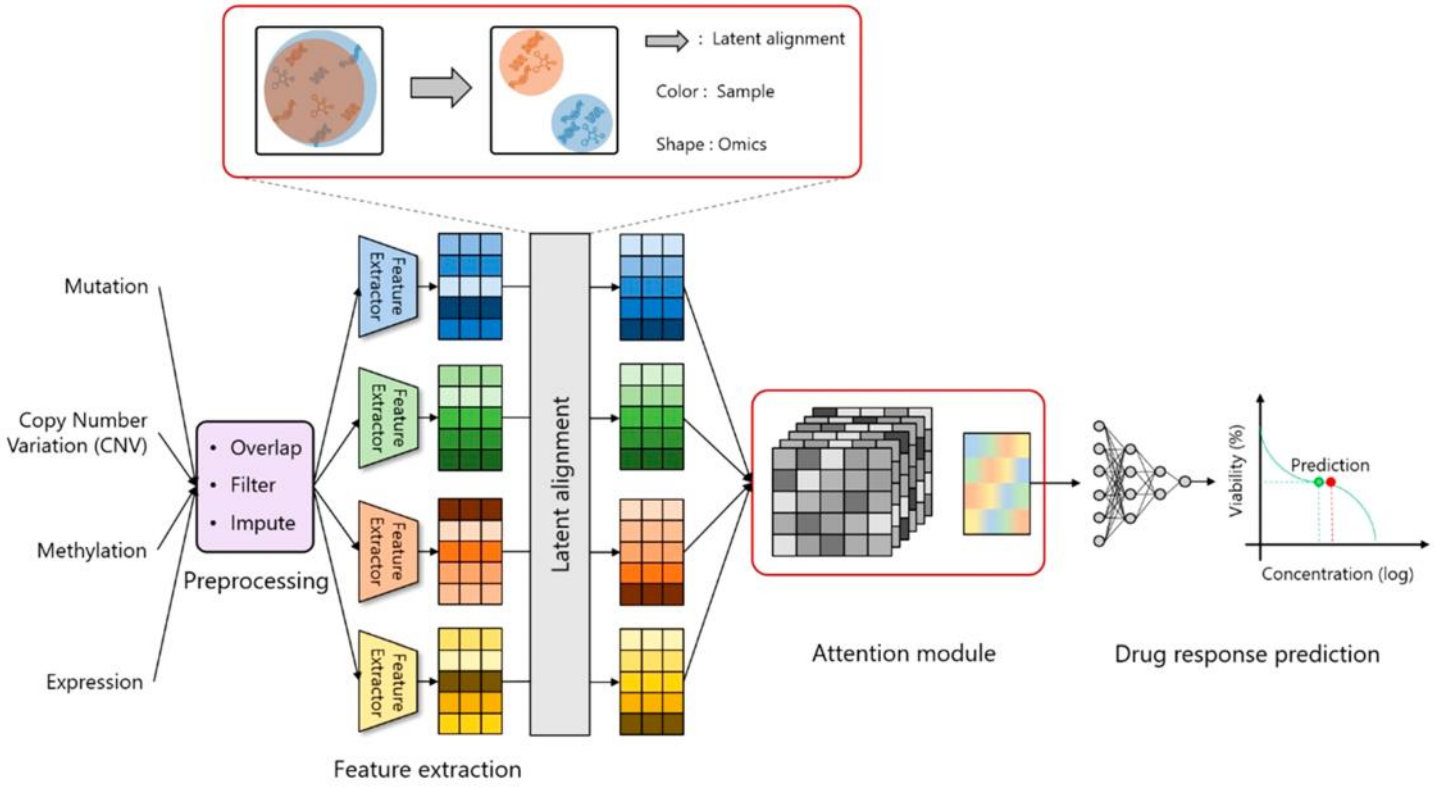
MOLI (2019)

Thursday, December 11, 2025 1:39 AM

Το MOLI (**M**ulti-**O**mics **L**ate **I**ntegration) είναι ένα από τα πρώτα *deep learning* μοντέλα που εφαρμόζουν **late integration** για drug response prediction, χρησιμοποιώντας **mutations**, **copy-number variations** και **gene expression**.

Multi-omics = συνδυασμός **πολλών διαφορετικών βιολογικών στρωμάτων πληροφορίας** για να κατανοήσουμε την κατάσταση ενός κυττάρου ή ενός οργανισμού.



1. Είσοδοι (Omics input types)

Το μοντέλο παίρνει τρία omics modalities:

- **Somatic mutations** → binary vectors για gene-level mutation status
- **Copy-number aberrations (CNA)** → continuous genomic features
  Πόσα αντίγραφα ενός συγκεκριμένου γονιδίου υπάρχουν στο DNA του κυττάρου.
- **Gene expression** → transcriptomic profiles

2. Omics-specific subnetworks (Late integration)

Κάθε modality περνάει από **ξεχωριστό MLP subnetwork**:

- MLP για mutations
- MLP για CNA
- MLP για gene expression

Κάθε subnetwork μαθαίνει μια **latent embedding representation**.

3. Training objective = Triplet loss + Binary cross-entropy

Τι θελει να πετυχει το Moli στη διαρκεια του training:

Να ταξινομεί σωστά → Responder / Non-responder
Αυτό επιτυγχάνεται με **Binary Cross-Entropy Loss (BCE)**.

Να δημιουργήσει έναν embedding space με καθαρό διαχωρισμό
όπου:

- οι responders "μαζεύονται" σε μια περιοχή,
- οι non-responders βρίσκονται μακριά , ώστε το μοντέλο να γενικεύει καλύτερα.

Αυτό **δεν** μπορεί να το πετύχει η BCE μόνη της.
Άρα χρησιμοποιούμε **Triplet Loss**.

Triplet Loss.

$$\mathcal{L}_{\text{tri}}(\theta) = \sum_{\substack{a,p,n \\ y_a = y_p \neq y_n}} [m + D_{a,p} - D_{a,n}]_+ .$$

Η triplet loss αποδεικνυει ότι δίνοντας ενα **anchor point Xa** , η προβολη ενος **positive point Xp** που ανηκει στην ίδια class (person) με την Ya ειναι πιο κοντά στην προβολή του projection από την προβολή ενός άλλου negative point που ανήκει σε μια αλλη class Yn τουλαχιστον κατα margin m

Την χρειαζομαστε γιατι:

Μονο της η bca δεν μπορει να οργανωσει τον embedding space γεωμετρικα,
απλα προβλεπει 0/1.

Ενω με την tiplet loss οργανωνουμε τον χωρο , μειωνουμε το Overlap αναμεσα στις κλασεις ,βελτιωνουμε το generalization και επιτρεπει και το transfer learning σε PDX/TCGA

4. Datasets

- GDSC (cell lines): κύριο dataset εκπαίδευσης (training) επειδή έχει πολλά screened drugs και αρκετά δείγματα ανά drug.
- PDX Encyclopedia (mouse xenografts): external validation σε πιο “in vivo-like” μοντέλα.
- TCGA με διαθέσιμο drug response (από clinical annotations): external validation σε ασθενείς.
- TCGA χωρίς drug response: δεν χρησιμοποιείται για αξιολόγηση accuracy, αλλά για biological/clinical association analysis (π.χ. αν οι προβλέψεις για EGFR inhibitors σχετίζονται με EGFR pathway genes).

Συμπερασμα:

Το MOLI αποτελεί χαρακτηριστικό baseline της πρόσφατης βιβλιογραφίας, καθώς συνδυάζει multi-omics late integration με deep learning και αξιολογείται με external validation σε PDX και TCGA δεδομένα.
Αρα ειδαμε οτι στο MOLI το drug response αντιμετωπίζεται ως πρόβλημα δυαδικής ταξινόμησης (responder / non-responder), μέσω binarization του IC50.