Graph DRP (2021)

Monday, December 15, 2025 6:50 PM

Προβλημα

Drug Response Prediction =
προβλέπουμε πόσο καλά *ένα συγκεκριμένο φάρμακο* δουλεύει σε *ένα συγκεκριμένο κύτταρο* (cell line).

Μεχρι το 2021:
ML & DL μοντελα:
Παριστανανε τα drugs ως string (SMILES)
Δεν ξερανε ποια γονιδια των Cells επηρεαζουν την προβλεψη

Ζητηματα που προσπαθει να λυσει το GraphDrp
Το SMILES δεν είναι φυσική αναπαράσταση μορίων
Έλλειψη interpretability

Βασικη Ιδεα

Βλεπουμε το drug σαν γραφο
Προσθέτουμε interpretability με saliency maps.

Drug representation

Από SMILES → Molecular Graph

Για καθε φαρμακο εχουμε ενα : **SMILES string**
CC(=O)NC1=CC=C(O)C=C1
Χρησιμοποιούν RDKit για να το μετατρέψουν σε:
Graph G = (V, E)

- V (nodes) = άτομα
- E (edges) = χημικοί δεσμοί

Τι πληροφορία έχει κάθε node (Atom Features)

Κάθε άτομο **δεν είναι απλώς "C" ή "O"**.
Περιγράφεται με **feature vector** από το **DeepChem atom featurizer**.

5 atom features: (**binary / categorical features**)

multi-dimensional binary feature vector

1) Atom symbol
2) **Atom degree**
3) **Total number of Hydrogens**
4) **Implicit valence**
5) **Is aromatic**

Πώς συνδέονται τα άτομα

Adjacency matrix

Αν υπάρχει δεσμός μεταξύ atom i και j:
A[i, j] = 1
Αλλιώς :
A[i, j] = 0

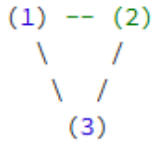
GCN

Η λύση: GCN του Kipf & Welling (2017)

$$H^{(l+1)} = \sigma(\hat{A}H^{(l)}W^{(l)})$$

H(l)= ο node feature matrix στο επιπεδο l
A = ο adjancy matrix
W = ο πινακας βαρων των παραμετρων
σ= Μη γραμμικη activation function(ReLU)

Εστω οτι εχουμε αυττο τον γραφο:



Adjacency matrix (A)

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Προσθετουμε self-loops προσθετωντας τον identity matrix (I)

$$A' = A + I = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Degree Matrix

Ενας πινακας οπου το καθε στοιχειο στην διαγωνιο του αναπαριστα το αθροισμα των connections για αυτο τον node.

$$D' = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Normalizing the Adjacency Matrix

Υπολογιζουμε τον αντιστροφο της τετραγωνικης ριζας του degree matrix

$$D'^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{3}} \end{bmatrix}$$

Επειτα υπολογιζουμε

$$\hat{A} = D'^{-\frac{1}{2}} A' D'^{-\frac{1}{2}}$$

Applying Features and Weights X matrix

Για να υλοποιησουμε graph convolution πολλαπλασιαζουμε τον X με τον Normalized Adjancy matrix A

A'X

Applying the Weight Matrix W

A'XW

Τελικα καθε κομβος ειναι ενημερωμενος με βαση :

Τα δικα του χαρακτηριστικα

Των γειτονων του

Τον πινακα βαρων που εκπαιδευτηκε.

Επειτα προσθετουμε Relu και Global max pooling.

GAT

Το ισο βαρος σε ολους δεν ειναι απολυτα σωστο

Με το GCN:

Όλοι οι γείτονες:

- συμβάλλουν ισότιμα
- μόνο η κανονικοποίηση αλλάζει το scale

Ενω στην πραγματικοτητα σε ενα μοριο:

Δεν ειναι ολοι οι δεσμοι εξισου σημαντικοι

π.χ.

- functional group
- aromatic ring
- side chain

GAT = > Attention σε γραφο

Attention coefficient aij

Για καθε edje i-----j

aij=attention(Wxi,Wxj)

Δηλαδη συγκρινει το ατομο i με το ατομο j

Μετα

αij=softmaxj(aij)

Επειτα κανουμε update τον X

X = $$\sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mathbf{W}\mathbf{x}_j\right),$$

Multi-head attention (10 heads)

10 attention heads στο 1ο GAT layer

Το ίδιο neighborhood εξετάζεται από 10 διαφορετικές “οπτικές”
Κάθε head μπορεί να εστιάζει σε άλλο chemical pattern
Παράδειγμα:

- head 1 → aromatic rings
- head 2 → polar groups
- head 3 → side chains

Στο τέλος τα heads συνδυάζονται

GIN

Προβλημα του GCN & GAT

Και τα δύο κάνουν weighted sum aggregation
Αυτό μπορεί να:

- χάσει πληροφορία
- διαφορετικοί γράφοι → ίδιο embedding

GIN

X' = $$MLP\left((1+\mu)\mathbf{x}_i + \sum_{j\in N(i)} \mathbf{x}_i\right),$$

Εχουμε :

Το άθροισμα γειτόνων (χωρίς normalization)

Γιατί το άθροισμα διατηρεί πληροφορία πλήθους και ταυτότητας

(1+ε)xi

Ο κόμβος δεν συγχέεται με τους γείτονες

Το **μ** επιτρέπει στο μοντέλο να δώσει *περισσότερο ή λιγότερο βάρος* στο ίδιο το atom

Διαισθητικά:

"Άλλο είμαι εγώ, άλλο οι γείτονές μου"

Γιατί MLP αντί για γραμμικό W

Στο GCN:

x'=AXW

Στο GIN:

x'=MLP(·)

Αυτό σημαίνει:

βαθιά μη-γραμμική συνάρτηση

όχι απλή γραμμική προβολή

Το paper αποδεικνύει ότι το GIN έχει την ίδια διακριτική ικανότητα με το **Weisfeiler–Lehman graph isomorphism test**

Τελικη ροη

SMILES
→ RDKit
→ Molecular Graph (X, A)
→ Drug GNN Encoder ← GCN / GAT / GIN ή/GAT+GCN
→ Global Max Pooling
→ FC → 128-dim drug embedding

Cell representation

Καθε κυτταρο αναπαρισταται με εναν 735-dim binary vector

C ε [0,1]^735

Κάθε θέση Ck αντιστοιχεί σε ένα συγκεκριμένο aberration (π.χ. "TP53 mutation", "MLL mutation", ή κάποιο CNV event — ανάλογα πώς ορίζεται στο dataset).

Ck = 1 , το κυτταρο εχει αυτο το aberration

Ck = 0, δεν το εχει

Το GraphDRP χρησιμοποιεί binary genomic aberrations ,σε αντιθεση με το task μας που , το

cell line αναπαρίσταται μέσω συνεχών gene expression τιμες , τα οποία απαιτούν διαφορετική αρχιτεκτονική επεξεργασίας

Ποιο aberration επηρεαζει το κυττραρο;

Saliency map

Saliency maps are visual tools, often heatmaps, that highlight the most important parts (pixels, words) of an input that influence a machine learning model's decision, acting as a key part of Explainable AI.

From <https://www.google.com/search?q=Saliency+map&oq=Saliency+map&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIGCAEQLhhA0gEHNDQ2ajBqMagCCLACAfEFsXC6HsEL74E&sourceid=chrome&ie=UTF-8>

Ορίζουν όλο το μοντέλο ως συνάρτηση:

$$\hat{Y} = f(C, D)$$

- C: cell vector (735)
- D: drug graph (μέσω GNN)
- Y^: predicted response (IC50 normalized 0–1)

Saliency

$$S = \frac{\partial \hat{Y}}{\partial C}$$

Παίρνεις την παράγωγο της πρόβλεψης ως προς **κάθε** είσοδο-feature του cell line.
Μεγαλο Sk = το feature CkC_kCk (ένα συγκεκριμένο aberration) έχει μεγάλη επίδραση στην πρόβλεψη
Μικρο Sk = μικρή επίδραση

Συμπερασμα

Το GraphDRP έδειξε ότι η αναπαράσταση των φαρμάκων ως μοριακών γράφων μέσω graph neural networks υπερέχει έναντι των SMILES-based προσεγγίσεων στην πρόβλεψη drug response. Ωστόσο, το μοντέλο περιορίζεται στη χρήση binary genomic aberrations για την αναπαράσταση των cell lines. Οι ίδιοι οι συγγραφείς αναγνωρίζουν ότι η ενσωμάτωση gene expression δεδομένων αποτελεί σημαντική κατεύθυνση μελλοντικής έρευνας, γεγονός που ευθυγραμμίζεται με το gene expression–based setup που υιοθετείται στην παρουσα πτυχιακη εργασια.