Combining bottom up and top down approaches for Drug-Target-Interaction prediction

Tilman Hinnerichs

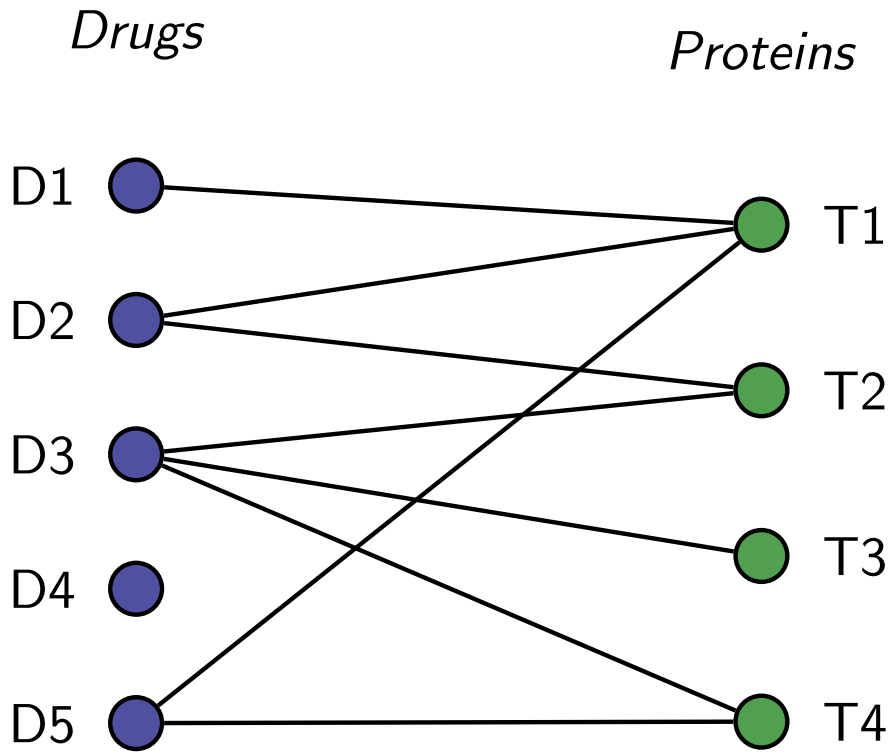
BORG - KAUST

December 09, 2019

Outline

Problem Description

Prediction over bipartite graph:



Classification of recent approaches[1][2]

	Drugs	Protein
bottom-up	▶ GCN over molecules ▶ drug similarity	▶ secondary structure prediction ▶ contact prediction ▶ convolution over amino acid sequences
top-down	▶ network approaches ▶ drug similarity	▶ protein similarity

[1] Chen Wang et al., Briefings in Bioinformatics, 2018
[2] Yu Ding et al., Briefings in Bioinformatics, 2019

Problems of recent approaches

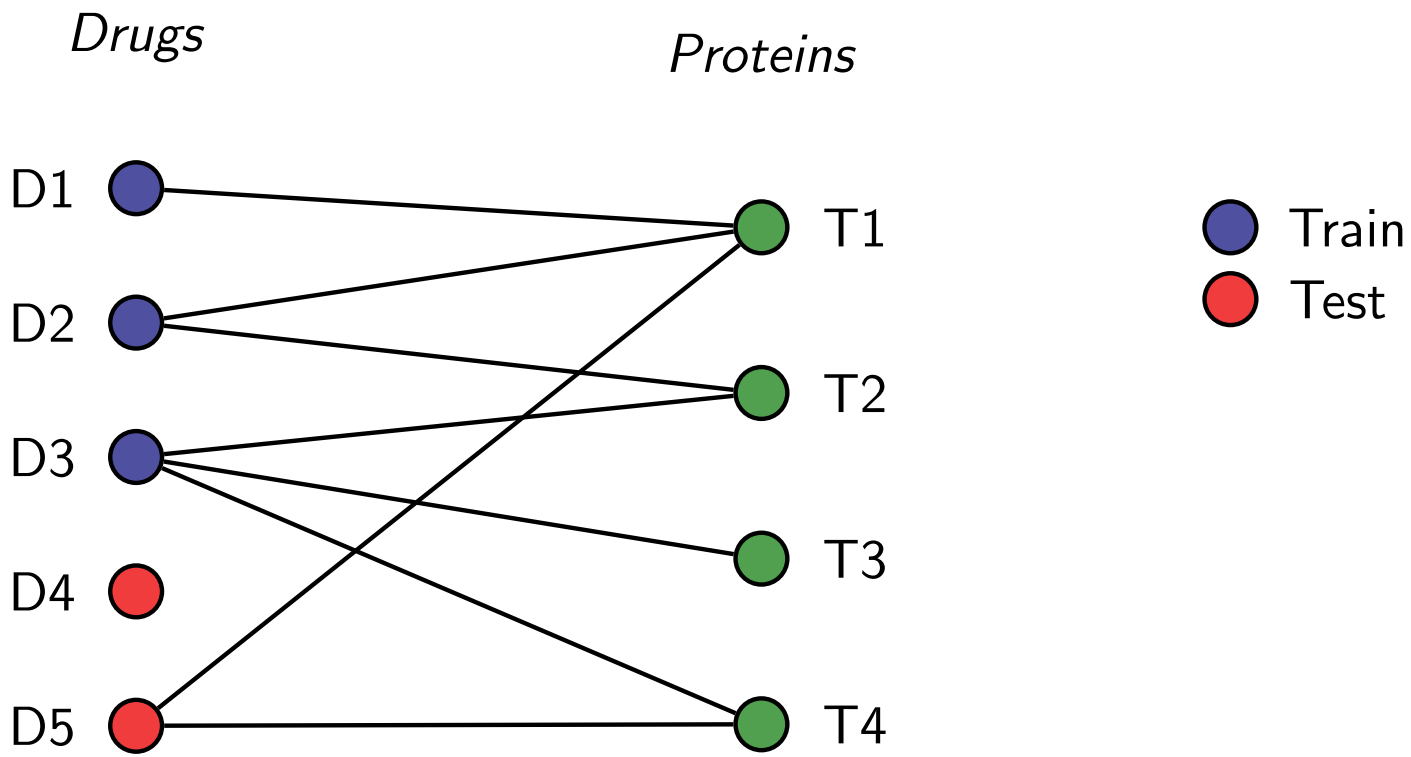
Main issues:

- Lack ability to generalize or are unable to spot small differences
- Usually only top-down or bottom-up
- Not making use of interaction networks

Problems of recent approaches

Main issues:

- Lack ability to generalize or are unable to spot small differences
- Usually only top-down or bottom-up
- Not making use of interaction networks



How to bring both approaches together?

Algorithm

1. For each drug, filter possible targets according to protein based measurement
2. Bring those together with the interaction networks for the prediction

How to bring both approaches together?

Algorithm

1. For each drug, filter possible targets according to protein based measurement
2. Bring those together with the interaction networks for the prediction

Two hypotheses to be tested:

- Interaction only network sufficient?
- Does filter method increase performance?

How to build such a filter?

STITCH database got

- 390 000 chemicals, and
- 3.6 million proteins
- from over 2000 organisms

How to build such a filter?

STITCH database got

- 390 000 chemicals, and
- 3.6 million proteins
- from over 2000 organisms

For each drug, find a motif in the target amino acid sequences and query that against all possible proteins and use the results as features for the prediction

How to find motifs?

1. Build multi sequence alignment with alignment tool of choice over non-human proteins
   - FAMSA, MAFFT, Kalign, MSA Props, Muscle, ...
2. Build representation of that alignment
   - Hidden Markov model with HMMER
   - PWM
3. Query representation against database

How to find motifs?

1. Build multi sequence alignment with alignment tool of choice over non-human proteins
   - FAMSA, MAFFT, Kalign, MSA Props, Muscle, ...
2. Build representation of that alignment
   - Hidden Markov model with HMMER
   - PWM
3. Query representation against database

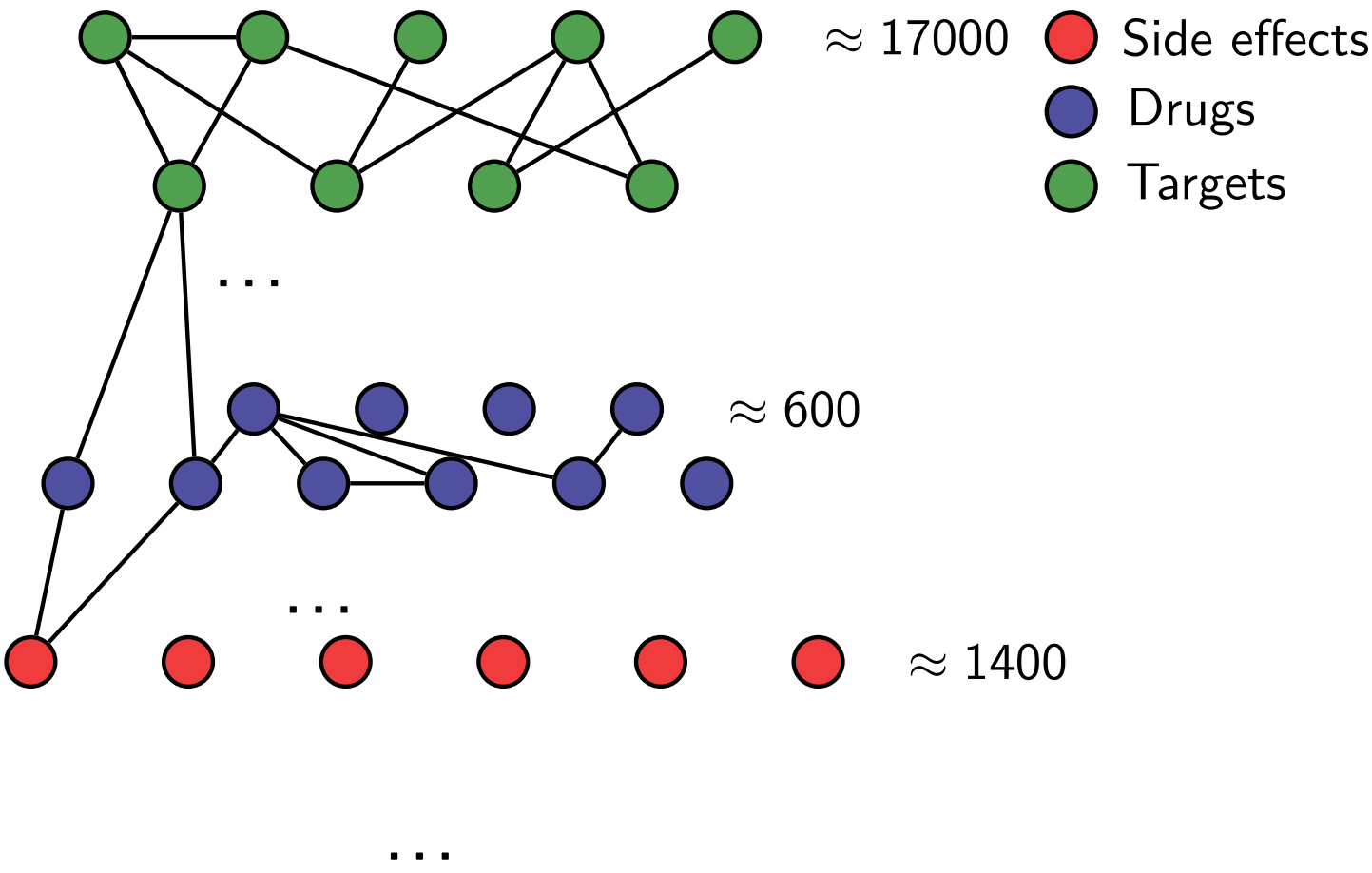
This is quite expensive to compute ... (8000 hours for alignments of 614 drugs + IBEX queue times for such big jobs)

Other features for the model

- Protein-protein-interaction network from STRING (only human)
- Drug-drug-interaction network from Boyce[3] and Drugbank
- Side-effect data for drugs from SIDER, annotated with MedDRA hierarchy (semantic similarity)

[3] Boyce et al., 2015

Features in context



How to learn a sufficient embedding?

- Use Filtered results as node features
- Run different models for classification of proteins/nodes in the graph:
  - GraphSage
  - HinSage
  - Attri2Vec
  - Graph Attention Network (GAT)
  - Graph Convolutional Network (GCN) for node classification
  - GCNs for graph classification of induced subgraph
  - SimpleGCN
  - Personalized Propagation of Neural Predictions (PPNP)
  - Approximate Personalized Propagation of Neural Predictions (APPNP)

First results

Used data				layer sizes	model	Acc	AUC
PPI	DDI	SE	HMM				
1	1	0	0	[32, 32]	graphsage	80.28	70.51
1	1	0	0	[64]	hinsage	83.18	70.33
1	1	0	0	[32, 64]	gcn	88.61	67.27
1	1	0	0	[32, 64]	gat	82.65	71.69
1	1	0	0	[64, 64, 128]	ppnp	89.80	67.78
1	1	0	0	[64, 64, 128]	appnp	84.35	71.82

First results

Used data				layer sizes	model	Acc	AUC
PPI	DDI	SE	HMM				
1	1	1	0	[32, 64]	graphsage	93.27	64.81
1	1	1	0	[64]	hinsage	94.06	63.52
1	1	1	0	[32, 64]	gcn	97.11	56.02
1	1	1	0	[32, 64]	gat	59.77	50.06
1	1	1	0	[64, 64, 128]	ppnp	77.98	55.85
1	1	1	0	[64, 64, 128]	appnp	94.43	66.43
1	1	1	0	[64, 64, 128]	sgc	86.50	68.45
1	0	1	0	[64]	hinsage	59.77	50.06
1	0	1	0	[32, 64]	gcn	59.84	50.10
1	0	1	0	[32, 64]	gat	55.48	55.93
1	0	1	0	[64, 64, 128]	ppnp	21.54	50.15
1	0	1	0	[64, 64, 128]	appnp	40.71	50.13
1	0	1	0	[64, 64, 128]	sgc	2.27	50.18

Future Work

- Prove improvement of generalization
- Test performance against recent approaches
- Tweak model properly
- If performance and generalization is good enough:
  - Test some predicted interactions in vitro
  - Use DTI for Side-effect-phenotype mapping

Full citations

- Yu Ding, Hong Wang, Hewei Zheng, Lianzong Wang, Guosi Zhang, Jiaxin Yang, Xiaoyan Lu, Yu Bai, Haotian Zhang, Jing Li, Wenyan Gao, Fukun Chen, Shui Hu, Jingqi Wu, Liangde Xu, Evaluation of drug efficacy based on the spatial position comparison of drug–target interaction centers, Briefings in Bioinformatics, , bbz024, https://doi.org/10.1093/bib/bbz024
- Chen Wang, Lukasz Kurgan, Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome, Briefings in Bioinformatics, , bby069, https://doi.org/10.1093/bib/bby069
- Toward a complete dataset of drug–drug interaction information from publicly available sources, SerkanAyvazaJohnHornbOktieHassanzadehcQianZhudJohannStaneNicholas P.TatonettifSantiagoVilarfMathiasBrochhausengMatthiasSamwaldhMajidRastegar-MojaradiMichelDumontierjRichard D.Boyce, 2015

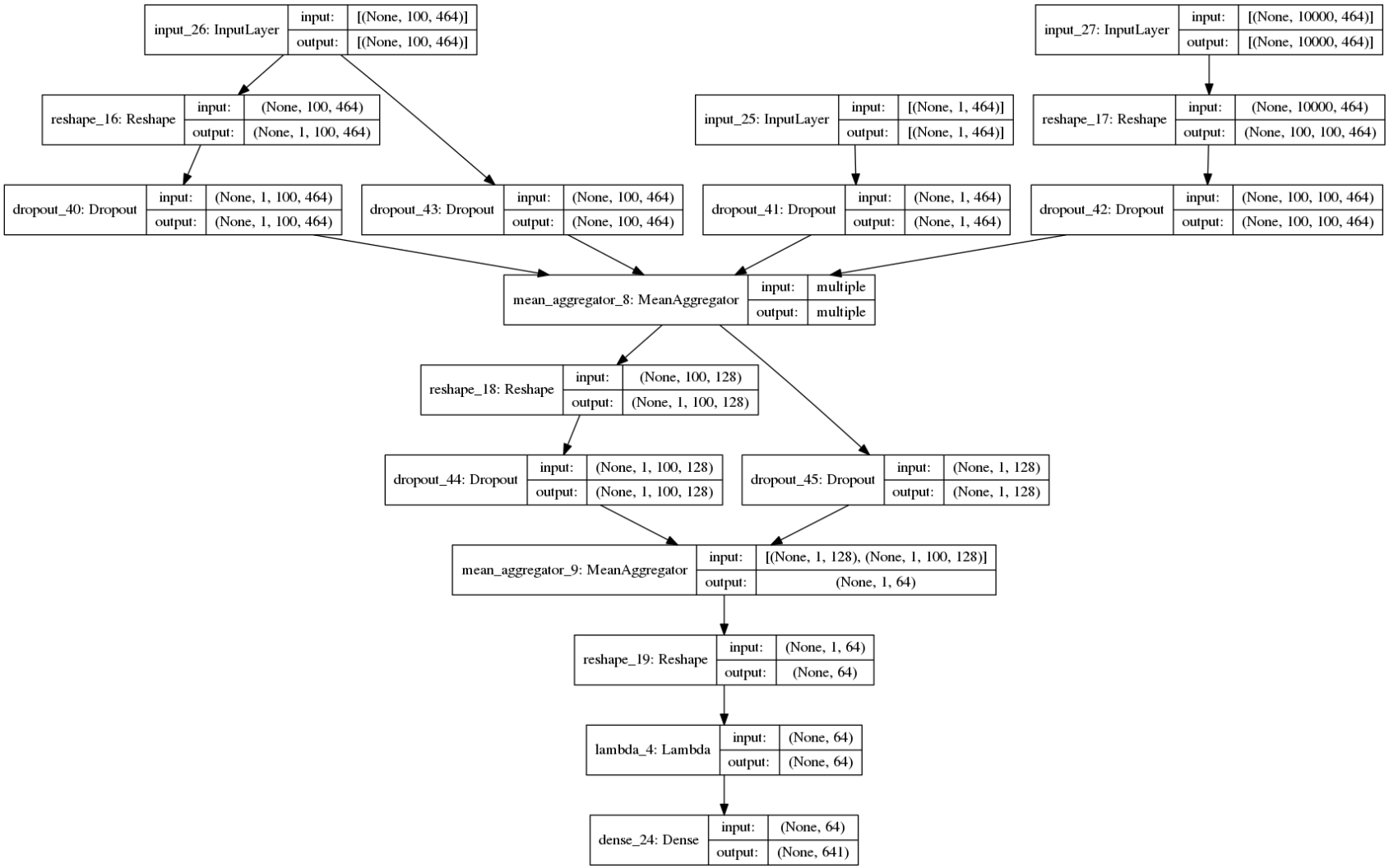


Figure: Graphsage model

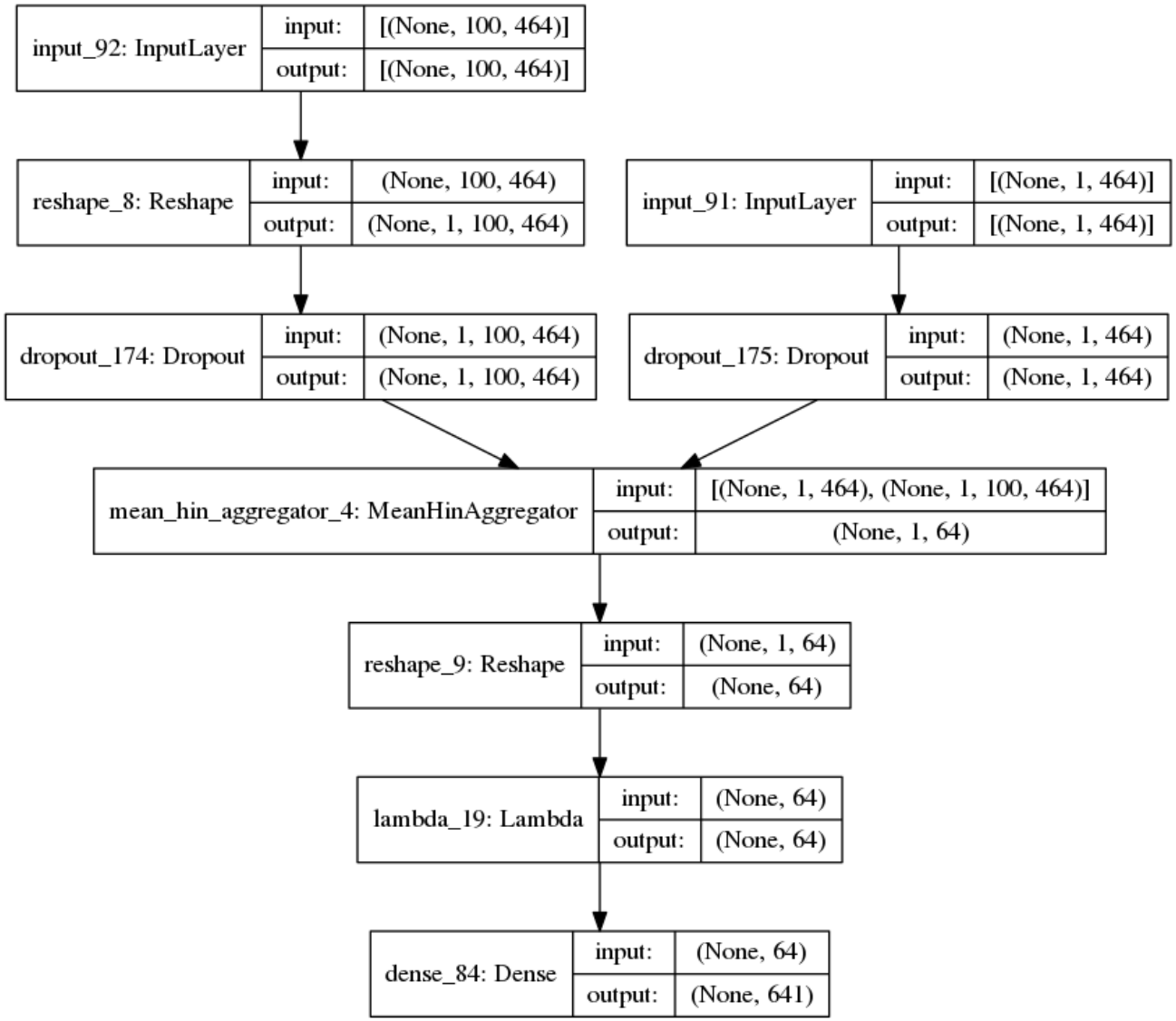


Figure: Hinsage model

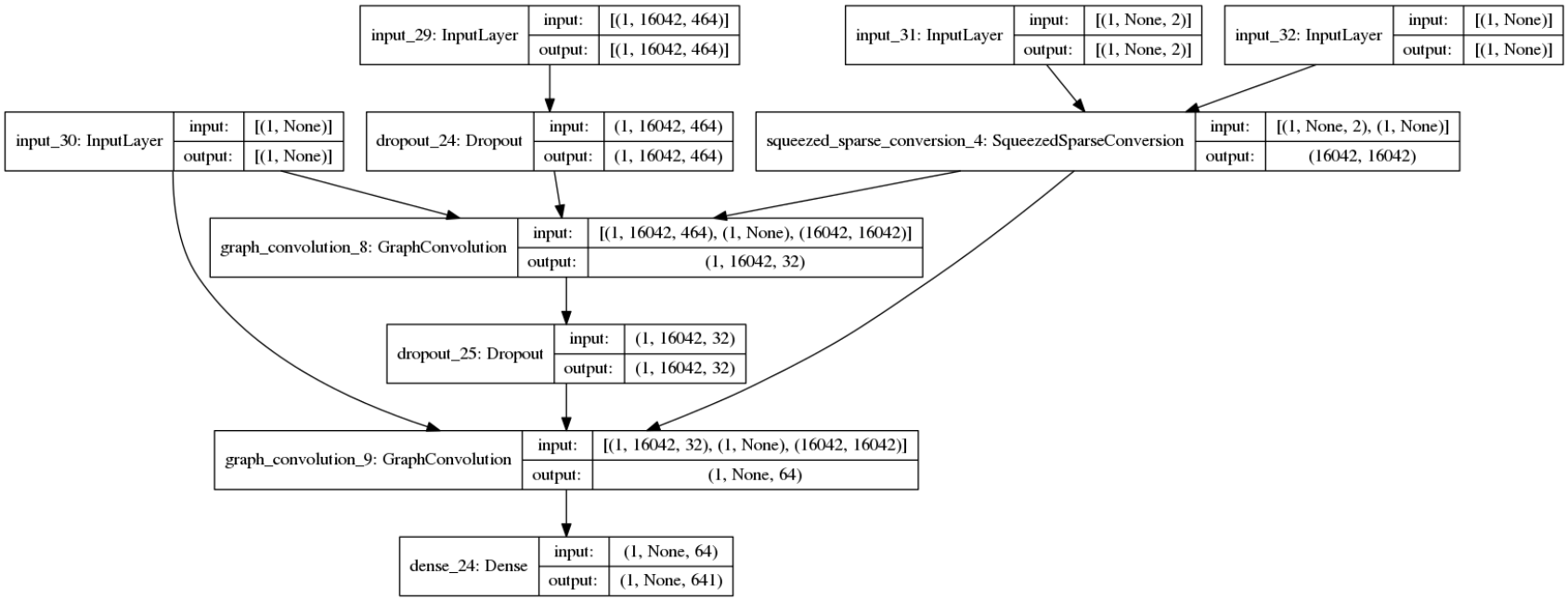


Figure: GCN model

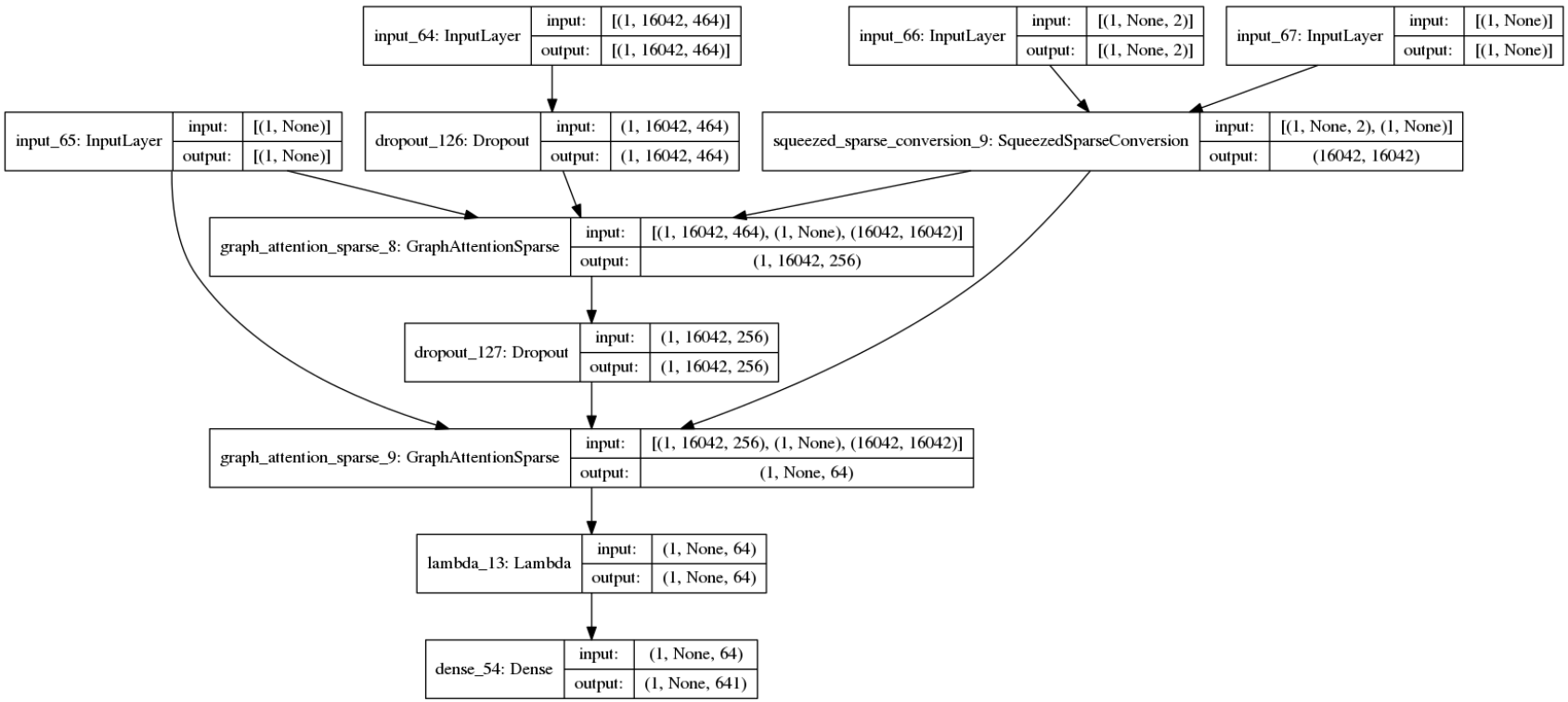


Figure: GAT model

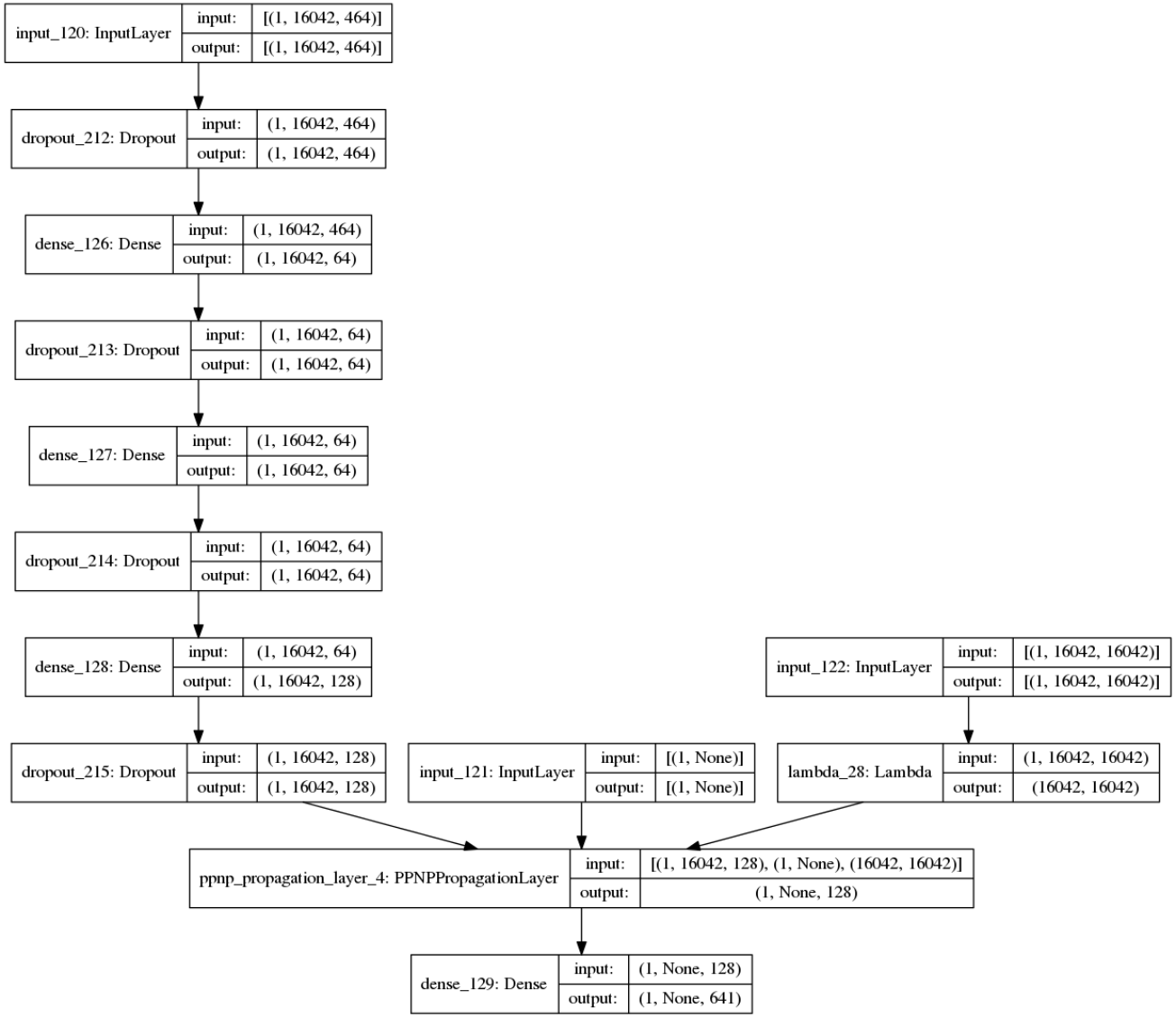


Figure: PPNP model

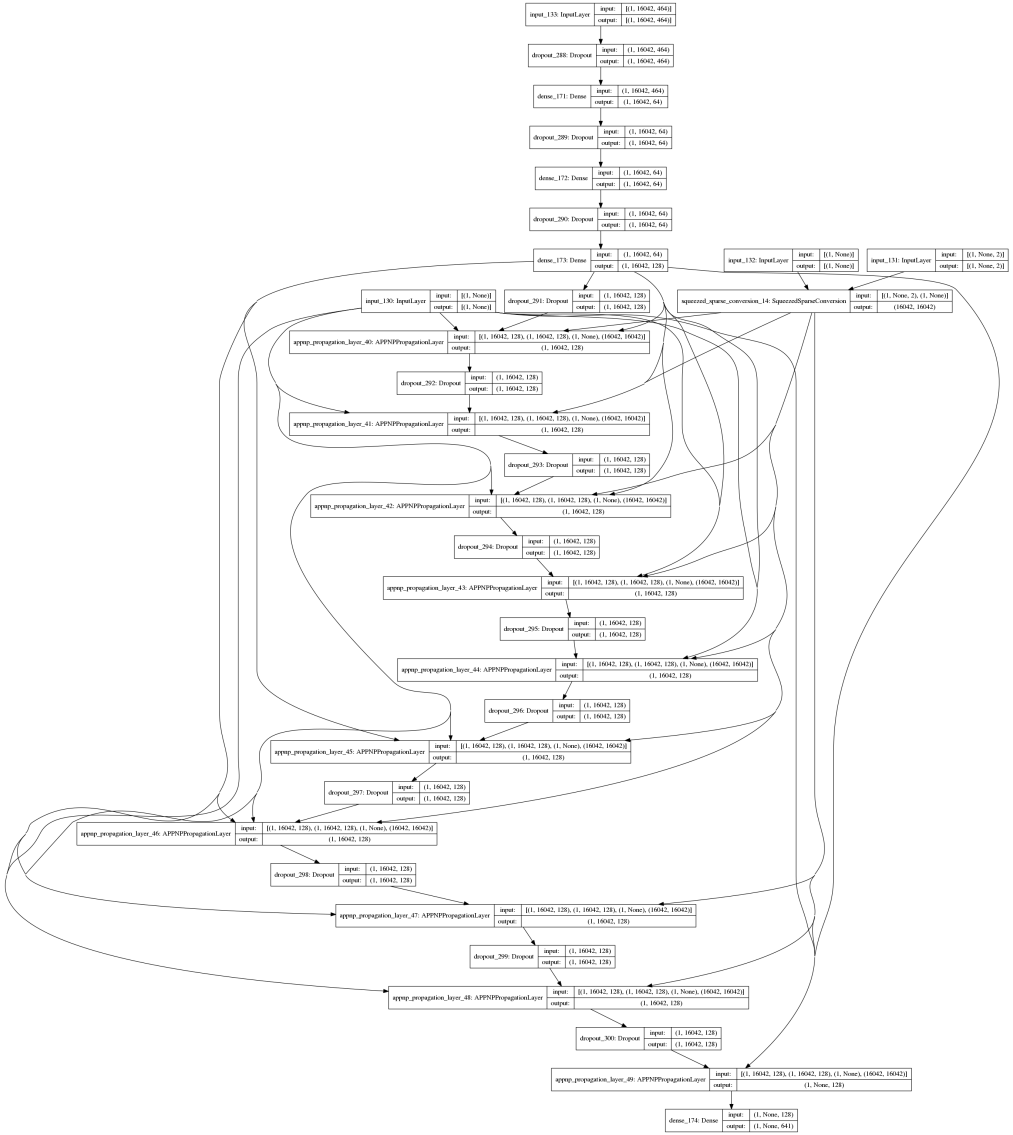


Figure: APPNP model

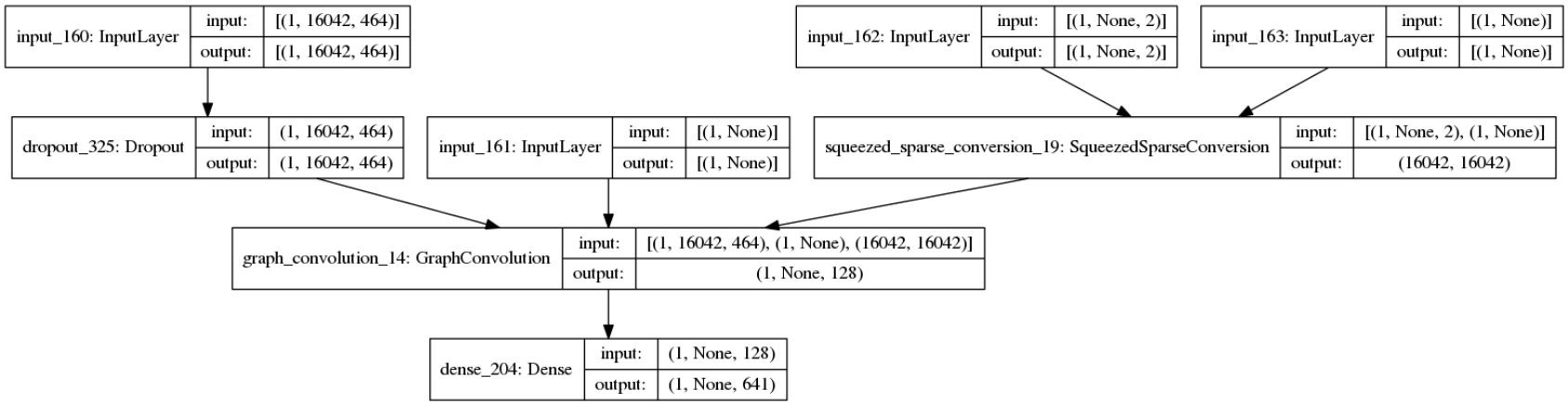


Figure: SGC model

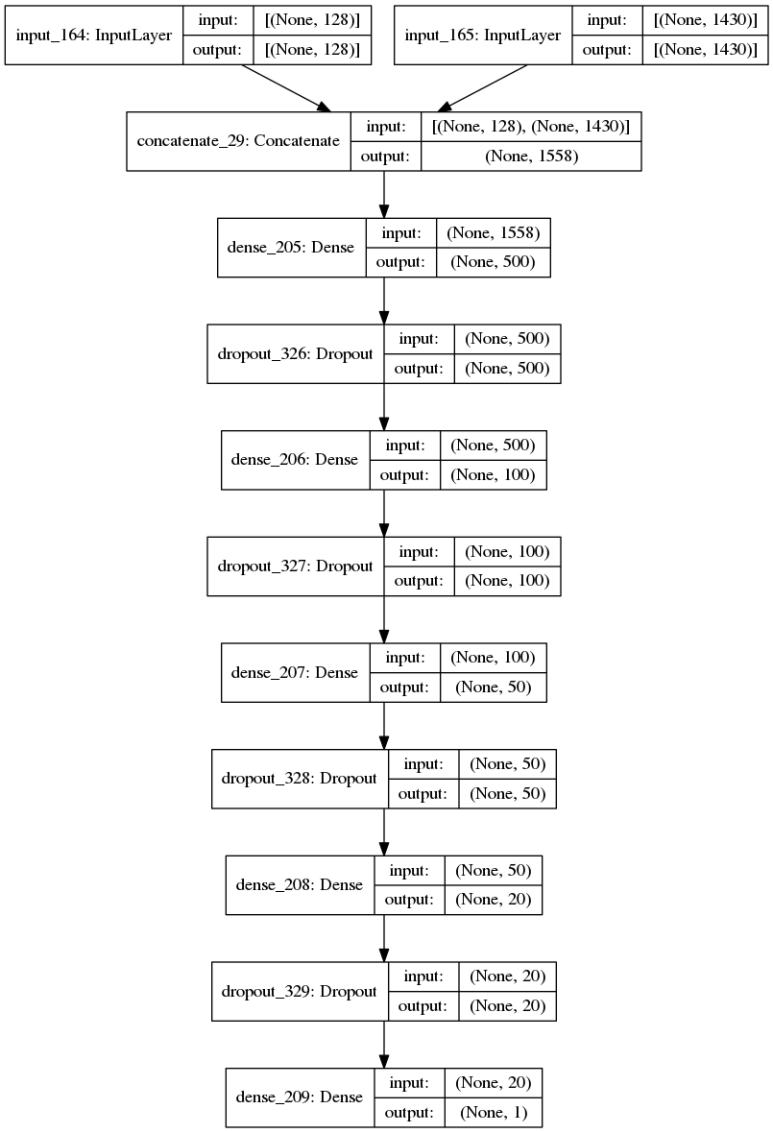


Figure: Overall model