BMC Part of Springer Nature
BMC Bioinformatics

Pang *et al. BMC Bioinformatics* (2024) 25:182
https://doi.org/10.1186/s12859-024-05669-x

Prediction of anticancer drug sensitivity using an interpretable model guided by deep learning

Weixiong Pang[1,2], Ming Chen[1,2] and Yufang Qin[1,2*]

BIBS seminar

2023.05.23

YEONJU SEO

BIBS

Contents

1. Introduction

Abstract of project

- **Object**

  - **Prediction of cancer treatment response is a**n important topic in clinical and pharmacological research, as people expect it **to customize effective treatment plans for individual patients.**
  - **However,** (1) due to tumor heterogeneity, patients with the same tumor type may have different treatment responses → it making drug selection important.

    (2) Most deep learning models are black boxes, making it difficult to understand the underlying mechanisms of drug therapy → it is challenging to explain the relationship between network models and cellular molecular feature functions without understanding or paying attention to the biological mechanisms behind the predicted results.

    Therefore, it is imperative **to establish an interpretable model that receives various cell line and drug feature data** to learn drug response mechanisms and achieve stable predictions between available datasets.

Abstract of project

- **Object**

- So, this study proposes **DrugGene, a new interpretable deep learning model.**
- DrugGene integrates **(1)gene expressions, (3)gene mutations, (2)copy number variations(CNV)** of cancer cells, and **(4)chemical characteristics of anticancer drugs** to predict their sensitivity.

→ in order to predict their sensitivity.

- Also, they employ two branches model: **a visual neural network (VNN)** that models the hierarchical structure of biological subsystems, and **a traditional artificial neural network (ANN)**

→ in order to capture the chemical structural features of drugs for **establishing interpretable model** .

Why this study need?

- **Key points**

- **Enhancing the interpretability of the model** and **understanding the molecular pathways** that control or reflect drug sensitivity **can help determine which cancer patients should receive treatment and which specific drugs have actual positive catalytic effects.**

- **Utilizing biological pathways to construct neural networks,** which can use genotypes to monitor changes in the state of network subsystems, **can help interpret the prediction results in the model and achieve satisfactory prediction accuracy.**

→ So, using proposed approach, we can help explore new directions in cancer treatment.

2. Materials & methods

Materials

< Datasets >

- **From Cancer Treatment Response Portal (CTRP)**
  1. Links genetic and cellular characteristics with drug sensitivity.
  2. Morgan fingerprint *SMILES notation based on the drug names provided in the dataset.
- **From Cancer Drug Sensitivity Genome (GDSC)**
  1. Provides genomic data for cancer cell lines. (gene mutation, gene expression, gene copy number variation)
  2. Morgan fingerprint *SMILES notation based on the drug names provided in the dataset.
- **From Cancer Cell Line Encyclopedia (CCLE)**
  1. Provides genomic data for cancer cell lines. (gene mutation, gene expression, gene copy number variation)
- **Gene Ontology (GO)**
  1. information on molecular function, cellular components, and biological processes

How to get data ?

< From Cancer Treatment Response Portal (CTRP)>

: links the genetic, lineage, and other cellular characteristics of cancer cell lines with small molecule sensitivity

	Name	Quantity	Data form	Database
Cell lines	Cell line	684	gene mutation, gene expression, and gene copy number variation	CCLE GDSC
Drugs	SMILES	942	Morgan fingerprint	GDSC, CTRP
Cell line-drug pairs	AUC	8969	AUC	CTRP
Gene ontology	Biological process	2086	GO Term	GO

1. **Dataset Composition:**
   a. Drugs: 684
   b. Cell Lines: 942
   c. Cell Line-Drug Pairs: 8969
2. **Target Value:** Area Under the Dose-Response Curve (AUC), which measures the effectiveness of a drug on a specific cell line.(On the x-axis, plot the concentrations, and on the y-axis, plot the response rates. Connect these points to form a curve.)

- Example:
  - Anticancer Drug A is tested on Breast Cancer Cell Line X.
  - The dose-response relationship is measured and graphed.
  - The AUC is calculated from this graph, indicating the drug's effectiveness on this cell line.

*This setup helps researchers identify which drugs are most effective for specific cancer types, facilitating personalized cancer treatment development

2-1. Materials

How to get data ?

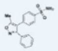
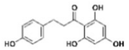
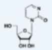
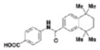
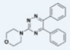
Cancer Therapeutics Response Portal v2

CENTER FOR THE SCIENCE OF THERAPEUTICS

COMPOUNDS FEATURES TARGETS CLUSTER

show 10 entries search all columns:

clear filters

compound structure	compound name	compound synonyms	target or activity of compound	gene symbol of protein target	status	tested concentrations	max conc for AUC
	valdecoxib	valdecoxib, valdecoxib, SC 65872, SC-65872, Bextra, Bextra, celecoxib, BRD-K12994359	inhibitor of cyclooxygenase-2 (COX2)	PTGS2	approved	0.0180 - 600 uM	600
	phloretin	phloretol, phloretin, dihydronaringenin, naringenin chalcone, Phloretol, phloretin, BRD-K15563106	natural product; inhibitor of glucose uptake	SLC5A1	probe	0.0180 - 590 uM	590
	zebularine	zebularine, BRD-A01145011	inhibitor of DNA methyltransferases	DNMT1	probe	0.0180 - 590 uM	590
	AM-580	AM-580, BRD-K06854232	agonist of retinoic acid receptor alpha	RARA	probe	0.0160 - 530 uM	530
	ML006	ML006, T6748133, BRD2698, BRD-K89692698	agonist of sphingosine 1-phosphate receptor 3	S1PR3	probe	0.0160 - 530 uM	530

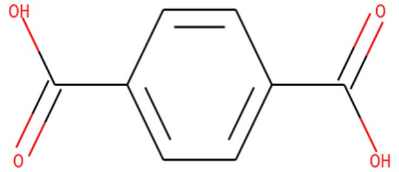
How to get data ?

* SMILES : notation for representing the structure of a chemical in ASCII sentences.
→ Complex chemicals can be described in a single line.
ex) CO_2 : O=C=O, hydrogen cyanide(HCN) : C#N

< From GDSC & CTRP>

- Compound Data: From GDSC and CTRP, converted to SMILES notation

```
smile = 'O=C(C1=CC=C(C=C1)C(O)=O)O'
molecules = []
molecules.append(Chem.MolFromSmiles(smile))
featurizer = dc.feat.graph_features.ConvMolFeaturizer()
mol_object = featurizer.featurize(mols=molecules)
```



< From GDSC & CTRP>

- Genomic Data: From CCLE and GDSC, including:
  a. Gene mutation data
  b. Gene expression level data
  c. Gene copy number data

Table 1 Experimental datasets on cell lines, drugs, and gene ontology

	Name	Quantity	Data form	Database
Cell lines	Cell line	684	gene mutation, gene expression, and gene copy number variation	CCLE GDSC
Drugs	SMILES	942	Morgan fingerprint	GDSC, CTRP
Cell line-drug pairs	AUC	8969	AUC	CTRP
Gene ontology	Biological process	2086	GO Term	GO

< From GO>

- Selected Information: 2086 biological processes for model branch modeling.
  → To enhance the understanding of molecular functions and biological processes in cancer research.

How to get data ?

+ < Data preprocessing >

Drug Data Preprocessing

- Software: alvaDesc, RDKit.
- Process:
  - Molecular descriptors and Morgan fingerprint encoding.
  - Decomposition into molecular fragments represented as a 2048-bit vector.

Genomic Data Preprocessing

- **Gene Selection:** Top 15% most commonly mutated genes (3008 genes).
- **Handling Missing Data:** Average genotype data used to replace missing data.
- **Encoding:**
  - Gene mutations: One-hot encoding.
  - Gene expression and copy number variation: Normalization to 0-1 range.

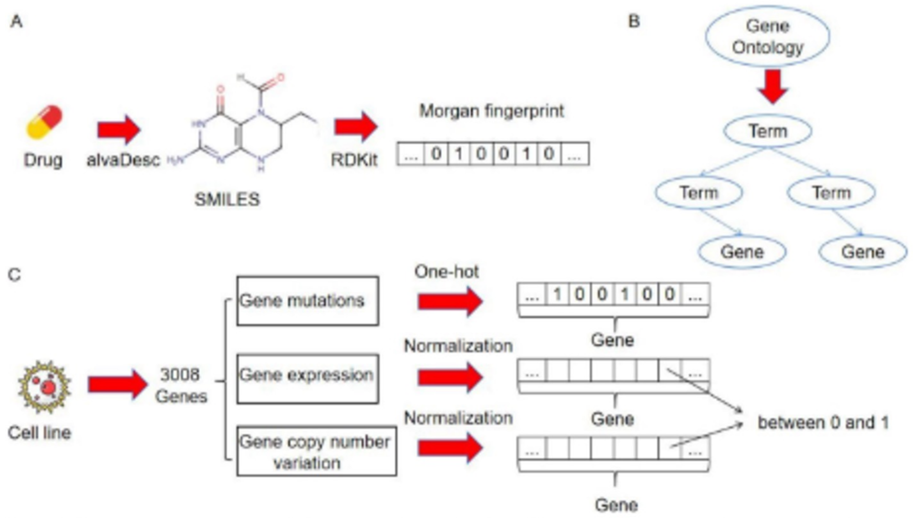


Fig. 1 **A** The processing process of drug data. Obtain the SMILES symbol and Morgan fingerprint code with a length of 2048 for each drug, **B** Select available biological process information from Gene Ontology (GO), **C** The preprocessing process of cell line data, from which available gene mutation, gene expression, and gene copy number variation data can be obtained

*Finally, combine medicinal chemistry characteristics and cancer cell lines to ensure the data format conforms to deep learning model specifications.

About model architecture

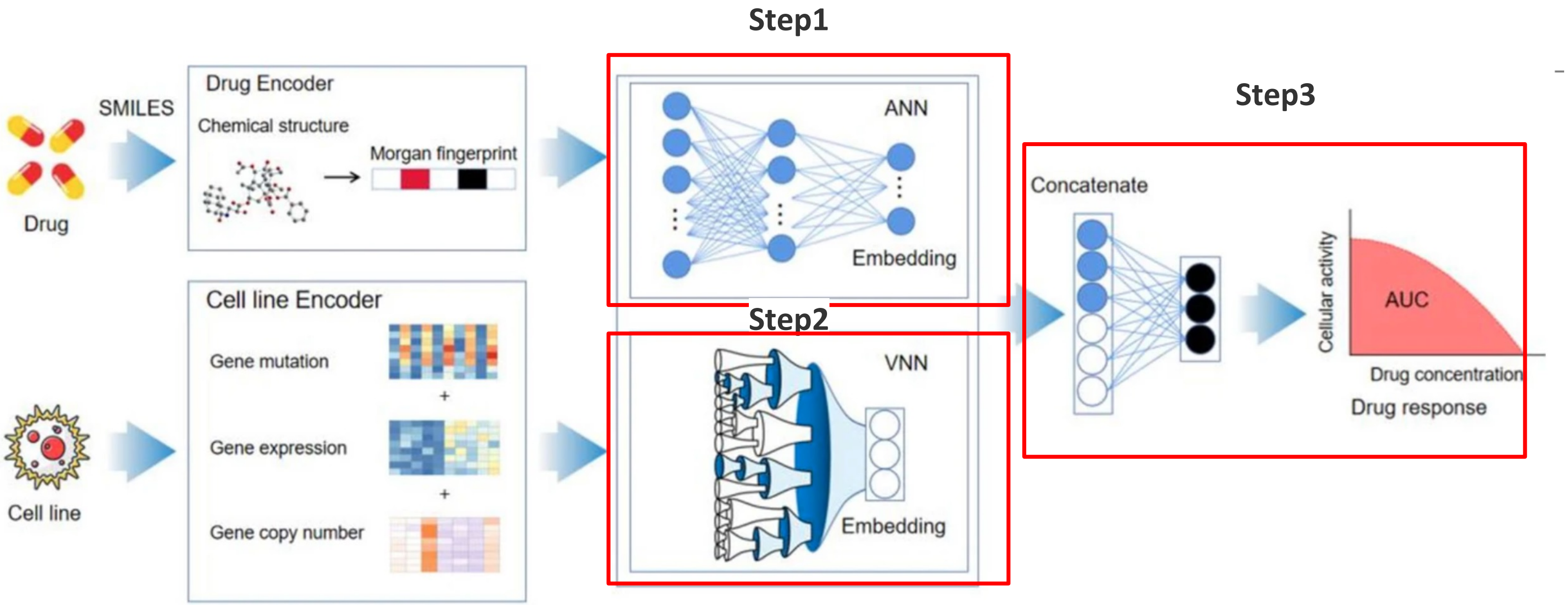
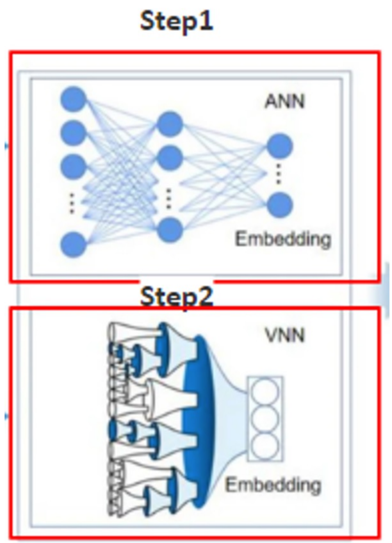


Fig. 2 Workflow of DrugGene. DrugGene uses visible neural networks (VNN) and traditional artificial neural networks (ANN) as sub-modules and combines their outputs for drug response prediction

About model architecture

▪ Two-branch model

Step1) Visual Neural Network (VNN)
Step2) Artificial Neural Network (ANN)



Visual Neural Network (VNN):

- Models hierarchical structure of molecular subsystems in cancer cells. Inputs: Gene mutation, gene expression, and gene copy number variation data.
- Data is fused into a new matrix without changing dimensions for VNN input.

Artificial Neural Network (ANN):

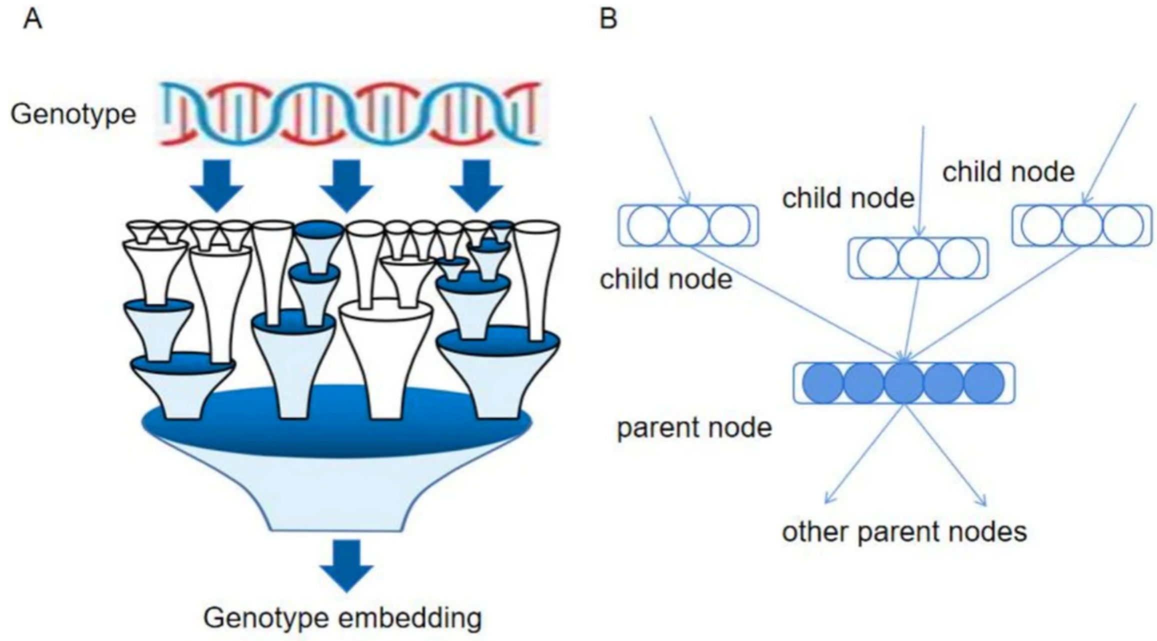
- Inputs: Morgan fingerprint encoding for drugs.

Training and Integration:

- VNN and ANN sub-models are trained independently during the training phase.
- Outputs are combined into a neuron layer.

→ Final output: Predicted drug sensitivity response.

Step 1) VNN (Visual neural network)



Step 1) VNN (Visual neural network)

▪ VNN

1. **Hierarchical Structure Modeling:**

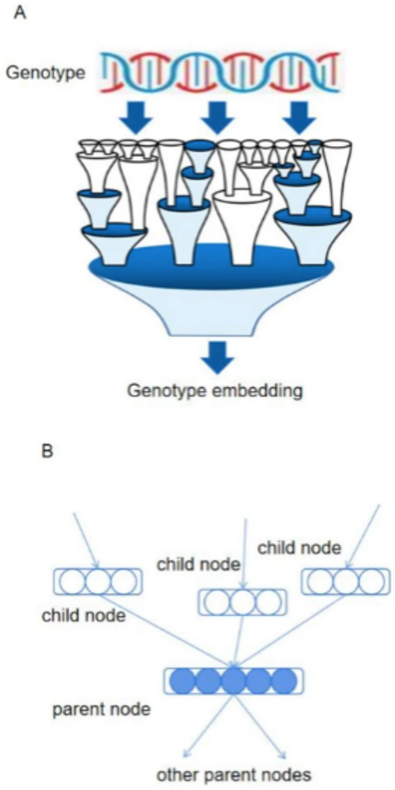
- Based on Gene Ontology database.
- Constructs cellular subsystems using 2086 biological processes.
- **Subsystems** are nodes in a neural network connected through hierarchical relationships.

***Subsystem Representation:**

➔ Neurons represent the functional state of each subsystem.

➔ Connectivity follows the hierarchical structure from small reactions to overall cell functions.

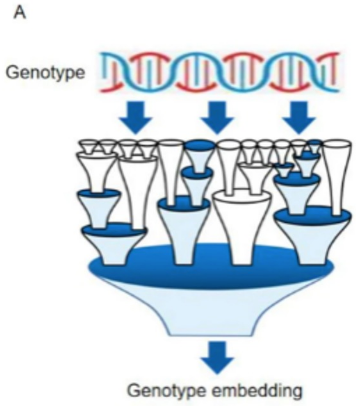
➔ Neurons receive input from child nodes and send output to parent nodes.



Step 1) VNN (Visual neural network)

Network Design:

- 2086 subsystems with a maximum depth of six layers.
- Bridges genotype changes to cell activity or drug sensitivity.



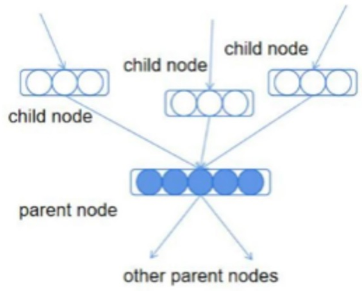
step 1) Genomic Data Input:

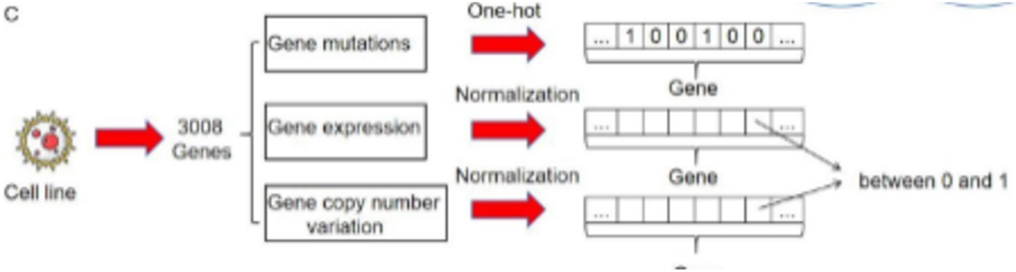
➔ Includes gene mutations, gene expression, and gene copy number variation.

➔ Data represented by 3008-length two-dimensional tensors.

* **selected the top 15% of genes most commonly mutated in human cancer based on CCLE and the genes annotated in the GO database**

➔ Tensors are merged, normalized, and scaled between 0 and 1.





Step 1) VNN (Visual neural network)

step 2) Encoding:

➔ Gene mutations converted from binary to Gray code to minimize errors.

***Gray code : a binary numeral system where two successive values differ in only one bit.**

→ This property is useful for minimizing errors during the transition between consecutive values, which is beneficial in applications like digital encoding and error correction.

Conversion Process:

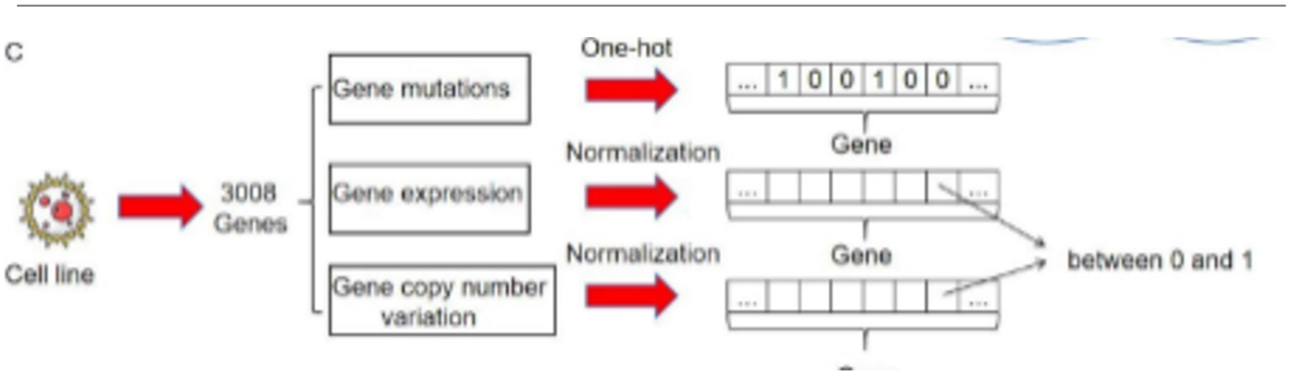
1. Binary to Gray Code Conversion:
   - The first bit of the Gray code is the same as the first bit of the binary number.
   - Each subsequent Gray code bit is found by XORing the current binary bit with the previous binary bit.
2. For example, converting the binary number 0100 (decimal 4) to Gray code:
   - First bit: 0 (same as the first bit of the binary number)
   - Second bit: 1 (XOR of the first and second binary bits: 0 ^ 1 = 1)
   - Third bit: 1 (XOR of the second and third binary bits: 1 ^ 0 = 1)
   - Fourth bit: 0 (XOR of the third and fourth binary bits: 0 ^ 0 = 0)
3. Result: 0100 (binary) -> 0110 (Gray code)

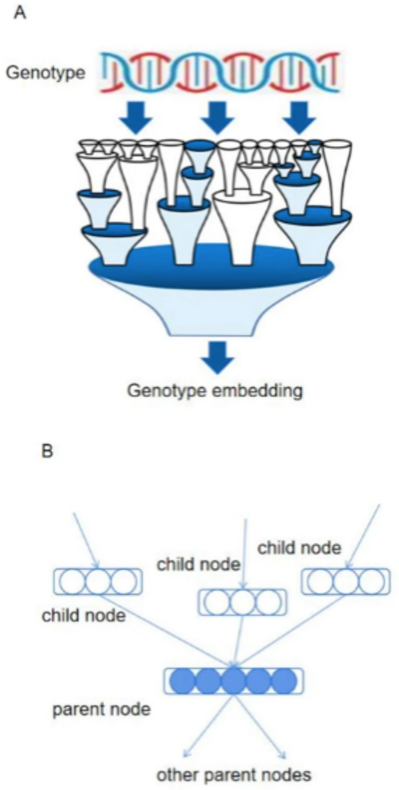
Example of Gray Code and Binary Conversion:

4-bit Gray Code and Binary Example:

Decimal	Binary	Gray Code
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
4	0100	0110
5	0101	0111

Step 1) VNN (Visual neural network)





→ Gene perturbations propagate through the hierarchical structure of subsystems, leading to functional changes and predictive responses in cell activity.

→ Embedding the structure of deep neural networks into the biological hierarchy allows VNN to monitor changes in network subsystems, interpret prediction results, and improve model performance.

Step 1) VNN (Visual neural network)

step 3) Training Process:

- Minimizes objective function, initializes weights between -0.01 and 0.01.
- Uses Batch Normalization to reduce internal covariate shifts caused by different weight scales.
- We set the training dataset to **D = {(X1, Y1), ... ,(XN , YN)},** where **N is the number of samples, for each sample i,Xi ∈ R^M represents genotype through a binary vector of states on M genes,** and **Yi ∈ R is a numerical value representing the observed drug response.**

- **The multidimensional state of each subsystem t is represented by the output vector Oi(t), denoted by a linear function of all its subsystems and annotated gene states, connected to the input vector Vi(t).**

$$O_i(t) = BatchNormalization(Tanh(W(t)V_i(t) + b(t))) \quad (1)$$

** BatchNormalization () : a regularization of model weights, which can solve gradient vanishing and reduces traditional drop out steps in deep learning*

**Tanh : a nonlinear transformation hyperbolic tangent function.*

Step 1) VNN (Visual neural network)

step 3) Training Process:

- Then , perform the training process by minimizing the objective function:

$$\frac{1}{N}\sum_{i=1}^{N} Loss(linear(O_i(r), Y_i)) + \alpha \Sigma_{t\neq r} Loss(linear(O_i(t), Y_i)) + \lambda \| W(t) \|_2 \qquad (2)$$

** Loss () : the squared error loss function, and r is the root of the hierarchy.*
Oi(r) : the output of the root
Oi(t) : the output of other subsystems.
α: selecting appropriate learning rate parameters

- ADAM optimizer with batch size of 10,000.
- Learning rate determined through grid search (10^{-1} to 10^{-4}).
- Standard backpropagation for gradient calculation.

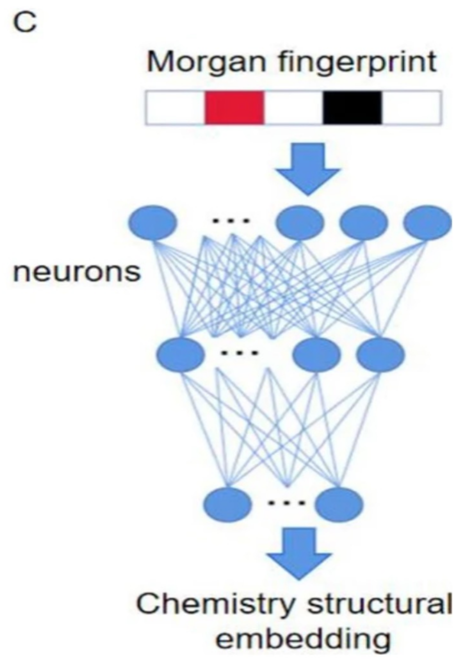
Step 1) VNN (Visual neural network)

step 4) Model Output:

- Represents the embedding state of the entire cell.
- Predicts drug sensitivity responses.

⇒ Enhances interpretability by embedding neural network structure into biological hierarchy.

Step 2) ANN (Artificial Neural Network)



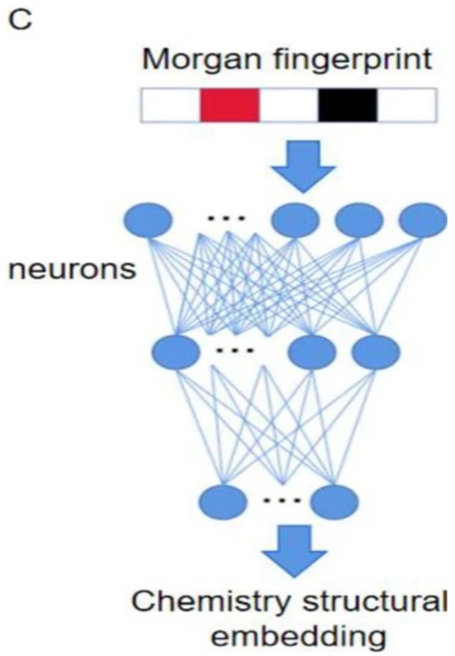
Step 2) ANN (Artificial Neural Network)

Network Design:

- Three layers with a specific number of neurons in each layer.
- Processes high-dimensional drug data encoded by Morgan fingerprints to predict drug sensitivity.

step 1) Drug data Input:

- ➔ **Encoding:** Morgan molecular fingerprint code.
- ➔ **Representation:** 2048-length binary vectors.
- ➔ **Input Format:** Each element represents an activation state (0 = inactive, 1 = activated).



Step 2) ANN (Artificial Neural Network)

Step 2) Training Process:

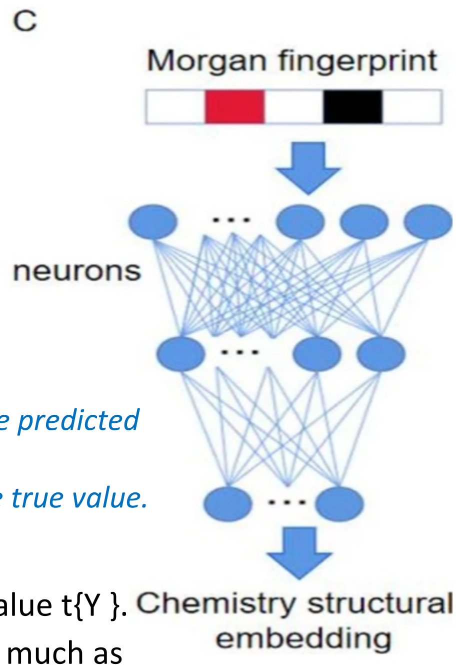
- **Objective:** Minimize the loss function by adjusting weights and biases.

basic formula : $Y = WX + b$

** X : features or known conditions*
Y : labels or results
W : weight vector
Y : labels or results
b : bias

loss function : $Loss = (p\{Y\} - t\{Y\})^2$

** p{Y } : numerical value representing the predicted value of the sample*
t{Y } : numerical value representing the true value.



➔ The goal is to make the predicted value p{Y } as close as possible to the true value t{Y }.
➔ Loss function is to minimize the sum of the loss values of a neural network as much as possible.
➔ Training should be terminated and the parameters of the trained neural network saved when the loss function reaches a certain convergence threshold

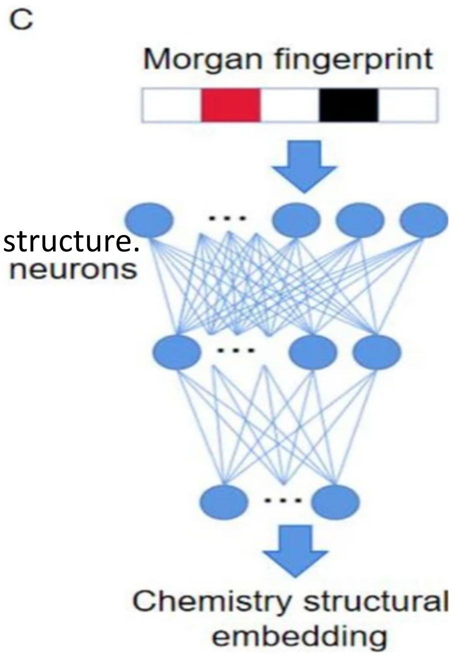
Step 2) ANN (Artificial Neural Network)

Step 3) Data Propagation:

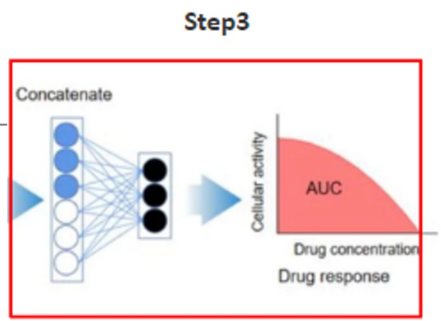
- **Input Layer:** Receives the Morgan fingerprint encoding.
- **Hidden Layers:** Data propagates through layers, neurons process the input.
- **Output Layer:** Generates an embedded representation of the drug's chemical structure.

Step 4) Prediction Output:

- **Result:** Final layer provides the prediction output for drug sensitivity.



Step 3) Full connection between VNN and ANN



Input Data: Output vectors from VNN and ANN.

- **VNN:** Generates genotypic embeddings.
- **ANN:** Generates medicinal chemistry structural embeddings.

→ Fully connected layer (with concatenate) combines these embeddings to establish a complete model network.

Vector Concatenation: Combine the output vectors from VNN and ANN to create a new high-dimensional vector.

Fully Connected Layer: Use the combined vector as input to generate the prediction results.

Final Output:

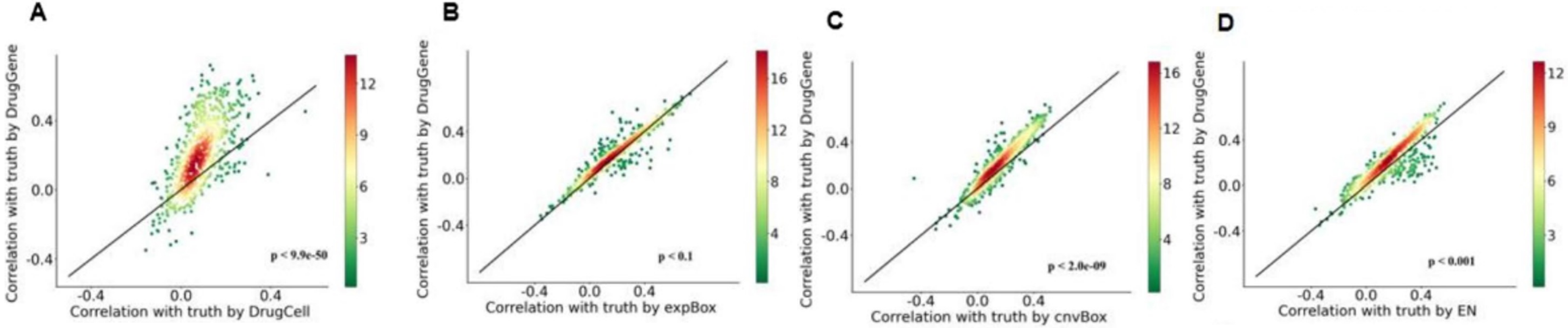
➔ Area Under the Curve (AUC) of the normalized dose-response curve.
➔ **AUC=0:** Indicates complete cell killing , **AUC=1:** Indicates no effect.

3. Results

1) Performance evaluation of DrugGene in predicting drug sensitivity

To compare its performance against current models on the same dataset
→ so let's evaluates the predictive accuracy of DrugGene using a tenfold cross-validation method

Training: 684 drugs, 942 cell lines, 8969 cell line-drug pairs. , **Testing:** Same cell line-drug pairs to evaluate models.



Most points are shifted more towards the Y-axis, meaning the DrugGene model has a higher correlation with the actual values compared to the other model. → This suggests that DrugGene has better predictive performance.

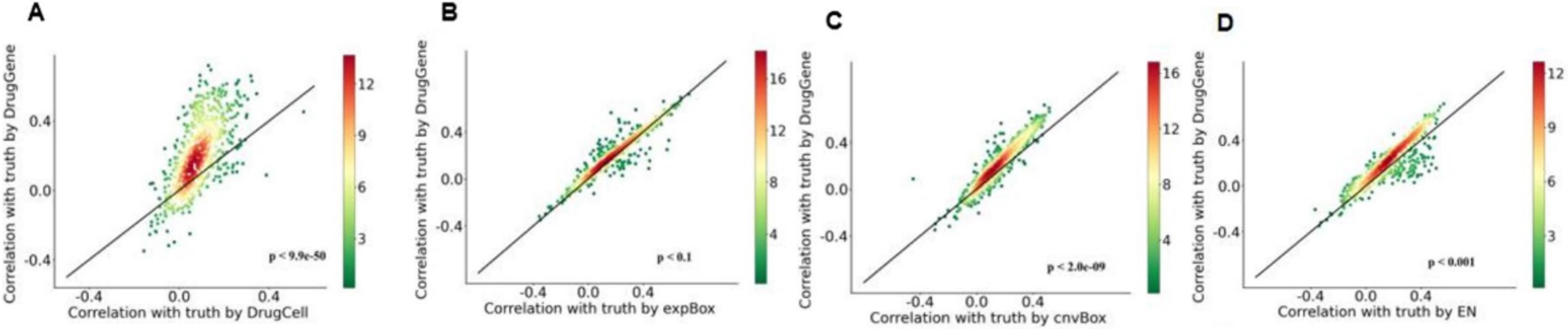
Most points are clustered around the diagonal line, indicating that the correlations of the DrugGene and expBox models with the actual values are similar. → However, the points are more often above the diagonal, suggesting that DrugGene generally has a slightly better predictive performance.

1) Performance evaluation of DrugGene in predicting drug sensitivity

To compare its performance against current models on the same dataset

→ so let's evaluates the predictive accuracy of DrugGene using a tenfold cross-validation method

Training: 684 drugs, 942 cell lines, 8969 cell line-drug pairs. , **Testing:** Same cell line-drug pairs to evaluate models.



***DrugCell:** uses gene mutations and drug characteristics for drug sensitivity prediction

***expBox:** only uses gene expression and medicinal chemistry features

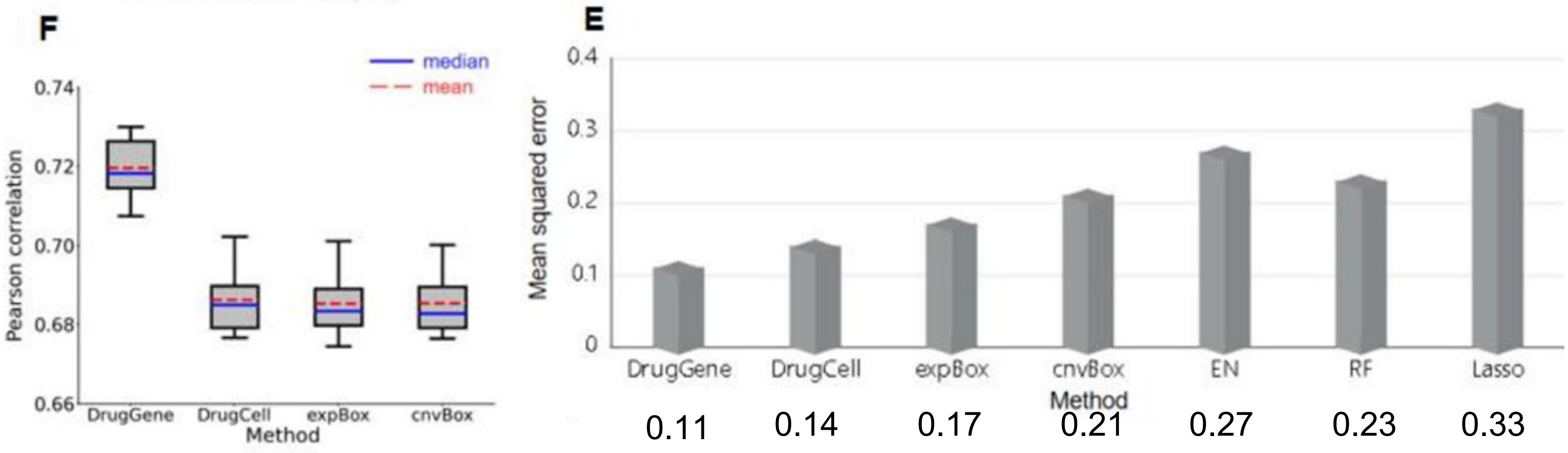
***cnvBox:** only uses copy number variation and drug coding data

***elastic network (EN)**

DrugGene can effectively improve the prediction results by integrating gene mutation, gene expression, gene copy number variation, and Medicinal chemistry characteristics.

1) Performance evaluation of DrugGene in predicting drug sensitivity

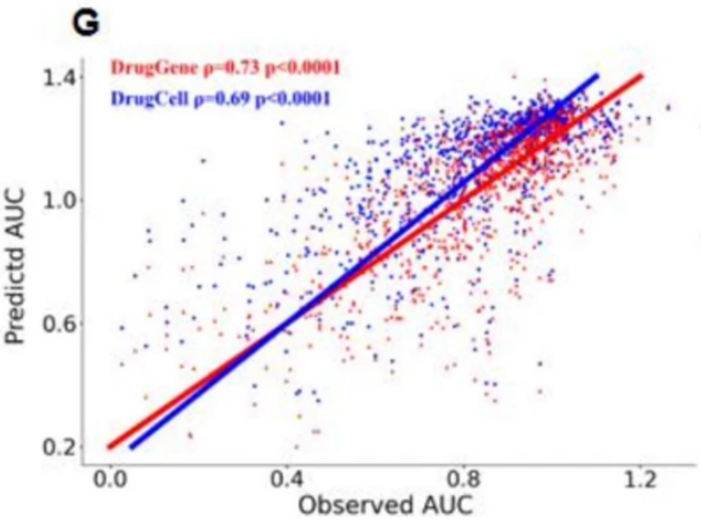
To compare its performance against current models on the same dataset
→ so let's evaluates the predictive accuracy of DrugGene using a tenfold cross-validation method

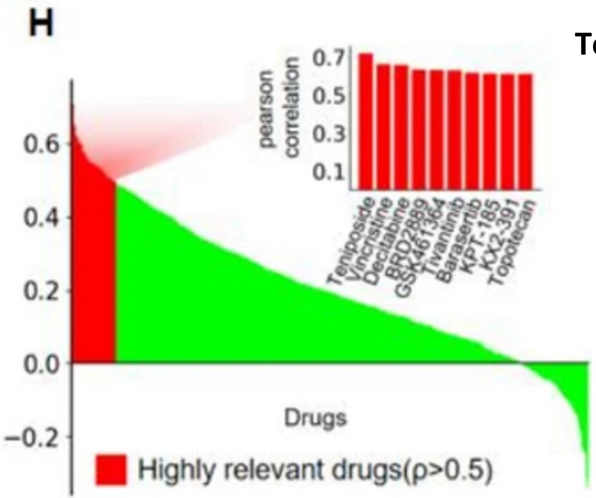


DrugGene's predictive correlation is significantly higher than the competitor models, which have relatively close median values.
MSE results infer that our method has the best predictive performance, followed by DrugCell.

1) Performance evaluation of DrugGene in predicting drug sensitivity

To compare its performance against current models on the same dataset
→ so let's evaluates the predictive accuracy of DrugGene using a tenfold cross-validation method





: top ten drugs with the highest prediction accuracy.

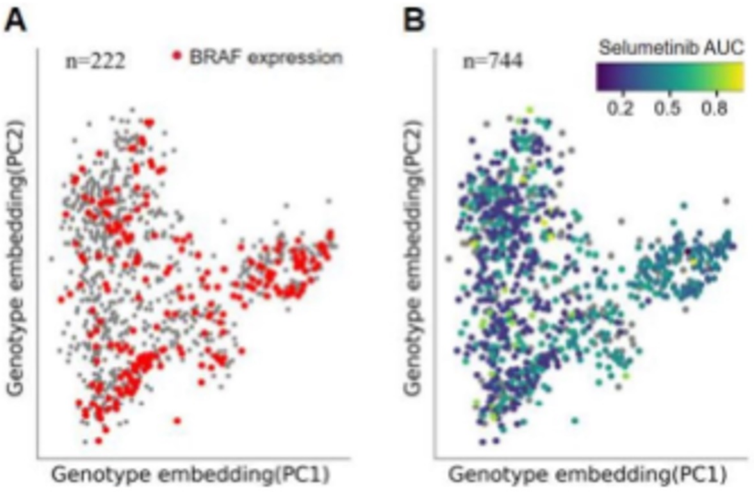
Top Performing Drugs:

- **Teniposide:** Highest prediction accuracy, used to treat acute lymphocytic leukemia in children.
- **Vincristine:** Second highest prediction accuracy, used as a clinical anti-tumor drug, especially for treating acute leukemia in children.

They plotted a visual scatter using the predicted values of DrugGene and DrugCell on the test set, revealing that DrugGene has a better fitting performance than DrugCell.
The predicted results of the model can reflect the therapeutic effects of specific targeted drugs

2) Learning the mechanisms of drug reactions through DrugGene

After evaluating DrugGene's predictive ability based on the treatment response of each drug
→ let's discuss the model's interpretability



Analysis of Drug Sensitivity:

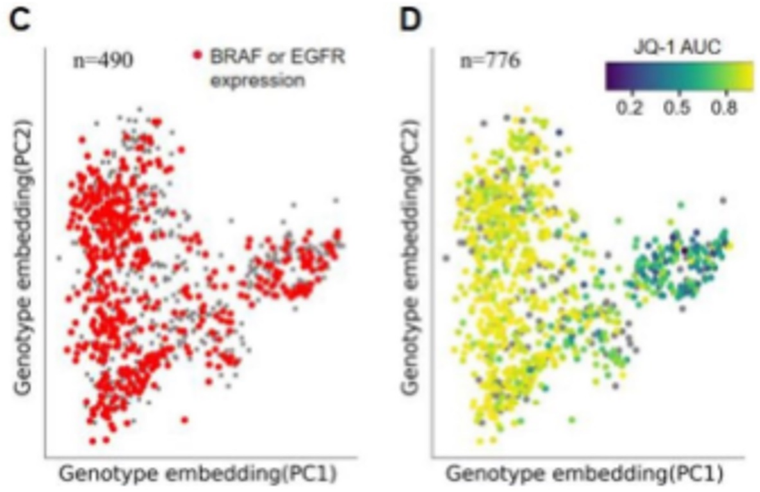
Two main components from genotype data are visualized.

- **Fig. 5A:** Points represent cell lines, colored by BRAF expression levels (red for high, gray for low).
- **Result:** High BRAF expression levels promote sensitivity to the MEK inhibitor selumetinib (Fig. 5B).
- **AUC Values:** Smaller AUC values indicate sensitivity, higher values indicate resistance.
- **Conclusion:** Sensitive cell lines (Fig. 5B) correspond to red dots (high BRAF) in Fig. 5A.

The two-dimensional visualization results of each cell line can be intuitively observed by extracting the two main components from all genotype data generated by VNN.
Two-dimensional visualization shows that high BRAF expression levels correlate with sensitivity to the MEK inhibitor selumetinib → find that Selumetinib is an inhibitor for BRAF mutations in clinical treatment

2) Learning the mechanisms of drug reactions through DrugGene

After evaluating DrugGene's predictive ability based on the treatment response of each drug
→ let's discuss the model's interpretability



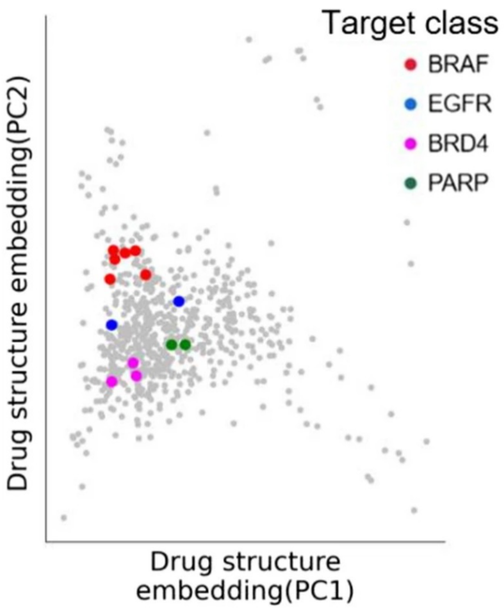
Analysis of Drug Resistance:

- **Fig. 5C:** Points colored by EGFR or BRAF expression levels.
- **Result:** High EGFR or BRAF expression levels confer resistance to the BET family inhibitor JQ1 (Fig. 5D).
- **Conclusion:** Resistant cell lines (Fig. 5D) correspond to red dots (high EGFR or BRAF) in Fig. 5C.

They also analyze the interpretability of the model when the cell lines exhibit drug resistance. Figure 5C distinguishes the distribution of EGFR or BRAF expression levels.
For EGFR or BRAF, high expression levels can confer resistance to the BET family inhibitor JQ1 (Fig. 5D).
The points presented as drug resistance mostly correspond to the red points in Fig. 5C. In clinical treatment, JQ1 is often used as an inhibitor for EGFR or BRAF mutations.

2) Learning the mechanisms of drug reactions through DrugGene

After evaluating DrugGene's predictive ability based on the treatment response of each drug
→ let's discuss the model's interpretability



- ➔ Selected two main components from ANN embeddings.
- ➔ Each point represents a drug.
- ➔ Drugs are layered based on targeted genes' mechanisms of action.
- ➔ Clustering phenomenon observed for different targeted genes.

Key Target Genes:

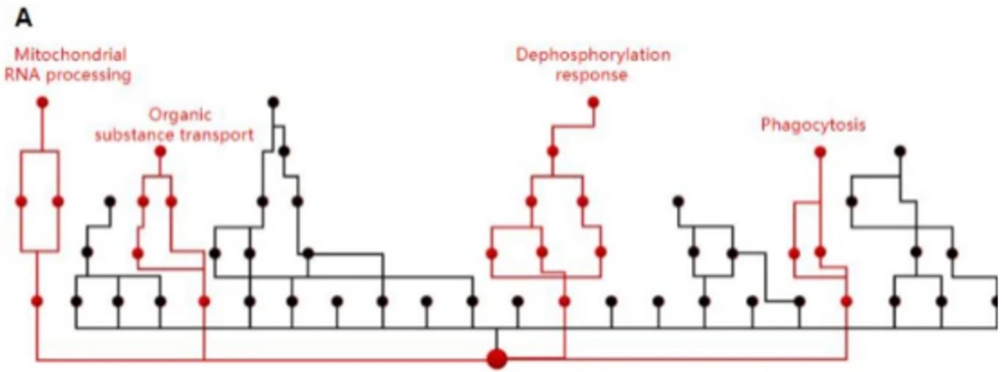
- Notable clustering for BRAF, BRD4, and PARP.

  → These drugs act as inhibitors for their respective genes in clinical trials.

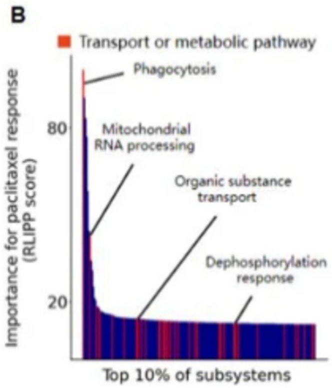
DrugGene distinguishes key features leading to drug sensitivity and resistance.
In summary, DrugGene is able to distinguish key features of genotypes that lead to drug sensitivity and resistance, as well as understand the chemical structural characteristics of drug biological activity

3) The role of subsystems in neural networks

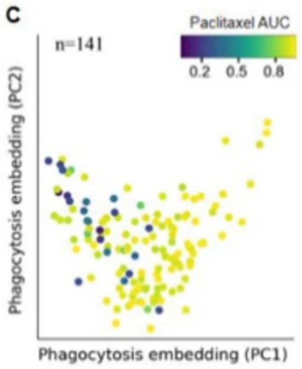
To further explore the effectiveness of identifying key subsystems using VNN's genotypic output, we
→ so let's performed ablation experiments by evaluating the predictive performance of subsystems using different metrics



: 2D visualization of VNN with key pathways highlighted in red



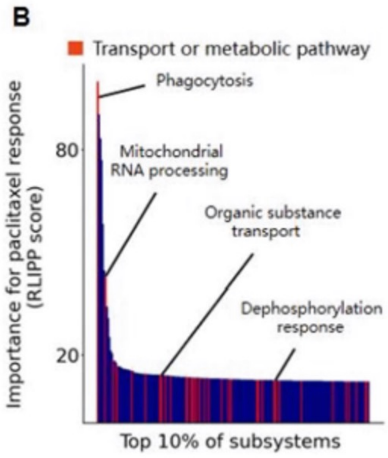
: Plot of top 10% subsystems by RLIPP score



: Phagocytosis subsystem's effectiveness in distinguishing sensitivity (low AUC) and resistance (high AUC) to paclitaxel

3) The role of subsystems in neural networks

To further explore the effectiveness of identifying key subsystems using VNN's genotypic output, we
→ so let's performed ablation experiments by evaluating the predictive performance of subsystems using different metrics



: Plot of top 10% subsystems by RLIPP score

***Top Subsystems:**

- ➔ Phagocytosis
- ➔ Mitochondrial RNA processing
- ➔ Organic substance transport
- ➔ Dephosphorylation response

B. RLIPP Score

- **Purpose:** Evaluate the performance of subsystems based on predicted drug response of parent node relative to child node in VNN.
- **Method:** Used neuron values representing states of parent and child nodes to predict drug response. The Pearson correlation coefficient between predicted values and actual target values was used to compute RLIPP.

$$RLIPP = \frac{P_2 - P_1}{P_1}$$

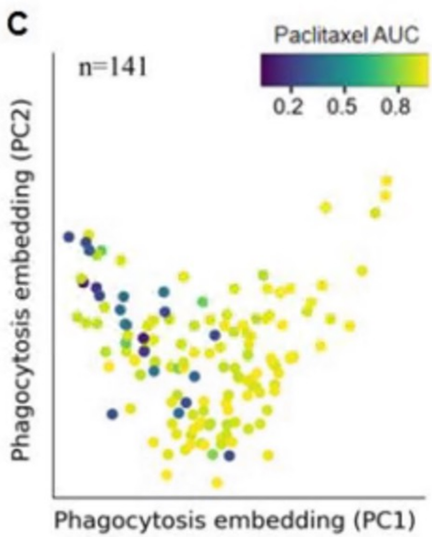
** P1 :Pearson correlation coefficient predicted by the child node*
P2 : predicted result of the parent node.

: Higher RLIPP scores indicate subsystems with better predictive performance.

RLIPP score indicates the importance of the parent–child system during prediction.
We chose paclitaxel to react with cells and used the RLIPP score to evaluate the important subsystems in this reaction process (Fig. 7B).

3) The role of subsystems in neural networks

To further explore the effectiveness of identifying key subsystems using VNN’s genotypic output, we
→ so let’s performed ablation experiments by evaluating the predictive performance of subsystems using different metrics



: Phagocytosis subsystem’s effectiveness in distinguishing sensitivity (low AUC) and resistance (high AUC) to paclitaxel

In the reaction process of paclitaxel, we used the state changes of the subsystems with the highest scores to represent the predicted values of drug reactions → found that the higher-ranked Phagocytosis subsystem could distinguish the sensitivity and resistance of cell lines reacting with paclitaxel (Fig. 7C).
The lower the AUC value, the more sensitive the response, while the opposite indicates the drug resistance response.

4. Conclusion

Conclusion

Results show that

- The DrugGene method integrates reference data from cell lines and drugs, utilizing partial reference information to align with practical clinical practice.
- It incorporates gene ontology data to construct part of the network, making the model interpretable.
- The method achieves satisfactory drug sensitivity prediction accuracy, which helps in reducing medical costs, analyzing new cancer drug treatment strategies, and supporting cancer immunotherapy.

Novelties in their model

1) Enhances model interpretability.
2) Achieves accurate drug sensitivity prediction.
3) Reduces medical costs.
4) Aids in analyzing new cancer drug treatment strategies.
5) Supports cancer immunotherapy.

→ This approach provides a great predictive power of anti-cancer drug responses, together with an insight of the potential reactions between cell lines and drugs.

Thank you for listening ☺️