Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπλογιστών
Τομέας Ηλεκτρονικής και Υπλογιστών

Κείμενο υποστήριξης διπλωματικής εργασίας

Μάθηση πολλαπλών εργασιών στη μοντελοποίηση διαταραχών με εφαρμογή σε μονοκυτταρικά δεδομένα

Θεόδωρος Κατζάλης

Επιβλέπων:
Περικλής Μήτκας
Καθηγητής Α.Π.Θ

Συνεπιβλέπων:
Φώτης Ε. Ψωμόπουλος
Ερευνητής στο Ινστιτούτο Εφαρμοσμένων Βιοεπιστημών (INAB) του
Εθνικού Κέντρου Έρευνας και Τεχνολογικής Ανάπτυξης (ΕΚΕΤΑ)

Θεσσαλονίκη, Νοέμβριος 2025

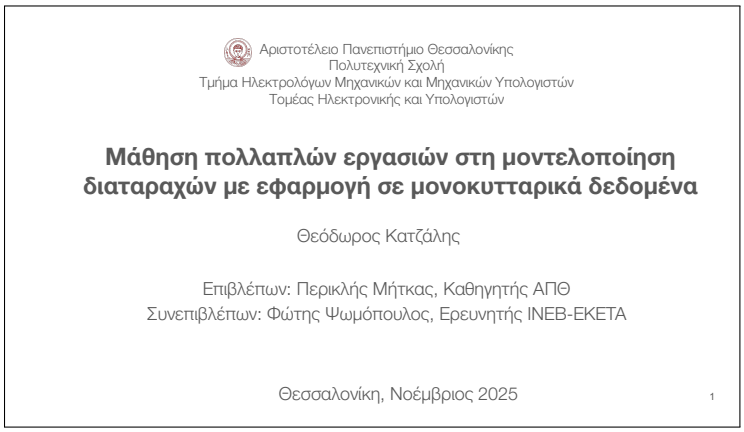
Περιεχόμενα

1η διαφάνεια

Καλημέρα/Καλησπέρα σας,

Ονομάζομαι Θεόδωρος Κατζάλης και θα σας παρουσιάσω την διπλωματική μου εργασία με τίτλο ”Μάθηση πολλαπλών εργασιών στη μοντελοποίηση διαταραχών με εφαρμογή σε μονοκυτταρικά δεδομένα”.

Η διπλωματική αυτή είναι προϊόν συνεργασίας μεταξύ του ΑΠΘ, με επιβλέποντα τον κ. Περικλή Μήτκα, και της ομάδας βιοπληροφορικής του Ινστιτούτου Εφαρμοσμένων Βιοεπιστημών (INAB) του Εθνικού Κέντρου Έρευνας και Τεχνολογικής Ανάπτυξης (ΕΚΕΤΑ), με συνεπιβλέποντα τον κ. Φώτη Ψωμόπουλο.

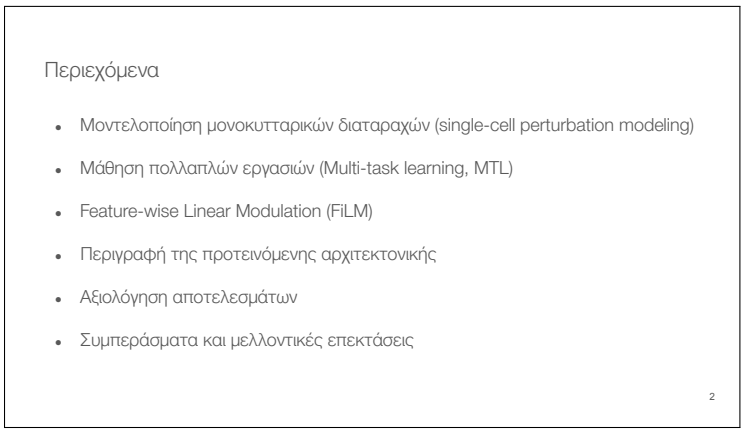


2η διαφάνεια

’Οσον αφορά τη δομή της παρουσίασης, αρχικά θα δώσουμε τους ορισμούς της μοντελοποίησης μονοκυτταρικών διαταραχών και της μάθησης πολλαπλών εργασιών.

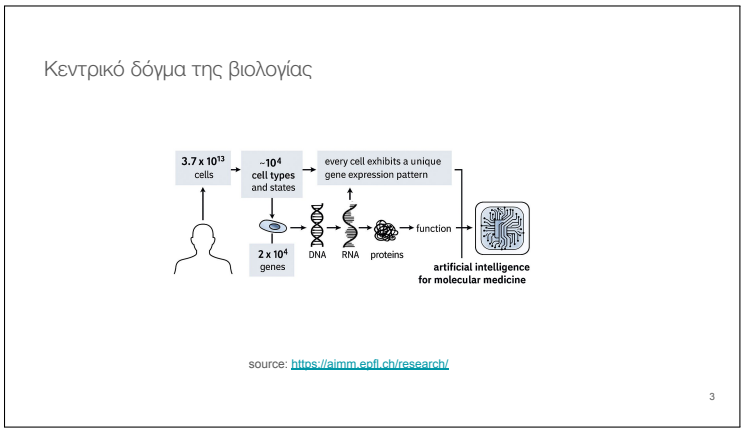
Στη συνέχεια, θα αναλύσουμε την αρχιτεκτονική που σχεδίασαμε και υλοποιήσαμε, αξιοποιώντας τη μέθοδο της βιβλιογραφίας, ονόματι Feature-wise Linear Modulation, με στόχο να καταφέρουμε να συνδυάσουμε πολλαπλές εργασίες σε ένα ενιαίο μοντέλο.

Τέλος, θα παρουσιάσουμε τα αποτελέσματα της μεθόδου μας σε σχέση με άλλες προσεγγίσεις και θα κλείσουμε με τα συμπεράσματα και τις μελλοντικές κατευθύνσεις της έρευνας.



3η διαφάνεια

Πριν μιλήσουμε για τη μοντελοποίηση μονοκυτταρικών διαταραχών, είναι πολύ σημαντικό να αντιληφθούμε την αφετηρία της μοριακής βιολογίας, και συγκεκριμένα το κεντρικό δόγμα της βιολογίας.

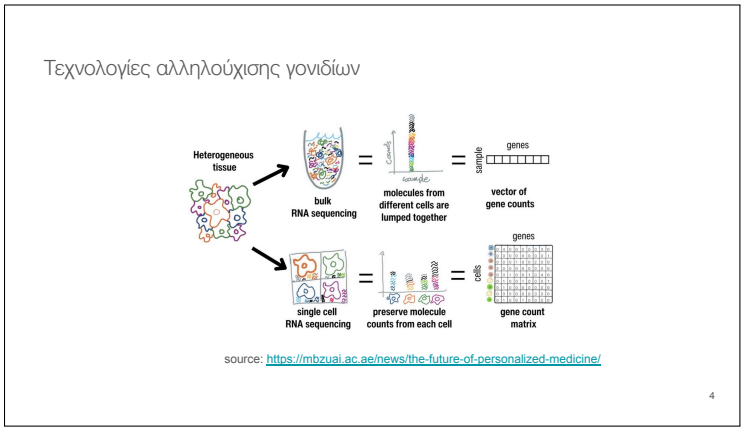


Το 1958, ανακαλύφθηκε ότι η ροή της γενετικής πληροφορίας σε έναν οργανισμό ακολουθεί την πορεία DNA ⟶ RNA ⟶ Πρωτεΐνη. Εναλλακτικά, ο κόσμος της βιοχημείας και της ανάλυσης των πρωτεϊνών γεφυρώθηκε με την γενετική, και διαπιστώθηκε ότι το RNA, ο αγγελιοφόρος του DNA, είναι η συνταγή για την παραγωγή των πρωτεϊνών, και κατ'επέκταση, για την έκφραση των χαρακτηριστικών και των λειτουργιών ενός κυττάρου.

Ωστόσο, αυτή η ροή είναι ιδιαίτερη περίπλοκη, εξαιτίας της ποικιλομορφίας των κυττάρων, των μοριακών μηχανισμών, και των περιβαλλοντικών συνθηκών που την επηρεάζουν.

4η διαφάνεια

Στην προσπάθεια κατανόησης αυτής της πολυπλοκότητας, αναπτύχθηκαν τεχνικές αλληλούχησης, οι οποίες έχουν τη δυνατότητα να καταγράφουν την γονιδιακή έκφραση σε διάφορα βιολογικά δείγματα (επίπεδο ιστών, κυτταρικών πληθυσμών κλπ.).

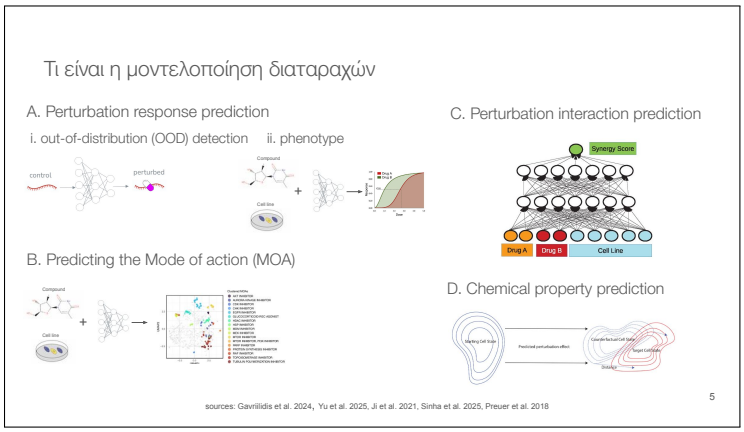


Αρχικά, έχουμε τη λεγόμενη bulk RNA sequencing μέθοδο, στην οποία το εκάστοτε βιολογικό δείγμα αντιμετωπίζεται ως ένα ομοιογενές σύνολο κυττάρων, και η γονιδιακή έκφραση μετριέται ως ο μέσος όρος της έκφρασης όλων των κυττάρων στο δείγμα.

Με την πρόοδο της τεχνολογίας, γεννήθηκε μια πιο εξελιγμένη μονοκυτταρική τεχνολογία, η "next-generation sequencing", η οποία επιτρέπει τη μέτρηση της έκφρασης γονιδίων σε επίπεδο μεμονωμένου κυττάρου γνωστή ως single cell RNA sequencing. Αυτός ο τρόπος αλληλούχισης μας δίνει τη δυνατότητα να μελετήσουμε την ετερογένεια των κυττάρων (μέσα σε έναν ιστό ή έναν οργανισμό), και να κατανοήσουμε πώς διαφορετικοί τύποι κυττάρων αντιδρούν σε διάφορες διαταραχές, όπως ασθένειες ή φαρμακευτικές παρεμβάσεις.

5η διαφάνεια

Έχοντας κάνει αυτή την εισαγωγή, μπορούμε τώρα να μιλήσουμε για τη μοντελοποίηση διαταραχών.



Κεντρικός στόχος της μοντελοποίησης αυτής είναι η ανάπτυξη υπολογιστικών μοντέλων που μπορούν να προβλέψουν πώς τα κύτταρα ανταποκρίνονται σε διάφορες διαταραχές. Οι διαταραχές αυτές μπορεί να περιλαμβάνουν τη χορήγηση φαρμάκων, γενετικές τροποποιήσεις, ή περιβαλλοντικούς παράγοντες.

Συγκεκριμένα η μοντελοποίηση διαταραχών έχει τους εξής τέσσερις βασικούς στόχους:

Ο πρώτος στόχος έχει δύο υποκατηγορίες. Η πρώτη υποκατηγορία σχετίζεται με την πρόβλεψη της γονιδιακής έκφρασης μετά από μια διαταραχή, ενώ η δεύτερη αφορά την πρόβλεψη φαινοτυπικών αλλαγών, δηλαδή την πρόβλεψη οποιουδήποτε εμφανούς/παρατηρήσιμου χαρακτηριστικού, όπως η κυτταρική βιωσιμότητα.

Ο δεύτερος στόχος εστιάζει στην πρόβλεψη του μηχανισμού δράσης μιας διαταραχής (mode of action). Αυτό συνεπάγεται την αναγνώριση των σηματοδοτικών οδών (signaling pathways) και των πρωτεϊνών-στόχων (target proteins) που ενεργοποιούνται ή αναστέλλονται ως απόκριση σε μια δεδομένη διαταραχή.

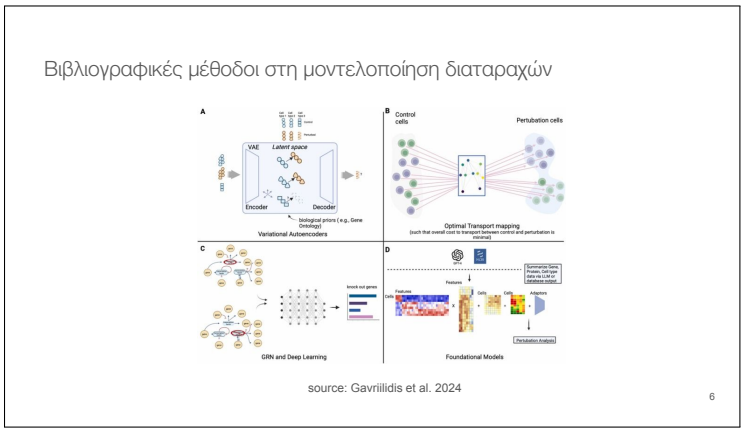
Ο τρίτος στόχος σχετίζεται με την αλληλεπίδραση μεταξύ διαταραχών, και με το αν η εφαρμογή δύο ή περισσότερων διαταραχών ταυτόχρονα έχει συνεργιστικά ή ανταγωνιστικά αποτελέσματα.

Τέλος, ο τέταρτος στόχος αφορά τη σχεδίαση χημικών ενώσεων με επιθυμητά χαρακτηριστικά, θεωρώντας ως δεδομένο το επιθυμητό γονιδιακό αποτέλεσμα πριν και μετά τη διαταραχή.

Αξίζει να σημειωθεί ότι τα περισσότερα dataset, ειδικά στην περίπτωση της πρόβλεψης φαινοτυπικών αλλαγών, και στον μηχανισμό δράσης, προέρχονται από bulk RNA-seq δεδομένα, καθώς η μονοκυτταρική αλληλούχιση είναι μια σχετικά νέα τεχνολογία με περιορισμένη διαθεσιμότητα δεδομένων. Συγκεκριμένα, οι περισσότερες μέθοδοι αιχμής εστιάζουν στην πρόβλεψη εκτός κατανομής, με τη χρήση μονοκυτταρικών δεδομένων. Όπως θα αναφερθεί στη συνέχεια, η μέθοδός μας επικεντρώνεται σε αυτήν την κατηγορία.

6η διαφάνεια

Στη βιβλιογραφία, έχουν προταθεί διάφορες μέθοδοι για τη μοντελοποίηση διαταραχών σε μονοκυτταρικά δεδομένα και έχουν κατηγοριοποιηθεί στις εξής τέσσερις βασικές προσεγγίσεις:



Η πρώτη βασίζεται σε autoencoders και στην αξιοποίηση του λανθάνοντος χώρου (latent space) για την μοντελοποίηση της διαταραχής και της εξαγωγής χαρακτηριστικών.

Η δεύτερη χρησιμοποιεί την προσέγγιση του optimal transport για τη μοντελοποίηση της μετάβασης μεταξύ καταστάσεων κυττάρων πριν και μετά από μια διαταραχή.

Η τρίτη μοντελοποιεί τη γονιδιακή ρύθμιση με τη χρήση γραφικών νευρωνικών δικτύων (graph neural networks), λαμβάνοντας υπόψη τις αλληλεπιδράσεις μεταξύ γονιδίων.

Τέλος, η τέταρτη κατηγορία περιλαμβάνει foundational models που έχουν εκπαιδευτεί σε πολύ μεγάλα σύνολα δεδομένων.

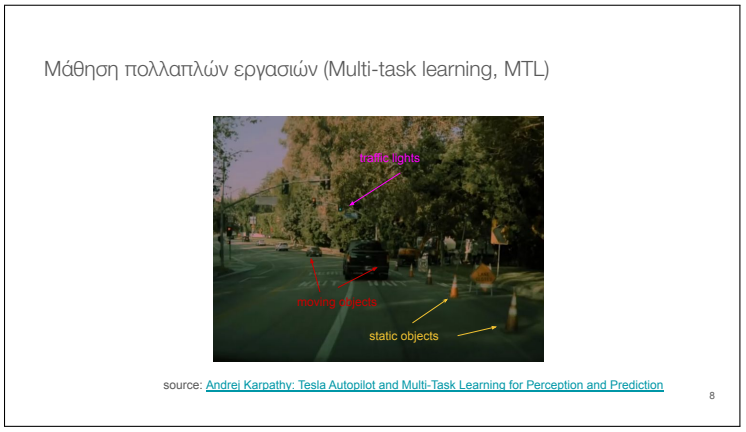
7η διαφάνεια

Έχοντας αποκτήσει μια εικόνα των προκλήσεων της μοντελοποίησης διαταραχών, αλλά και των μεθόδων που έχουν εφαρμοστεί, καταλαβαίνουμε ότι υπάρχει μια σημαντική πολυπλοκότητα και ποικιλία προσεγγίσεων.

Μία από τις σύνθετες και προηγμένες μεθόδους, που θα μπορούσε να θεωρηθεί πολλά υποσχόμενη για να διαχειριστεί αυτήν την πολυπλοκότητα, χωρίς ταυτόχρονα να έχει μελετηθεί εκτενώς στη βιβλιογραφία, είναι η μάθηση πολλαπλών εργασιών (multi-task learning).

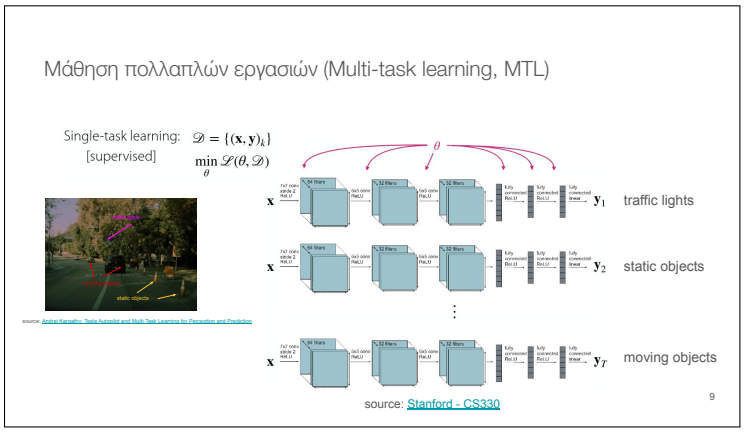
8η διαφάνεια

Για να κατανοήσουμε την έννοια της μάθησης πολλαπλών εργασιών, ας εξετάσουμε αρχικά την αναγνώριση αντικειμένων σε μια εικόνα. Για παράδειγμα, σε αυτήν την εικόνα που βλέπουμε εδώ επιθυμούμε να αναγνωρίσουμε στατικά αντικείμενα στον δρόμο όπως οι κώνοι, δυναμικά αντικείμενα όπως τα αυτοκίνητα, αλλά και φανάρια.



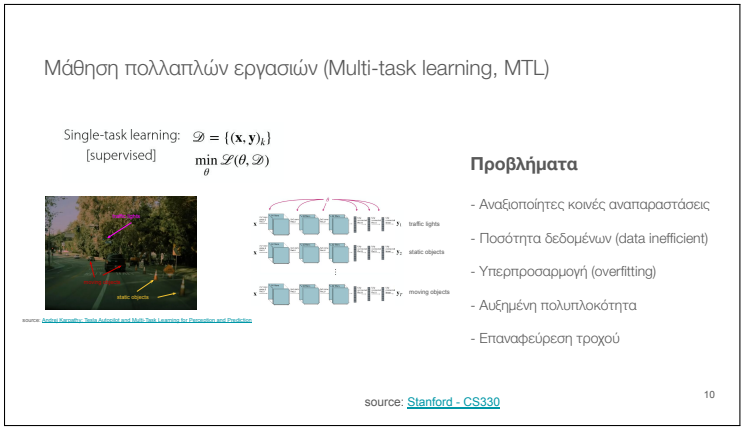
9η διαφάνεια

Στην περίπτωση αυτήν, θα μπορούσαμε να εκπαιδεύσουμε τρία ξεχωριστά μοντέλα, ένα για κάθε εργασία.



10η διαφάνεια

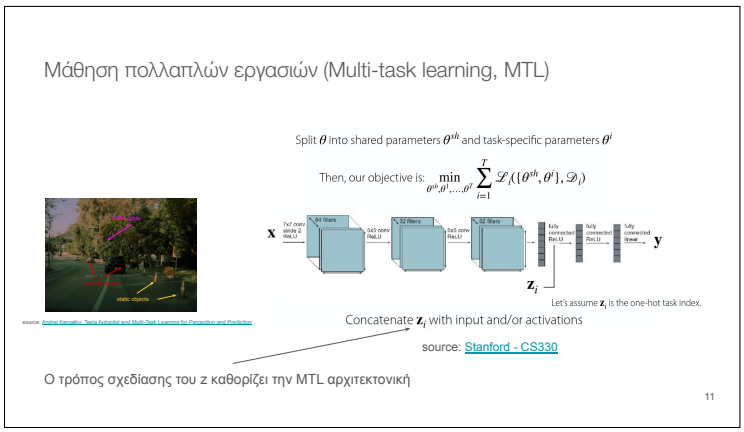
Η προσέγγιση αυτή, ωστόσο, μπορεί να είναι αναποτελεσματική, καθώς τα μοντέλα δεν μοιράζονται πληροφορίες μεταξύ τους, και το εκάστοτε μοντέλο ξεκινάει από το μηδέν, απαιτώντας μεγάλο όγκο δεδομένων, υπολογιστικούς πόρους, και χρόνο εκπαίδευσης.



11η διαφάνεια

Μια πιο αποδοτική προσέγγιση είναι η μάθηση πολλαπλών εργασιών, όπου ένα ενιαίο μοντέλο εκπαιδεύεται ταυτόχρονα σε όλες τις εργασίες, μοιράζοντας κοινές παραστάσεις και χαρακτηριστικά.

Για να επιτευχθεί αυτό, καίριας σημασίας είναι η σχεδίαση του διανύσματος που καθορίζει την εκάστοτε εργασία και την αρχιτεκτονική του μοντέλου, ώστε αυτό να μπορεί να προσαρμόζεται δυναμικά ανάλογα με την εργασία που εκτελείται.



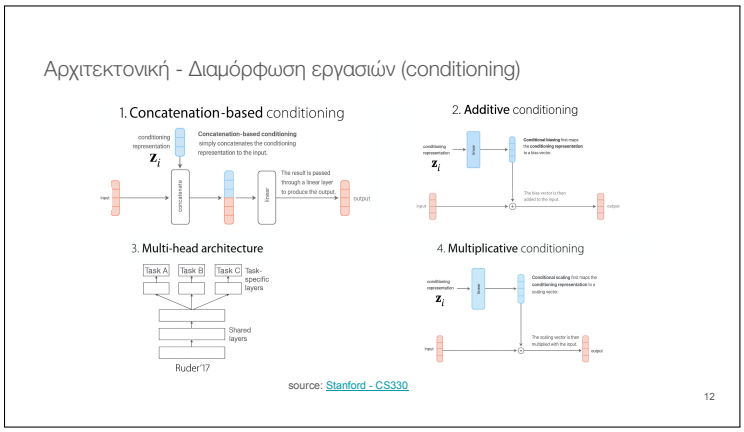
12η διαφάνεια

Το z, το σήμα δηλαδή διαμόρφωσης των εργασιών, μπορεί να μετασχηματιστεί με διάφορους τρόπους.

Για παράδειγμα, με τον μετασχηματισμό της συνένωσης (concatenation), όπου το z συνενώνεται με το διάνυσμα εισόδου πριν υποστεί γραμμικό μετασχηματισμό.

Στη συνέχεια, με τον αθροιστικό μετασχηματισμό (additive transformation), όπου το z προστίθεται με το διάνυσμα εισόδου.

αλλά και παρόμοια, έχουμε τον πολλαπλασιαστικό μετασχηματισμό (multiplicative transformation), όπου το z πολλαπλασιάζεται με το διάνυσμα εισόδου.



Τέλος, υπάρχει και η αρχιτεκτονική πολλαπλής κεφαλής, όπου κάθε εργασία έχει πλέον το δικό της ξεχωριστό κλάδο μέσα στο μοντέλο, χωρίς την παρουσία κάποιου διανύσματος κωδικοποίησης της εργασίας (z).

13η διαφάνεια

Έχοντας παρουσιάσει και εξηγήσει τη μάθηση πολλαπλών εργασιών, ας δούμε πώς αυτή μπορεί να εφαρμοστεί στη μοντελοποίηση διαταραχών.

Αρχικά δεν ήταν πολύ ξεκάθαρο ποιες εργασίες θα μπορούσαν να συνδυαστούν αποτελεσματικά σε ένα ενιαίο μοντέλο, διότι η μάθηση πολλαπλών εργασιών παρουσιάζει αρκετές προκλήσεις.

Μία από αυτές για παράδειγμα είναι η διαθεσιμότητα ενός διευρυμένου συνόλου δεδομένων, έτσι ώστε να μπορεί να καλύψει όλες τις εργασίες.

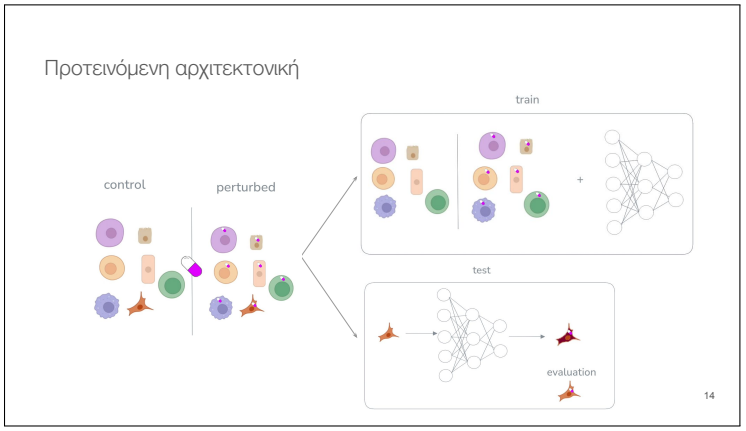
Αυτό είναι ιδιαίτερα απαιτητικό, αν σκεφτούμε ότι τα περισσότερα datasets που χρησιμοποιούνται για τους στόχους της πρόβλεψης αλληλεπίδρασης διαταραχών και του μηχανισμού δράσης (δηλαδή ο δεύτερος και τρίτος στόχος που προανέφερα), προέρχονται απο bulk-RNA seq δεδομένα, και η συσχέτιση τους με μονοκυτταρικά δεδομένα δεν είναι άμεση και έγκυρη λόγω διαφορετικών πειραματικών πρωτοκόλλων και συνθηκών. Λειτουργούν δηλαδή σε διαφορετικό επίπεδο λεπτομέρειας (granurality).

Επιπλέον, η επίλυση των προβλημάτων στη μοντελοποίηση διαταραχών μπορεί να μην γίνεται αποκλειστικά παράλληλα (όπως στην αναγνώριση στάσιμων και δυναμικών αντικειμένων στην εικόνα που είδαμε), με αποτέλεσμα η ταυτόχρονη επίλυση πολλαπλών εργασιών να απαιτεί μια πιο εξειδικευμένη προσέγγιση διάταξης αυτών, αυξάνοντας την πολυπλοκότητα ενός υποψήφιου μοντέλου μάθησης πολλαπλών εργασιών.

Για τον λόγο αυτόν, εστιάσαμε κυρίως στο πρώτο πρόβλημα, και συγκεκριμένα στην πρόβλεψη εκτός κατανομής (out-of-distribution, OOD) της γονιδιακής έκφρασης μετά από μια δαταραχή.

14η διαφάνεια

Ας δούμε τώρα με λεπτομέρεια τι σημαίνει πρόβλεψη εκτός κατανομής. Ένα τυπικό σύνολο δεδομένων για τη μοντελοποίηση διαταραχών περιλαμβάνει δείγματα γονιδιακής έκφρασης πριν και μετά από μια διαταραχή, μαζί με πληροφορίες για τον τύπο κυττάρου και τη διαταραχή που εφαρμόστηκε.

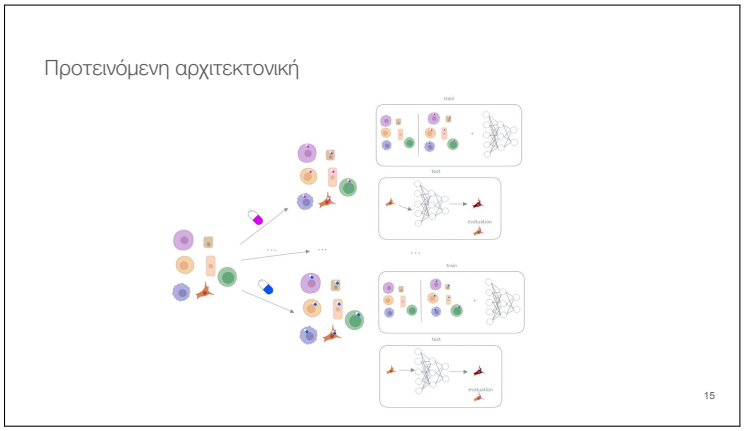


Στόχος είναι να προβλέψουμε τη γονιδιακή έκφραση μετά από τη διαταραχή, βασιζόμενοι στον τύπο της και στην αρχική γονιδιακή έκφραση.

Για να αξιολογήσουμε την ικανότητα ενός μοντέλου να γενικεύει σε άγνωστες καταστάσεις, χωρίζουμε τα δεδομένα σε εκπαιδευτικό και δοκιμαστικό σύνολο, όπου στο τελευταίο χρησιμοποιούμε ένα διαταραγμένο τύπο κυττάρου που δεν έχει εμφανιστεί στο εκπαιδευτικό σύνολο.

15η διαφάνεια

Βέβαια, υπάρχουν σύνολα δεδομένων, τα οποία περιλαμβάνουν περισσότερες από μία διαταραχές, και για τις μεθόδους αιχμής που είναι σχεδιασμένες για έναν τύπο διαταραχής, θα πρέπει να δημιουργηθούν ξεχωριστά μοντέλα για κάθε μία από αυτές.

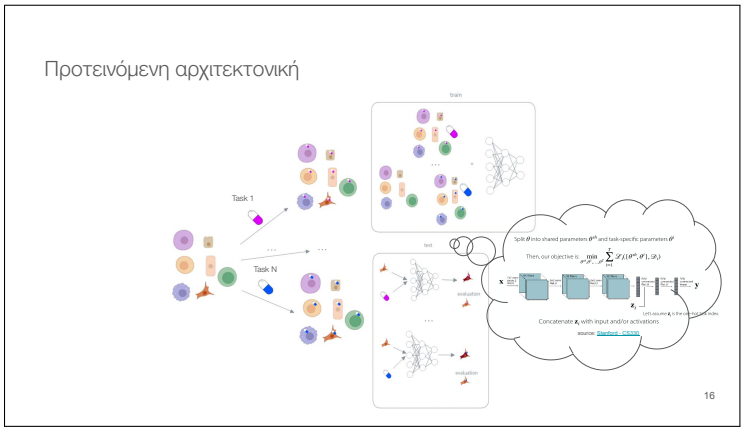


16η διαφάνεια

Σε ένα τέτοιο πλαίσιο, θα μπορούσαμε να αξιοποιήσουμε τη μάθηση πολλαπλών εργασιών, ώστε να εκπαιδεύσουμε ένα ενιαίο μοντέλο που να μπορεί να προβλέψει τη γονιδιακή έκφραση μετά από διάφορες διαταραχές, μοιράζοντας κοινές παραστάσεις μεταξύ τους.

Για να το επιτύχουμε αυτό, αναγνωρίζουμε κάθε διαταραχή ως μια ξεχωριστή εργασία.

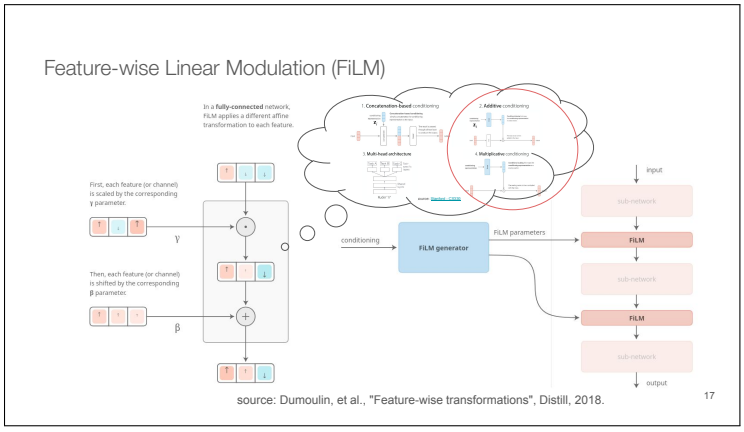
Ανατρέχοντας στις αρχιτεκτονικές που παρουσιάσαμε για τον σχεδιασμό μοντέλων μάθησης πολλαπλών εργασιών, καθοριστικός παράγοντας αποτελεί η σχεδίαση μετασχηματισμού του διανύσματος διαμόρφωσης των εργασιών (z).



17η διαφάνεια

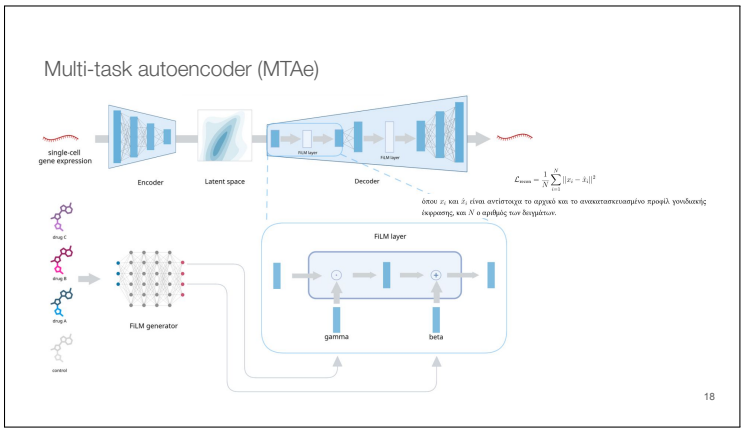
Για αυτό στη συνέχεια, θα μιλήσουμε για τη μέθοδο που επιλέξαμε, ονόματι Feature-wise Linear Modulation, ως έναν συνδυαστικό μετασχηματισμό της αθροιστικής και της πολλαπλασιαστικής περίπτωσης.

Αρχικά, έχουμε ένα δίκτυο, τον FiLM generator, ο οποίος λαμβάνει ως είσοδο το διάνυσμα διαταραχής (z) και παράγει δύο σύνολα παραμέτρων, τα γάμμα (γ) και βήτα (β), τα οποία χρησιμοποιούνται για τον γραμμικό μετασχηματισμό της εισόδου ενός επιπέδου του δικτύου.



18η διαφάνεια

Πλέον μπορούμε να αναλύσουμε την αρχιτεκτονική που σχεδιάσαμε και υλοποιήσαμε. Αποτελείται από έναν autoencoder, ο οποίος λαμβάνει ως είσοδο τη γονιδιακή έκφραση πριν από τη διαταραχή, και χρησιμοποιεί FiLM layers, στον αποκωδικοποιητή (decoder) για να ενσωματώσει την πληροφορία της διαταραχής.



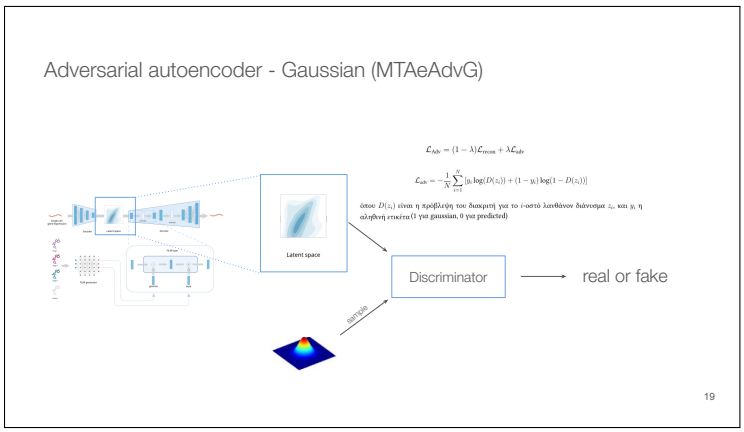
Η φιλοσοφία της αρχιτεκτονικής μας βασίζεται στη σχεδίαση ενός λανθάνοντα χώρου που αντιπροσωπεύει τη γονιδιακή έκφραση ανεξάρτητα από τη διαταραχή, ανιχνεύοντας/φιλτράροντας μέσω της συμπίεσης τα ουσιώδη βιολογικά χαρακτηριστικά.

Στη συνέχεια χρησιμοποιούμε τα FiLM layers για να προσαρμόσουμε την έξοδο του αποκωδικοποιητή ανάλογα με τη διαταραχή που εφαρμόζεται.

Αυτή η αρχιτεκτονική αποτελεί τη βάση, επάνω στην οποία επιχειρήσαμε να σχεδιάσουμε διάφορες παραλλαγές, με στόχο τη βελτίωση της απόδοσής της.

19η διαφάνεια

Μία από αυτές τις παραλλαγές είναι η χρήση ενός adversarial autoencoder.



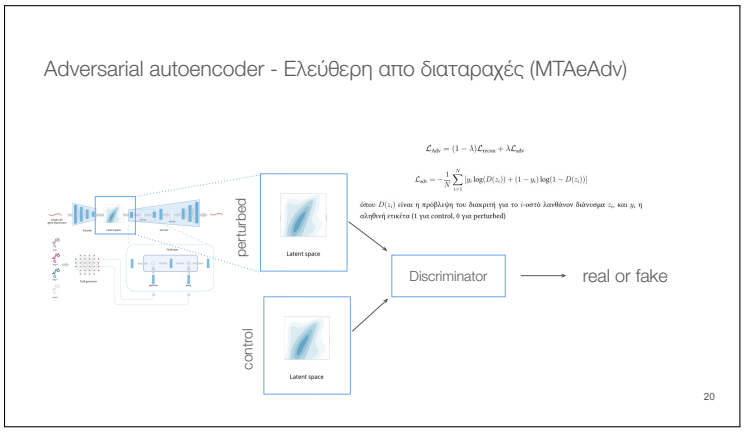
Σε αυτήν, προσθέσαμε έναν διακριτή (discriminator) ο οποίος προσπαθεί να διακρίνει αν το λανθάνον διάνυσμα προέρχεται από τον κωδικοποιητή (encoder) ή από μια προκαθορισμένη κατανομή αναφοράς (prior distribution).

Έτσι, εισαγάγαμε έναν ανταγωνιστικό μηχανισμό που ενθαρρύνει τον κωδικοποιητή να παράγει λανθάνοντα διανύσματα που ακολουθούν αυτήν την κατανομή αναφοράς.

Στη συγκεκριμένη περίπτωση, η προκαθορισμένη κατανομή αναφοράς είναι η κανονική κατανομή (normal distribution).

20η διαφάνεια

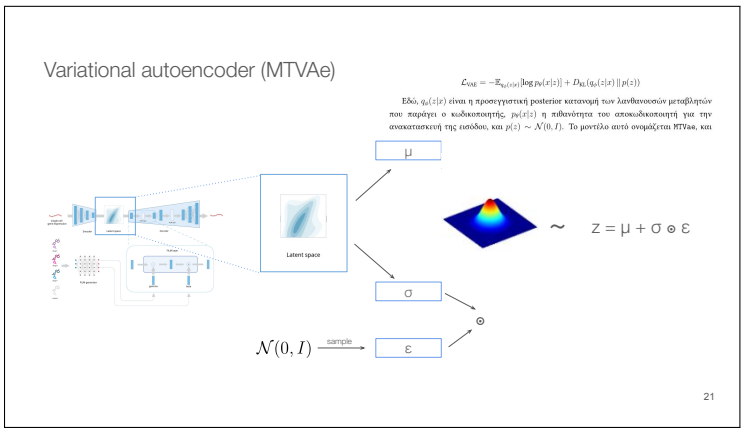
Δοκιμάσαμε ακόμη και την περίπτωση όπου η προκαθορισμένη κατανομή αναφοράς είναι η κατανομή των δειγμάτων ελέγχου, ώστε το λανθάνον διάνυσμα να μην περιέχει πληροφορία σχετικά με τη διαταραχή. Στόχος μας ήταν να διαμορφώσουμε ρητά έναν λανθάνοντα χώρο ελεύθερο διαταραχών, ώστε να κάνουμε αποσύζευξη της πληροφορίας της διαταραχής αποκλειστικά μέσω των FiLM layers.



21η διαφάνεια

Μια ακόμη παραλλαγή που δοκιμάσαμε ήταν η χρήση ενός variational autoencoder (VAE) στη θέση του συμβατικού/απλού autoencoder, διατηρώντας την ίδια αρχιτεκτονική με τα FiLM layers στον αποκωδικοποιητή.

Ο λανθάνοντας χώρος σε αυτήν την περίπτωση είναι στοχαστικός, και κάθε είσοδος χαρτογραφείται σε μια κατανομή παρά σε ένα σημείο, επιτρέποντας έτσι τη δειγματοληψία από αυτήν την κατανομή κατά τη διαδικασία της ανακατασκευής.

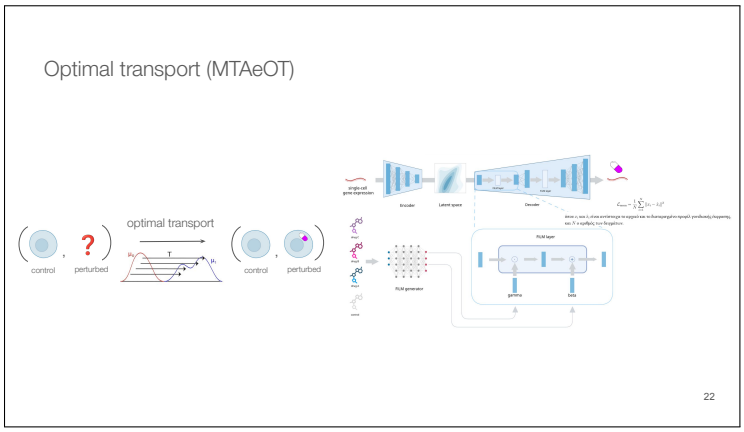


22η διαφάνεια

Επιπλέον, στις μεθόδους μας συμπεριλάβαμε και τη χρήση του optimal transport.

Στο πεδίο της μοντελοποίησης μονοκυτταρικών διαταραχών, τα δεδομένα που χρησιμοποιούνται συνήθως δεν περιλαμβάνουν ζεύγη κυττάρων πριν και μετά από μια διαταραχή, δηλαδή δεν έχουμε την πληροφορία για το πώς ένα συγκεκριμένο κύτταρο αντιδρά σε μια διαταραχή. Τα σύνολα δεδομένων περιέχουν ξεχωριστά δείγματα γονιδιακής έκφρασης πριν και μετά από τη διαταραχή, αλλά δεν υπάρχει άμεση αντιστοίχιση μεταξύ των κυττάρων σε αυτές τις δύο καταστάσεις.

Στόχος μας ήταν να αντιμετωπίσουμε αυτό το πρόβλημα ζευγαριών, μέσω της εκτίμησης της βέλτιστης μεταφοράς μεταξύ των κατανομών γονιδιακής έκφρασης πριν και μετά από τη διαταραχή.

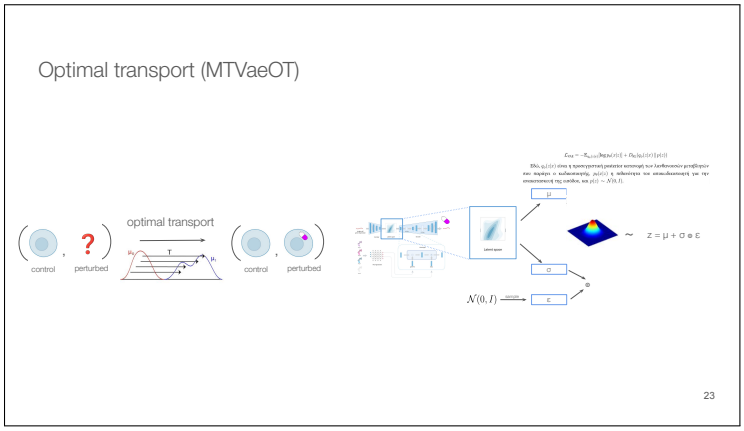


Έχοντας πλέον αυτήν την αντιστοιχία, μπορέσαμε να κατασκευάσουμε ένα μοντέλο που αντί να ανακατασκευάζει τη γονιδιακή έκφραση, να προβλέπει τη γονιδιακή έκφραση του κυττάρου-ζεύγους μετά

από τη διαταραχή, βασιζόμενο στη γονιδιακή έκφραση πριν από τη διαταραχή.

23η διαφάνεια

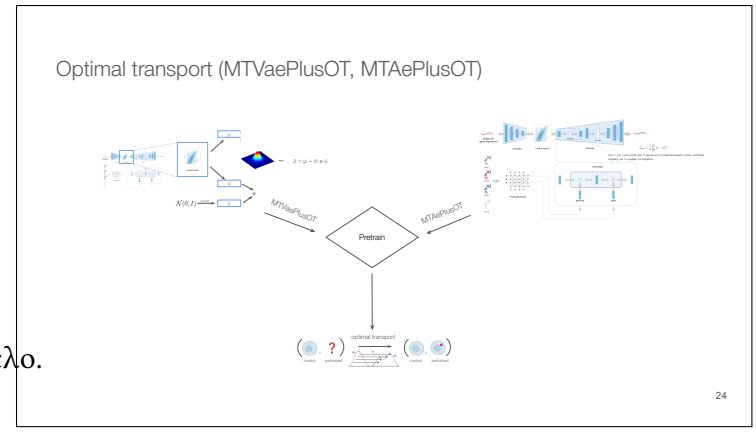
Παρομοίως, δοκιμάσαμε και την παραλλαγή όπου ο autoencoder αντικαταστάθηκε από έναν variational autoencoder, διατηρώντας την ίδια φιλοσοφία με τη χρήση του optimal transport για τη δημιουργία ζευγών κυττάρων πριν και μετά από μια διαταραχή.



24η διαφάνεια

Τέλος, επιχειρήσαμε να συνδυάσουμε τις προηγούμενες παραλλαγές, κάνοντας προ-εκπαίδευση (pretrain), είτε την αρχική βασική αρχιτεκτονική, είτε αυτήν με τον variational autoencoder.

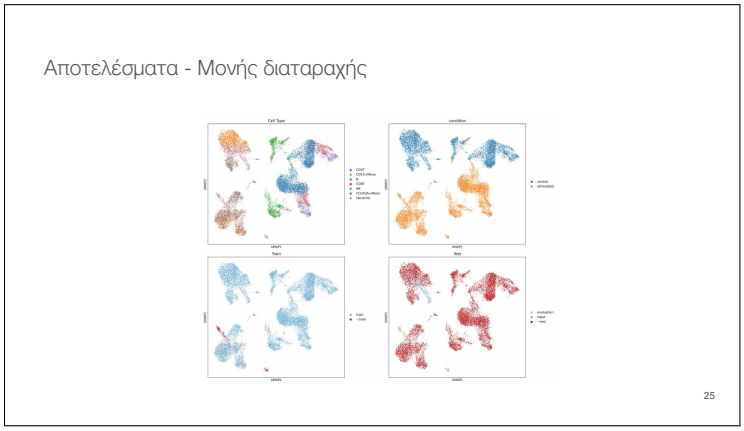
Στη συνέχεια, χρησιμοποιώντας τον optimal transport για τη δημιουργία ζευγών κυττάρων πριν και μετά από μια διαταραχή, κάναμε fine-tuning το μοντέλο.



25η διαφάνεια

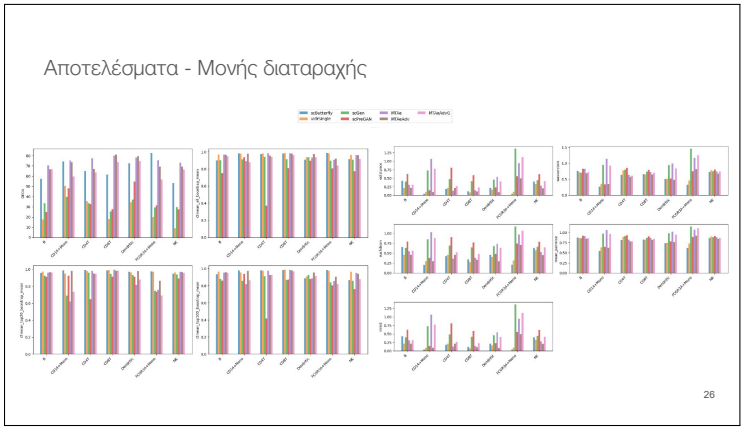
Έχοντας παρουσίασει όλες τις παραλλαγές της μεθόδου μας, θα δούμε τώρα τα αποτελέσματα που πετύχαμε. Για την αξιολόγηση των αρχιτεκτονικών μας, χρησιμοποιήσαμε τρία διαφορετικά σύνολα δεδομένων, και κάναμε σύγκριση αυτών με μεθόδους αιχμής.

Αρχικά, στο πρώτο σύνολο εξετάστηκε η απόδοση σε περίπτωση μονής διαταραχής, της ιντερφερόνης βήτα (IFN-β), σε ανθρώπινα μονοπύρηνα κύτταρα περιφερικού αίματος (PBMCs).



Όπως είπαμε δει προηγουμένως, γνωρίζουμε τους κυτταρικούς τύπους των γονιδιακών εκφράσεων, και κρατούμε έναν τύπο κυττάρου εκτός για αξιολόγηση. Για παράδειγμα σε αυτήν την UMAP αναπαράσταση, κρατήσαμε εκτός τα δενδριτικά κύτταρα (dendritic cells).

26η διαφάνεια



Επαναλάβαμε την ίδια διαδικασία για όλους τους τύπους κυττάρων, δηλαδή για κάθε τύπο κυττάρου κρατήσαμε εκτός τα δείγματα που τον περιείχαν, και εκπαιδεύσαμε τα μοντέλα με τα υπόλοιπα δείγματα.

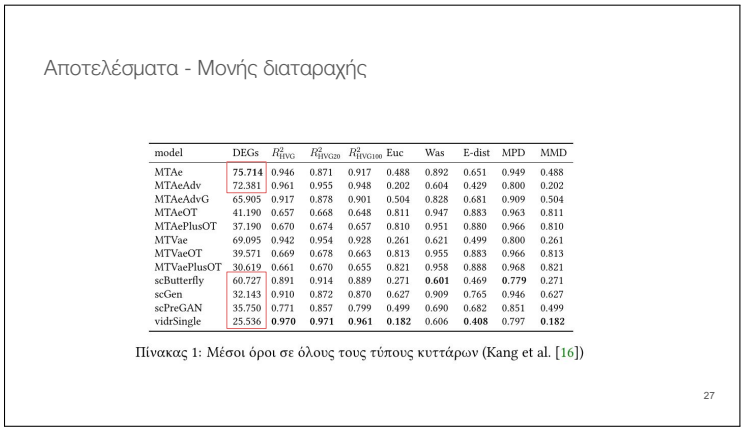
Χρησιμοποιήσαμε δύο κατηγορίες μετρικών αξιολόγησης για τη σύγκριση των προβλεπόμενων προφίλ γονιδιακής έκφρασης με τα πραγματικά διαταραγμένα: α) βασικές μετρικές και β) μετρικές απόστασης.

Οι βασικές μετρικές περιλαμβάνουν τον αριθμό κοινών διαφορετικών εκφραζόμενων γονιδίων (DEGs), τον τετραγωνισμένο συντελεστή Pearson R2 για τα μέσα επίπεδα έκφρασης υπολογισμένα πάνω σε όλα τα γονίδια υψηλής μεταβλητότητας (HVGs), καθώς και το R 2 για τα κορυφαία 100 HVGs. Ένα διαφοροποιημένο γονίδιο (DEG) ορίζεται ως γονίδιο του οποίου η κατανομή έκφρασης διαφέρει σημαντικά μεταξύ συνθήκης ελέγχου και διαταραχής.

Οι μετρικές απόστασης καταγράφουν τόσο σημειακές όσο και διαφορές στις κατανομές μεταξύ προβλεπόμενων και πραγματικών προφίλ και περιλαμβάνουν: (α) Ευκλείδεια απόσταση, (β) E-distance, (γ) Wasserstein απόσταση, (δ) mean pairwise distance (MPD) και (ε) maximum mean discrepancy (MMD).

'Οσο υψηλότερες είναι οι τιμές για τις βασικές μετρικές, τόσο καλύτερη είναι η απόδοση, ενώ για τις μετρικές απόστασης, όσο χαμηλότερες είναι οι τιμές, τόσο το καλύτερο.

27η διαφάνεια



model	DEGs	R^2_{HVG}	R^2_{HVG20}	R^2_{HVG100}	Euc	Was	E-dist	MPD	MMD
MTAe	**75.714**	0.946	0.871	0.917	0.488	0.892	0.651	0.949	0.488
MTAeAdv	72.381	0.961	0.955	0.948	0.202	0.604	0.429	0.800	0.202
MTAeAdvG	65.905	0.917	0.878	0.901	0.504	0.828	0.681	0.909	0.504
MTAeOT	41.190	0.657	0.668	0.648	0.811	0.947	0.883	0.963	0.811
MTAePlusOT	37.190	0.670	0.674	0.657	0.810	0.951	0.880	0.966	0.810
MTVae	69.095	0.942	0.954	0.928	0.261	0.621	0.499	0.800	0.261
MTVaeOT	39.571	0.669	0.678	0.663	0.813	0.955	0.883	0.966	0.813
MTVaePlusOT	30.619	0.661	0.670	0.655	0.821	0.958	0.888	0.968	0.821
scButterfly	60.727	0.891	0.914	0.889	0.271	**0.601**	0.469	**0.779**	0.271
scGen	32.143	0.910	0.872	0.870	0.627	0.909	0.765	0.946	0.627
scPreGAN	35.750	0.771	0.857	0.799	0.499	0.690	0.682	0.851	0.499
vidrSingle	25.536	**0.970**	**0.971**	**0.961**	**0.182**	0.606	**0.408**	0.797	**0.182**

Πίνακας 1: Μέσοι όροι σε όλους τους τύπους κυττάρων (Kang et al. [16])

Επειδή δεν είναι τόσο εύκολο να συγκρίνουμε όλες αυτές τις μετρικές για τον εκάστοτε τύπο κυττάρου, υπολογίσαμε τον μέσο όρο των τιμών για κάθε μετρική, και έτσι έχουμε το ακόλουθο πινακάκι.

Αρχικά βλέπουμε τα μοντέλα μηχανικής μάθησης υπερέχουν σημαντικά στη μετρική DEG, η οποία καταγράφει τα διαφορετικά εκφραζόμενα γονίδια.

28η διαφάνεια

Όσον αφορά τις υπόλοιπες μετρικές, παρατηρούμε ότι το μοντέλο μας με την παραλλαγή του adversarial autoencoder, για παράδειγμα, έχει πολύ μικρές διαφορές στις τιμές με τα καλύτερα μοντέλα αιχμής.

Αποτελέσματα - Μονής διαταραχής

model	DEGs	R^2_{HVG}	R^2_{HVG20}	R^2_{HVG100}	Euc	Was	E-dist	MPD	MMD
MTAe	**75.714**	0.946	0.871	0.917	0.488	0.892	0.651	0.949	0.488
MTAeAdv	72.381	0.961	0.955	0.948	0.202	0.604	0.429	0.800	0.202
MTAeAdvG	65.905	0.917	0.878	0.901	0.504	0.828	0.681	0.909	0.504
MTAeOT	41.190	0.657	0.668	0.648	0.811	0.947	0.883	0.963	0.811
MTAePlusOT	37.190	0.670	0.674	0.657	0.810	0.951	0.880	0.966	0.810
MTVae	69.095	0.942	0.954	0.928	0.261	0.621	0.499	0.800	0.261
MTVaeOT	39.571	0.669	0.678	0.663	0.813	0.955	0.883	0.966	0.813
MTVaePlusOT	30.619	0.661	0.670	0.655	0.821	0.958	0.888	0.968	0.821
scButterfly	60.727	0.891	0.914	0.889	0.271	**0.601**	0.469	**0.779**	0.271
scGen	32.143	0.910	0.872	0.870	0.627	0.909	0.765	0.946	0.627
scPreGAN	35.750	0.771	0.857	0.799	0.499	0.690	0.682	0.851	0.499
vidrSingle	25.536	**0.970**	**0.971**	**0.961**	**0.182**	0.606	**0.408**	0.797	**0.182**

Πίνακας 1: Μέσοι όροι σε όλους τους τύπους κυττάρων (Kang et al. [16])

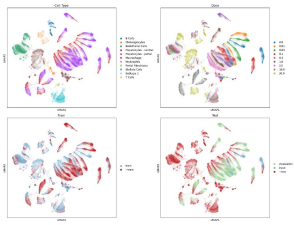
28

29η διαφάνεια

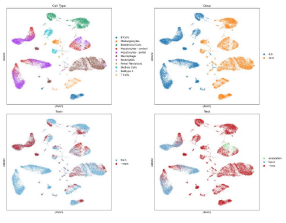
Στη συνέχεια, θα περάσουμε στο δεύτερο και ένα πιο ουσιαστικό σύνολο δεδομένων, όπου τα μοντέλα μάθησης πολλαπλών εργασιών θα μπορέσουν να αξιοποιήσουν ένα ευρύτερο σύνολο δεδομένων με πολλαπλές διαταραχές.

Συγκεκριμένα, αξιολογήθηκε στην περίπτωση πολλαπλών διαταραχών σε ποντίκια, μέσω της εφαρμογής της χημικής ουσίας TCDD (2,3,7,8 τετραχλωρο-διβενζο-παρα-διοξίνη) σε εννέα διαφορετικές δοσολογίες.

Αποτελέσματα - Πολλαπλών διαταραχών



Σχήμα 17: Απεικόνιση UMAP του διαχωρισμού των δεδομένων για όλες τις δοσολογίες (Nault et al. [26,27])



Σχήμα 18: Απεικόνιση UMAP του διαχωρισμού των δεδομένων για τη δοσολογία $30\mu g/kg$ (Nault et al. [26,27])

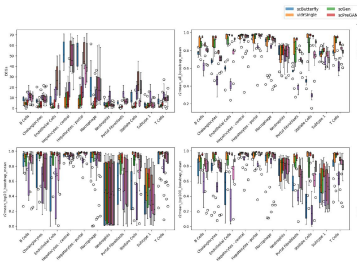
29

Σε αυτήν την περίπτωση, για τις μεθόδους αιχμής που είναι σχεδιασμένες να διαχειρίζονται μία μονή διαταραχή, χρησιμοποιήσουμε για κάθε δοσολογία ένα ξεχωριστό μοντέλο.

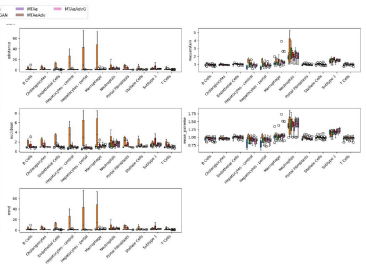
30η διαφάνεια

Σχετικά με τις μετρικές ανά τύπο κυττάρου, πλέον χρησιμοποιούμε θηκογράμματα (boxplots) για να παρουσιάσουμε την κατανομή των τιμών για κάθε μετρική για τις εννέα διαφορετικές δοσολογίες.

Αποτελέσματα - Πολλαπλών διαταραχών



Σχήμα 13: Βασικές μετρικές ανά τύπο κυττάρου (Nault et al. [26,27])



Σχήμα 14: Μετρικές απόστασης ανά τύπο κυττάρου (Nault et al. [26,27])

30

31η διαφάνεια

’Οπως κάναμε και προηγουμένως, υπολογίσαμε τον μέσο όρο των τιμών για κάθε μετρική, και έτσι έχουμε τον ακόλουθο πινακάκι.

Για μία ακόμη φορά, τα μοντέλα μηχανικής μάθησης υπερέχουν σημαντικά στη μετρική DEG, η οποία καταγράφει τα διαφορετικά εκφραζόμενα γονίδια και παρουσιάζουν ανταγωνιστική απόδοση στις υπόλοιπες μετρικές.

Αποτελέσματα - Πολλαπλών διαταραχών

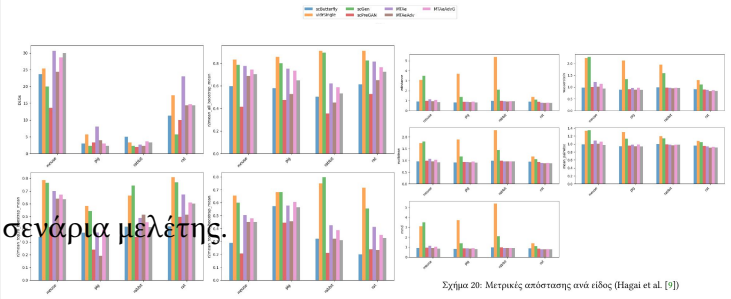
model	DEGs	R^2_{HVG}	R^2_{HVG20}	R^2_{HVG100}	Euc	Was	E-dist	MPD	MMD
MTAe	**20.341**	0.862	0.792	0.833	1.386	1.217	1.116	1.050	1.386
MTAeAdv	13.716	0.792	0.725	0.743	1.128	1.091	1.011	1.017	1.128
MTAeAdvG	18.307	0.808	0.736	0.764	1.164	1.107	1.030	1.029	1.164
MTAeOT	8.652	0.608	0.642	0.590	0.925	1.006	0.951	0.998	0.925
MTAePlusOT	8.519	0.613	0.644	0.596	**0.917**	**1.004**	0.948	0.996	0.917
MTVae	18.981	0.808	0.724	0.753	1.124	1.100	1.005	1.016	1.124
MTVaeOT	8.163	0.614	0.642	0.593	0.929	1.009	0.952	0.998	0.929
MTVaePlusOT	8.701	0.615	0.645	0.597	0.919	1.006	0.948	0.997	**0.919**
scButterfly	16.818	0.740	0.694	0.696	0.984	1.014	**0.944**	**0.990**	0.984
scGen	6.288	**0.915**	**0.863**	**0.897**	2.408	1.229	1.387	1.041	2.408
scPreGAN	14.511	0.599	0.596	0.562	0.972	1.019	0.969	1.000	0.972
vidrMult	2.352	0.870	0.837	0.852	9.295	1.358	2.425	1.054	9.295
vidrSingle	3.795	0.855	0.797	0.824	1.431	1.174	1.118	1.025	1.431

Πίνακας 2: Μέσοι όροι σε όλους τους τύπους κυττάρων Nault et al. [26, 27]

32η διαφάνεια

Τέλος, αναφορικά με το τρίτο σύνολο δεδομένων, την τρίτη δηλαδή περίπτωση αξιολόγησης έγινε εξέταση του σεναρίου εφαρμογής μονής διαταραχής, του λιποπολυσακχαρίτη (LPS), μεταξύ διαφορετικών ειδών (ποντίκι, λαγός, αρουραίος, χοίρος), εισάγοντας έναν επιπλέον άξονα μεταβλητότητας ως προς το είδος, ο οποίος επεκτείνει τα προηγούμενα σενάρια μελέτης.

Αποτελέσματα - Μεταξύ διαφορετικών ειδών



Σχήμα 19: Βασικές μετρικές ανά είδος (Hagai et al. [9])

Σχήμα 20: Μετρικές απόστασης ανά είδος (Hagai et al. [9])

33η διαφάνεια

Παρατηρούμε ότι τα μοντέλα μηχανικής μάθησης υπερέχουν και πάλι σημαντικά στην μετρική DEG, και μάλιστα το MTVae, παρουσιάζει καλύτερη απόδοση στις μετρικές απόστασης.

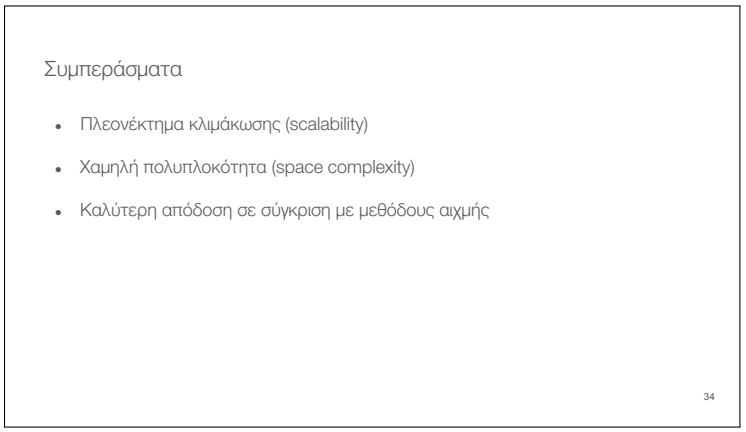
Αποτελέσματα - Μεταξύ διαφορετικών ειδών

model	DEGs	R^2_{HVG}	R^2_{HVG20}	R^2_{HVG100}	Euc	Was	E-dist	MPD	MMD
MTAe	**16.083**	0.740	0.559	0.481	0.930	1.008	0.962	0.995	0.930
MTAeAdv	11.250	0.579	0.465	0.365	0.865	0.919	0.929	0.957	0.865
MTAeAdvG	12.500	0.708	0.533	0.456	0.921	0.985	0.958	0.987	0.921
MTAeOT	7.500	0.483	0.432	0.304	0.899	0.932	0.948	0.966	0.899
MTAePlusOT	8.000	0.480	0.436	0.309	0.876	0.913	0.936	0.956	0.876
MTVae	12.500	0.652	0.498	0.413	**0.840**	**0.903**	**0.916**	**0.951**	0.840
MTVaeOT	7.417	0.479	0.423	0.302	0.895	0.929	0.946	0.964	0.895
MTVaePlusOT	7.833	0.473	0.431	0.301	0.883	0.919	0.940	0.959	0.883
scButterfly	10.750	0.574	0.389	0.346	0.899	0.942	0.948	0.971	0.899
scGen	7.583	0.826	0.705	0.658	2.014	1.576	1.367	1.165	2.014
scPreGAN	7.250	0.443	0.374	0.276	0.914	0.945	0.955	0.973	0.914
vidrSingle	12.917	**0.878**	**0.711**	**0.701**	3.386	1.905	1.769	1.225	3.386

Πίνακας 3: Μέσοι όροι σε όλους τους τύπους κυττάρων Hagel et al. [9]

34η διαφάνεια

Βάσει των αποτελεσμάτων που παρουσίασαμε, μπορούμε να συμπεράνουμε ότι η αρχιτεκτονική μάθησης πολλαπλών εργασιών που σχεδιάσαμε και υλοποιήσαμε είναι μια πολλά υποσχόμενη μέθοδος για τη μοντελοποίηση διαταραχών σε μονοκυτταρικά δεδομένα γονιδιακής έκφρασης.

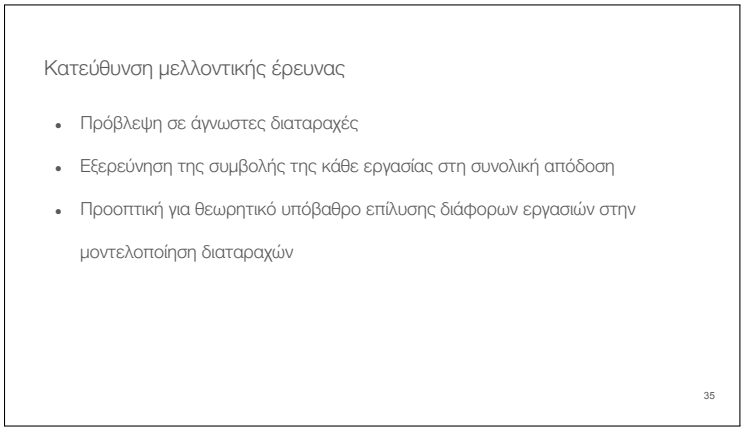


Η απόδοση της μεθόδου είναι ιδιαίτερα ανταγωνιστική και σε αρκετές μετρικές υπερέχει έναντι των μεθόδων αιχμής, ενώ παράλληλα προσφέρει την ευελιξία να διαχειρίζεται πολλαπλές διαταραχές με ένα ενιαίο μοντέλο.

Έτσι, η αρχιτεκτονική μας διαθέτει χαμηλή πολυπλοκότητα και παρέχει προοπτική για την εφαρμογή της σε πληθώρα διαταραχών, έχοντας καλύτερη κλιμάκωση.

35η διαφάνεια

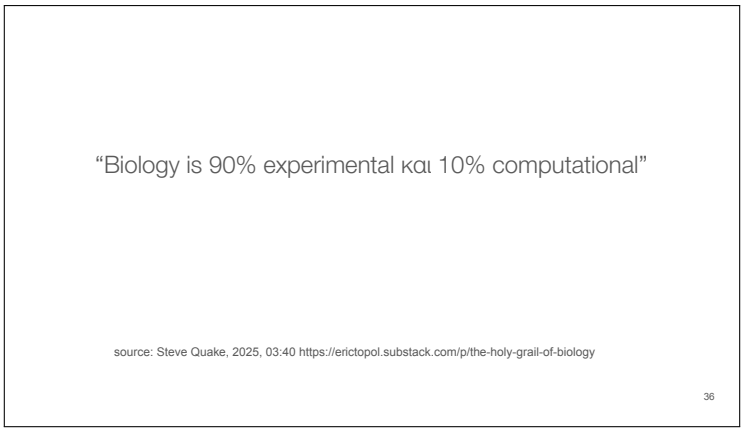
Ως προς τους περιορισμούς της μεθόδου μας, θα θέλαμε να επισημάνουμε την αδυναμία πρόβλεψης σε άγνωστες διαταραχές. Όπως είδαμε, το μοντέλο κάνει προβλέψεις για τύπους διαταραχών που έχουν ήδη παρουσιαστεί κατά την εκπαίδευση, και δεν μπορεί να γενικεύσει σε νέες διαταραχές, αλλά σε νέους τύπους κυττάρων ή ειδών.



Σχετικά με τις μελλοντικές προεκτάσεις της έρευνας μας, θα είχε ιδιαίτερο ενδιαφέρον να εξετάσουμε τη συμβολή της κάθε εργασίας στη συνολική απόδοση του μοντέλου, και να αποτίμησουμε το πλεονέκτημα της κλιμάκωσης σε ακόμη μεγαλύτερο αριθμό διαταραχών.

Τέλος, εκτιμούμε ότι η μέθοδός μας θα μπορούσε να αποτελέσει αφετηρία για τον σχεδιασμό ενός θεωρητικού πλαισίου για την επίλυση διάφορων εργασιών στη μοντελοποίηση διαταραχών και στην βιοπληροφορική ευρύτερα.

36η διαφάνεια



Θα ήθελα να κλείσω με την εξής αναφορά, η οποία αναδεικνύει το μέλλον και την προοπτική της χρήσης υπολογιστικών μεθόδων στη βιολογία. ”Η βιολογία είναι 90% πειραματική, και 10% υπολογιστική.”

Συνεπώς, η χρήση της τεχνητής νοημοσύνης μπορεί να αποτελέσει ένα κομβικής σημασίας πεδίο για την αυτοματοποίηση και την επιτάχυνση της βιολογικής έρευνας.

Σας ευχαριστώ πολύ για την προσοχή σας!