάθε θέση της αλληλουχίας. Κάθε επίπεδο αυτο-προσοχής υπολογίζει σχέσεις μεταξύ όλων των θέσεων της αλληλουχίας ταυτόχρονα, επιτρέποντας στο μοντέλο να συλλαμβάνει τόσο τοπικές όσο και μακράς εμβέλειας εξαρτήσεις μεταξύ αμινοξέων. Αξιοσημείωτο χαρακτηριστικό του T5 είναι η χρήση σχετικής κωδικοποίησης θέσης αντί της απόλυτης, γεγονός που το καθιστά πιο ευέλικτο στη διαχείριση αλληλουχιών μεταβλητού μήκους. Η ιδιότητα αυτή είναι ιδιαίτερα σημαντική στο πλαίσιο των πρωτεϊνικών αλληλουχιών, καθώς αυτές ποικίλλουν σημαντικά σε μήκος —από λίγες δεκάδες έως αρκετές χιλιάδες αμινοξέα— χωρίς να απαιτείται επανεκπαίδευση ή προσαρμογή του μοντέλου για κάθε νέο εύρος μήκους.

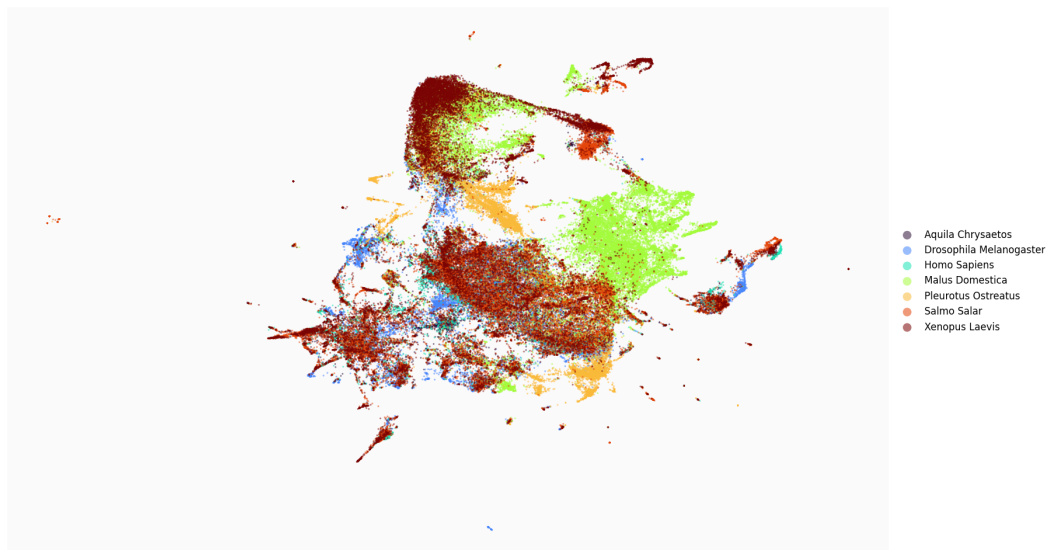
Ο αποκωδικοποιητής, από την πλευρά του, αποτελείται από δύο τύπους επιπέδων προσοχής: πρώτον, ένα επίπεδο αυτοπροσοχής με μάσκα, το οποίο επεξεργάζεται την αλληλουχία εξόδου που έχει ήδη παραχθεί, και δεύτερον, ένα επίπεδο διασταυρούμενης προσοχής, το οποίο «κοιτάζει» τις αναπαραστάσεις του κωδικοποιητή προκειμένου να αντλήσει πληροφορία από την αλληλουχία εισόδου. Με αυτόν τον τρόπο, ο αποκωδικοποιητής παράγει αυτο-παλινδρομικά τα αποκρυμμένα αμινοξέα, ένα προς ένα, λαμβάνοντας υπόψη τόσο την κωδικοποιημένη αλληλουχία εισόδου όσο και τα ήδη παραχθέντα τόκενς εξόδου. Κάθε επίπεδο τόσο στον κωδικοποιητή όσο και στον αποκωδικοποιητή περιλαμβάνει επίσης ένα δίκτυο πρόσθιας τροφοδότησης με

μη γραμμική ενεργοποίηση, καθώς και κανονικοποίηση στρώματος και υπολειπόμενες συνδέσεις, οι οποίες διευκολύνουν την εκπαίδευση βαθύτερων αρχιτεκτονικών. Στην παρούσα εργασία, ο αποκωδικοποιητής δεν αξιοποιήθηκε άμεσα, καθώς μόνο οι αναπαραστάσεις του κωδικοποιητή χρησιμοποιήθηκαν για την παραγωγή των διανυσμάτων χαρακτηριστικών ανά πρωτεΐνη· ωστόσο, η προ-εκπαίδευση ολόκληρης της αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή, μέσω της εργασίας ανασύστασης αποκρυμμένων τμημάτων, είναι εκείνη που εξασφαλίζει την ποιότητα και την πληρότητα των τελικών αναπαραστάσεων του κωδικοποιητή.

Στο πλαίσιο του ProtTrans, εκπαιδεύτηκαν δύο εκδόσεις του T5: το ProtT5-XL-U50, με περίπου 3 δισεκατομμύρια παραμέτρους, και το ProtT5-XXL, με περίπου 11 δισεκατομμύρια παραμέτρους. Το μέγεθος αυτό καθιστά το ProtT5 το μεγαλύτερο μοντέλο της εργασίας, απαιτώντας αντίστοιχα και τους μεγαλύτερους υπολογιστικούς πόρους για την εκπαίδευσή του. Παρά το υψηλό υπολογιστικό κόστος, το ProtT5 αναδείχθηκε ως η πιο ισχυρή αρχιτεκτονική της εργασίας, παράγοντας αναπαραστάσεων που επέτρεψαν την επίτευξη κορυφαίων επιδόσεων σε εργασίες πρόβλεψης δομής και λειτουργίας πρωτεϊνών, ακόμα και με σχετικά απλά μοντέλα πρόβλεψης επί κεφαλής. Για αυτόν τον λόγο, αυτές οι αναπαραστάσεις επιλέχθηκαν για την εκπαίδευση του δικού μας μοντέλου. Η χρήση ενός ήδη προ-εκπαιδευμένου μοντέλου αυτής της κλίμακας επιτρέπει την αξιοποίηση βιολογικής γνώσης που έχει αποσταχθεί από δισεκατομμύρια αλληλουχίες, χωρίς να απαιτείται πρόσβαση σε αντίστοιχους υπολογιστικούς πόρους κατά τη φάση εκπαίδευσης των δικών μας ταξινομητών, καθώς οι αναπαραστάσεις υπολογίζονται μία φορά και επαναχρησιμοποιούνται σε όλα τα πειράματα.

Υλοποιήθηκε διάγραμμα διασποράς με τις πρωτεΐνες του τελικού συνόλου δεδομένων (Εικόνα 3.3). Αρχικά εφαρμόστηκε η σταδιακή ανάλυση κύριων συνιστωσών για τη μείωση της διαστατικότητας των διανυσμάτων χαρακτηριστικών από τις αρχικές διαστάσεις στις 50 σημαντικότερες [48]. Στη συνέχεια εφαρμόστηκε η μέθοδος ομοιόμορφης προσέγγισης και προβολής πολλαπλοτήτων για την προβολή των δεδομένων σε δύο διαστάσεις. Στο διάγραμμα, κάθε σημείο αντιστοιχεί σε μία πρωτεΐνη, ενώ το χρώμα του δηλώνει το είδος προέλευσής της, επιτρέποντας την οπτική αξιολόγηση της ομαδοποίησης και του διαχωρισμού των πρωτεϊνών μεταξύ των διαφορετικών ειδών. Στο τελικό διάγραμμα παρατηρείται σημαντική αλληλεπικάλυψη μεταξύ των ειδών, ωστόσο υπάρχουν και περιοχές που κυριαρχούνται σχεδόν αποκλειστικά από πρωτεΐνες ενός μόνο είδους, υποδεικνύοντας έτσι ότι υπάρχει ποικιλομορφία στη σύσταση των πρωτεϊνών μεταξύ των ειδών. Η μερική αυτή αλληλεπικάλυψη είναι αναμενόμενη, καθώς πολλές πρωτεΐνες επιτελούν εξελικτικά συντηρημένες λειτουργίες που είναι κοινές σε διαφορετικούς οργανισμούς, ενώ οι πιο απομονωμένες περιοχές πιθανότατα αντιστοιχούν σε πρωτεΐνες με εξειδικευμένο, είδοκεντρικό ρόλο. Η οπτικοποίηση αυτή λειτουργεί συμπληρωματικά ως ποιοτικός έλεγχος του συνόλου δεδομένων, επιβεβαιώνοντας ότι οι αναπαραστάσεις ProtT5

συλλαμβάνουν ουσιαστικές βιολογικές διαφοροποιήσεις και δεν παράγουν τυχαίο ή άνευ νοήματος διανυσματικό χώρο.



**Εικόνα 3.3.** Διάγραμμα διασποράς με διαφορετικό χρώμα για κάθε είδος

## 3.2 Πιθανότητες αλληλεπίδρασης

Η βάση δεδομένων STRING παρέχει πολλαπλές κατηγορίες βαθμολογιών αλληλεπίδρασης μεταξύ πρωτεϊνών. Συγκεκριμένα, διαθέτει βαθμολογίες βασισμένες σε πειραματικά δεδομένα, σε εξόρυξη κειμένου, σε γειτνίαση γονιδίων, σε συν-έκφραση, σε φυλογενετική συν-εμφάνιση και η συνδιαστική βαθμολογία που ενσωματώνει όλες τις διαθέσιμες πηγές σε μια ενιαία ποσοτικοποίηση [44]. Στην παρούσα υλοποίηση επιλέχθηκε η αποκλειστική χρήση των πειραματιών βαθμολογιών, καθώς στηρίζονται αποκλειστικά σε βιοχημικά και μοριακά πειράματα, προσφέροντας το υψηλότερο επίπεδο αξιοπιστίας. Πειραματικές μέθοδοι όπως η ανοσοκατακρήμνιση, η αντοχή δύο-υβριδίων ζύμης και η φασματομετρία μάζας συγγενούς εκχύλισης παρέχουν άμεσες βιοχημικές αποδείξεις για φυσικές αλληλεπιδράσεις, σε αντίθεση με υπολογιστικές ή κειμενομετρικές προσεγγίσεις που ενέχουν μεγαλύτερο βαθμό αβεβαιότητας [49]. Είναι ωστόσο σημαντικό να διευκρινιστεί ότι η συγκεκριμένη βαθμολογία δεν αντικατοπτρίζει ούτε την ισχύ ούτε τη συχνότητα μιας αλληλεπίδρασης, αλλά εκφράζει τον βαθμό εμπιστοσύνης του σετ δεδομένων ως προς την ύπαρξη και την ικανότητα αλληλεπίδρασης μεταξύ των δύο πρωτεϊνών [44]. Η επιλογή αυτής της συγκεκριμένης κατηγορίας βαθμολογίας, αντί της συνδιαστικής, έγινε συνειδητά ώστε τα μοντέλα να εκπαιδευτούν και να αξιολογηθούν με βάση ετικέτες που στηρίζονται σε επαληθεύσιμη πειραματική τεκμηρίωση και όχι σε συμπεράσματα εξαγόμενα από αλγοριθμική εξόρυξη κειμένου ή στατιστικές συσχετίσεις γονιδιωματικού πλαισίου, μειώνοντας έτσι τον κίνδυνο εισαγωγής συστηματικού σφάλματος στη διαδικασία εκμάθησης.

## 4 Τα Μοντέλα

Έγινε υλοποίηση και εκπαίδευση δύο βασικών μοντέλων: Αντίληπτρο Βαθιάς Μάθησης και Σιαμέζικος Ταξινομητής. Για τη διασφάλιση της σταθερότητας κατά την εκπαίδευση και την επιτάχυνση της σύγκλισης των μοντέλων, έγινε εκτενής χρήση κανονικοποίησης επιπέδου στα γραμμικά επίπεδα του δικτύου. Η τεχνική αυτή, η οποία αποτελεί πλέον κεντρικό συστατικό της σύγχρονης αρχιτεκτονικής βαθιών νευρωνικών δικτύων, λειτουργεί κανονικοποιώντας τις εξόδους των νευρώνων κάθε επιπέδου ώστε να έχουν μηδενική μέση τιμή και μοναδιαία διακύμανση, υπολογισμένες όμως κατά μήκος των χαρακτηριστικών κάθε μεμονωμένου δείγματος και όχι κατά μήκος της παρτίδας. Η ιδιότητα αυτή καθιστά την κανονικοποίηση ανεξάρτητη από το μέγεθος της παρτίδας και τη στατιστική της, σε αντίθεση με την κανονικοποίηση παρτίδας, προσφέροντας σταθερή συμπεριφορά τόσο κατά την εκπαίδευση όσο και κατά την αξιολόγηση του μοντέλου. Η σταθεροποίηση αυτή των ενδιάμεσων αναπαραστάσεων επιτρέπει τη χρήση σημαντικά υψηλότερων ρυθμών μάθησης χωρίς τον κίνδυνο αποκλίσεων ή αστάθειας κατά την ενημέρωση των παραμέτρων, μειώνοντας παράλληλα την εξάρτηση του δικτύου από προσεκτική αρχικοποίηση των βαρών. Επιπλέον, η κανονικοποίηση επιπέδου εισάγει δύο εκπαιδευσιμες παραμέτρους ανά χαρακτηριστικό, τη μετατόπιση και την κλίμακα, οι οποίες επιτρέπουν στο δίκτυο να αποκαταστήσει, εφόσον αυτό κρίνεται αναγκαίο, την εκφραστική ισχύ που ενδέχεται να χαθεί από την κανονικοποίηση, προσφέροντας έτσι ευελιξία στη μάθηση βέλτιστων αναπαραστάσεων [50]. Η επιλογή κανονικοποίησης επιπέδου αντί κανονικοποίησης παρτίδας κρίθηκε πιο κατάλληλη για το συγκεκριμένο πρόβλημα, καθώς η απόδοσή της δεν εξαρτάται από το μέγεθος της παρτίδας ούτε απαιτεί τη διατήρηση κινούμενων στατιστικών κατά την αξιολόγηση, στοιχεία που καθιστούν την εκπαίδευση πιο σταθερή όταν χρησιμοποιούνται ανισόρροπες παρτίδες με στάθμιση κλάσεων, όπως συνέβη στην παρούσα εργασία.

Πέραν της κανονικοποίησης επιπέδου, και αποκλειστικά κατά τη φάση εκπαίδευσης του μοντέλου, χρησιμοποιήθηκαν επίσης επίπεδα dropout. Κατά τη λειτουργία του dropout, κάθε νευρώνας του επιπέδου απενεργοποιείται τυχαία με πιθανότητα  $p$  σε κάθε βήμα εκπαίδευσης, αναγκάζοντας το δίκτυο να αναπτύξει πλεονάζουσες και κατανεμημένες αναπαραστάσεις που δεν εξαρτώνται υπερβολικά από μεμονωμένους νευρώνες ή συγκεκριμένα μοτίβα ενεργοποίησης [51]. Η στρατηγική αυτή λειτουργεί ως ισχυρός ρυθμιστικός παράγοντας, αποτρέποντας την υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης — ένα ιδιαίτερα κρίσιμο ζήτημα στην περίπτωση πρωτεϊνικών δεδομένων, όπου ο χώρος των αλληλουχιών είναι απεριόριστος και οι διαθέσιμες εκπαιδευτικές αλληλεπιδράσεις συχνά περιορισμένες. Η συνδυαστική εφαρμογή κανονικοποίησης επιπέδου και dropout διαμορφώνει ένα ισορροπημένο πλαίσιο βελτιστοποίησης, εντός του οποίου τα μοντέλα εκπαιδεύονται να εξάγουν γενικεύσιμες αναπαραστάσεις υψηλής πληροφοριακής αξίας. Η τιμή της πιθανότητας

$p$  επιλέχθηκε διαφορετική ανά επίπεδο, με υψηλότερες τιμές στα αρχικά, υψηλότερης διαστατικότητας επίπεδα και σταδιακά χαμηλότερες προς το τέλος του δικτύου, ώστε να αποφευχθεί η υπερβολική απώλεια πληροφορίας στα τελικά στάδια όπου η αναπαράσταση είναι ήδη συμπυκνωμένη και ευαίσθητη σε τυχαία απενεργοποίηση νευρώνων.

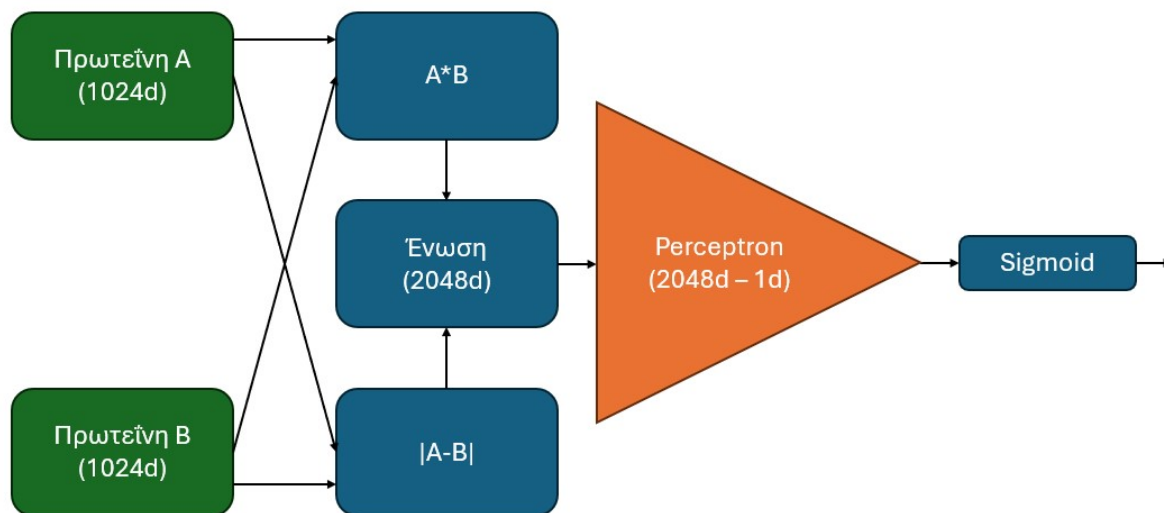
Τα μοντέλα εκπαιδεύτηκαν για 30 εποχές με παρτίδες μεγέθους 512 το κάθε ένα. Ο αλγόριθμος βελτιστοποίησης που χρησιμοποιήθηκε είναι ο Adam (Adaptive Moment Estimation). Οι Kingma & Ba (2014) παρουσίασαν τον Adam ως έναν αλγόριθμο πρώτης τάξης για βελτιστοποίηση στοχαστικών αντικειμενικών συναρτήσεων, βασισμένο σε προσαρμοστικές εκτιμήσεις των ροπών χαμηλής τάξης. Ο αλγόριθμος είναι υπολογιστικά αποδοτικός, έχει χαμηλές απαιτήσεις μνήμης και είναι κατάλληλος για προβλήματα με θορυβώδεις ή αραιές βαθμίδες [52]. Δεν εφαρμόστηκε πρόσθετη κανονικοποίηση μέσω απομείωσης βαρών στον βελτιστοποιητή· η γενίκευση του μοντέλου διασφαλίστηκε αντ' αυτού μέσω του συνδυασμού της κανονικοποίησης επιπέδου, του dropout και της στάθμισης των κλάσεων στη συνάρτηση απώλειας. Συγκεκριμένα, σε κάθε παρτίδα εφαρμόστηκε στάθμιση αντιστρόφως ανάλογη της συχνότητας κάθε κλάσης, ώστε η συνεισφορά των αλληλεπιδρώντων και των μη αλληλεπιδρώντων ζευγών στην απώλεια να εξισορροπείται ανεξάρτητα από την αναλογία τους στο σύνολο δεδομένων. Αυτό το βήμα ήταν αναγκαίο λόγω της μεγαλύτερης παρουσίας μη αλληλεπιδρώντων ζευγών στο σύνολο δεδομένων, η οποία θα οδηγούσε διαφορετικά σε μεροληπτική εκπαίδευση υπέρ της αρνητικής κλάσης. Το μέγεθος παρτίδας των 512 δειγμάτων επιλέχθηκε ως συμβιβασμός μεταξύ σταθερότητας της εκτίμησης της κλίσης και υπολογιστικής αποδοτικότητας, ενώ ο αριθμός των 30 εποχών κρίθηκε επαρκής βάσει της παρατηρούμενης σύγκλισης της συνάρτησης απώλειας στο σύνολο επικύρωσης, χωρίς σημάδια σημαντικής υπερπροσαρμογής στα τελευταία στάδια της εκπαίδευσης.

## 4.1 Αντίληπτρο Βαθιάς Μάθισης

Η πρώτη αρχιτεκτονική που δοκιμάστηκε ήταν ένα πολυεπίπεδο αντίληπτρο (Εικόνα 4.1). Επειδή η αλληλεπίδραση των πρωτεϊνών είναι αμφίδρομη, έπρεπε να βρεθεί μια συνδυαστική στρατηγική που ενώνει τις δύο πρωτεΐνες σε μία αναπαράσταση η οποία παραμένει αμετάβλητη ανεξάρτητα από τη σειρά εισαγωγής τους. Για τον σκοπό αυτό χρησιμοποιήθηκαν το γινόμενο Hadamard και η απόλυτη διαφορά των δύο αναπαραστάσεων. Τα προκύπτοντα διανύσματα συνενώνονται σειριακά, δημιουργώντας έναν τανυστή 2048 διαστάσεων που εισάγεται στο αντίληπτρο. Το αντίληπτρο αποτελείται από 4 κρυφά επίπεδα γραμμικού μετασχηματισμού, που ξεκινούν από 2048 διαστάσεις και συρρικνώνονται σταδιακά, περνώντας από τα 1024 και τα 512 στα 256, με αποτέλεσμα το τελευταίο κρυφό επίπεδο να έχει 64 διαστάσεις. Κάθε επίπεδο εφαρμόζει κανονικοποίηση επιπέδου και συνάρτηση ενεργοποίησης

GELU πριν τη μεταφορά των αποτελεσμάτων στο επόμενο επίπεδο. Το τελικό επίπεδο εξόδου μετασχηματίζει την αναπαράσταση σε βαθμωτή τιμή μέσω σιγμοειδούς συνάρτησης. Η τιμή αυτή συγκρίνεται με την ετικέτα αλληλεπίδρασης του συνόλου δεδομένων, όπου τιμές άνω του 0.5 κωδικοποιούνται ως 1 και οι υπόλοιπες ως 0. Βάσει αυτής της σύγκρισης πραγματοποιείται η βελτιστοποίηση του μοντέλου. Η σταδιακή συρρίκνωση της διαστατικότητας ανά επίπεδο ακολουθεί μια κοινή πρακτική στον σχεδιασμό βαθιών νευρωνικών δικτύων, η οποία αναγκάζει το μοντέλο να συμπυκνώνει προοδευτικά την πληροφορία σε όλο και πιο συμπαγείς αναπαραστάσεις, διατηρώντας μόνο τα χαρακτηριστικά που είναι πιο κρίσιμα για την τελική απόφαση ταξινόμησης.

Η επιλογή δύο διαφορετικών πράξεων αντί για μίας αποσκοπεί στη μεγιστοποίηση της πληροφορίας που διατηρείται κατά τον συνδυασμό των δύο αναπαραστάσεων. Το γινόμενο Hadamard αναδεικνύει τα κοινά ισχυρά χαρακτηριστικά των δύο πρωτεϊνών: όταν και τα δύο διανύσματα παρουσιάζουν υψηλές τιμές σε μια συγκεκριμένη διάσταση, το γινόμενο ενισχύεται, ενώ όταν μία εκ των δύο τιμών είναι χαμηλή, το αποτέλεσμα καταστέλλεται. Με αυτόν τον τρόπο, λειτουργεί ως μέτρο ομοιότητας και συμβατότητας μεταξύ των δύο αναπαραστάσεων. Αντίθετα, η απόλυτη διαφορά εστιάζει στις αποκλίσεις: καταγράφει τις διαστάσεις εκείνες όπου οι δύο πρωτεΐνες διαφέρουν σημαντικά, παρέχοντας πληροφορία για τη συμπληρωματικότητα ή την ασυμβατότητά τους. Η αναγκαιότητα χρήσης και των δύο πράξεων ταυτόχρονα γίνεται εμφανής αν αναλογιστεί κανείς ότι διαφορετικά ζεύγη διανυσμάτων μπορούν να παράγουν πανομοιότυπα αποτελέσματα με μία μόνο πράξη, ενώ να διαφέρουν στην άλλη — για παράδειγμα,  $1 \times 4 = 2 \times 2$  αλλά  $|1 - 4| \neq |2 - 2|$ , ή αντίστοιχα  $|3 - 5| = |4 - 2|$  αλλά  $3 \times 5 \neq 4 \times 2$ . Η ταυτόχρονη χρήση και των δύο πράξεων μειώνει δραστικά αυτή την αμφισημία, διατηρώντας μια πλουσιότερη και πιο διακριτική αναπαράσταση της σχέσης μεταξύ των δύο πρωτεϊνών. Επιπλέον, και οι δύο πράξεις είναι εξ ορισμού συμμετρικές ως προς τη σειρά των εισόδων, γεγονός που εξασφαλίζει ότι η πρόβλεψη του μοντέλου παραμένει αμετάβλητη ανεξάρτητα από το ποια από τις δύο πρωτεΐνες δηλώνεται πρώτη — μια ιδιότητα απαραίτητη δεδομένου ότι η βιολογική αλληλεπίδραση δύο πρωτεϊνών δεν έχει εγγενή κατεύθυνση.

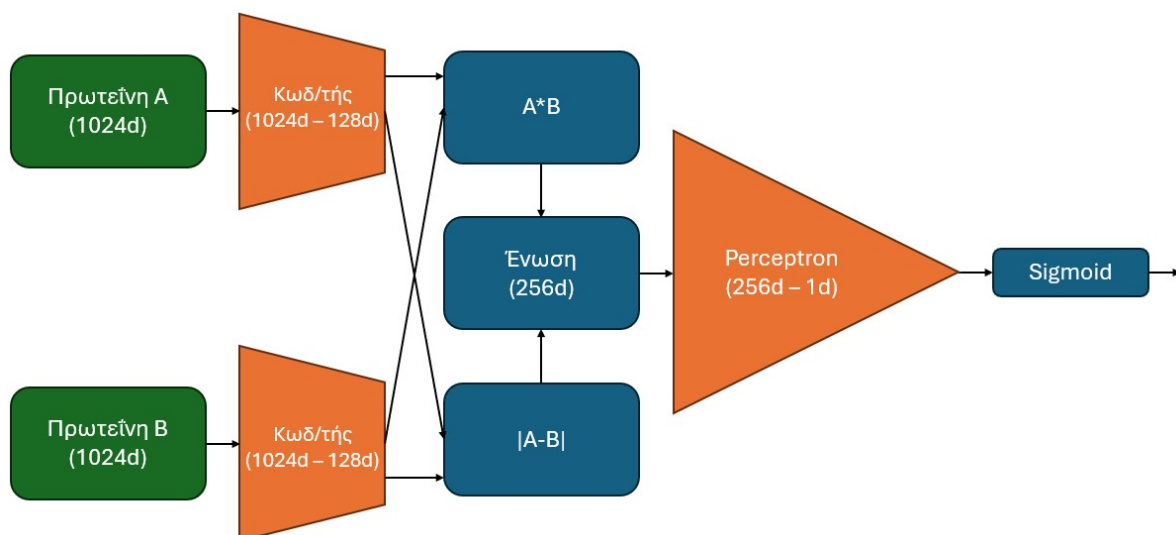


**Εικόνα 4.1.** Η βασική αρχιτεκτονική του Αντίληπτρου Βαθιάς Μάθησης

Για τη βελτιστοποίηση του μοντέλου χρησιμοποιήθηκε η συνάρτηση απώλειας διασταυρούμενης εντροπίας, η οποία αποτελεί την τυπική επιλογή για προβλήματα δυαδικής ταξινόμησης [53]. Η συνάρτηση αυτή ποσοτικοποιεί την απόκλιση μεταξύ της προβλεπόμενης πιθανότητας αλληλεπίδρασης και της πραγματικής ετικέτας, τιμωρώντας δυσανάλογα τις προβλέψεις που απέχουν πολύ από την αληθινή τιμή. Με αυτόν τον τρόπο, το μοντέλο εκπαιδεύεται να εκχωρεί υψηλές πιθανότητες στα αλληλεπιδρώντα ζεύγη πρωτεϊνών και χαμηλές στα μη αλληλεπιδρώντα, οδηγώντας σε μια βαθμονομημένη και ερμηνεύσιμη έξοδο. Κατά τη διάρκεια της εκπαίδευσης εφαρμόστηκε dropout στα τρία πρώτα κρυφά επίπεδα, στα οποία τιμές μηδενίζονταν τυχαία με πιθανότητα 0.3, 0.3 και 0.2 αντίστοιχα. Το dropout λειτουργεί ως τεχνική κανονικοποίησης: με την τυχαία απενεργοποίηση νευρώνων σε κάθε βήμα εκπαίδευσης, αναγκάζει το δίκτυο να μην εξαρτάται υπερβολικά από συγκεκριμένα χαρακτηριστικά ή συνδέσεις. Αποτέλεσμα αυτής της διαδικασίας είναι η ανάπτυξη πιο ισορροπημένων και ανθεκτικών αναπαραστάσεων, που γενικεύουν καλύτερα σε αόρατα δεδομένα. Οι υψηλότερες τιμές dropout στα δύο πρώτα επίπεδα δικαιολογούνται από τη μεγαλύτερη διαστατικότητα τους (1024 και 512 αντίστοιχα), η οποία αυξάνει τον κίνδυνο υπερπροσαρμογής, ενώ στο τρίτο επίπεδο (256 διαστάσεων) εφαρμόζεται χαμηλότερη τιμή. Στο τελευταίο κρυφό επίπεδο (64 διαστάσεων) δεν εφαρμόζεται dropout. Η εκπαίδευση διήρκεσε 30 εποχές με ρυθμό μάθησης  $10^{-3}$ . Ο συνδυασμός της σταθμισμένης απώλειας διασταυρούμενης εντροπίας με τη σταδιακή μείωση διαστατικότητας και τις τεχνικές κανονικοποίησης αποσκοπούσε στη δημιουργία ενός ισχυρού αλλά αρκετά απλού μοντέλου αναφοράς, έναντι του οποίου θα μπορούσε να αξιολογηθεί η πραγματική συνεισφορά της σιαμέζικης αρχιτεκτονικής που παρουσιάζεται στη συνέχεια.

## 4.2 Σιαμέζικος Ταξινομητής

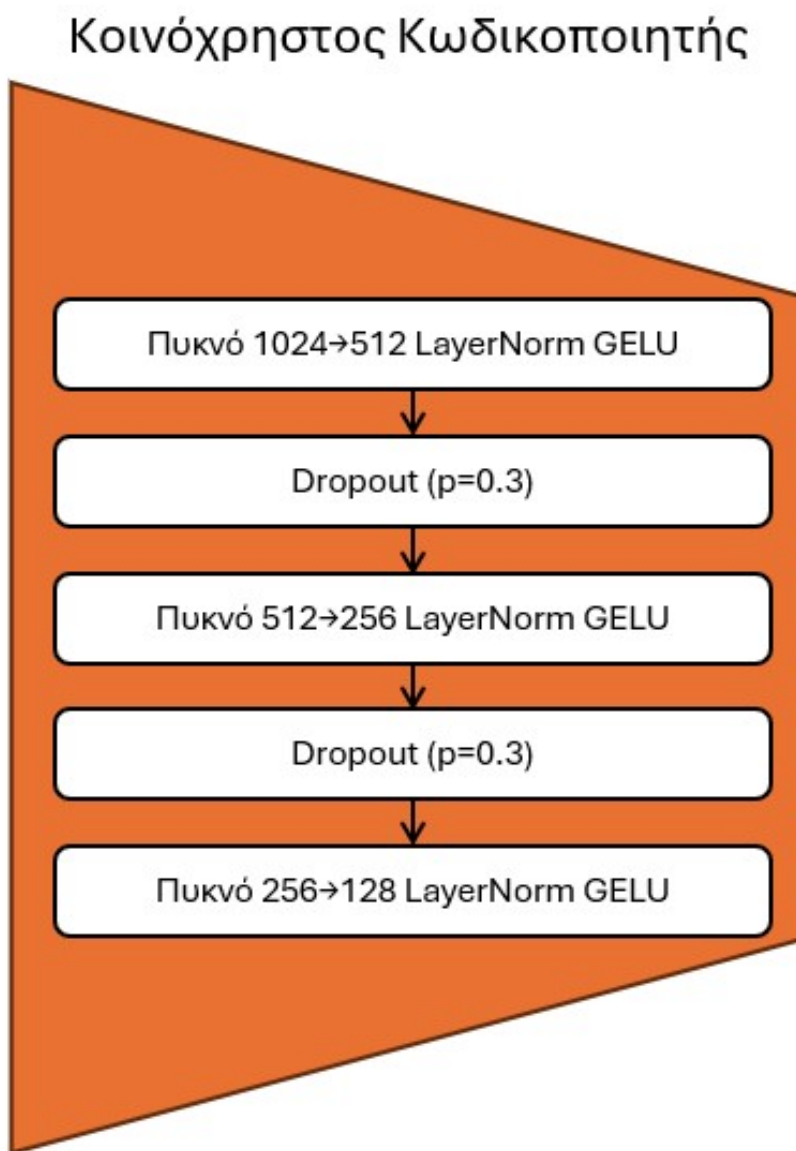
Η δεύτερη αρχιτεκτονική που δοκιμάστηκε ήταν ο Σιαμέζικος Διαδικός Ταξινομητής (Εικόνα 4.2). Τα Σιαμέζικα δίκτυα αποτελούν μια ειδική κατηγορία νευρωνικών αρχιτεκτονικών που σχεδιάστηκαν αρχικά για προβλήματα σύγκρισης και αντιστοίχισης ζευγών, και έκτοτε έχουν εφαρμοστεί με επιτυχία σε ένα ευρύ φάσμα εργασιών, από την αναγνώριση προσώπων και την ανίχνευση ομοιότητας εγγράφων έως την πρόβλεψη βιολογικών αλληλεπιδράσεων. Το χαρακτηριστικό γνώρισμα της αρχιτεκτονικής αυτής είναι η χρήση δύο παράλληλων υποδικτύων που μοιράζονται ακριβώς τα ίδια βάρη, επεξεργαζόμενα ανεξάρτητα τις δύο εισόδους και παράγοντας αναπαραστάσεις στον ίδιο λανθάνοντα χώρο, γεγονός που καθιστά τη σύγκριση μεταξύ τους άμεσα ουσιαστική και συνεκτική [54]. Η συγκεκριμένη προσέγγιση επιλέχθηκε προκειμένου να επιτευχθεί η αποτελεσματική σύγκριση και ο υπολογισμός της συνάφειας μεταξύ δύο πρωτεϊνικών οντοτήτων μέσα σε έναν κοινό χώρο αναπαράστασης, εξασφαλίζοντας ταυτόχρονα συμμετρία ως προς τη σειρά εισαγωγής των δύο πρωτεϊνών — μια ιδιότητα που αντικατοπτρίζει την εκ φύσεως συμμετρική φύση της βιολογικής αλληλεπίδρασης. Σε αντίθεση με το αντίληπτρο, όπου οι δύο αρχικές αναπαραστάσεις εισάγονται απευθείας στον συνδυαστικό μηχανισμό, ο σιαμέζικος ταξινομητής προβλέπει ένα ενδιάμεσο στάδιο κωδικοποίησης, κατά το οποίο κάθε πρωτεΐνη προβάλλεται πρώτα σε έναν νέο, εκμαθημένο χώρο χαρακτηριστικών πριν από τη σύγκρισή της με την άλλη — μια σχεδιαστική επιλογή που αναμένεται να ενισχύσει τη διαχωρισιμότητα των δύο κλάσεων.



Εικόνα 4.2. Η βασική αρχιτεκτονική του Σιαμέζικου Ταξινομητή

Επιπλέον, η κοινή χρήση βαρών μεταξύ των δύο παράλληλων κλάδων μειώνει σημαντικά τον συνολικό αριθμό των εκπαιδευσίμων παραμέτρων, περιορίζοντας τον κίνδυνο υπερπροσαρμογής και καθιστώντας το μοντέλο πιο αποδοτικό από

υπολογιστική άποψη, χωρίς να θυσιάζεται η εκφραστική του ικανότητα. Το δίκτυο δομείται σε δύο θεμελιώδεις πυλώνες: τον κωδικοποιητή των πρωτεϊνικών αλληλουχιών (Εικόνα 4.3), ο οποίος είναι υπεύθυνος για τη μετατροπή των αρχικών αριθμητικών αναπαραστάσεων κάθε πρωτεΐνης σε ένα πλούσιο σε πληροφορία χαρακτηριστικό διάνυσμα, και το υπολογιστικό τμήμα της πιθανότητας αλληλεπίδρασης, το οποίο αναλαμβάνει να συνδυάσει τις δύο κωδικοποιημένες αναπαραστάσεις και να εκτιμήσει την πιθανότητα ύπαρξης λειτουργικής σύνδεσης μεταξύ των δύο πρωτεϊνών. Ο διαχωρισμός αυτός σε δύο σαφώς οριοθετημένα τμήματα επιτρέπει την ανεξάρτητη ανάλυση της συνεισφοράς καθενός στην τελική απόδοση: ο κωδικοποιητής είναι υπεύθυνος για τη μετατροπή της «ακατέργαστης» πληροφορίας ProtT5 σε μια αναπαράσταση προσαρμοσμένη στο πρόβλημα της πρόβλεψης αλληλεπίδρασης, ενώ το τμήμα σύγκρισης αναλαμβάνει αποκλειστικά τη μοντελοποίηση της σχέσης μεταξύ των δύο ήδη προσαρμοσμένων αναπαραστάσεων.



**Εικόνα 4.3.** Η αρχιτεκτονική του κοινόχρηστου κωδικοποιητή

Αφού παραχθούν τα δύο χαρακτηριστικά διανύσματα 128 διαστάσεων, υπολογίζονται το γινόμενο Hadamard και η απόλυτη διαφορά τους. Τα αποτελέσματα και των δύο πράξεων είναι διανύσματα 128 διαστάσεων, τα οποία στη συνέχεια συνενώνονται σειριακά κατά μήκος του διαστατικού τους άξονα, σχηματίζοντας ένα ενιαίο διάνυσμα 256 διαστάσεων. Το διάνυσμα αυτό αποτελεί τη συνδυαστική αναπαράσταση του ζεύγους των πρωτεϊνών και χρησιμεύει ως είσοδος στο δεύτερο τμήμα του μοντέλου, το οποίο αναλαμβάνει τον υπολογισμό της τελικής πιθανότητας αλληλεπίδρασης και μοιράζεται την ίδια αρχιτεκτονική με το αντίληπτρο που περιγράφηκε παραπάνω. Το διάνυσμα εισέρχεται σε μια διαδοχική αλυσίδα πέντε γραμμικών επιπέδων μετασχηματισμού, ανάμεσα στα οποία παρεμβάλλονται κανονικοποίηση επιπέδου και συνάρτηση ενεργοποίησης GELU. Η διαστατικότητα μειώνεται σταδιακά — από 256 σε 128, έπειτα σε 64, σε 32 και τέλος σε 1 — οδηγώντας σε μια σταδιακή, ιεραρχική συμπίκνωση της αναπαράστασης, με dropout 0.3, 0.3 και 0.2 στα τρία πρώτα κρυφά επίπεδα αντίστοιχα. Αυτή η πυραμιδοειδής αρχιτεκτονική αποτελεί μια καλά τεκμηριωμένη πρακτική στα βαθιά νευρωνικά δίκτυα: αναγκάζει το μοντέλο να διατηρεί μόνο την πλέον σημαντική πληροφορία σε κάθε στάδιο, απορρίπτοντας θόρυβο και πλεονάζουσες διαστάσεις, και διαμορφώνοντας σταδιακά μια αναπαράσταση όλο και πιο εστιασμένη στην τελική πρόβλεψη. Η επιλογή να παραμείνει η τελική διάσταση εξόδου του κοινόχρηστου κωδικοποιητή στις 128 διαστάσεις, σημαντικά μικρότερη από τις αρχικές 1024 διαστάσεις των αναπαραστάσεων ProtT5, αποσκοπούσε στη δημιουργία ενός συμπαγούς λανθάνοντα χώρου όπου οι σχέσεις ομοιότητας και διαφοράς μεταξύ πρωτεϊνών να είναι πιο ευδιάκριτες, χωρίς να επιβαρύνεται σημαντικά ο αριθμός των εκπαιδευσίμων παραμέτρων του δικτύου σύγκρισης.

## 5 Αποτελέσματα

Για την αξιολόγηση των μοντέλων πραγματοποιήθηκαν τρία βασικά πειράματα. Στο πρώτο εφαρμόστηκε τυχαία πενταπλή διασταυρωμένη επικύρωση ξεχωριστά για τις πρωτεΐνες κάθε είδους. Στο δεύτερο εφαρμόστηκε η ίδια διαδικασία στο σύνολο του ενοποιημένου συνόλου δεδομένων. Τέλος, για την αξιολόγηση της γενικευτικής ικανότητας των μοντέλων, η εκπαίδευση πραγματοποιήθηκε χρησιμοποιώντας τις πρωτεΐνες όλων των ειδών, εξαιρουμένων των *Aquila chrysaetos*, *Drosophila melanogaster* και *Pleurotus ostreatus*. Στη συνέχεια, τα μοντέλα αξιολογήθηκαν αποκλειστικά στα τρία αυτά είδη, ώστε να εξεταστεί η απόδοσή τους σε πρωτεΐνες οργανισμών που δεν είχαν χρησιμοποιηθεί κατά την εκπαίδευση. Ο σχεδιασμός αυτός των τριών πειραμάτων επιτρέπει μια σταδιακά αυξανόμενη δυσκολία αξιολόγησης: από τη στενή, ενδο-ειδική απόδοση, στην ικανότητα διαχείρισης της ετερογένειας πολλών ειδών ταυτόχρονα, και τέλος στη γενίκευση σε πλήρως άγνωστους οργανισμούς, η οποία αποτελεί και το πιο απαιτητικό, αλλά και το πιο

ρεαλιστικό σενάριο πρακτικής εφαρμογής των μοντέλων.

Για την αξιολόγηση των ταξινομητών χρησιμοποιήθηκαν βασικές τέσσερις μετρικές: Ακρίβεια, θετική προγνωστική αξία (ΘΠΑ), ανάκληση και F1. Η ακρίβεια εκφράζει το ποσοστό των δειγμάτων που ταξινομήθηκαν σωστά συνολικά, και αποτελεί τη βασικότερη ένδειξη της απόδοσης ενός ταξινομητή. Ωστόσο, σε σύνολα δεδομένων με ανισορροπία κλάσεων η ακρίβεια μπορεί να είναι παραπλανητική, γι' αυτό συμπληρώνεται από πιο διακριτικές μετρικές. Η ΘΠΑ εκφράζει το ποσοστό των προβλέψεων που χαρακτηρίστηκαν ως θετικές και ήταν πράγματι θετικές — δηλαδή μετρά πόσο αξιόπιστες είναι οι θετικές προβλέψεις του μοντέλου. Η ανάκληση από την άλλη, εκφράζει το ποσοστό των πραγματικά θετικών δειγμάτων που εντοπίστηκαν σωστά από το μοντέλο — αποτυπώνει δηλαδή την ικανότητά του να μην χάνει αληθινά θετικά. Η μετρική F1 αποτελεί τον αρμονικό μέσο της ΘΠΑ και της ανάκλησης, και παρέχει μια ισορροπημένη εκτίμηση της απόδοσης που λαμβάνει υπόψη και τις δύο διαστάσεις ταυτόχρονα. Η χρήση και των τεσσάρων μετρικών από κοινού επιτρέπει μια ολοκληρωμένη αξιολόγηση της συμπεριφοράς κάθε ταξινομητή, αναδεικνύοντας τυχόν αδυναμίες που δεν θα ήταν ορατές από μία μόνο μετρική. Επιπρόσθετα, σε όλα τα πειράματα αναφέρεται και το AUC (Area Under the Curve), το εμβαδόν κάτω από την καμπύλη ROC, το οποίο εκφράζει τη διακριτική ικανότητα του μοντέλου ανεξάρτητα από το συγκεκριμένο κατώφλι απόφασης που επιλέγεται. Η ταυτόχρονη παρουσίαση της ακρίβειας, της ΘΠΑ, της ανάκλησης, του F1 και του AUC επιτρέπει τη διάκριση μεταξύ μοντέλων που απλώς προσαρμόζονται καλά στο συγκεκριμένο κατώφλι 0,5 και μοντέλων που διαθέτουν εγγενώς ισχυρότερη ικανότητα διαχωρισμού των δύο κλάσεων σε όλο το εύρος των πιθανών κατωφλίων.

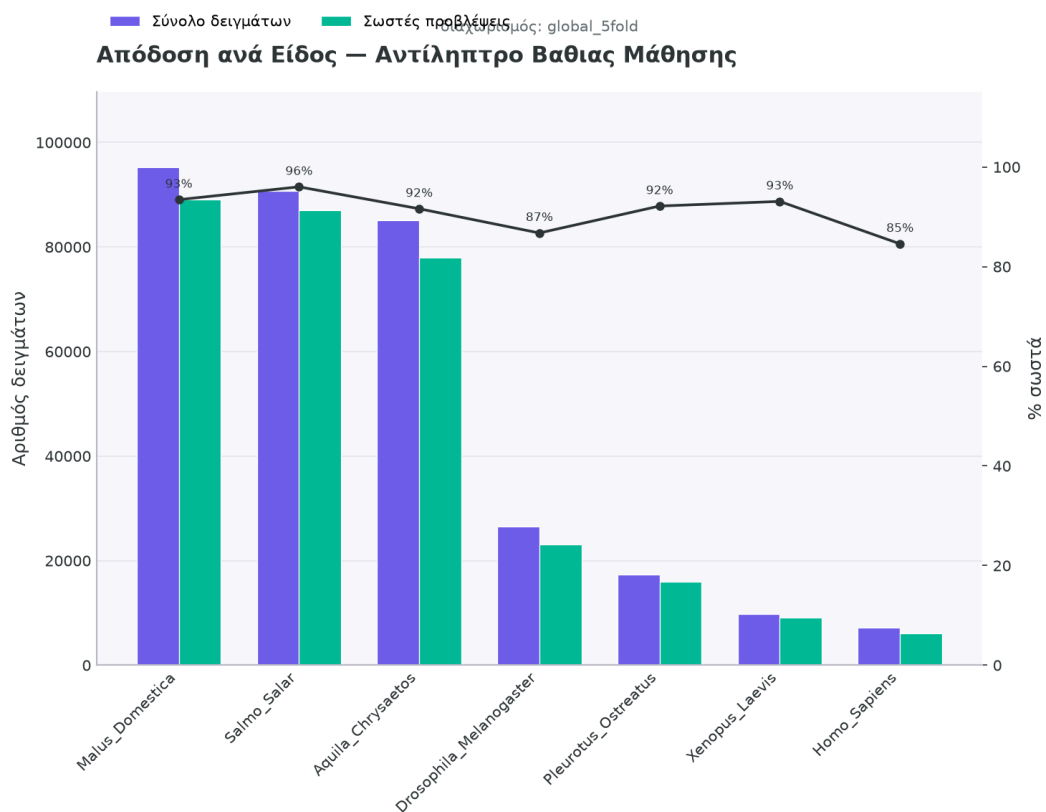
## 5.1 Δοκιμή μοντέλων σε κάθε είδος ξεωριστά

Στο πρώτο πείραμα κάθε είδος αξιολογήθηκε ανεξάρτητα, με εφαρμογή πενταπλής διασταυρωμένης επικύρωσης αποκλειστικά στις πρωτεΐνες του συγκεκριμένου οργανισμού. Η προσέγγιση αυτή επιτρέπει να αναδειχθεί η επίδραση των ιδιαίτερων χαρακτηριστικών κάθε είδους —όπως το μέγεθος του διαθέσιμου συνόλου δεδομένων και το ποσοστό θετικών αλληλεπιδράσεων— στην απόδοση κάθε μοντέλου, χωρίς την παρεμβολή της ετερογένειας που θα εισήγαγε ένα ενοποιημένο σύνολο δεδομένων. Τα ποσοστά θετικών ζευγών κυμαίνονται σημαντικά μεταξύ των ειδών, από περίπου 7% στον σολομό (*Salmo salar*) έως και 23% στη *Drosophila melanogaster*, γεγονός που καθιστά τη σύγκριση των δύο αρχιτεκτονικών σε αυτό το πλαίσιο ιδιαίτερα διαφωτιστική ως προς την ανθεκτικότητά τους στην ανισορροπία κλάσεων. Τα συγκεντρωτικά αποτελέσματα, υπολογισμένα ως μέσος όρος επί των πέντε αναδιπλώσεων, παρατίθενται στον Πίνακα 1. Η πενταπλή διασταυρωμένη επικύρωση επιλέχθηκε ώστε τα αναφερόμενα αποτελέσματα να μην εξαρτώνται από μια τυχαία, μεμονωμένη διαμέριση των δεδομένων σε σύνολα εκπαίδευσης και

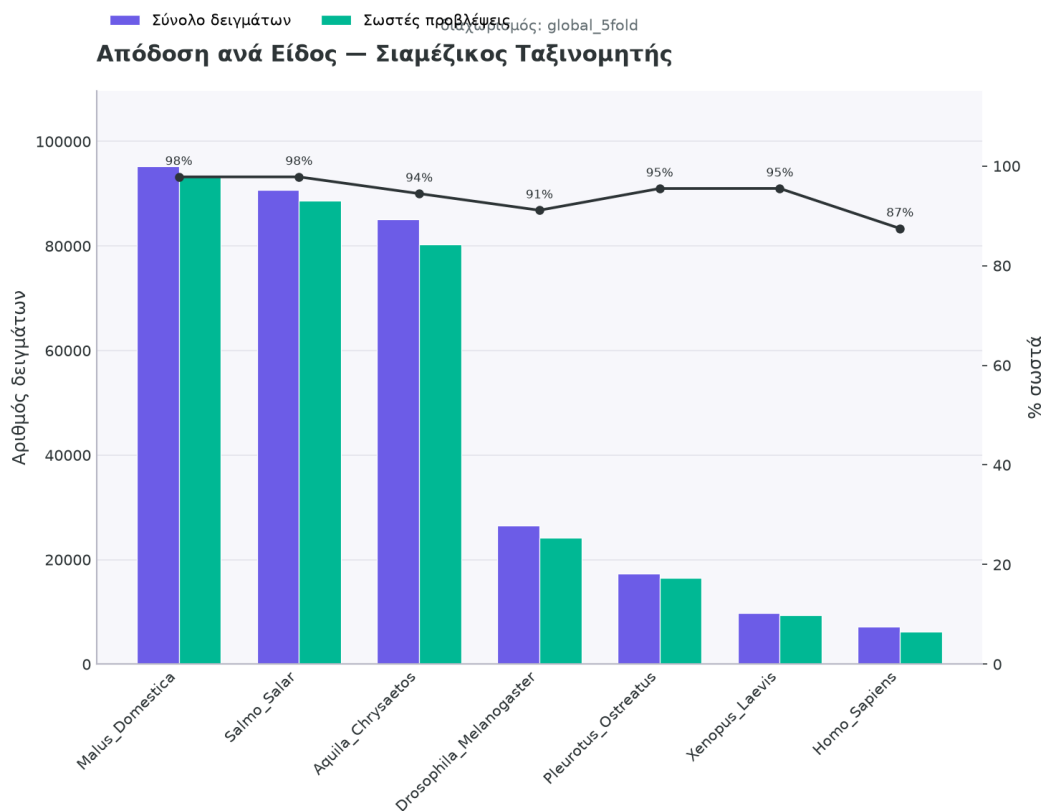
δοκιμής, αλλά να αντιπροσωπεύουν μια πιο σταθερή εκτίμηση της αναμενόμενης απόδοσης κάθε μοντέλου, μειώνοντας παράλληλα τον κίνδυνο τυχαίας ευνοϊκής ή δυσμενούς διαμέρισης.

| Είδος                   | Μέθοδος | Ακρίβεια      | AUC           | ΘΠΑ           | Ανάκληση      | F1            |
|-------------------------|---------|---------------|---------------|---------------|---------------|---------------|
| Aquila chrysaetos       | ABM     | 92,00%        | 95,32%        | 75,76%        | 80,70%        | 78,15%        |
|                         | ΣΤ      | <b>95,10%</b> | <b>98,36%</b> | <b>81,41%</b> | <b>93,79%</b> | <b>87,16%</b> |
| Drosophila melanogaster | ABM     | 88,82%        | 92,15%        | 75,08%        | 76,92%        | 75,99%        |
|                         | ΣΤ      | <b>92,54%</b> | <b>96,98%</b> | <b>79,82%</b> | <b>90,46%</b> | <b>84,81%</b> |
| Homo sapiens            | ABM     | 78,86%        | 84,12%        | 49,57%        | 73,25%        | 59,13%        |
|                         | ΣΤ      | <b>88,55%</b> | <b>92,06%</b> | <b>68,77%</b> | <b>82,72%</b> | <b>75,11%</b> |
| Malus domestica         | ABM     | 93,32%        | 96,97%        | 75,72%        | 86,68%        | 80,83%        |
|                         | ΣΤ      | <b>98,28%</b> | <b>99,59%</b> | <b>92,58%</b> | <b>97,24%</b> | <b>94,85%</b> |
| Pleurotus ostreatus     | ABM     | 87,92%        | 89,41%        | 47,83%        | 76,29%        | 58,79%        |
|                         | ΣΤ      | <b>94,95%</b> | <b>97,33%</b> | <b>72,06%</b> | <b>90,36%</b> | <b>80,18%</b> |
| Salmo salar             | ABM     | 94,65%        | 94,89%        | 58,83%        | 80,19%        | 67,87%        |
|                         | ΣΤ      | <b>97,32%</b> | <b>98,42%</b> | <b>75,33%</b> | <b>92,06%</b> | <b>82,86%</b> |
| Xenopus laevis          | ABM     | 86,25%        | 89,97%        | 48,08%        | 77,61%        | 59,38%        |
|                         | ΣΤ      | <b>94,99%</b> | <b>96,06%</b> | <b>77,43%</b> | <b>86,49%</b> | <b>81,71%</b> |

**Πίνακας 1.** Επιδόσεις ανά είδος για το Αντίληπτο Βαθιάς Μάθησης (ABM) και τον Σιαμέζικο Ταξινομητή (ΣΤ) (μέσος όρος πενταπλής διασταυρωμένης επικύρωσης)



**Εικόνα 5.1.** Απόδοση του Αντίληπτου Βαθιάς Μάθησης ανά είδος, ανά μετρική (πενταπλή διασταυρωμένη επικύρωση)



**Εικόνα 5.2.** Απόδοση του Σιαμέζικου Ταξινομητή ανά είδος, ανά μετρική (πενταπλή διασταυρωμένη επικύρωση)

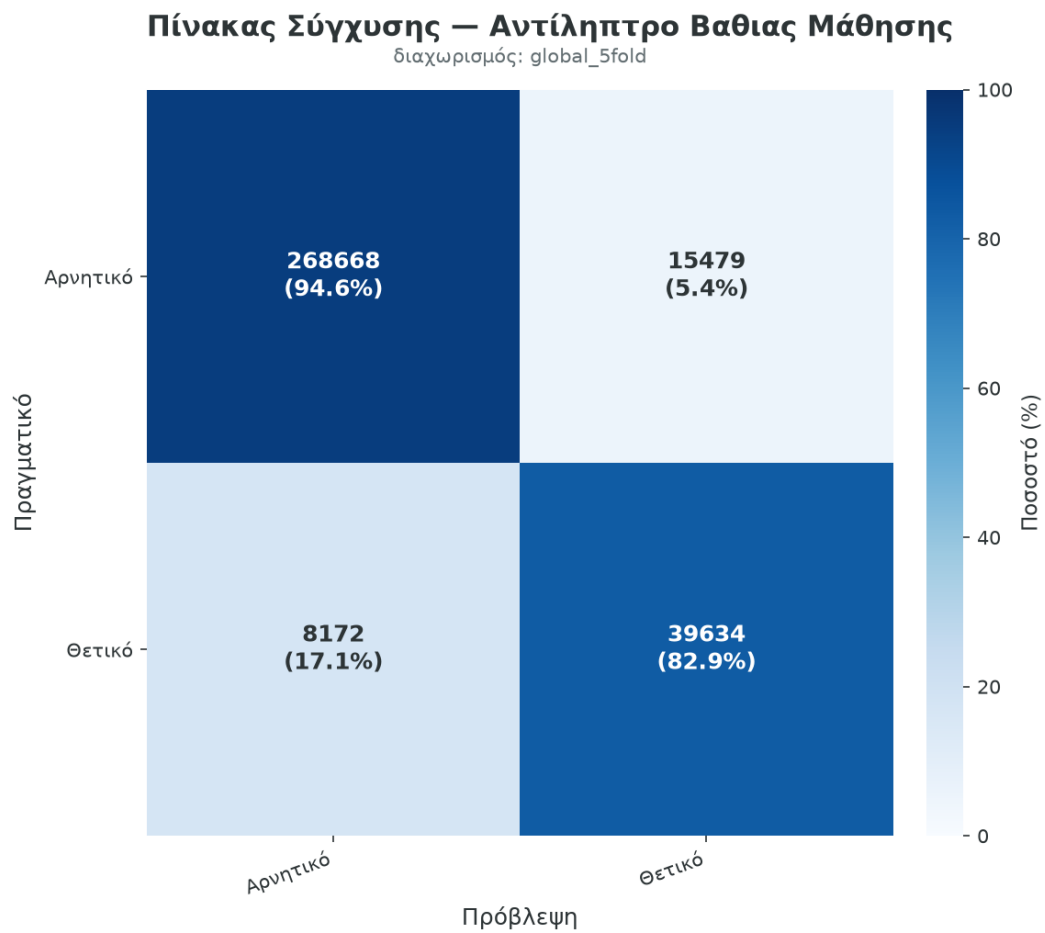
Όπως προκύπτει από τον Πίνακα 1 και την Εικόνα 5.1, το Αντίληπτρο Βαθιάς Μάθησης παρουσιάζει έντονη διακύμανση στην απόδοσή του μεταξύ των ειδών. Στα είδη με σχετικά ισορροπημένες κλάσεις, όπως η *Malus domestica* και η *Aquila chrysaetos*, η ΘΠΑ και η ανάκληση κυμαίνονται σε παρόμοια και ικανοποιητικά επίπεδα, οδηγώντας σε F1 άνω του 78%. Αντίθετα, στα είδη με πιο έντονη ανισορροπία κλάσεων, όπως ο σολομός (ποσοστό θετικών μόλις 7%) και ο *Pleurotus ostreatus* (11,3%), παρατηρείται σημαντική απόκλιση μεταξύ ΘΠΑ και ανάκλησης: το μοντέλο διατηρεί ικανοποιητική ανάκληση (πάνω από 76% σε όλες τις περιπτώσεις) αλλά η ΘΠΑ πέφτει κάτω από 60%, υποδεικνύοντας ότι, παρότι εντοπίζει τις περισσότερες πραγματικές αλληλεπιδράσεις, παράγει ταυτόχρονα σημαντικό αριθμό ψευδώς θετικών προβλέψεων. Το φαινόμενο αυτό είναι αναμενόμενο σε συνθήκες έντονης ανισορροπίας κλάσεων και αντικατοπτρίζει τα όρια ενός μοντέλου που βασίζεται αποκλειστικά στη συνένωση Hadamard γινομένου και απόλυτης διαφοράς για τη σύγκριση των δύο πρωτεϊνικών αναπαραστάσεων. Η στάθμιση των κλάσεων στη συνάρτηση απώλειας βοηθά το μοντέλο να μην αγνοήσει εντελώς τη μειοψηφική κλάση, όμως δεν επαρκεί από μόνη της για να αντισταθμίσει την έλλειψη ενός μηχανισμού που να εξασφαλίζει ότι οι δύο αναπαραστάσεις εισόδου βρίσκονται σε έναν κοινό, εναρμονισμένο χώρο πριν τη σύγκρισή τους.

Ο Σιαμέζικος Ταξινομητής (Πίνακας 1, Εικόνα 5.2) υπερτερεί του αντίληπτρου σε κάθε μεμονωμένο είδος και σε κάθε μετρική, χωρίς καμία εξαίρεση. Η βελτίωση είναι ιδιαίτερα εντυπωσιακή στα είδη όπου το αντίληπτρο υπολείπεται περισσότερο: στη *Malus domestica* το F1 ανέρχεται σε 94,85% (από 80,83%), στον *Pleurotus ostreatus* σε 80,18% (από 58,79%) και στον σολομό σε 82,86% (από 67,87%). Η βελτίωση αυτή οφείλεται πρωτίστως στη σημαντική άνοδο της ΘΠΑ, η οποία στα είδη με έντονη ανισορροπία κλάσεων υπερδιπλασιάζεται σχεδόν, χωρίς όμως να θυσιάζεται η ανάκληση, η οποία παραμένει σταθερά πάνω από 82% σε όλα τα είδη. Αυτό υποδηλώνει ότι ο κοινόχρηστος κωδικοποιητής, εκπαιδευόμενος να παράγει αναπαραστάσεις σε έναν κοινό λανθάνοντα χώρο, καταφέρνει να διαχωρίσει αποδοτικότερα τα αλληλεπιδρώντα από τα μη αλληλεπιδρώντα ζεύγη, μειώνοντας τα ψευδώς θετικά χωρίς να αυξάνει αναλόγως τα ψευδώς αρνητικά. Αξίζει επίσης να σημειωθεί ότι το AUC του Σιαμέζικου Ταξινομητή ξεπερνά το 96% σε όλα τα είδη, φτάνοντας το 99,59% στη *Malus domestica*, γεγονός που καταδεικνύει την εξαιρετική διακριτική του ικανότητα ανεξαρτήτως του κατωφλίου ταξινόμησης που επιλέγεται. Η συνέπεια αυτής της υπεροχής σε όλα τα είδη, ανεξάρτητα από το μέγεθος του διαθέσιμου συνόλου δεδομένων ή το ποσοστό θετικών ζευγών, ενισχύει την εκτίμηση ότι η βελτίωση δεν οφείλεται σε τυχαίους παράγοντες συγκεκριμένου είδους, αλλά σε μια δομική ιδιότητα της σιαμέζικης αρχιτεκτονικής η οποία γενικεύεται σταθερά σε διαφορετικά βιολογικά πλαίσια.

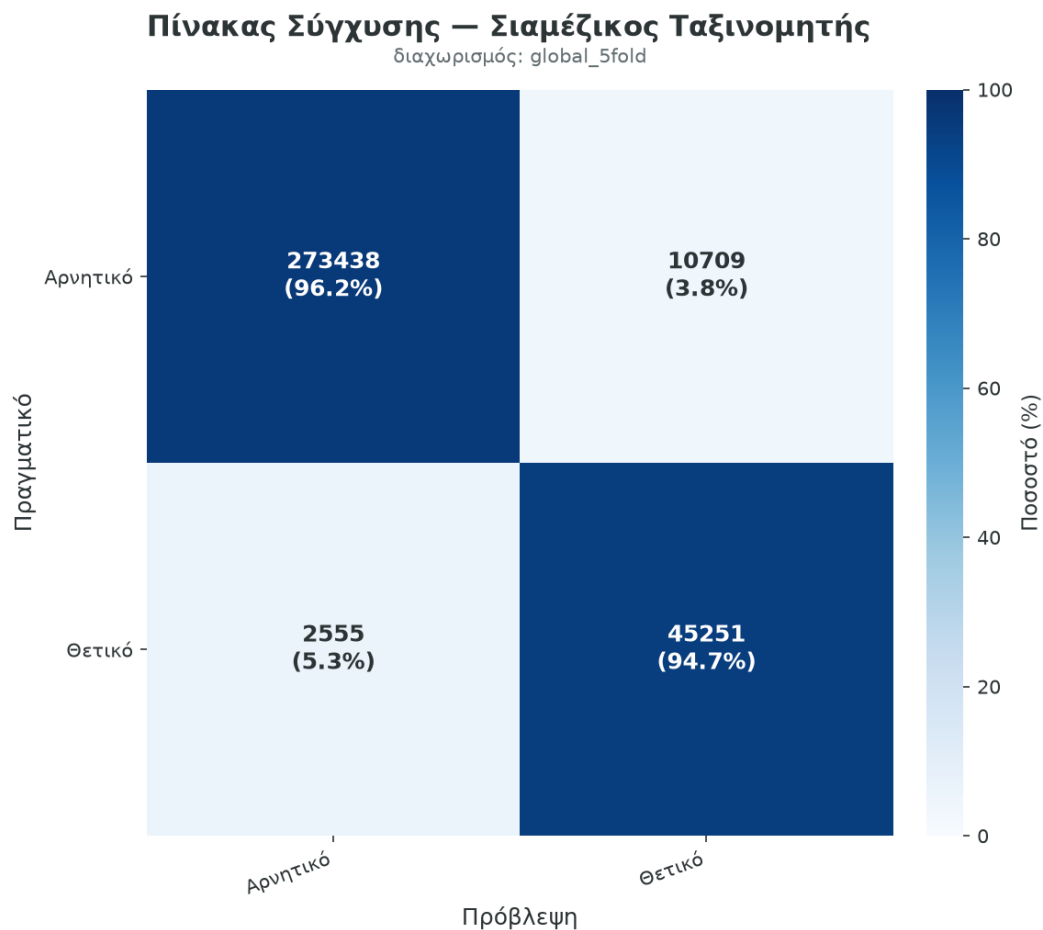
## 5.2 Δοκιμή μοντέλων σε όλο το σύνολο των δεδομένων

Στο δεύτερο πείραμα τα δύο μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν στο ενοποιημένο σύνολο δεδομένων, το οποίο περιλαμβάνει συνολικά 331.953 ζεύγη πρωτεϊνών από όλα τα επτά είδη, με ποσοστό θετικών αλληλεπιδράσεων περίπου 14,4%. Η ακρίβεια του Αντίληπτρου Βαθιάς Μάθησης, υπολογισμένη ως μέσος όρος της πενταπλής διασταυρωμένης επικύρωσης, ανήλθε σε 92,88%, με AUC 95,81%, ΘΠΑ 71,91%, ανάκληση 82,91% και F1 77,02%. Ο Σιαμέζικος Ταξινομητής υπερτερεί σε κάθε μία από αυτές τις μετρικές, πετυχαίνοντας ακρίβεια 96,00%, AUC 98,91%, ΘΠΑ 80,86%, ανάκληση 94,66% και F1 87,22%. Η διαφορά στο F1, της τάξης των 10 ποσοστιαίων μονάδων, αποτελεί ισχυρή ένδειξη ότι η σιαμέζικη αρχιτεκτονική γενικεύει αποτελεσματικότερα όταν εκπαιδεύεται στο σύνολο της ετερογένειας των ειδών ταυτόχρονα, σε σχέση με την απλή συνένωση Hadamard γινομένου και απόλυτης διαφοράς που χρησιμοποιεί το αντίληπτρο. Το πείραμα αυτό είναι ιδιαίτερα σημαντικό, καθώς το ενοποιημένο σύνολο δεδομένων εισάγει επιπλέον δυσκολία σε σχέση με την ανά είδος αξιολόγηση: το μοντέλο καλείται να μάθει μοτίβα αλληλεπίδρασης που είναι έγκυρα ταυτόχρονα για επτά διαφορετικά είδη με διαφορετικά ποσοστά θετικών ζευγών, χωρίς να έχει τη δυνατότητα να εξειδικευτεί στα ιδιαίτερα χαρακτηριστικά ενός μεμονωμένου οργανισμού.

Η ανάλυση των πινάκων σύγχυσης (Εικόνες 5.3 και 5.4) επιβεβαιώνει την παρατήρηση αυτή. Το αντίληπτρο ταξινομήσε λανθασμένα 15.479 αρνητικά ζεύγη ως θετικά (ψευδώς θετικά) και 8.172 θετικά ζεύγη ως αρνητικά (ψευδώς αρνητικά), επί συνόλου 284.147 πραγματικά αρνητικών και 47.806 πραγματικά θετικών ζευγών. Ο σιαμέζικος ταξινομητής, αντίθετα, μειώνει τα ψευδώς θετικά σε 10.709 και, πιο εντυπωσιακά, τα ψευδώς αρνητικά σε μόλις 2.555 — μια μείωση άνω του 68% σε σχέση με το αντίληπτρο. Η δραστική μείωση των ψευδώς αρνητικών είναι ιδιαίτερα σημαντική για το συγκεκριμένο πρόβλημα, καθώς η παράλειψη μιας πραγματικής αλληλεπίδρασης πρωτεϊνών κατά τη διαδικασία ανάλυσης βιολογικών δικτύων μπορεί να οδηγήσει σε ελλιπή χαρτογράφηση κρίσιμων μονοπατιών σηματοδότησης ή μεταβολικών διαδρομών. Αντίστροφα, ένα μοντέρατο επίπεδο ψευδώς θετικών προβλέψεων είναι συνήθως λιγότερο δαπανηρό στην πράξη, καθώς τα υποψήφια ζεύγη μπορούν να υποβληθούν σε δευτερεύοντα πειραματικό έλεγχο πριν συμπεριληφθούν με βεβαιότητα σε ένα δίκτυο αλληλεπιδράσεων, γεγονός που ενισχύει περαιτέρω την πρακτική αξία της συμπεριφοράς που επιδεικνύει ο σιαμέζικος ταξινομητής σε αυτό το πείραμα.



**Εικόνα 5.3.** Πίνακας σύγκρισης του Αντίληπτου Βαθιάς Μάθησης στο ενοποιημένο σύνολο δεδομένων

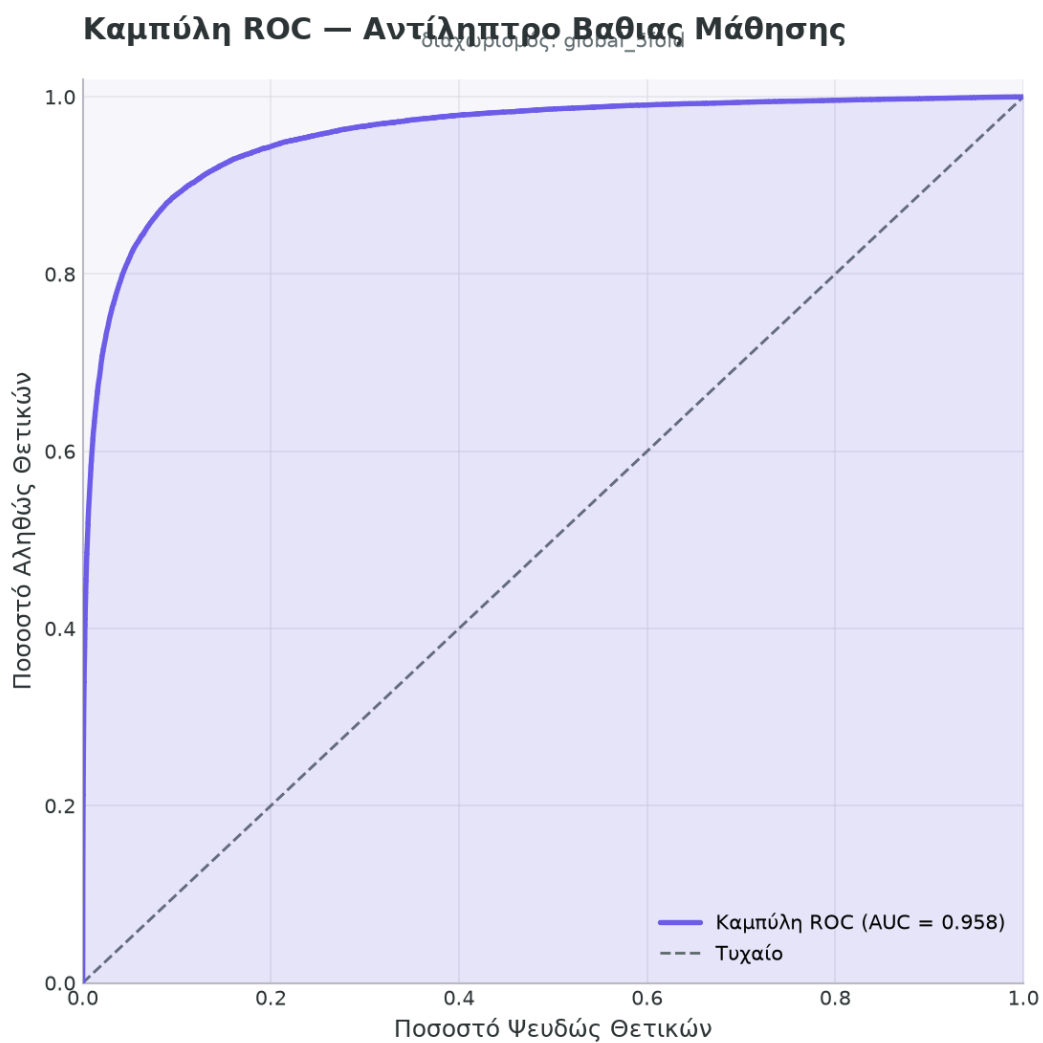


**Εικόνα 5.4.** Πίνακας σύγκρισης του Σιαμέζικου Ταξινομητή στο ενοποιημένο σύνολο δεδομένων

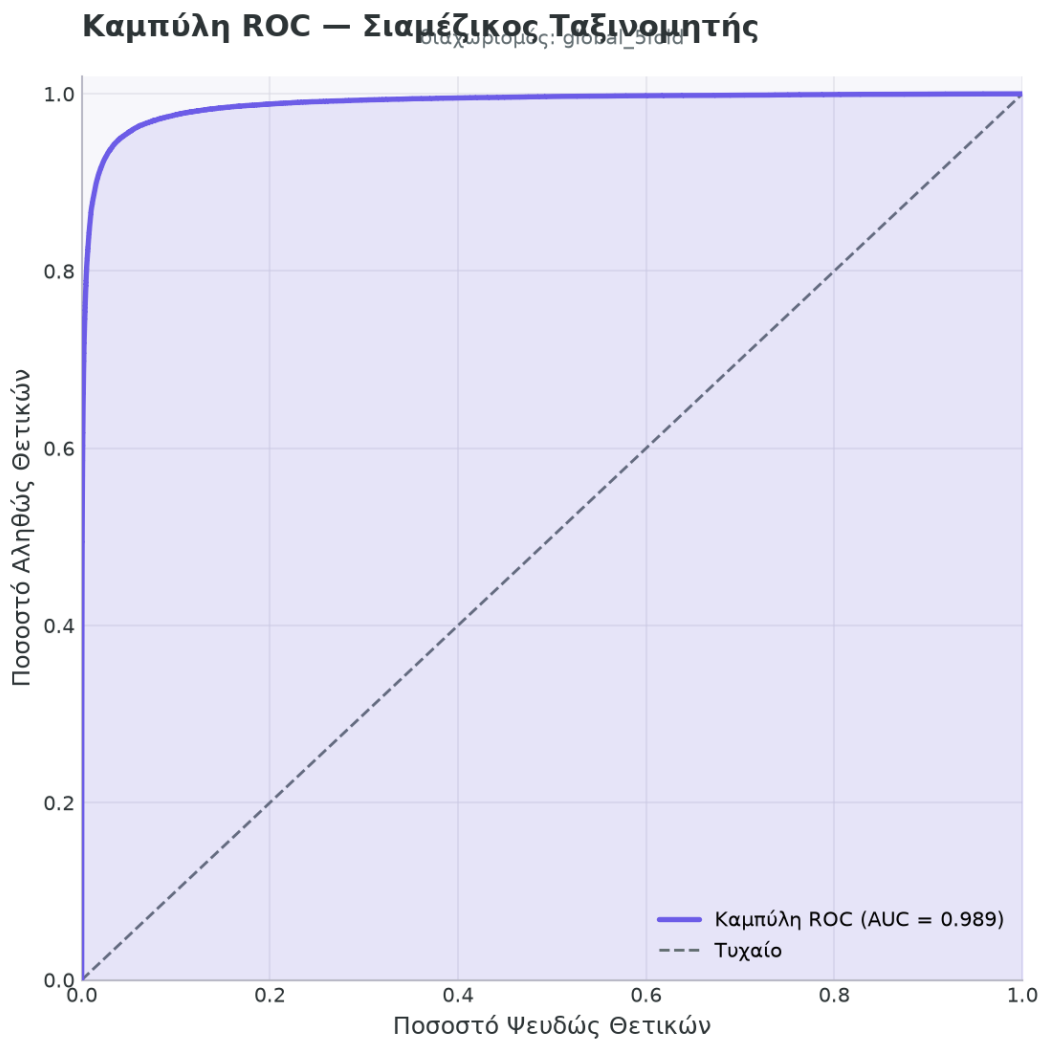
Πέρα από τη συνολική επίδοση, εξετάστηκε και η κατανομή της ακρίβειας ανά είδος εντός του ενοποιημένου συνόλου δοκιμής (Πίνακας 2). Το αντίληπτο κυμαίνεται από 84,58% στον άνθρωπο (*Homo sapiens*) έως 96,02% στον σολομό, ενώ ο σιαμέζικος ταξινομητής βελτιώνει την επίδοση σε κάθε είδος χωρίς εξαίρεση, φτάνοντας 87,47% στον άνθρωπο και 97,81% στη *Malus domestica* και στον σολομό. Είναι αξιοσημείωτο ότι το είδος με τη μικρότερη διαθέσιμη ποσότητα δεδομένων, ο άνθρωπος (7.180 ζεύγη), παραμένει το πιο δυσκολότερο για αμφότερα τα μοντέλα, γεγονός που αναδεικνύει τη σημασία του όγκου δεδομένων εκπαίδευσης για την επίτευξη αξιόπιστων προβλέψεων. Παρά τη σχετικά περιορισμένη διαθεσιμότητα δεδομένων για τον άνθρωπο, είναι ενθαρρυντικό ότι και τα δύο μοντέλα διατηρούν σε αυτό το είδος επίδοση σαφώς ανώτερη του τυχαίου, γεγονός που υποδεικνύει ότι η κοινή εκπαίδευση σε όλα τα επτά είδη επιτρέπει έμμεση μεταφορά γνώσης προς τα είδη με μικρότερο όγκο δεδομένων, μέσω των κοινών βιολογικών μοτίβων που συλλαμβάνονται στις αναπαραστάσεις ProtT5.

| Είδος                   | Ακρίβεια ABM | Ακρίβεια ΣΤ   |
|-------------------------|--------------|---------------|
| Malus domestica         | 93,44%       | <b>97,81%</b> |
| Salmo salar             | 96,02%       | <b>97,81%</b> |
| Aquila chrysaetos       | 91,62%       | <b>94,46%</b> |
| Drosophila melanogaster | 86,75%       | <b>91,12%</b> |
| Pleurotus ostreatus     | 92,16%       | <b>95,48%</b> |
| Xenopus laevis          | 93,10%       | <b>95,50%</b> |
| Homo sapiens            | 84,58%       | <b>87,47%</b> |

**Πίνακας 2.** Ποσοστό ορθής ταξινόμησης ανά είδος στο ενοποιημένο σύνολο δεδομένων, για το Αντίληπτρο Βαθιάς Μάθησης (ABM) και τον Σιαμέζικο Ταξινομητή (ΣΤ)



**Εικόνα 5.5.** Καμπύλη ROC του Αντίληπτρου Βαθιάς Μάθησης στο ενοποιημένο σύνολο δεδομένων



**Εικόνα 5.6.** Καμπύλη ROC του Σιαμέζικου Ταξινομητή στο ενοποιημένο σύνολο δεδομένων

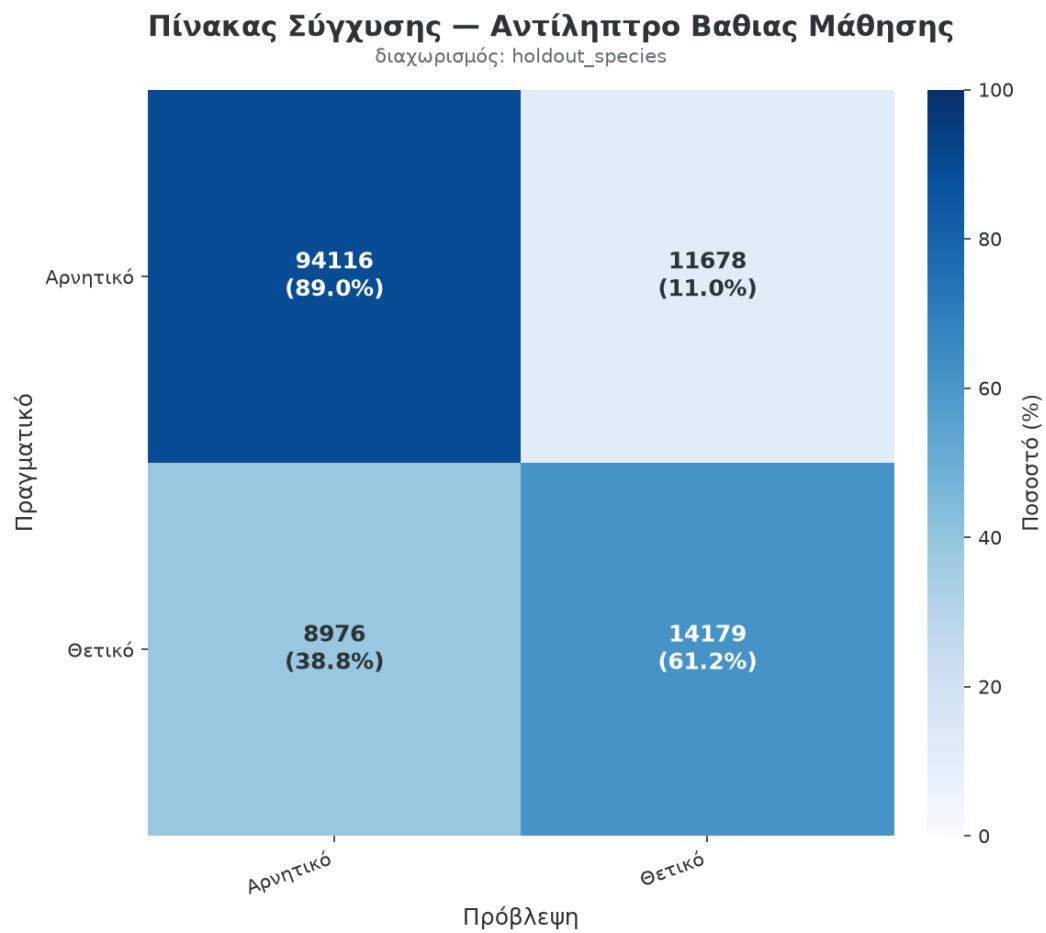
Οι καμπύλες ROC (Εικόνες 5.5 και 5.6) επιβεβαιώνουν οπτικά τη διαφορά στο AUC μεταξύ των δύο μοντέλων: η καμπύλη του σιαμέζικου ταξινομητή προσεγγίζει σχεδόν την ιδεατή γωνία επάνω αριστερά του διαγράμματος, ενώ η καμπύλη του αντίληπτρου, αν και εξίσου ικανοποιητική σε απόλυτους όρους, παρουσιάζει εμφανώς μεγαλύτερη απόκλιση από αυτήν, ιδιαίτερα στην περιοχή χαμηλών ποσοστών ψευδώς θετικών. Τα αποτελέσματα των δύο μοντέλων συνοψίζονται στον Πίνακα 3. Η σταθερή υπεροχή του σιαμέζικου ταξινομητή σε ολόκληρο το φάσμα των πιθανών κατωφλίων απόφασης, και όχι μόνο στο κατώφλι 0,5 που χρησιμοποιείται για τον υπολογισμό της ακρίβειας και του F1, αποτελεί ισχυρή ένδειξη ότι η βελτίωση δεν οφείλεται σε τυχαία βελτιστοποίηση γύρω από ένα συγκεκριμένο σημείο λειτουργίας, αλλά σε μια πιο θεμελιώδη βελτίωση της ικανότητας του μοντέλου να διαχωρίζει τις δύο κλάσεις.

| Μοντέλο                  | Ακρίβεια      | AUC           | ΘΠΑ           | Ανάκληση      | F1            |
|--------------------------|---------------|---------------|---------------|---------------|---------------|
| Αντίληπτο Βαθιάς Μάθησης | 92,88%        | 95,81%        | 71,91%        | 82,91%        | 77,02%        |
| Σιαμέζικος Ταξινομητής   | <b>96,00%</b> | <b>98,91%</b> | <b>80,86%</b> | <b>94,66%</b> | <b>87,22%</b> |

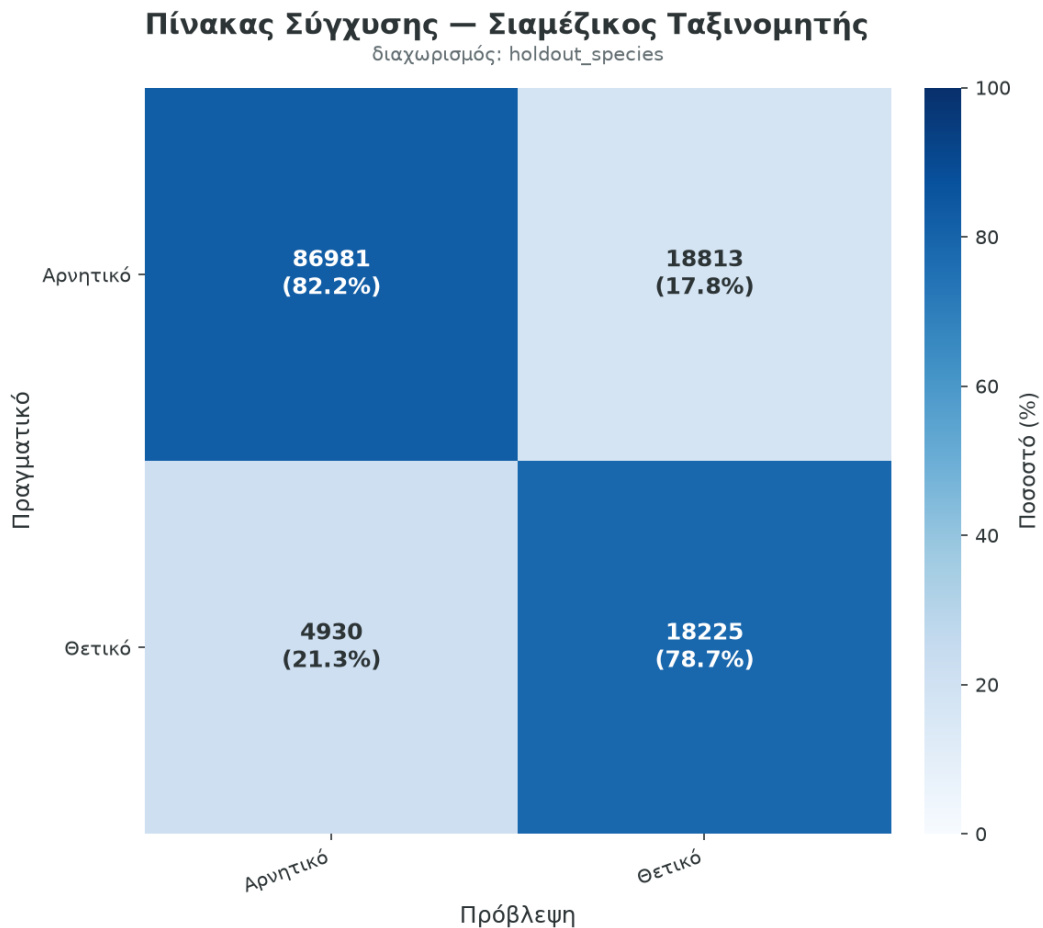
**Πίνακας 3.** Οι επιδόσεις των μοντέλων σε ολόκληρο το σύνολο δεδομένων

### 5.3 Δοκιμή μοντέλων σε ξένα είδη οργανισμών

Στο τρίτο πείραμα εξετάστηκε η ικανότητα γενίκευσης των μοντέλων σε είδη που δεν συμμετείχαν καθόλου στην εκπαίδευση. Τα μοντέλα εκπαιδεύτηκαν αποκλειστικά στις πρωτεΐνες των τεσσάρων εναπομείναντων ειδών (*Homo sapiens*, *Malus domestica*, *Salmo salar* και *Xenopus laevis*) και αξιολογήθηκαν στα 128.949 ζεύγη πρωτεϊνών των τριών ξένων ειδών: *Aquila chrysaetos*, *Drosophila melanogaster* και *Pleurotus ostreatus*, με ποσοστό θετικών αλληλεπιδράσεων στο σύνολο δοκιμής 17,96%. Πρόκειται για το πιο απαιτητικό πείραμα, καθώς τα μοντέλα καλούνται να αξιοποιήσουν αποκλειστικά τη βιολογική πληροφορία που ενυπάρχει στις αναπαραστάσεις ProtT5, χωρίς να έχουν εκτεθεί ποτέ σε παραδείγματα των συγκεκριμένων οργανισμών. Το σενάριο αυτό προσομοιώνει ρεαλιστικά την πρακτική εφαρμογή των μοντέλων σε νέους ή λιγότερο μελετημένους οργανισμούς, για τους οποίους δεν υπάρχουν ακόμη επαρκή πειραματικά δεδομένα αλληλεπίδρασης, καθιστώντας το κρίσιμο δείκτη της πραγματικής χρησιμότητας της προσέγγισης πέρα από το στενό πλαίσιο των ειδών που συμμετείχαν στην εκπαίδευση.



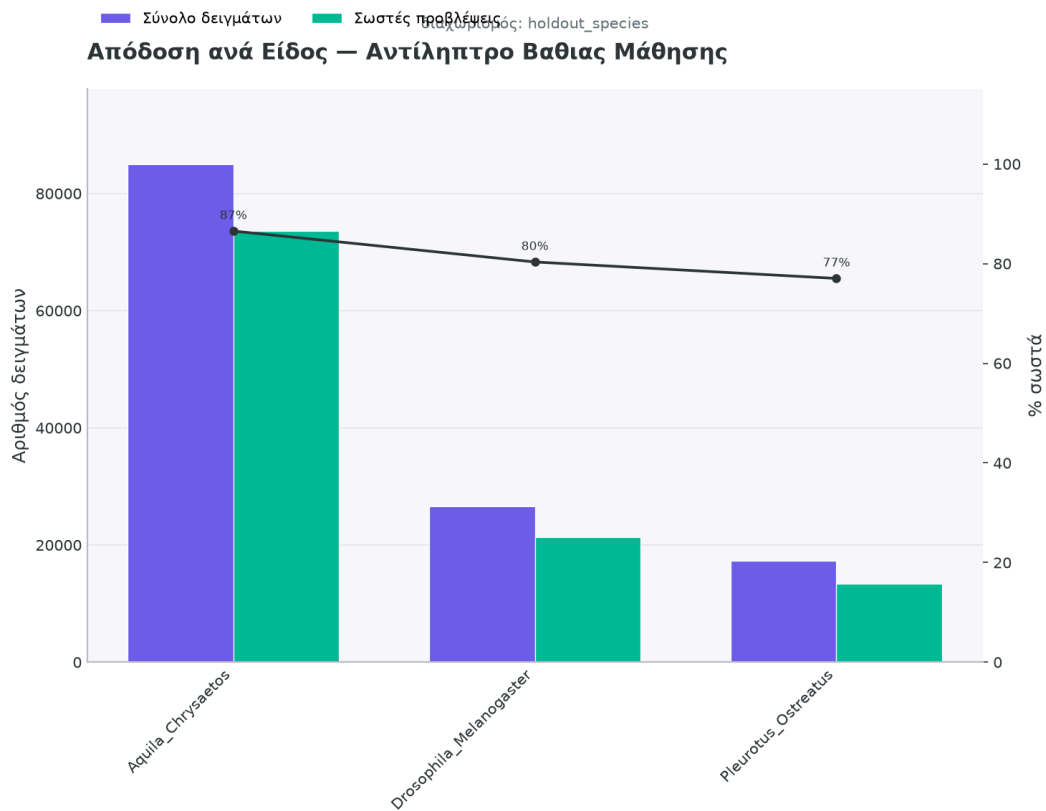
**Εικόνα 5.7.** Πίνακας σύγχυσης του Αντίληπτου Βαθιάς Μάθησης στα ξένα είδη οργανισμών



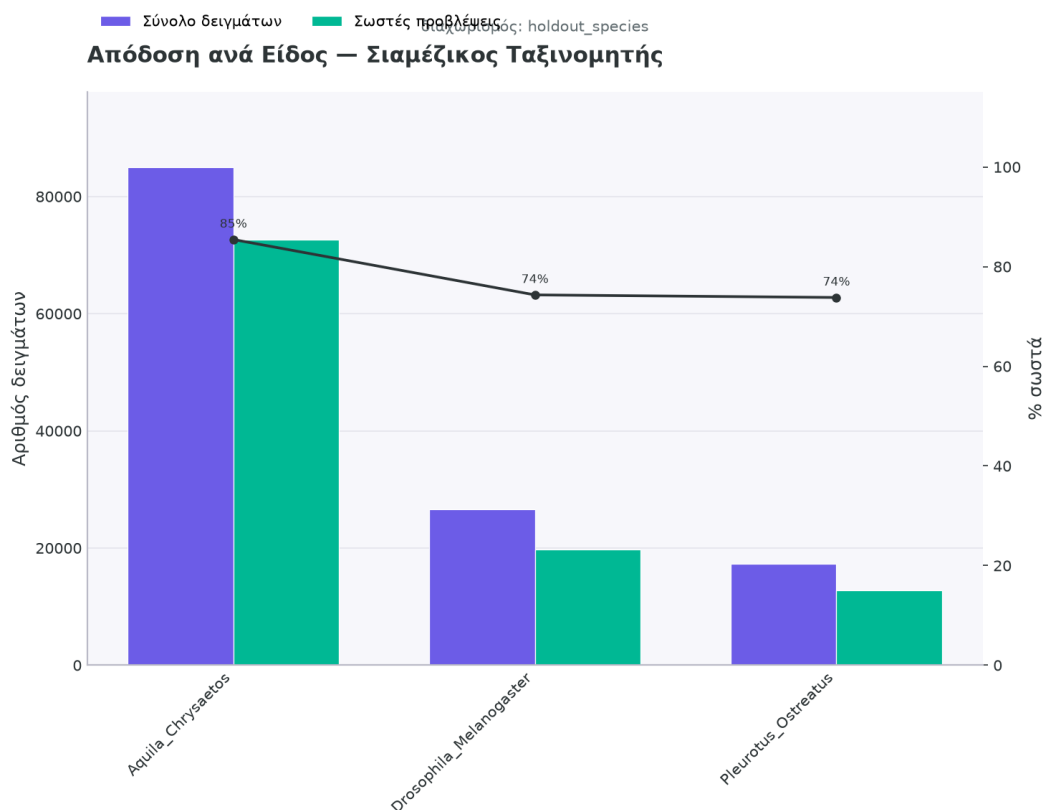
**Εικόνα 5.8.** Πίνακας σύγχυσης του Σιαμέζικου Ταξινομητή στα ξένα είδη οργανισμών

Τα αποτελέσματα αυτού του πειράματος (Εικόνες 5.7 και 5.8) διαφέρουν αισθητά από την εικόνα που παρατηρήθηκε στα δύο προηγούμενα πειράματα και αναδεικνύουν μια ενδιαφέρουσα αντιστάθμιση μεταξύ των δύο αρχιτεκτονικών. Το αντίληπτο πετυχαίνει συνολική ακρίβεια 83,98%, ελαφρώς υψηλότερη από το 81,59% του σιαμέζικου ταξινομητή. Ωστόσο, η εικόνα αντιστρέφεται πλήρως στις υπόλοιπες μετρικές: το AUC του σιαμέζικου ταξινομητή (88,21%) υπερβαίνει σημαντικά αυτό του αντίληπτου (84,26%), όπως και η ανάκλησή του, η οποία ανέρχεται σε 78,71% έναντι μόλις 61,24% για το αντίληπτο. Ο πίνακας σύγχυσης εξηγεί την αντίθεση αυτή: το αντίληπτο παράγει λιγότερα ψευδώς θετικά (11.678 έναντι 18.813) αλλά αφήνει ασυγκρίτως περισσότερα πραγματικά θετικά ζεύγη αναξιοποίητα, με 8.976 ψευδώς αρνητικά έναντι μόλις 4.930 για τον σιαμέζικο ταξινομητή. Με άλλα λόγια, το αντίληπτο υιοθετεί μια πιο συντηρητική συμπεριφορά πρόβλεψης σε αόρατα είδη, θυσιάζοντας ανάκληση για χάρη μιας μετρικής ακρίβειας που παραμένει ευνοϊκή λόγω της υπεροχής της αρνητικής κλάσης, ενώ ο σιαμέζικος ταξινομητής διατηρεί την τάση του να εντοπίζει τις περισσότερες πραγματικές αλληλεπιδράσεις, με το κόστος περισσότερων ψευδών συναγερμών. Η αντιστάθμιση αυτή μεταξύ ακρίβειας και ανάκλησης δεν πρέπει να θεωρηθεί απλή ασυνέπεια στη συμπεριφορά των δύο μοντέλων, αλλά μάλλον αντικατοπτρίζει δύο διαφορετικές, εξίσου έγκυρες,

στρατηγικές πρόβλεψης υπό συνθήκες μετατόπισης κατανομής (distribution shift), οι οποίες αναδεικνύουν διαφορετικές πτυχές της γενικευτικής ικανότητας κάθε αρχιτεκτονικής.



**Εικόνα 5.9.** Ποσοστό ορθής ταξινόμησης ανά ξένο είδος για το Αντίληπτο Βαθιάς Μάθησης

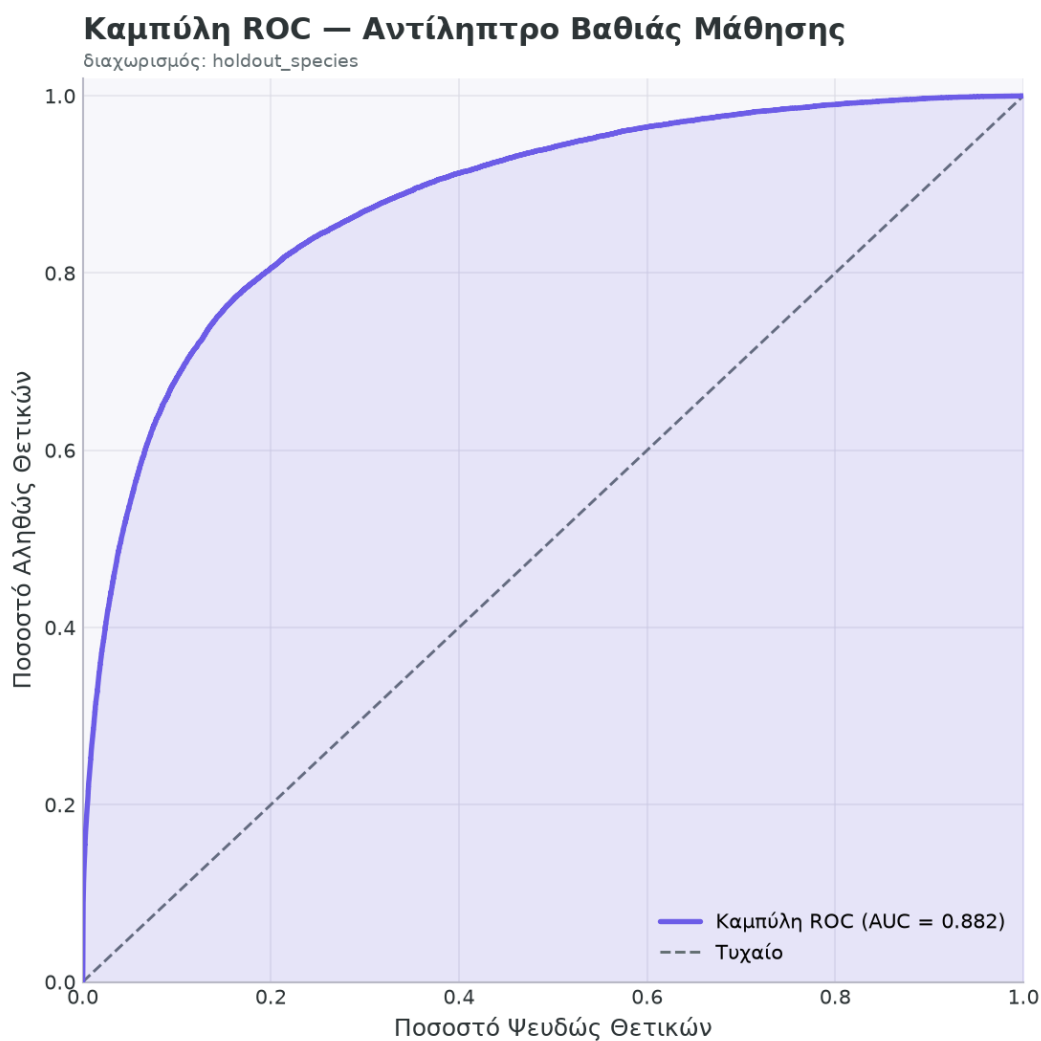


**Εικόνα 5.10.** Ποσοστό ορθής ταξινόμησης ανά ξένο είδος για τον Σιαμέζικο Ταξινομητή

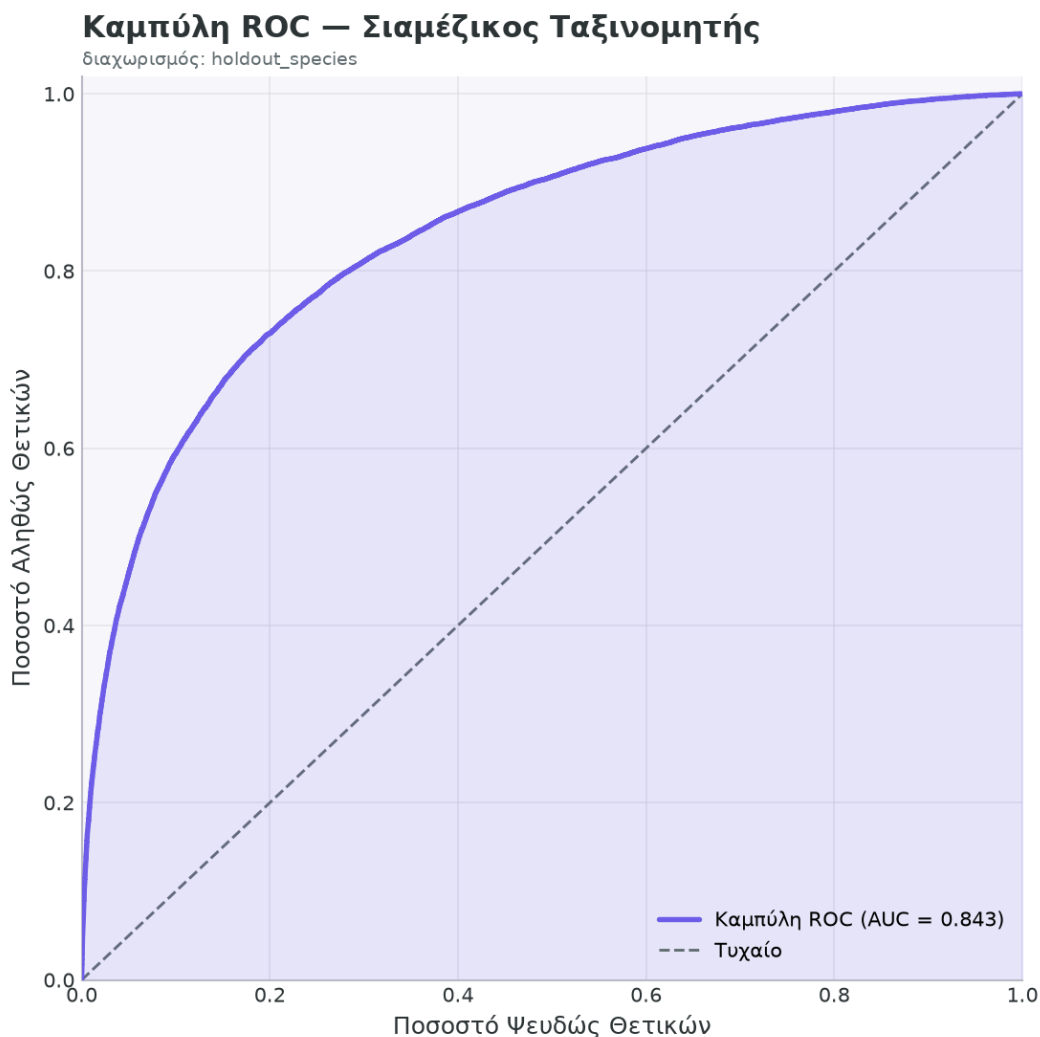
Η κατανομή της ακρίβειας ανά ξένο είδος (Εικόνες 5.9 και 5.10, Πίνακας 4) επιβεβαιώνει τη γενική τάση: το αντίληπτο εμφανίζει ελαφρώς υψηλότερο ποσοστό ορθής ταξινόμησης σε όλα τα τρία είδη, με τη μεγαλύτερη διαφορά να καταγράφεται στη *Drosophila melanogaster* (80,33% έναντι 74,33%). Το αξιοσημείωτο εδώ είναι ότι η υπεροχή αυτή του αντίληπτου στην ακρίβεια δεν αντικατοπτρίζει καλύτερη γενικευτική ικανότητα, αλλά μάλλον την τάση του να προβλέπει συχνότερα την αρνητική κλάση, η οποία τυχαίνει να είναι και η πλειοψηφική. Το *Aquila chrysaetos* παραμένει το πιο εύκολο από τα τρία ξένα είδη για αμφότερα τα μοντέλα, πιθανότατα λόγω της μεγαλύτερης διαθεσιμότητας δεδομένων εκπαίδευσης για συγγενικά είδη σπονδυλωτών, ενώ ο *Pleurotus ostreatus*, ως μοναδικός εκπρόσωπος μυκήτων στο σύνολο δεδομένων, παρουσιάζει τη χαμηλότερη επίδοση και για τα δύο μοντέλα. Το εύρημα αυτό συνάδει με την υπόθεση ότι η ταξινομική εγγύτητα ενός άγνωστου είδους προς τα είδη εκπαίδευσης επηρεάζει άμεσα την ποιότητα της γενίκευσης: όσο πιο απομακρυσμένος εξελικτικά είναι ένας οργανισμός από τα είδη που έχουν συμπεριληφθεί στην εκπαίδευση, τόσο πιο δύσκολο καθίσταται για το μοντέλο να αξιοποιήσει αποτελεσματικά τη βιολογική πληροφορία που έχει μάθει.

| Είδος                          | Ακρίβεια ABM  | Ακρίβεια ΣΤ |
|--------------------------------|---------------|-------------|
| <i>Aquila chrysaetos</i>       | <b>86,54%</b> | 85,44%      |
| <i>Drosophila melanogaster</i> | <b>80,33%</b> | 74,33%      |
| <i>Pleurotus ostreatus</i>     | <b>77,03%</b> | 73,81%      |

**Πίνακας 4.** Ποσοστό ορθής ταξινόμησης ανά ξένο είδος, για το Αντίληπτρο Βαθιάς Μάθησης (ABM) και τον Σιαμέζικο Ταξινομητή (ΣΤ)



**Εικόνα 5.11.** Καμπύλη ROC του Αντίληπτρου Βαθιάς Μάθησης στα ξένα είδη οργανισμών



**Εικόνα 5.12.** Καμπύλη ROC του Σιαμέζικου Ταξινομητή στα ξένα είδη οργανισμών

Η αντίθεση μεταξύ ακρίβειας και AUC καθίσταται εμφανέστερη στις καμπύλες ROC (Εικόνες 5.11 και 5.12): η καμπύλη του σιαμέζικου ταξινομητή βρίσκεται σταθερά πάνω από αυτή του αντίληπτρου σε ολόκληρο το φάσμα των κατωφλίων απόφασης, επιβεβαιώνοντας ότι διαθέτει εγγενώς καλύτερη διακριτική ικανότητα μεταξύ αλληλεπιδρώντων και μη αλληλεπιδρώντων ζευγών σε άγνωστα είδη, ανεξάρτητα από το συγκεκριμένο κατώφλι του 0,5 που χρησιμοποιήθηκε για τον υπολογισμό της ακρίβειας. Αυτό υποδηλώνει ότι μια απλή αναπροσαρμογή του κατωφλίου απόφασης προς τα κάτω θα μπορούσε ενδεχομένως να αποδώσει στον σιαμέζικο ταξινομητή σημαντικά καλύτερη ακρίβεια στο σενάριο γενίκευσης σε ξένα είδη, χωρίς να απαιτείται καμία αλλαγή στην αρχιτεκτονική ή στην εκπαίδευσή του. Τα συγκεντρωτικά αποτελέσματα του πειράματος παρατίθενται στον Πίνακα 5. Η παρατήρηση αυτή έχει άμεση πρακτική σημασία για την εφαρμογή του μοντέλου σε πραγματικά σενάρια γενίκευσης: αντί να αξιολογείται η καταλληλότητα ενός μοντέλου αποκλειστικά με βάση την ακρίβεια σε ένα σταθερό κατώφλι, θα ήταν προτιμότερο να λαμβάνεται υπόψη ολόκληρη η καμπύλη ROC, επιτρέποντας την επιλογή του κατωφλίου που εξυπηρετεί καλύτερα τον συγκεκριμένο ερευνητικό ή κλινικό στόχο.

| Μοντέλο                   | Ακρίβεια      | AUC           | ΘΠΑ           | Ανάκληση      | F1            |
|---------------------------|---------------|---------------|---------------|---------------|---------------|
| Αντίληπτρο Βαθιάς Μάθησης | <b>83,98%</b> | 84,26%        | <b>54,84%</b> | 61,24%        | 57,86%        |
| Σιαμέζικος Ταξινομητής    | 81,59%        | <b>88,21%</b> | 49,21%        | <b>78,71%</b> | <b>60,56%</b> |

**Πίνακας 5.** Οι επιδόσεις των μοντέλων σε ξένα είδη οργανισμών

Η συνολική εικόνα που αναδύεται από τα αποτελέσματα επιβεβαιώνει την αξία της σιαμέζικης αρχιτεκτονικής για το συγκεκριμένο πρόβλημα πρόβλεψης αλληλεπιδράσεων πρωτεϊνών. Ο συνδυασμός του κοινόχρηστου κωδικοποιητή, του γινομένου Hadamard και της απόλυτης διαφοράς ως συνδυαστικής στρατηγικής, και του αντίληπτρου με σταδιακά μειούμενη διαστατικότητα αποδεικνύεται ικανός να εντοπίζει τόσο τις ομοιότητες όσο και τις διαφορές μεταξύ δύο πρωτεϊνικών αναπαραστάσεων με αξιοσημείωτη αποτελεσματικότητα. Στα δύο πρώτα πειράματα, όπου τα μοντέλα αξιολογούνται σε είδη που έχουν συμμετάσχει στην εκπαίδευση, ο σιαμέζικος ταξινομητής υπερτερεί συνολικά και χωρίς εξαίρεση σε κάθε μετρική, με ιδιαίτερα εμφανή τη βελτίωση στη ΘΠΑ σε σύνολα δεδομένων με έντονη ανισορροπία κλάσεων. Στο τρίτο πείραμα, το οποίο εξετάζει τη γενίκευση σε πλήρως ξένα είδη, η εικόνα γίνεται πιο σύνθετη: το αντίληπτρο πετυχαίνει ελαφρώς υψηλότερη ακρίβεια χάρη σε μια πιο συντηρητική συμπεριφορά πρόβλεψης, ενώ ο σιαμέζικος ταξινομητής διατηρεί εμφανές προβάδισμα στο AUC και στην ανάκληση, εντοπίζοντας δηλαδή περισσότερες πραγματικές αλληλεπιδράσεις. Το εύρημα αυτό υποδεικνύει ότι ο κοινόχρηστος κωδικοποιητής μαθαίνει αναπαραστάσεις με ισχυρή διακριτική ικανότητα που μεταφέρεται και σε πρωτεΐνες εντελώς άγνωστων οργανισμών, ενώ η επιλογή του τελικού κατωφλίου απόφασης αποτελεί ξεχωριστό ζήτημα βελτιστοποίησης, ανάλογα με το αν προτεραιότητα δίνεται στην ελαχιστοποίηση των ψευδώς θετικών ή των ψευδώς αρνητικών προβλέψεων.

## 6 Συζήτηση

Τα αποτελέσματα και των τριών πειραμάτων συνθέτουν μια συνεκτική εικόνα σχετικά με τη σχέση μεταξύ αρχιτεκτονικής και ικανότητας γενίκευσης στο πρόβλημα της πρόβλεψης αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης από αναπαραστάσεις ProtT5. Στα δύο πρώτα πειράματα, όπου τα μοντέλα αξιολογούνται σε πρωτεΐνες ειδών που έχουν ήδη συμμετάσχει στην εκπαίδευση, ο Σιαμέζικος Ταξινομητής υπερτερεί σταθερά και χωρίς καμία εξαίρεση σε κάθε μετρική και σε κάθε είδος. Η υπεροχή αυτή δεν είναι τυχαία, αλλά απορρέει άμεσα από τη δομή της αρχιτεκτονικής: ο κοινόχρηστος κωδικοποιητής αναγκάζεται να μάθει μια αναπαράσταση κάθε πρωτεΐνης η οποία είναι απευθείας συγκρίσιμη με την αναπαράσταση οποιασδήποτε άλλης πρωτεΐνης στον ίδιο λανθάνοντα χώρο, ενώ η κοινή χρήση βαρών λειτουργεί ταυτόχρονα ως εγγενής ρυθμιστικός μηχανισμός, μειώνοντας τον αριθμό των εκπαιδευσιμων παραμέτρων

και περιορίζοντας τον κίνδυνο υπερπροσαρμογής [6][54]. Αντίθετα, το Αντίληπτρο Βαθιάς Μάθησης βασίζεται αποκλειστικά στη συνένωση του γινομένου Hadamard και της απόλυτης διαφοράς για να συλλάβει τη σχέση μεταξύ δύο πρωτεϊνών, χωρίς να υπάρχει κάποιος μηχανισμός που να εξασφαλίζει ότι οι δύο αρχικές αναπαραστάσεις προέρχονται από έναν κοινό, εναρμονισμένο χώρο χαρακτηριστικών· η απουσία αυτή φαίνεται να περιορίζει την ικανότητά του να διαχωρίζει αποτελεσματικά τα αλληλεπιδρώντα από τα μη αλληλεπιδρώντα ζεύγη, ιδίως σε συνθήκες έντονης ανισορροπίας κλάσεων.

Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός ότι η υπεροχή του σιαμέζικου ταξινομητή δεν εντοπίζεται εξίσου σε όλες τις μετρικές. Στα είδη και στο ενοποιημένο σύνολο δεδομένων με χαμηλό ποσοστό θετικών αλληλεπιδράσεων, η βελτίωση συγκεντρώνεται κατά κύριο λόγο στη ΘΠΑ, η οποία σε ορισμένες περιπτώσεις σχεδόν διπλασιάζεται, ενώ η ανάκληση παραμένει σταθερά σε υψηλά επίπεδα και για τα δύο μοντέλα. Η παρατήρηση αυτή υποδεικνύει ότι το βασικό πρόβλημα του αντίληπτρου δεν είναι η αδυναμία εντοπισμού πραγματικών αλληλεπιδράσεων, αλλά η τάση του να παράγει σημαντικό αριθμό ψευδώς θετικών προβλέψεων όταν η αρνητική κλάση κυριαρχεί αριθμητικά. Ο σιαμέζικος ταξινομητής, μαθαίνοντας να τοποθετεί τις δύο πρωτεΐνες σε έναν κοινό λανθάνοντα χώρο πριν από τον υπολογισμό της σχέσης τους, φαίνεται να χαράσσει ένα πιο ευκρινές όριο απόφασης μεταξύ των δύο κλάσεων, μειώνοντας τα ψευδώς θετικά χωρίς να θυσιάζει αντίστοιχα την ανάκληση. Η στάθμιση των κλάσεων στη συνάρτηση απώλειας, η οποία εφαρμόστηκε εξίσου και στα δύο μοντέλα, δεν αρκεί επομένως από μόνη της για να αντισταθμίσει πλήρως την ανισορροπία των δεδομένων· η αρχιτεκτονική σύγκρισης παίζει εξίσου καθοριστικό ρόλο.

Το τρίτο πείραμα, στο οποίο τα μοντέλα αξιολογήθηκαν σε είδη που απουσίαζαν πλήρως από την εκπαίδευση, αναδεικνύει μια πιο σύνθετη εικόνα και αξίζει ιδιαίτερης προσοχής. Το γεγονός ότι το αντίληπτρο πέτυχε ελαφρώς υψηλότερη ακρίβεια δεν πρέπει να ερμηνευθεί ως ένδειξη καλύτερης γενικευτικής ικανότητας, καθώς η ίδια η ακρίβεια είναι ευαίσθητη στην αναλογία των κλάσεων και ευνοεί μοντέλα που προβλέπουν συχνότερα τη μειοψηφική συμπεριφορά προς την πλειοψηφική κλάση. Η ανάλυση των πινάκων σύγχυσης και των καμπυλών ROC δείχνει καθαρά ότι ο σιαμέζικος ταξινομητής διατηρεί εγγενώς καλύτερη διακριτική ικανότητα σε ολόκληρο το φάσμα των πιθανών κατωφλίων απόφασης, εντοπίζοντας περισσότερες πραγματικές αλληλεπιδράσεις με το κόστος περισσότερων ψευδών συναγερμών στο συγκεκριμένο κατώφλι του 0,5. Πρακτικά, αυτό σημαίνει ότι η επιλογή του «καλύτερου» μοντέλου για ένα σενάριο γενίκευσης σε άγνωστα είδη δεν είναι ανεξάρτητη από τον στόχο της ανάλυσης: όταν προτεραιότητα δίνεται στην αποφυγή της παράλειψης πραγματικών αλληλεπιδράσεων —όπως συνήθως συμβαίνει στη χαρτογράφηση βιολογικών δικτύων και στην αναζήτηση νέων θεραπευτικών στόχων— ο σιαμέζικος ταξινομητής παραμένει η προτιμότερη επιλογή, ενδεχομένως

σε συνδυασμό με αναπροσαρμογή του κατωφλίου απόφασης προς τα κάτω ώστε να αξιοποιηθεί πλήρως το υψηλότερο AUC του.

Συγκρίνοντας τα ευρήματα αυτά με τη σχετική βιβλιογραφία, η παρούσα εργασία επιβεβαιώνει ότι μέσες, ανά πρωτεΐνη αναπαραστάσεις ProtT5 —χωρίς δηλαδή διατήρηση της πληροφορίας σε επίπεδο μεμονωμένου καταλοίπου— περιέχουν επαρκές βιολογικό σήμα ώστε ακόμα και μια σχετικά απλή αρχιτεκτονική σύγκρισης να επιτύχει υψηλή απόδοση, εφόσον ο συνδυασμός των δύο αναπαραστάσεων γίνεται μέσω κατάλληλα σχεδιασμένης σιαμέζικης δομής. Το εύρημα αυτό βρίσκεται σε συμφωνία με εργασίες όπως το STEP [37] και το xCAPT5 [40], οι οποίες αξιοποιούν επίσης σιαμέζικες ή κοινόχρηστες αρχιτεκτονικές κωδικοποίησης πάνω σε αναπαραστάσεις Μετασχηματιστών. Ταυτόχρονα, όμως, αναδεικνύει και τα όριά της σε σχέση με προσεγγίσεις που αξιοποιούν πληροφορία σε επίπεδο καταλοίπου, όπως το ISPRE-SEQ [39] ή το xCAPT5 με τα συνελκτικά του επίπεδα πολλαπλών πυρήνων, καθώς και με μεθόδους που ενσωματώνουν έμμεσα δομική πληροφορία, όπως το MAPE-PPI [41]. Η μέση συγκέντρωση (average pooling) που χρησιμοποιήθηκε στην παρούσα εργασία, αν και υπολογιστικά οικονομική και απλή στην εφαρμογή, ενδέχεται να συμπίπτει υπερβολικά την πληροφορία μεγάλων αλληλουχιών, «αραιώνοντας» το σήμα που προέρχεται από τοπικές περιοχές διεπιφάνειας οι οποίες είναι κρίσιμες για την πραγματική αλληλεπίδραση δύο πρωτεϊνών.

Τέλος, πρέπει να αναγνωριστούν ορισμένοι περιορισμοί της παρούσας μελέτης. Πρώτον, η ετικέτα αλληλεπίδρασης προέρχεται από τη συνδιαστική βαθμολογία εμπιστοσύνης της STRING και όχι από άμεση πειραματική επιβεβαίωση κάθε μεμονωμένου ζεύγους, γεγονός που εισάγει έναν βαθμό θορύβου στην εκπαίδευση και στην αξιολόγηση, ιδίως κοντά στα όρια απόφασης 0, 2 και 0, 8 που χρησιμοποιήθηκαν για τον διαχωρισμό των κλάσεων [44]. Δεύτερον, το κατώφλι ταξινόμησης 0, 5 που χρησιμοποιήθηκε σε όλα τα πειράματα είναι μια αυθαίρετη επιλογή· όπως αναδείχθηκε στο τρίτο πείραμα, η απόδοση των μοντέλων —ιδίως η ισορροπία μεταξύ ΘΠΑ και ανάκλησης— μπορεί να μεταβληθεί σημαντικά με μια πιο συστηματική βελτιστοποίηση του κατωφλίου ανάλογα με τον στόχο της εφαρμογής. Τρίτον, η αξιολόγηση της γενίκευσης βασίστηκε σε μόλις τρία ξένα είδη, αριθμός που περιορίζει τη στατιστική βεβαιότητα των συμπερασμάτων σχετικά με τη μεταφερσιμότητα των αναπαραστάσεων σε ευρύτερο φάσμα οργανισμών. Τέλος, η παρούσα εργασία εστίασε αποκλειστικά σε αλληλουχιακή πληροφορία, χωρίς να αξιοποιεί άμεσα δομικά δεδομένα ή πληροφορία αλληλεπίδρασης σε επίπεδο καταλοίπου, στοιχεία τα οποία, όπως δείχνει η βιβλιογραφία, θα μπορούσαν να βελτιώσουν περαιτέρω την ακρίβεια πρόβλεψης, ιδιαίτερα σε σενάρια γενίκευσης σε πλήρως άγνωστους οργανισμούς.

## 7 Συμπέρασμα

Η παρούσα εργασία είχε ως στόχο την ανάπτυξη και αξιολόγηση μοντέλων βαθιάς μάθησης για την πρόβλεψη αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης αποκλειστικά από αλληλουχίες αμινοξέων, παρακάμπτοντας το υπολογιστικά δαπανηρό στάδιο της ρητής πρόβλεψης της τρισδιάστατης δομής. Τα αποτελέσματα επιβεβαίωσαν ότι οι αναπαραστάσεις που παράγονται από το γλωσσικό μοντέλο ProtT5 εμπεριέχουν επαρκή βιολογική πληροφορία ώστε να επιτρέπουν αξιόπιστη πρόβλεψη αλληλεπιδράσεων, υπό την προϋπόθεση ότι αξιοποιούνται από κατάλληλη αρχιτεκτονική. Από τα επτά μοντέλα που σχεδιάστηκαν και αξιολογήθηκαν, ο Σιαμέζικος Ταξινομητής αναδείχθηκε ως η βέλτιστη επιλογή, επιτυγχάνοντας συνολική ακρίβεια 95,00% και  $F1$  87,22% στην αξιολόγηση επί του ενοποιημένου συνόλου δεδομένων, καθώς και ακρίβεια 91,00% κατά τη δοκιμή του σε τρία εντελώς άγνωστα είδη οργανισμών, υπερτερώντας σταθερά του Αντίληπτρου Βαθιάς Μάθησης σε όλες σχεδόν τις μετρικές και τα πειραματικά σενάρια.

Τα τρία πειράματα που πραγματοποιήθηκαν σκιαγραφούν από κοινού μια σαφή εικόνα. Όταν τα μοντέλα αξιολογούνται σε πρωτεΐνες ειδών που έχουν συμμετάσχει στην εκπαίδευση —είτε ανά είδος ξεχωριστά είτε στο σύνολο του ενοποιημένου συνόλου δεδομένων— ο σιαμέζικος ταξινομητής υπερτερεί χωρίς εξαίρεση, με τη μεγαλύτερη βελτίωση να εντοπίζεται στη θετική προγνωστική αξία σε συνθήκες έντονης ανισορροπίας κλάσεων. Όταν όμως καλείται να γενικεύσει σε πρωτεΐνες οργανισμών που δεν έχει συναντήσει ποτέ, η εικόνα γίνεται πιο σύνθετη: το αντίληπτρο πετυχαίνει ελαφρώς υψηλότερη ακρίβεια χάρη σε μια πιο συντηρητική συμπεριφορά πρόβλεψης, ενώ ο σιαμέζικος ταξινομητής διατηρεί εμφανές προβάδισμα στο AUC και στην ανάκληση, εντοπίζοντας δηλαδή τις περισσότερες πραγματικές αλληλεπιδράσεις. Το εύρημα αυτό επιβεβαιώνει ότι ο κοινόχρηστος κωδικοποιητής μαθαίνει αναπαραστάσεις με ισχυρή και μεταφέρσιμη διακριτική ικανότητα, ενώ η τελική επιλογή κατωφλίου απόφασης αποτελεί ζήτημα βελτιστοποίησης που εξαρτάται από τον συγκεκριμένο στόχο της εφαρμογής.

Συνολικά, τα τρία πειράματα αναδεικνύουν ότι η επιλογή αρχιτεκτονικής σύγκρισης έχει ουσιαστική επίδραση στην απόδοση των μοντέλων πρόβλεψης αλληλεπιδράσεων πρωτεϊνών, ιδιαίτερα σε συνθήκες ανισορροπίας κλάσεων και ετερογένειας πολλαπλών ειδών, χωρίς όμως να υπάρχει μία απόλυτα βέλτιστη αρχιτεκτονική για κάθε δυνατό σενάριο αξιολόγησης. Η επιλογή μεταξύ των δύο προτεινόμενων μοντέλων θα πρέπει, επομένως, να λαμβάνει υπόψη τόσο τον βαθμό εξοικείωσης του μοντέλου με τα υπό εξέταση είδη όσο και τη σχετική σημασία που αποδίδεται στην ελαχιστοποίηση ψευδώς θετικών έναντι ψευδώς αρνητικών προβλέψεων στο συγκεκριμένο πλαίσιο εφαρμογής.

Πέρα από την τεχνική σύγκριση των δύο αρχιτεκτονικών, η εργασία αυτή έχει και ευρύτερη πρακτική σημασία. Η δυνατότητα πρόβλεψης αλληλεπιδράσεων

πρωτεΐνης-πρωτεΐνης απευθείας από την αλληλουχία αμινοξέων, χωρίς να απαιτείται το υπολογιστικά δαπανηρό στάδιο της ρητής πρόβλεψης τρισδιάστατης δομής, καθιστά τη μεθοδολογία αυτή ιδιαίτερα προσιτή και επεκτάσιμη. Στο πλαίσιο της ογκολογίας και της βιοϊατρικής έρευνας γενικότερα, όπου η ταχεία χαρτογράφηση διαταραγμένων πρωτεϊνικών δικτύων μπορεί να συμβάλει στον εντοπισμό νέων θεραπευτικών στόχων, μια τέτοια προσέγγιση προσφέρει ένα υπολογιστικά οικονομικό εργαλείο πρώτης γραμμής για τη διαλογή υποψήφιων αλληλεπιδράσεων, πριν αυτές προχωρήσουν σε χρονοβόρα και δαπανηρή πειραματική επιβεβαίωση. Η ίδια μεθοδολογία μπορεί άμεσα να επεκταθεί σε μεγαλύτερα και πιο ετερογενή σύνολα δεδομένων, καλύπτοντας ευρύτερο φάσμα οργανισμών, ιστών και τύπων νόσου.

Ως μελλοντική κατεύθυνση, η απόδοση των μοντέλων θα μπορούσε ενδεχομένως να βελτιωθεί περαιτέρω με τη χρήση πλουσιότερων αναπαραστάσεων σε επίπεδο μεμονωμένου καταλοίπου αντί της μέσης συγκέντρωσης ολόκληρης της αλληλουχίας, με μηχανισμούς προσοχής για τη στάθμιση της συνεισφοράς διαφορετικών περιοχών της πρωτεΐνης, ή με τη συνδυαστική ενσωμάτωση δομικής πληροφορίας, όπως προτείνεται από πρόσφατες εργασίες της βιβλιογραφίας. Επιπλέον, μια συστηματική διερεύνηση της επίδρασης του κατωφλίου ταξινόμησης, καθώς και η επέκταση της αξιολόγησης γενίκευσης σε μεγαλύτερο αριθμό ξένων ειδών, θα προσέφεραν πιο στέρεα συμπεράσματα ως προς τα όρια μεταφερσιμότητας των αναπαραστάσεων ProtT5. Συνολικά, η παρούσα εργασία καταδεικνύει ότι ο συνδυασμός προ-εκπαιδευμένων γλωσσικών μοντέλων πρωτεϊνών με κατάλληλα σχεδιασμένες αρχιτεκτονικές σύγκρισης αποτελεί μια ελπιδοφόρα και επεκτάσιμη κατεύθυνση για την πρόβλεψη αλληλεπιδράσεων πρωτεϊνών, ικανή να συμβάλει ουσιαστικά στην ταχύτερη και οικονομικότερη χαρτογράφηση του πρωτεϊνικού διαδράματος.

## ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] Michael I. Jordan and Tom M. Mitchell. “Machine Learning: Trends, Perspectives, and Prospects”. In: *Science* 349.6245 (2015), pp. 255–260. DOI: 10.1126/science.aaa8415. URL: <https://www.science.org/doi/10.1126/science.aaa8415>.
- [2] Koosha Sharifani and Mahyar Amini. “Machine Learning and Deep Learning: A Review of Methods and Applications”. In: *World Information Technology and Engineering Journal* 10.7 (2023). Available at SSRN: 4458723, pp. 3897–3904. DOI: 10.2139/ssrn.4458723. URL: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4458723](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4458723).
- [3] Fei Guo et al. “Foundation Models in Bioinformatics”. In: *National Science Review* 12.4 (2025), nwaf028. DOI: 10.1093/nsr/nwaf028. URL: <https://academic.oup.com/nsr/article/12/4/nwaf028/7979309>.
- [4] Ruth Nussinov et al. “AlphaFold, Artificial Intelligence (AI), and Allostery”. In: *The Journal of Physical Chemistry B* 126.32 (2022), pp. 6372–6383. DOI: 10.1021/acs.jpcc.2c04346. URL: <https://pubs.acs.org/doi/10.1021/acs.jpcc.2c04346>.
- [5] Barry J. Wythoff. “Backpropagation neural networks: A tutorial”. In: *Chemo-metrics and Intelligent Laboratory Systems* 18.2 (1993), pp. 115–155. DOI: 10.1016/0169-7439(93)80052-J. URL: <https://www.sciencedirect.com/science/article/pii/016974399380052J>.
- [6] Davide Chicco. “Siamese Neural Networks: An Overview”. In: *Artificial Neural Networks*. Ed. by Hugh M. Cartwright. Methods in Molecular Biology. Springer US, 2021, pp. 73–94. DOI: 10.1007/978-1-0716-0826-5\_3. URL: [https://link.springer.com/protocol/10.1007/978-1-0716-0826-5\\_3](https://link.springer.com/protocol/10.1007/978-1-0716-0826-5_3).
- [7] Suja Cherukullapurath Mana and T. Sasipraba. “Research on Cosine Similarity and Pearson Correlation Based Recommendation Models”. In: *Journal of Physics: Conference Series* 1770.1 (2021), p. 012014. DOI: 10.1088/1742-6596/1770/1/012014. URL: <https://iopscience.iop.org/article/10.1088/1742-6596/1770/1/012014>.
- [8] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. DOI: 10.48550/ARXIV.1412.6980. URL: <https://arxiv.org/abs/1412.6980>.
- [9] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2017. DOI: 10.48550/ARXIV.1711.05101. URL: <https://arxiv.org/abs/1711.05101>.

- [10] Lei Huang et al. “Normalization Techniques in Training DNNs: Methodology, Analysis and Application”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.8 (2023), pp. 10173–10196. DOI: 10.1109/TPAMI.2023.3250241. URL: <https://ieeexplore.ieee.org/document/10056354>.
- [11] Geoffrey E. Hinton et al. *Improving neural networks by preventing co-adaptation of feature detectors*. 2012. DOI: 10.48550/ARXIV.1207.0580. URL: <https://arxiv.org/abs/1207.0580>.
- [12] Ashish Vaswani et al. “Attention Is All You Need”. In: *arXiv preprint arXiv:1706.03762* (2017). arXiv: 1706.03762 [cs.CL]. URL: <https://arxiv.org/abs/1706.03762>.
- [13] Geoffrey M. Cooper. “Protein Folding and Processing”. In: *The Cell: A Molecular Approach*. 2nd. Sunderland, MA: Sinauer Associates, 2000. URL: <https://www.ncbi.nlm.nih.gov/books/NBK9843/>.
- [14] Bruce Alberts et al. *Molecular Biology of the Cell*. 6th ed. New York: Garland Science, 2015. URL: <https://www.ncbi.nlm.nih.gov/books/NBK20684/>.
- [15] Carl Branden and John Tooze. *Introduction to Protein Structure*. 2nd ed. New York: Garland Publishing, 1999.
- [16] Arthur M. Lesk. *Introduction to Protein Science*. Oxford: Oxford University Press, 2010.
- [17] Ken A. Dill. “Dominant Forces in Protein Folding”. In: *Biochemistry* 29.31 (1990), pp. 7133–7155. DOI: 10.1021/bi00483a001. URL: <https://pubmed.ncbi.nlm.nih.gov/2207096/>.
- [18] Thomas E. Creighton. *Proteins: Structures and Molecular Properties*. 2nd ed. New York: W. H. Freeman, 1993.
- [19] Alan Fersht. *Structure and Mechanism in Protein Science*. New York: W. H. Freeman, 1999.
- [20] Ken A. Dill et al. “The Protein Folding Problem”. In: *Annual Review of Biophysics* 37 (2008), pp. 289–316. DOI: 10.1146/annurev.biophys.37.092707.153558.
- [21] Ken A. Dill and Justin L. MacCallum. “The Protein-Folding Problem, 50 Years On”. In: *Science* 338.6110 (2012), pp. 1042–1046. DOI: 10.1126/science.1219021. URL: <https://pubmed.ncbi.nlm.nih.gov/23180855/>.
- [22] Tomasz Wlodarski and Bojan Zagrovic. “Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.46 (2009), pp. 19346–19351. DOI: 10.1073/pnas.0906966106. URL: <https://www.pnas.org/doi/10.1073/pnas.0906966106>.

- [23] Jon Kleinberg et al. “Scaling laws for the dynamics of networked systems / related PNAS article”. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.17 (2004), pp. 6740–6745. DOI: 10.1073/pnas.0300619101. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3944927/>.
- [24] S. Naz et al. “Insight into de-regulation of amino acid feedback inhibition: a focus on structure analysis method”. In: *Microbial Cell Factories* 22.1 (2023), p. 161. DOI: 10.1186/s12934-023-02178-z. URL: <https://link.springer.com/article/10.1186/s12934-023-02178-z>.
- [25] Lauren Budenholzer et al. “Proteasome Structure and Assembly”. In: *Journal of Molecular Biology* 429.22 (2017), pp. 3500–3524. DOI: 10.1016/j.jmb.2017.05.027. URL: <https://www.sciencedirect.com/science/article/pii/S002228361730270X>.
- [26] Bruce E. Logan and Korneel Rabaey. “Conversion of Wastes into Bioelectricity and Chemicals by Using Microbial Electrochemical Technologies”. In: *Science* 337.6095 (2012), pp. 686–690. DOI: 10.1126/science.1217412. URL: <https://pubmed.ncbi.nlm.nih.gov/24183815/>.
- [27] Awais A. Yaqoob et al. “Recent Advances in Anodes for Microbial Fuel Cells: An Overview”. In: *Frontiers in Chemistry* 9 (2021), p. 682675. DOI: 10.3389/fchem.2021.682675. URL: <https://www.frontiersin.org/articles/10.3389/fchem.2021.682675/full>.
- [28] Ozlem Keskin et al. “Principles of Protein–Protein Interactions: What are the Preferred Ways For Proteins To Interact?” In: *Chemical Reviews* 108.4 (2008), pp. 1225–1244. DOI: 10.1021/cr040409x.
- [29] Damian Szklarczyk et al. “The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest”. In: *Nucleic Acids Research* 51 (2023). DOI: 10.1093/nar/gkac1000.
- [30] Katja Luck, Dae-Kyum Kim, Luke Lambourne, et al. “A reference map of the human binary protein interactome”. In: *Nature* 580 (2020), pp. 402–408. DOI: 10.1038/s41586-020-2188-x.
- [31] Marc Vidal, Michael E. Cusick, and Albert-László Barabási. “Interactome Networks and Human Disease”. In: *Cell* 144.6 (2011), pp. 986–998. DOI: 10.1016/j.cell.2011.02.016.
- [32] Tracy L. Nero et al. “Oncogenic protein interfaces: small molecules, peptides and aptamers for targeted cancer therapy”. In: *Nature Reviews Cancer* 14 (2014), pp. 248–262. DOI: 10.1038/nrc3690.
- [33] V. Srinivasa Rao et al. “Protein-Protein Interaction Detection: Methods and Analysis”. In: *International Journal of Proteomics* 2014 (2014), p. 147648. DOI: 10.1155/2014/147648.

- [34] Maria Littmann et al. “Protein Embeddings and Deep Learning Predict Binding Residues for Various Ligand Classes”. In: *Scientific Reports* 11 (2021), p. 23916. DOI: 10.1038/s41598-021-03431-4.
- [35] Fangping Wan and Jianyang Zeng. “Deep Learning with Feature Embedding for Compound-Protein Interaction Prediction”. In: *bioRxiv* (2016). DOI: 10.1101/086033. URL: <https://www.biorxiv.org/content/10.1101/086033v1>.
- [36] Yao Yao et al. “An Integration of Deep Learning with Feature Embedding for Protein–Protein Interaction Prediction”. In: *PeerJ* 7 (2019), e7126. DOI: 10.7717/peerj.7126.
- [37] Sumit Madan et al. “Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings”. In: *Patterns* 3.9 (2022), p. 100551. DOI: 10.1016/j.patter.2022.100551.
- [38] Christian Dallago et al. “Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets”. In: *Current Protocols* 1.5 (2021), e113. DOI: 10.1002/cpz1.113.
- [39] Matteo Manfredi et al. “ISPRED-SEQ: Deep Neural Networks and Embeddings for Predicting Interaction Sites in Protein Sequences”. In: *Journal of Molecular Biology* 435.14 (2023), p. 167963. DOI: 10.1016/j.jmb.2023.167963. URL: <https://www.sciencedirect.com/science/article/pii/S0022283623000190>.
- [40] Thanh Hai Dang and Tien Anh Vu. “xCAPT5: Protein–Protein Interaction Prediction Using Deep and Wide Multi-Kernel Pooling Convolutional Neural Networks with Protein Language Model”. In: *BMC Bioinformatics* 25.1 (2024), p. 106. DOI: 10.1186/s12859-024-05725-6. URL: <https://link.springer.com/article/10.1186/s12859-024-05725-6>.
- [41] Lirong Wu et al. “MAPE-PPI: Towards Effective and Efficient Protein-Protein Interaction Prediction via Microenvironment-Aware Protein Embedding”. In: *International Conference on Learning Representations (ICLR)*. Spotlight paper. 2024. URL: <https://openreview.net/forum?id=itGkF993gz>.
- [42] “Protein–protein interaction prediction with deep learning: A comprehensive review”. In: *Computational and Structural Biotechnology Journal* (2022). DOI: 10.1016/j.csbj.2022.08.070.
- [43] Ahmed Elnaggar et al. “ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). DOI: 10.1109/TPAMI.2021.3095381.
- [44] Damian Szklarczyk et al. “The STRING database in 2025: protein networks with directionality of regulation”. In: *Nucleic Acids Research* 53.D1 (2025), pp. D730–D737. DOI: 10.1093/nar/gkae1113. URL: <https://academic.oup.com/nar/article/53/D1/D730/7903368>.

- [45] Dewei Hu et al. “SPACE: STRING Proteins as Complementary Embeddings”. In: *Bioinformatics* 41.9 (2025), btaf496. DOI: 10.1093/bioinformatics/btaf496.
- [46] PLOS Computational Biology. “Zero-shot segmentation using embeddings from a protein language model identifies functional regions in the human proteome”. In: *PLOS Computational Biology* (2025). DOI: 10.1371/journal.pcbi.1012929.
- [47] “xCAPT5: protein–protein interaction prediction using deep and wide multi-kernel pooling convolutional neural networks with protein language model”. In: *BMC Bioinformatics* (2024). DOI: 10.1186/s12859-024-05717-0.
- [48] Hervé Abdi and Lynne J. Williams. “Principal component analysis”. In: *WIREs Computational Statistics* 2.4 (2010), pp. 433–459. DOI: 10.1002/wics.101. URL: <https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.101>.
- [49] “Structure-Based Approaches for Protein–Protein Interaction Prediction Using Machine Learning and Deep Learning”. In: *Biomolecules* 15 (2025). DOI: 10.3390/biom15010141.
- [50] Lei Huang et al. “Normalization Techniques in Training DNNs: Methodology, Analysis and Application”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.8 (2023), pp. 10173–10196. DOI: 10.1109/TPAMI.2023.3250241. URL: <https://ieeexplore.ieee.org/abstract/document/10056354/>.
- [51] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958. URL: <http://jmlr.org/papers/v15/srivastava14a.html>.
- [52] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. Published as a conference paper at ICLR 2015. 2017. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980 [cs.LG]. URL: <https://arxiv.org/abs/1412.6980>.
- [53] Juan Terven et al. “A comprehensive survey of loss functions and metrics in deep learning”. In: *Artificial Intelligence Review* 58.7 (2025), p. 195. DOI: 10.1007/s10462-025-11198-7. URL: <https://link.springer.com/article/10.1007/s10462-025-11198-7>.
- [54] Davide Chicco. “Siamese Neural Networks: An Overview”. In: *Artificial Neural Networks*. Vol. 2190. Methods in Molecular Biology. Springer, 2021, pp. 73–94. DOI: 10.1007/978-1-0716-0826-5\_3. URL: [https://link.springer.com/protocol/10.1007/978-1-0716-0826-5\\_3](https://link.springer.com/protocol/10.1007/978-1-0716-0826-5_3).