DeepTTA: a transformer-based model for predicting cancer drug response

Likun Jiang † , Changzhi Jiang†, Xinyu Yu ,RaoFu ,ShutingJin and Xiangrong Liu

Published : 2022

Racha Bekhouche

Context and Motivation:

Cancer treatment remains a complex challenge due to several limitations of traditional approaches. DeepTTA, a Transformers-based model, emerges as an innovative solution by combining advances in pharmacogenomics and machine learning to predict drug response.

1. Limitations of traditional treatments

- Anticancer treatments often rely on well-defined protein targets (e.g. HER2 for some breast cancers).

However :

- Lack of exploitable targets: Many cancers do not have specific proteins to target.
- Variability of mutations: Tumors in the same patient or between patients are genetically different, making it difficult to develop universal treatments.

2. Complexity of drug responses

- Intra-tumor heterogeneity: A single tumor contains cells with variable sensitivities to drugs.
- Inter-tumor heterogeneity: Same type of cancer, but different response in different patients.

Context and Motivation:

Need a new approach:

Integrating transcriptomic data (gene expression) and AI advancements to better understand and predict interactions between drugs and cancer cells.

DeepTTA emerges as an innovative solution, combining Transformers and neural networks for precise and personalized predictions.

1. General Structure of DeepTTA

DeepTTA is an end-to-end deep learning model consisting of two main components:

- **Drug Feature Extraction** using a Transformer model.
- **Cancer Cell Feature Extraction** via a four-layer neural network.

These two components produce vector representations, which are then fused and passed through a model to predict the drug response, represented by the IC50 indicator.

Database

Database Used:

GDSC (Genomics of Drug Sensitivity in Cancer), a public resource for drug sensitivity data and molecular markers of response containing data pairs, cancer types and drug compounds.

Data Segmentation Methods:

To compare model performance, several data splitting strategies were employed:

- **'Random' Segmentation:** 95% of the data used for training and 5% for testing.
- **'According Cancer' Segmentation:**
  - 80% of instances covering 30 cancer types used for training.
  - 20% for testing the model's ability to predict across different cancers or drugs.
- **'Independent' Segmentation:**
    - 80% of drugs (123) for training.
    - Remaining 20% (31) for testing.

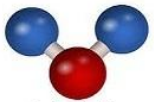
Drug Feature Extraction

1. Drug Representation

a. Data Used: SMILES

Drug features are extracted by representing the chemical structure of molecules using SMILES (Simplified Molecular Input Line Entry System), which encodes molecular structures as character sequences.

- Each string represents the chemical structure of a molecule.


Carbondioxide

O=C=O

SMILES Notation for Carbondioxide

b. Drug Segmentation into Substructures

DeepTTA employs the Explainable Substructure Partition Fingerprint (ESPF) algorithm to segment the SMILES sequence into chemical substructures. This approach is inspired by natural language processing, where words are split into subwords to capture detailed information.

- Example: A drug represented as "CCOCC" would be segmented into "CC," "O," and "CC."

Drug Feature Extraction

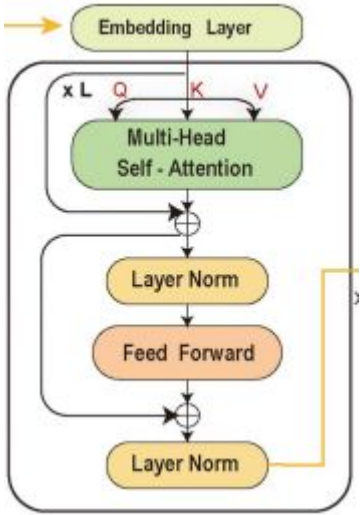
c. Vector Encoding of Substructures

- Substructures are converted into learnable numeric vectors using an encoding matrix.
- Positional information is embedded via a positional encoding matrix to preserve sequence order.
- The final substructure representation is calculated as : $\text{Final Representation} = C_i + P_i$
- where Ci represents the chemical information and Pi the positional information of the substructure in the molecule.

d. Encoding with Transformer

The substructure vectors are processed through a Transformer layer consisting of:

- **8 attention heads** and **6 stacked layers**.
- Each layer has two key components:
  1. **Multi-Attention Layers**:
     - Evaluate interactions between every pair of substructures.
     - Allow each substructure to "attend to" others, capturing their chemical interactions.
  2. **Fully-Connected Feed-Forward Networks**:
     - Extract complex, non-linear relationships between substructures.



Drug Feature Extraction

e. Final Output

- The Transformer generates a global vector summarizing the drug's chemical features and interactions, enabling prediction of cancer cell responses to the drug.

Why is it Effective?

- The Transformer captures complex contextual chemical relationships, such as interactions between functional groups, which directly affect therapeutic efficacy.
- It provides a more comprehensive and precise representation compared to traditional chemical encoding methods.
- By analyzing substructure interactions regardless of their sequence distance, it models intricate drug properties.

How Does it Work?

The attention mechanism establishes connections between all substructures, learning non-linear relationships that influence the drug’s properties.

Cancer Cell Representation

a. Data Used: Transcriptome

Each cancer cell is described by a vector containing **17,777 gene expression values**, extracted from the GDSC2 database.

b. Processing with a Neural Network

The data is processed through a **4-layer neural network**:

- **Input Layer:** Compresses the gene expression values into a dense vector.
- **Three Hidden Layers:**
  - Gradually reduce dimensions using hidden units of **1024, 256, and 64 neurons**.
  - Capture complex relationships between genes while reducing data size.
- **Output:** Extracts important biological features of the cells.

Each layer applies weights to:

- Identify recurring patterns (e.g., similarly expressed genes across multiple cancer cells).
- Reduce dimensions while retaining relevant information.

Representation Fusion

a. Concatenation

The vectors generated by the **Transformer** (drug processing) and the **neural network** (cell processing) are concatenated into a single global vector.

- This vector represents the hypothetical interaction between a drug and a cancer cell.
- Concatenation directly links the drug's chemical properties with the cell's biological features.

The model learns to identify relationships between chemical properties and gene activity.

- **Example:** If a drug contains a substructure that inhibits a highly expressed gene in the cell, the model can recognize this interaction and predict a low IC50 value (indicating high efficacy).

b. Passing Through a Classifier

The combined representation is fed into a **4-layer fully connected classifier** that predicts the IC50 value (indicating the drug's sensitivity or resistance to the cell).

- These layers learn the complex interactions between chemical and biological features.

How Does the Model Predict if a Drug Inhibits a Gene?

What is IC50?

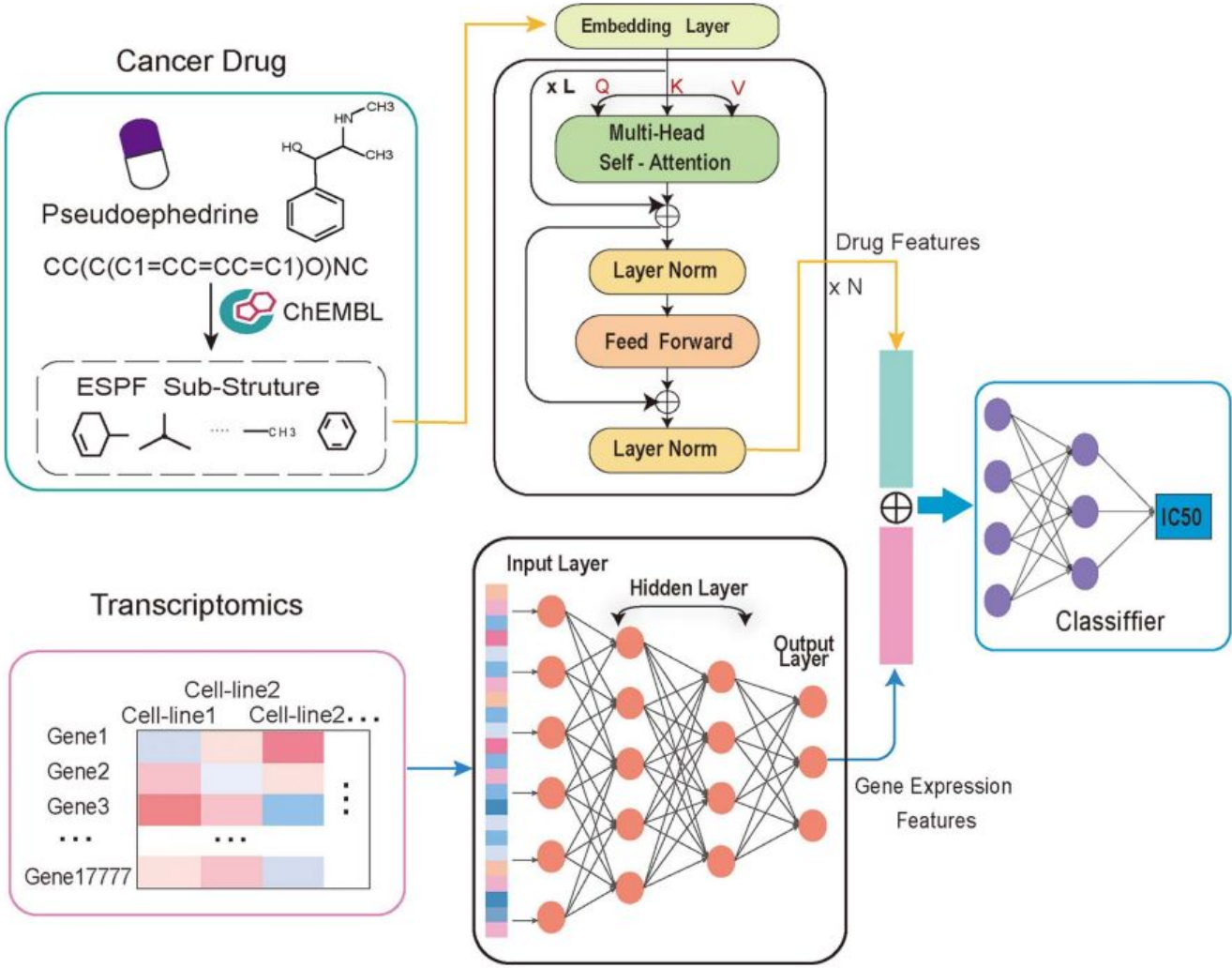
IC50 (Inhibitory Concentration 50) is a measure of a drug's effectiveness in inhibiting a specific biological activity, such as cancer cell growth.

1. **Significance:**
   - **Low IC50:** The drug is highly effective, requiring a low concentration to inhibit 50% of the activity.
   - **High IC50:** The drug is less effective, needing a higher concentration to achieve the same effect.
2. **Mechanism:**
   - The dataset contains pairs of drugs and cells, each associated with an experimentally measured IC50 value.
   - The model takes the concatenated vector (drug and cell features) as input.
   - Through mathematical transformations (matrix multiplications and activation functions), it predicts the IC50 value.
   - **Objective:** Minimize the error between the predicted and actual IC50 values, typically measured using the **Root Mean Square Error (RMSE)**.

Structure of DeepTTA:

Three Main Components:

- **Drug Feature Mining**: Processes drug data.
- **Gene Expression Feature Extraction**: Processes cell transcriptomic data.
- **Classifier**: Combines and predicts drug response.



Training, Optimisation and Prediction

4. Training and Optimization:

- During training, the model optimizes weights to reduce prediction error and improve correlation metrics like Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation Coefficient (SRCC).
- **Optimizer:** Uses algorithms like **Adam Optimizer** to adjust internal weights and minimize RMSE.

5. Prediction after training

Once the model is trained:

- After training, the model processes a concatenated vector of a new drug and an unknown cell.The model uses the learned weights to produce an IC50 prediction.
- This prediction is based on the relationships it discovered during training between the concatenated features and the observed IC50s.

Correlation Evaluation

The model's IC50 predictions are compared to experimental values. Metrics like Pearson Correlation Coefficient (PCC) and Spearman Correlation Coefficient (SRCC) assess whether the learned representations of drugs and cells effectively capture their biological interaction.

RMSE (Root Mean Square Error) :

- Measures the average quadratic difference between predicted IC50 values and true IC50 values.
- The lower the RMSE, the more accurate the model.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

PCC (Pearson Correlation Coefficient) :

- Evaluates the linear correlation between predicted and true IC50 values.
- DeepTTA uses PCC to quantify the linear relationship between predictions and experimental values on test datasets.
- A high PCC (close to 1) indicates the model can predict IC50 values with good accuracy.

SRCC (Spearman Rank Correlation Coefficient) :

- Measures monotonic correlation between the ranks of predicted and true IC50 values (useful for detecting non-linear relationships).
- The formula is based on the differences in IC50 ranks.
    -

Methods and Indicators for Classification

For binary classification of drug response (sensitivity or resistance), the following indicators are used:

- **AUROC (Area Under Receiver Operating Characteristic Curve):**
  Evaluates the model's ability to distinguish effective drugs from ineffective ones (ranges from 0 to 1).
- **AUPR (Area Under Precision-Recall Curve):**
  Measures precision-recall balance for imbalanced datasets.
- **F1-Score:**
  Combines precision and recall for overall classification performance in classification tasks
  -

Overview of Advanced Models

DeepTTA has been compared against four prominent models for drug response prediction:

CDRscan:

- The first deep learning model for drug response prediction.
- Utilizes a dual CNN architecture to process separately : **Genomic mutation fingerprints** of cells and **Molecular fingerprints** of drugs.
- Combines these features through **virtual docking**, simulating interactions between drugs and cells in silico.

- **Limitation**: Restricted to genomic mutation data, limiting its scope for broader biological insights.

tCNNS

- Employs CNNs to learn representations from:
    - Drug **SMILES sequences**.
    - **Genomic features** of cancer cells.

- **Limitation**: Demonstrates modest performance in drug response prediction.

MOLI

- Integrates multi-omics data, including: Somatic mutations, copy number aberrations, and gene expression profiles.
- Utilizes type-specific sub-networks for each data type and a **triplet loss** for improved learning.

- **Limitation**: Limited improvement in predictive accuracy despite multi-omics integration.

DeepCDR

- A hybrid **graph convolutional network (GNN)** integrating: Multi-omics profiles of cancer cells and Chemical structures of drugs.
- Designed to explore intrinsic relationships between molecular structures and cell biology.

- **Limitation**: Graph-based chemical representations are less effective for capturing intricate structural relationships.

Results of DeepTTA

1. Superior Performance on Key Metrics

DeepTTA outperforms all existing models on major regression and classification metrics:

- **Regression Metrics:**
  - **RMSE**: 0.952 (lowest, indicating high prediction accuracy).
  - **PCC**: 0.941 (strong linear correlation between predicted and actual IC50 values).
  - **SRCC**: 0.914 (high monotonic correlation).
- **Classification Metrics:**
  - **AUROC**: 0.884 (superior ability to distinguish effective drugs).
  - **AUPR**: 0.892 (better precision-recall balance for unbalanced data).

2. Performance Across Cancer Types and Drugs

- **Cancer Types:**
  - Performance varies by cancer type, with **PCC ranging from 0.905** (stomach adenocarcinoma) to **0.960** (chronic myelogenous leukemia).
- **Drugs:**
  - Prediction accuracy also varies across drugs, with **PCC ranging from 0.411** (AZD5582) to **0.909** (oxaliplatin).

3. Missing Data Prediction (Missing CDRs)

- DeepTTA successfully predicted **20,478 missing IC50 pairs** (17% of the GDSC dataset).

Limitations of DeepTTA

Despite its strong performance, DeepTTA has some limitations:

- **Variability in Drug Performance**:
  - The model struggles with certain drugs (e.g., AZD5582), resulting in lower PCC scores.
  - Drug-specific limitations may stem from inadequate training data or missing critical chemical features.
- **Dependence on Training Data**:
  - The model's accuracy is heavily reliant on the quality and quantity of available data. Missing or noisy data can affect its predictions.
- **Computational Complexity**:
  - Requires significant computational resources (e.g., multiple GPUs) for training due to the transformer-based architecture.

Summary

DeepTTA predicts drug response by combining **chemical features of drugs** with **biological data from cancer cells**, leveraging innovative AI techniques.

Key Features

- **Transformer for Drugs**: Analyzes SMILES sequences to capture complex chemical relationships.
- **Neural Network for Cells**: Extracts critical gene activity patterns from transcriptomic data.
- **Feature Fusion**: Models interactions between drugs and cells for precise IC50 predictions (measuring the concentration needed to inhibit 50% of cellular activity).

Model Design and Goal

- **Design**: Combines a transformer for drug features with a four-layer neural network for cell analysis.
- **Goal**: Simplify cancer drug discovery by enabling accurate and robust predictions.

Strengths

- **Integration**: Effectively models chemical and biological interactions.
- **Performance**: Surpasses existing models on key metrics (e.g., AUROC, AUPR).
- **Innovative Architecture**:
    - Transformer handles drug complexity.
    - Neural network captures biological insights.

Thanks for Listening

Questions ?