scientific reports

Check for updates

OPEN

Drug discovery and mechanism prediction with explainable graph neural networks

Conghao Wang, Gaurav Asok Kumar & Jagath C. Rajapakse✉

Apprehension of drug action mechanism is paramount for drug response prediction and precision medicine. The unprecedented development of machine learning and deep learning algorithms has expedited the drug response prediction research. However, existing methods mainly focus on forward encoding of drugs, which is to obtain an accurate prediction of the response levels, but omitted to decipher the reaction mechanism between drug molecules and genes. We propose the eXplainable Graph-based Drug response Prediction (XGDP) approach that achieves a precise drug response prediction and reveals the comprehensive mechanism of action between drugs and their targets. XGDP represents drugs with molecular graphs, which naturally preserve the structural information of molecules and a Graph Neural Network module is applied to learn the latent features of molecules. Gene expression data from cancer cell lines are incorporated and processed by a Convolutional Neural Network module. A couple of deep learning attribution algorithms are leveraged to interpret interactions between drug molecular features and genes. We demonstrate that XGDP not only enhances the prediction accuracy compared to pioneering works but is also capable of capturing the salient functional groups of drugs and interactions with significant genes of cancer cells.

Aiming at facilitating precision medicine in complex disease such as cancer, computational approaches have been increasingly proposed to delve into the reactions between drugs and cancer cells[1]. Recently, numerous machine learning[2,3] and deep learning[4,5] methods have been successfully applied to predict drug response levels precisely. However, most of them target at phenotypic screening[6] and do not come along with a reasonable interpretability, rendering drug reaction mechanism obscure. To expedite precision medicine, it is crucial to elucidate the mechanism of action of drugs and thereby promote novel drug discovery.

A proper representation of a drug molecule is pivotal to any drug response prediction methods. According to recent reviews of molecular representations of drugs[7], there are mainly three categories of representation: linear notations, molecular fingerprints (FPs), and graph notations. Linear notations encode the molecule with a vector of string. Two frequently used instances of linear notations are the IUPAC International Chemical Identifier (InChI)[8], and the Simplified Molecular-Input Line-Entry System (SMILES)[9]. SMILES strings are more widely used since it encodes the chemical structure into a string of ASCII characters. CaDRReS[10] applied Matrix Factorization to learn the latent features of drugs with the cell line gene expression data and drug sensitivity matrix, and compared the similarity scores derived from learned features and SMILES notations. tCNNs[11] and CDRScan[12] adopted Convolutional Neural Networks (CNN) to learn a latent representation of drugs' SMILES vector. CNN is a powerful deep learning approach to handle grid-like data in the domain of texts and images, which can be used to encode the linear notations of drugs as well. However, the SMILES notation does not possess the property of locality like texts and images since the physically adjacent atoms in the sequence of SMILES string can be far away from each other in the real molecular environment, and therefore dimisses the structural information of molecules.

Molecular fingerprints, such as Molecular Access System (MACCS)[13] and Chemically Advanced Template Search[14], identify the key structures of a molecule and represent them with a binary vector where each bit denotes the structure's existence. A drawback of this kind of representation is that only the pre-defined structure can be recognized, which might hamper the discovery of novel structures. To circumvent this problem, circular fingerprints such as Extended Connectivity FPs (ECFPs) based on Morgan algorithm[15] has been proposed to iteratively search the substructures of molecules rather than pre-define them. The information of these crucial structures is preserved in this kind of representation, whereas the positional information is lost, and we can hardly track where these sub-structures occur in the molecule. DeepDSC[16] combines Morgan fingerprints of drugs into the latent features of cancer cell lines learned by an auto-encoder. S2DV[17] applied word2vec[18] to

College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore. ✉email: ASJagath@ntu.edu.sg

tokenize ECFP features or SIMLES as drugs' representations. Ma et al. used Atom Pairs (AP), MACCS and circular fingerprints as the descriptor of drugs, and performed the quantitative structure activity relationship (QSAR) study with a Deep Neural Network[19].

Graph notations have recently been brought under the spotlight in the domain of drug representation. Previously, compromising on computational complexity of molecular structures and the confined power of graph learning, aforementioned methods are preferred to denote a molecule even at the cost of loss of information. However, with the advent of Graph Neural Networks (GNN) in the deep learning domain in recent years, it is now feasible to store and analyze the information from molecules in graphs[20].

Numerous variants of GNN models have been applied in the pharmaceutical domain[21,22] and demonstrated to learn the latent representation of the molecular graphs trading off the descriptive power against complexity. A Graph Convolution Network (GCN) model[23] was proposed to predict the chemical properties of molecules and discover porous materials. The typical message passing pattern of GNN intrinsically weakens the influence of distal nodes, which might contradict the real case in the molecule, where atoms from a long topological distance can still interact such as intramolecular hydrogen bonds. An Attentive FP model proposed by[24] leveraged the graph attention mechanism to learn the impact of a node to another. This model addressed the above issue by updating the nodes with a trade-off between the topological distance and the possibly intangible linkage with the attention mechanism. GraphDRP[25] enhanced the tCNN[11] prediction precision by substituting the drug-CNN module with GNN to better encapsulate the drug features. DeepCDR[26], TGSA[27] and DualGCN[28] further explored integrating multi-omics profiles for a better representation of cancer cell lines. Besides modeling drugs with GNN, SWNet[29] introduced a self-attention mechanism to bring drug similarity into the consideration when learning cell features. An algebraic graph-assisted bidirectional transformer (AGBT) model[30] was developed to encode the 3D structure of molecules into algebraic graphs. And Molecular Topographic Map (MTM) was generated from atom features by using Generative Topographic Mapping (GTM)[31] to represent drugs in graphs[32].

In this study, we propose a framework named eXplainable Graph-based Drug response Prediction (XGDP) for predicting anti-cancer drug responses and discovering the mechanism of action. The architecture of XGDP, as shown in Fig. 1, is composed of 3 modules. The GNN module learns the latent features of drugs denoted by molecular graphs. We propose to use a set of novel features adapted from ECFPs as the node features and incorporate chemical bond types as the edge features in our graph convolutional layers. And the CNN module learns the latent features of cancer cell lines from its gene expression profiles. Then, a cross-attention module is utilized to integrate latend features from drugs and cell lines, and thereafter predict the drug responses. The experimental results indicate that, with novel node and edge features, our model outperformed the previous drug response prediction methods[11,25]. Moreover, we leverage deep learning attribution approaches such as GNNExplainer[33] and Integrated Gradients[34] to interpret our model. It is demonstrated that our developed model is capable of identifying the active substructures of drugs and the significant genes in cancer cells, and thus revealing the mechanism of action of drugs.

Methods

Datasets

We propose a deep learning-based approach to predict the drug responses of cancer with molecular graphs of drugs and gene expression data from cancer cell lines. The dataset was acquired from Genomics of Drug

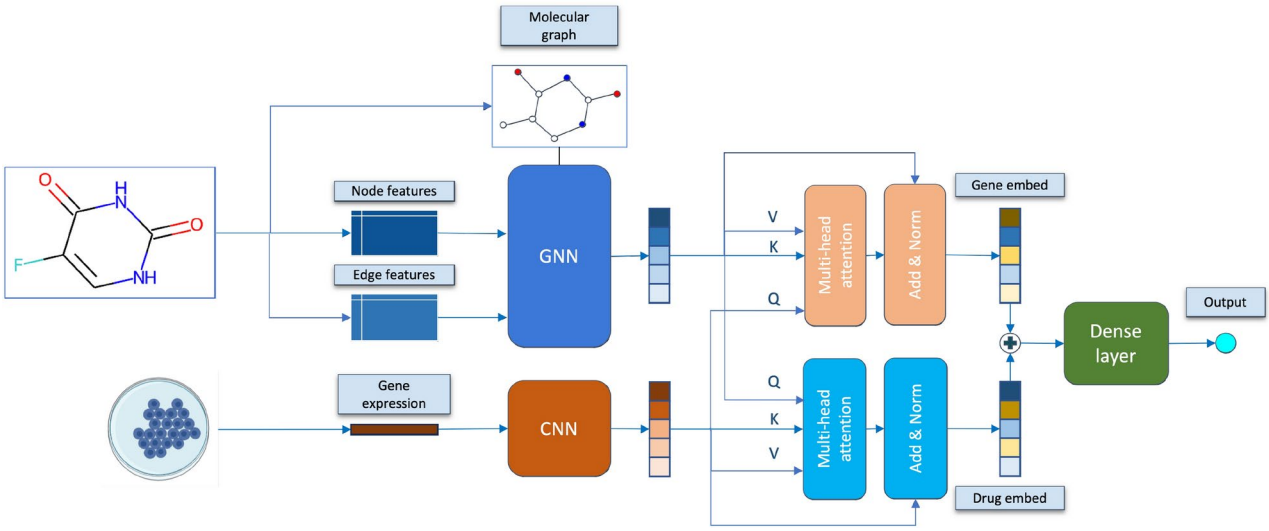

Fig. 1. The architecture of the proposed model XGDP for drug response and mechanism prediction. Molecular graph, node features and edge features are extracted from the drug molecule, and GNN is used for learning the latent features of drugs. CNN is applied to compress the gene expression features from cancer cell lines. Then two multi-head cross-attention layers are leveraged to combine drug and cell features, and the drug response is predicted with the integrated features.

Sensitivity in Cancer (GDSC) database[35], including response levels in IC50 formats, drug names, and cell line names. Gene expression data of cell lines are obtained from Cancer Cell Line Encyclopedia (CCLE)[36]. Drugs' names are retrieved in PubChem database[37] to obtain their SMILES vectors. Then the SMILES vectors are converted into molecular graphs with RDKit library[38].

We combine the GDSC and CCLE datasets by selecting cell lines whose drug responses and gene expression profiles are both recorded. In total, there are 223 drugs and 700 cell lines. After removing missing screening of drug responses, 133,212 pairs of data points are left for experiments. Each cell line is depicted by a transcriptomic profile of 13,142 genes. In order to reduce the dimensionality of the input features to avoid potential over-fitting in model training, we refer to the connectivity map proposed in LINCS L1000 research[39], and preserve only the expression values of the 956 landmark genes, since it is testified that the expression pattern of other genes can be precisely inferred by the landmark genes.

Drug representation

Previous research have demonstrated that representing drugs with molecular graphs provides better predictive power than compressed representations such as SMILES[11,25], since the structural information of a molecule can be naturally preserved in a graph. Specifically, by considering the atoms in a molecule as nodes and the chemical bonds between atoms as edges, an undirected unweighted graph is constructed to represent the drug molecule. From the molecular graphs, node features proposed by DeepChem[40] such as atom symbol, atom degree, etc., can be extracted.

In this chapter, we further enhance the predictive power of a drug's graph representation by incorporating properer node and edge features. In the previous work[25], there are five types of node features, i.e., atom symbol, atom degree, the total number of Hydrogen, implicit value of atom, and whether the atom is aromatic. Nevertheless, these features are intuitively restricted to depict an atom in a molecule. Inspired by the Morgan Algorithm and Extended-Connectivity Fingerprints (ECFP)[15], we present a circular algorithm to compute the feature of an atom, considering both the atom itself and its surrounding environment.

INITIALIZATION:
$X_i^0 := h(F_i)$ Identifier X_i of atom i in molecule M is initialized by hashing its chemical properties F_i
$r := 0$ Initialize interested radius from 0

UPDATING:
for $r \leq 3$ **do**
  for i in M **do**
    $X_i^r = \|_{j \in N_i}(b_j, X_j^r)$ Concatenate bond type b_j between atom i and j, and the identifier X_j^r of atom j. j belongs to the neighbour atoms N_i of i
    $X_i^r = h(X_i^r)$ Transfer the identifier into an integer with hashing function
  end for
  $X_i := X_i \| X_i^r$
  $r := r + 1$
end for

REDUCTION:
for i in M **do**
  for X_i^r in X_i **do**
    Convert the identifier X_i^r into binary vectors
    $B_i^r = binary(X_i^r)$
    for b in $range(length(B_i^r), 64)$ **do**
      Convert the identifier X_i^r into binary vectors
      $B_i^r(b) := 0$
    end for
    $B_i := B_i \| B_i^r$
  end for
end for

OUTPUT:
The final circular feature for atom i is a 256-bit binary vector B_i.

Algorithm 1. Circular atomic feature computation

In Circular Atomic Feature Computation Algorithm 1, F_i refers to the chemical properties of atom i to be encoded, which involves the seven Daylight atomic invariants as the initial chemical properties, including number of immediate neighbors who are non-hydrogen atoms, the valence minus the number of hydrogens (meaning total bond order ignoring bonds to hydrogens), the atomic number, the atomic mass, the atomic charge, the number of attached hydrogens, and aromaticity. X_i^r denotes the identifier of atom i after collecting features from its r-hop neighbour atoms. h is the hashing function used for feature compression and *binary* is the function to convert hashed integers back to binary features. Operator $\|$ refers to the concatenation operation.