



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 (EKETA)

Περιεχόμενα

1η διαφάνεια	2
2η διαφάνεια	2
3η διαφάνεια	2
4η διαφάνεια	3
5η διαφάνεια	3
6η διαφάνεια	4
7η διαφάνεια	4
8η διαφάνεια	5
9η διαφάνεια	5
10η διαφάνεια	5
11η διαφάνεια	6
12η διαφάνεια	6
13η διαφάνεια	6
14η διαφάνεια	7
15η διαφάνεια	7
16η διαφάνεια	8
17η διαφάνεια	8

1η διαφάνεια

Καλημέρα/Καλησπέρα σας,

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

Η διπλωματική αυτή ήταν συνεργασία μεταξύ του ΑΠΘ, με επιβλέπων 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.

1

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

Μάθηση πολλαπλών εργασιών στη μοντελοποίηση διαταραχών με εφαρμογή σε μονοκυτταρικά δεδομένα

Θεόδωρος Κατζάλης

Επιβλέπων: Περικλής Μήτκας, Καθηγητής ΑΠΘ
Συνεπιβλέπων: Φώτης Ψωμόπουλος, Ερευνητής INEB-ΕΚΕΤΑ

Θεσσαλονίκη, Νοέμβριος 2025

2η διαφάνεια

Όσον αφορά τη δομή της παρουσίασης, αρχικά θα εισάγουμε τους ορισμούς της μοντελοποίησης μονοκυτταρικών διαταραχών και της μάθησης πολλαπλών εργασιών.

Στη συνέχεια, θα αναλύσουμε την αρχιτεκτονική που υλοποιήσαμε και τη βασική μέθοδο της βιβλιογραφίας στην οποία αυτή η αρχιτεκτονική στηρίχθηκε, ονόματι Feature-wise Linear Modulation, για 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.

2

Περιεχόμενα

- Μοντελοποίηση μονοκυτταρικών διαταραχών (single-cell perturbation modeling)
- Μάθηση πολλαπλών εργασιών (Multi-task learning, MTL)
- Feature-wise Linear Modulation (FiLM)
- Περιγραφή της προτεινόμενης αρχιτεκτονικής
- Αξιολόγηση αποτελεσμάτων
- Συμπεράσματα και μελλοντικές επεκτάσεις

3η διαφάνεια

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

Σχετικά πρόσφατα, 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.

3

Κεντρικό δόγμα της βιολογίας

source: <https://aimm.epfl.ch/research/>

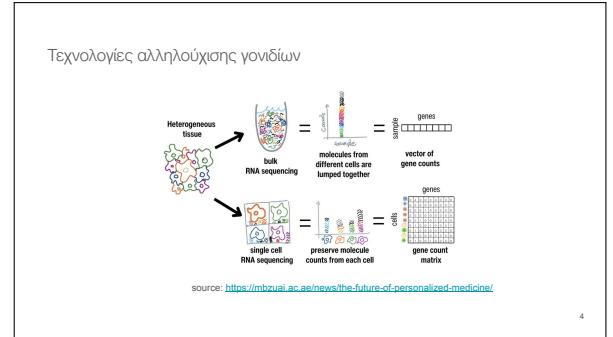
Ωστόσο αυτή η ροή είναι ιδιαίτερη περίπλοκη, εξαιτίας της ποικιλομορφίας των κυττάρων, των μοριακών μηχανισμών, και των περιβαλλοντικών συνθηκών που την επηρεάζουν.

4η διαφάνεια

Στην προσπάθεια να κατανοήσουμε αυτή την πολυπλοκότητα, αναπτύχθηκαν τεχνικές αλληλούχησης, οι οποίες έχουν την δυνατότητα να καταγράψουν 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.

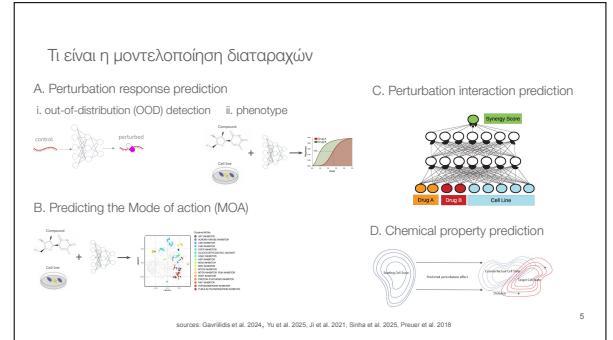


5η διαφάνεια

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

Κεντρικός της στόχος είναι η ανάπτυξη υπολογιστικών μοντέλων που μπορούν να προβλέψουν πως 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 τέσσερις βασικούς στόχους:



Ο πρώτος έχει δύο υποκατηγορίες. Η πρώτη σχετίζεται με την πρόβλεψη της γονιδιακής έκφρασης μετά από μια διαταραχή. 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 (signalling pathways) και των πρωτεΐνων-στόχων (target proteins) που ενεργοποιούνται ή αναστέλλονται ως απόκριση σε μια δεδομένη διαταραχή.

Ο τρίτος σχετίζεται με την αλληλεπίδραση μεταξύ διαταραχών, και 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 τον 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 μέθοδος μας επικεντρώνεται και αυτή σε αυτήν την κατηγορία.

6η διαφάνεια

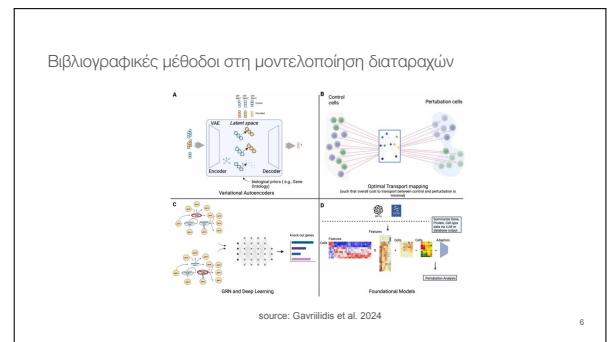
Στη βιβλιογραφία, έχουν προταθεί διάφορες μέθοδοι για τη μοντελοποίηση διαταραχών σε μονοκυτταρικά δεδομένα και έχουν κατηγοριοποιηθεί στις εξής τέσσερις βασικές προσεγγίσεις:

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 την χρήση γραφικών νευρωνικών δικτύων (graph neural networks), λαμβάνοντας υπόψη τις αλληλεπιδράσεις μεταξύ γονιδίων.

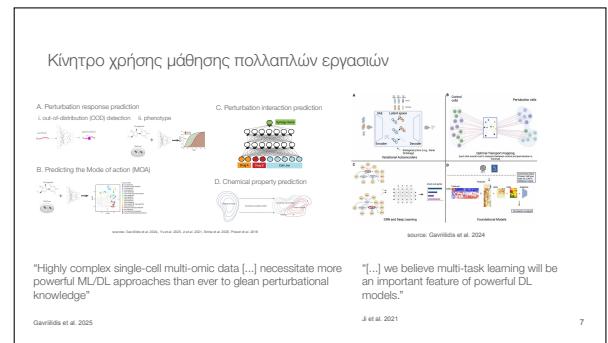
Τέλος, η τέταρτη κατηγορία περιλαμβάνει foundational models που έχουν εκπαιδευτεί σε πολύ μεγάλα σύνολα δεδομένων.



7η διαφάνεια

Έτσι, έχοντας δει μια γενική εικόνα των προκλήσεων της μοντελοποίησης διαταραχών, αλλά και των μεθόδων που έχουν εφαρμοστεί, καταλαβαίνουμε ότι υπάρχει μια σημαντική πολυπλοκότητα και ποικιλία προσεγγίσεων.

Ακόμη, μία από τις σύνθετες/προηγμένες μεθόδους, που θα μπορούσε να θεωρηθεί ως 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

Μάθηση πολλαπλών εργασιών (Multi-task learning, MTL)

source: Andrej Karpathy - Tesla Autopilot and Multi-Task Learning for Perception and Prediction

8

Για να εισάγουμε τη μάθηση πολλαπλών εργασιών, ας εξετάσουμε την αναγνώριση αντικειμένων σε μια εικόνα. Για παράδειγμα, σε αυτήν την εικόνα επιθυμούμε να αναγνωρίσουμε στατικά αντικείμενα στον δρόμο όπως οι κώνοι, δυναμικά αντικείμενα όπως τα αυτοκίνητα, αλλά μέχρι και τα φανάρια.

9η διαφάνεια

Μάθηση πολλαπλών εργασιών (Multi-task learning, MTL)

Single-task learning: $\mathcal{D} = \{(x, y_i)\}$
[supervised] $\min_{\theta} \mathcal{L}(\theta, \mathcal{D})$

source: https://cs.stanford.edu/people/karpathy/vis-lab/multi-task-learning-for-perception-and-prediction.html

9

Σε μια τέτοια περίπτωση, θα μπορούσαμε να εκπαιδεύσουμε τρία ξεχωριστά μοντέλα, ένα για κάθε εργασία.

10η διαφάνεια

Μάθηση πολλαπλών εργασιών (Multi-task learning, MTL)

Single-task learning: $\mathcal{D} = \{(x, y_i)\}$
[supervised] $\min_{\theta} \mathcal{L}(\theta, \mathcal{D})$

source: https://cs.stanford.edu/people/karpathy/vis-lab/multi-task-learning-for-perception-and-prediction.html

Προβλήματα

- Αναβούτητες κοινές αναπαραστάσεις
- Ποσότητα δεδομένων (data inefficient)
- Υπερπροσαρμογή (overfitting)
- Αυξημένη πολυπλοκότητα
- Επαναφεύρεση τροχού

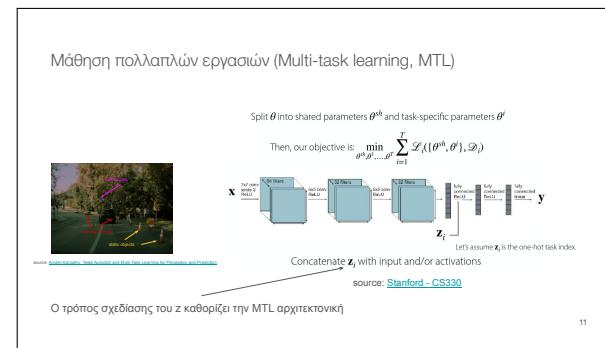
10

Ωστόσο, αυτή η προσέγγιση 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

Έτσι, 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 να μπορεί να προσαρμόζεται δυναμικά ανάλογα με την εργασία που εκτελείται.



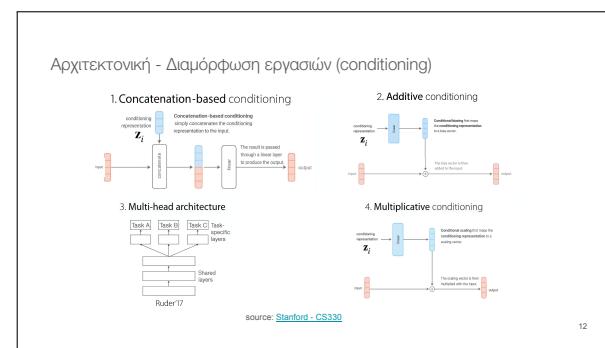
12η διαφάνεια

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

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

Στη συνέχεια, υπάρχει ο αθροιστικός μετασχηματισμός (additive transformation), όπου το z προστίθεται με το διάνυσμα εισόδου, και παρόμοια, έχουμε τον πολλαπλασιαστικό μετασχηματισμό (multiplicative transformation), όπου το z πολλαπλασιάζεται με το διάνυσμα εισόδου.

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



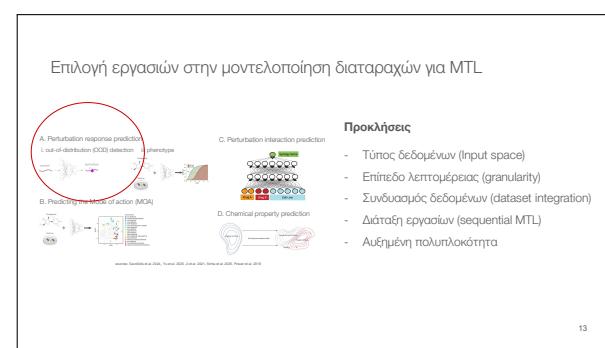
13η διαφάνεια

Έχοντας παρουσιάσει τη μάθηση πολλαπλών εργασιών, ας δούμε τώρα 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 dataset που χρησιμοποιούνται για τους στόχους της πρόβλεψης αλληλεπίδρασης διαταραχών και του μηχανισμού δράσης (δεύτερος και τρίτος στόχος δηλαδή), προέρχονται από 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 (granularity).

Ακόμη, η επίλυση των προβλημάτων στη μοντελοποίηση διαταραχών μπορεί να μην έχει αμιγώς παράλληλο χαρακτήρα (όπως έχει η αναγνώριση στάσιμων και δυναμικών αντικειμένων σε μια εικόνα όπως είδαμε προηγουμένως), με αποτέλεσμα η ταυτόχρονη επίλυση πολλαπλών εργασιών να απαιτεί μια πιο εξειδικευμένη προσέγγιση διάταξης αυτών, αυξάνοντας την πολυπλοκότητα ενός υποψήφιου μοντέλου μάθησης πολλαπλών εργασιών.

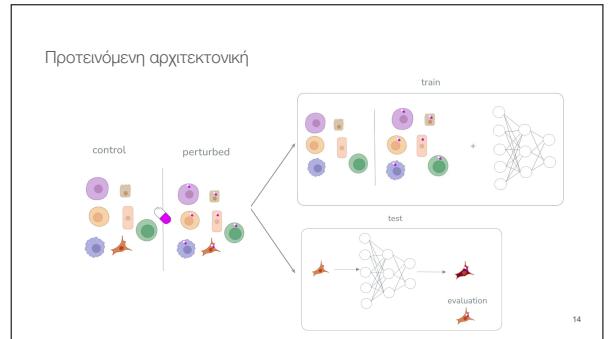
Για αυτό το λόγο, εστιάσαμε κυρίως στο πρώτο πρόβλημα, και συγκεκριμένα στην πρόβλεψη εκτός κατανομής (out-of-distribution, OOD) της γονιδιακής έκφρασης μετά από μια διαταραχή.

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.

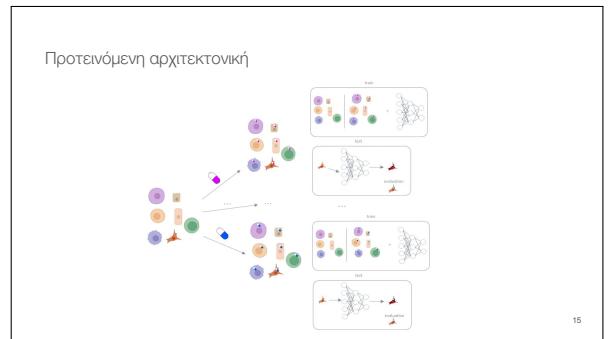
Στόχος είναι να προβλέψουμε την γονιδιακή έκφραση μετά από τη διαταραχή, βασιζόμενοι στην αρχική έκφραση και τη διαταραχή.

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

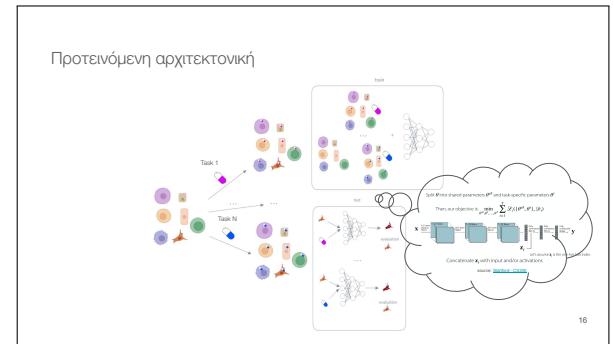
Βέβαια, υπάρχουν σύνολα δεδομένων, τα οποία περιλαμβάνουν περισσότερες από μία διαταραχής, και για τις μέθοδοι αιχμής που είναι σχεδιασμένες για έναν τύπο διαταραχής, θα πρέπει να δημιουργηθούν ξεχωριστά μοντέλα για κάθε διαταραχή.



16η διαφάνεια

Έτσι, 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, αναγνωρίζοντας κάθε διαταραχή ως μια ξεχωριστή εργασία.

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



17η διαφάνεια

Για αυτό λοιπόν στην συνέχεια, θα μιλήσουμε για την μέθοδο που επιλέξαμε, ονόματι Feature-wise Linear Modulation, ως ένας συνδυαστικός μετασχηματισμός της αθροιστικής και της πολλαπλασιαστικής περίπτωσης.

