


Multi-label learning for predicting drug resistance co-occurrence of MTB

25 Aug, 2017

Outline


- **Dataset overview**
- **Models**
- **Results and analysis**

Dataset overview



- **Total samples: 13660**
- **Genes: 23**
- **SNPs variable: 5219**
- **Drugs: 11**
- **Source: 26**
- **LinName: 8**

Challenges



- **High dimensionality**
- **Label interaction**
- **Imbalanced classes**
- **Missing labels**

Imbalanced classes and missing labels

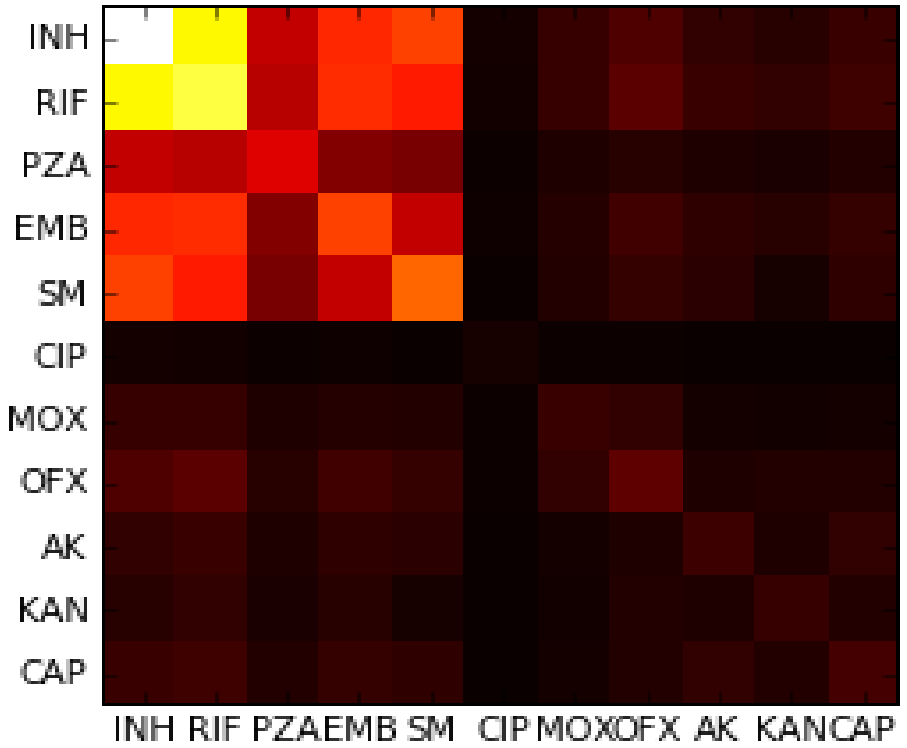

- **More susceptible samples than resistant samples**
- **More samples subject to four first-line**
- **More missing labels for second-line drugs**

Drugs	# Tot tested samples	# Susceptible samples	# Resistant samples	# Samples with missing label
EMB	12603	10933	1670	1057
RIF	12575	9660	2915	1085
INH	11729	8137	3592	1931
PZA	10414	9267	1147	3246
SM	7091	5231	1860	6569
OFX	3076	2618	458	10584
CAP	3056	2741	315	10604
AK	2936	2690	273	10687
KAN	2167	1925	242	11493
MOX	1511	1249	262	12149
CIP	606	529	77	13054

Label interaction



- **Resistance co-occurrence**
  - **irst-line drugs and SM**



Cope with label interaction

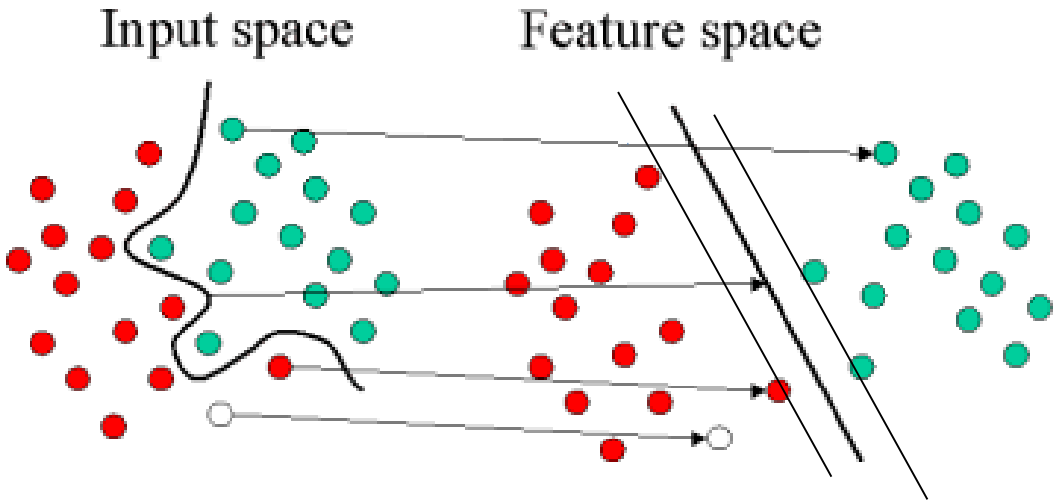


- **Baseline methods- single label learning**
  - –Logistic regression, support vector machine, random forest,etc.
- **Multi-label learning**
- **Multi-output deep neural network**

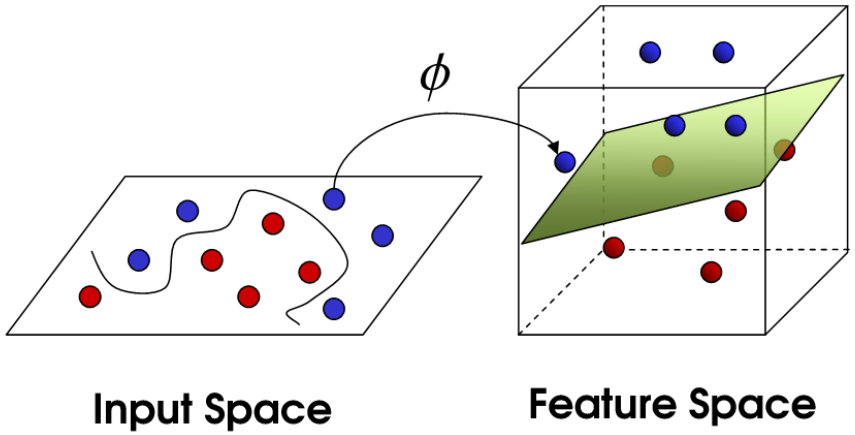
Support vector machine



Linear kernel



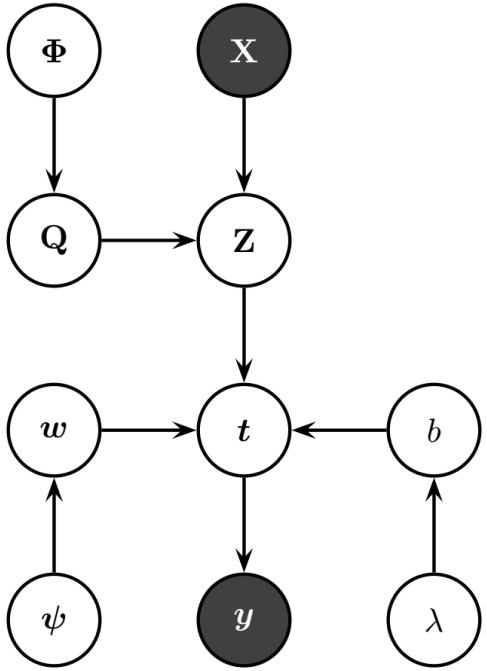
RBF kernel



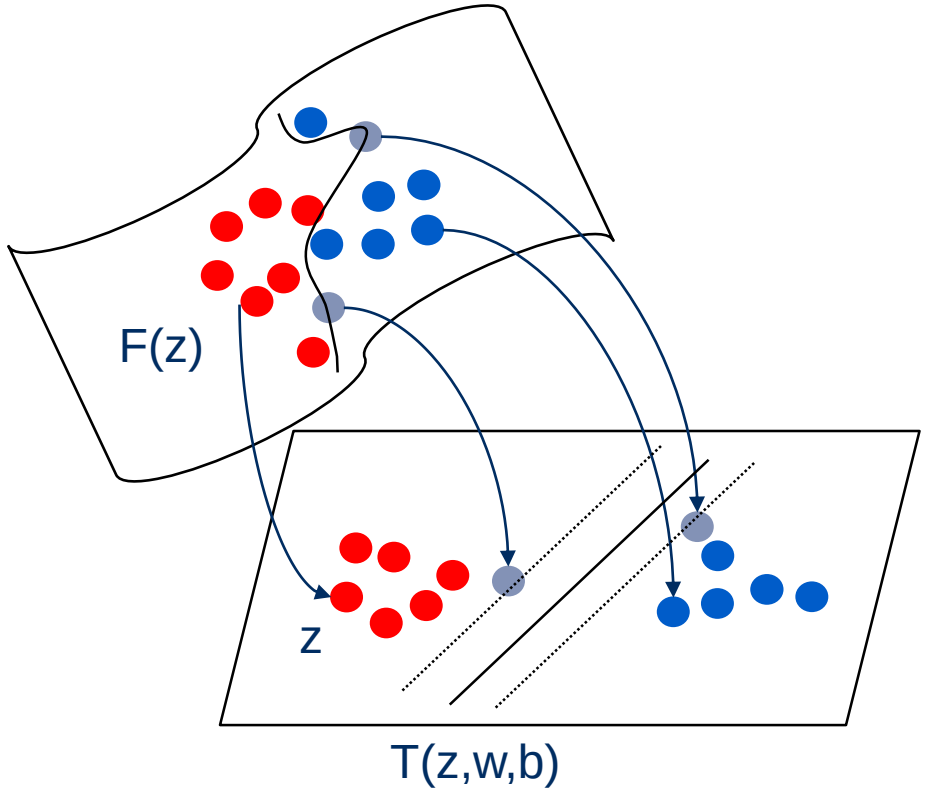
Bayesian supervised multi-label learning



- Coupled dimension reduction and classification

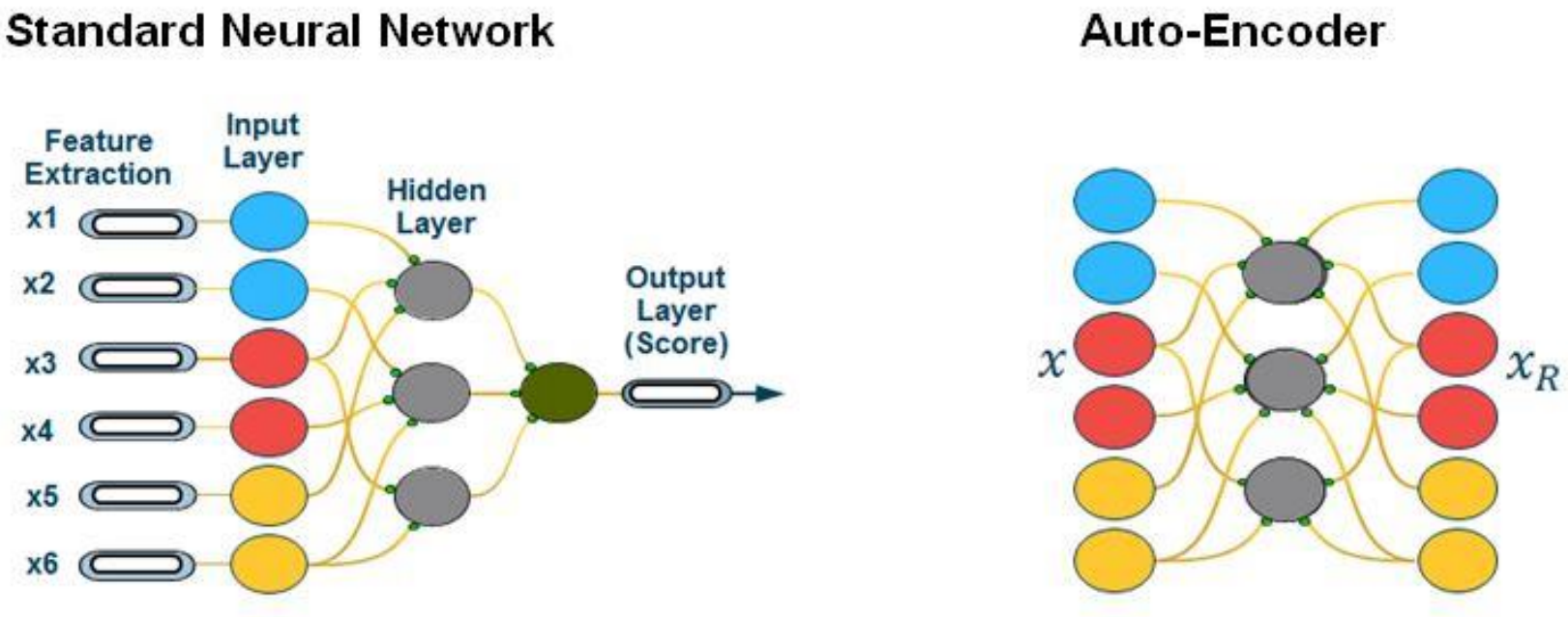


$$
\begin{aligned}
\phi_s^f &\sim \mathcal{G}(\phi_s^f; \alpha_\phi, \beta_\phi) && \forall (f,s) \\
q_s^f | \phi_s^f &\sim \mathcal{N}(q_s^f; 0, (\phi_s^f)^{-1}) && \forall (f,s) \\
z_i^s | \boldsymbol{q}_s, \boldsymbol{x}_i &\sim \mathcal{N}(z_i^s; \boldsymbol{q}_s^\top \boldsymbol{x}_i, 1) && \forall (s,i) \\
\lambda &\sim \mathcal{G}(\lambda; \alpha_\lambda, \beta_\lambda) \\
b | \lambda &\sim \mathcal{N}(b; 0, \lambda^{-1}) \\
\psi_s &\sim \mathcal{G}(\psi_s; \alpha_\psi, \beta_\psi) && \forall s \\
w_s | \psi_s &\sim \mathcal{N}(w_s; 0, \psi_s^{-1}) && \forall s \\
t_i | b, \boldsymbol{w}, \boldsymbol{z}_i &\sim \mathcal{N}(t_i; \boldsymbol{w}^\top \boldsymbol{z}_i + b, 1) && \forall i \\
y_i | t_i &\sim \delta(t_i y_i > 0) && \forall i
\end{aligned}
$$



Auto-encoder

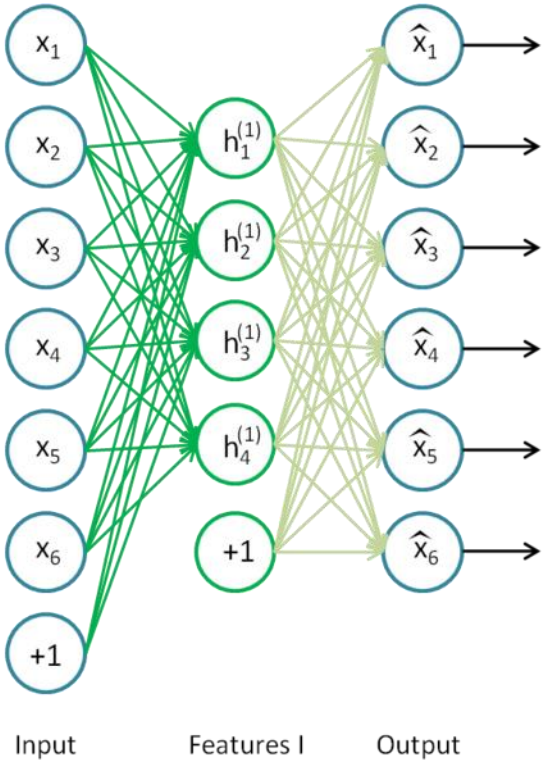




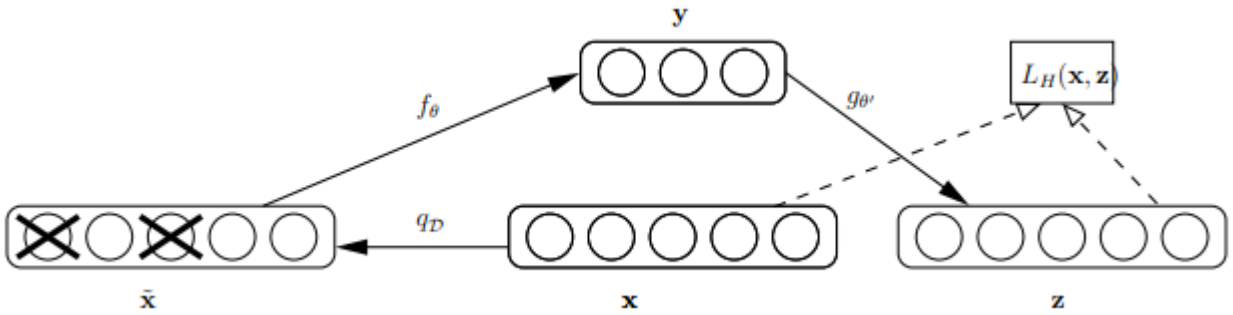
Source: FICO™ Blog. © 2015 Fair Isaac Corporation.

Auto-encoder family

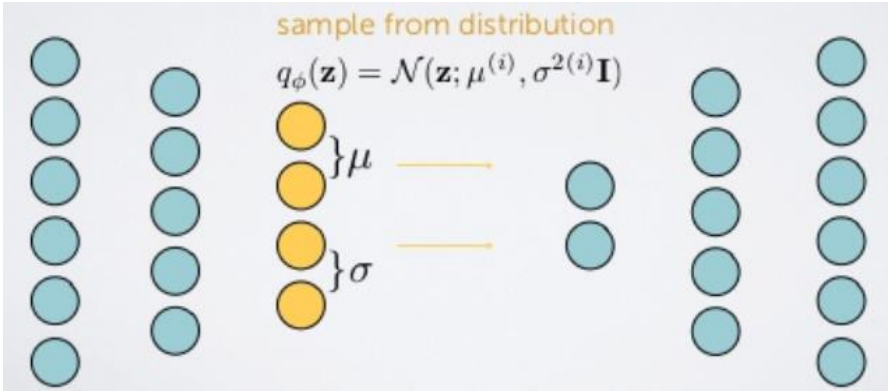

Auto-encoder



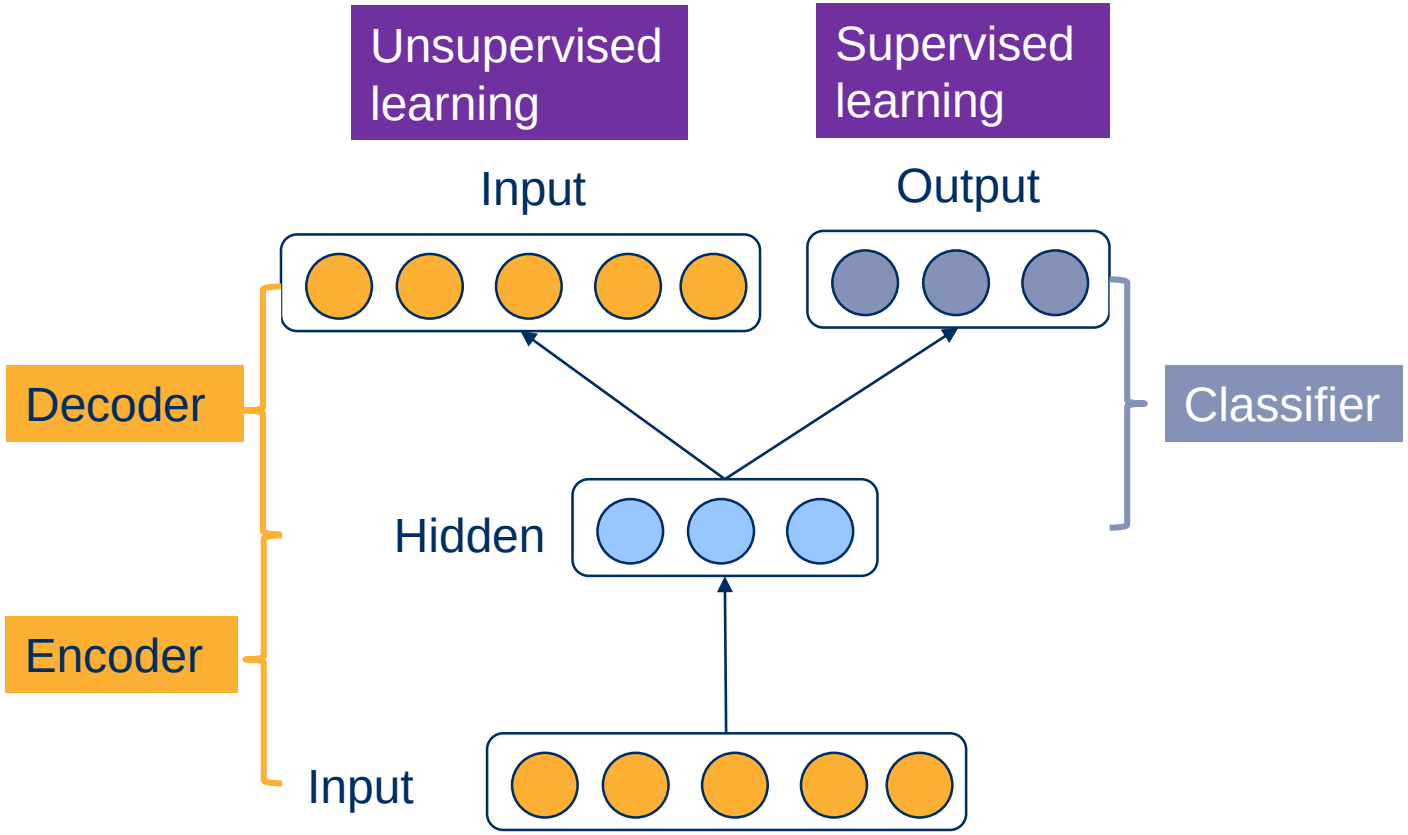
Denoising auto-encoder (DAE)

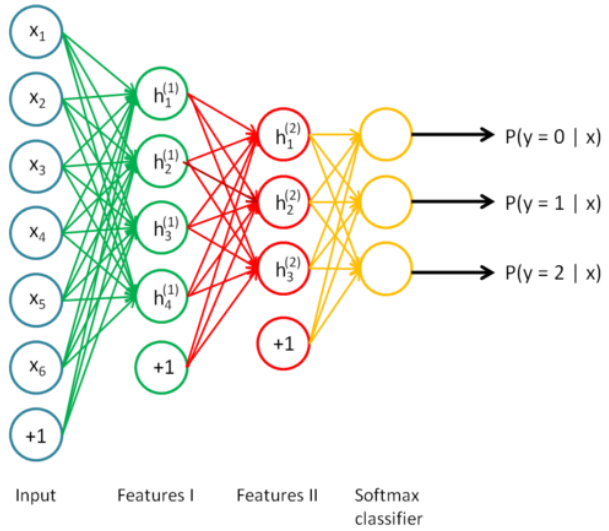


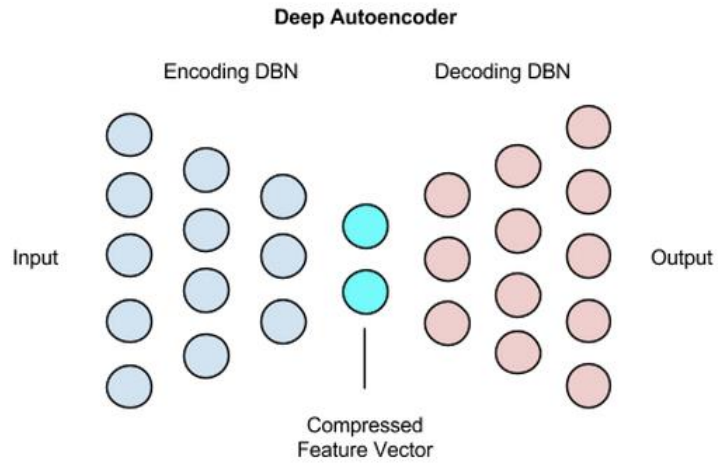
Variational auto-encoder



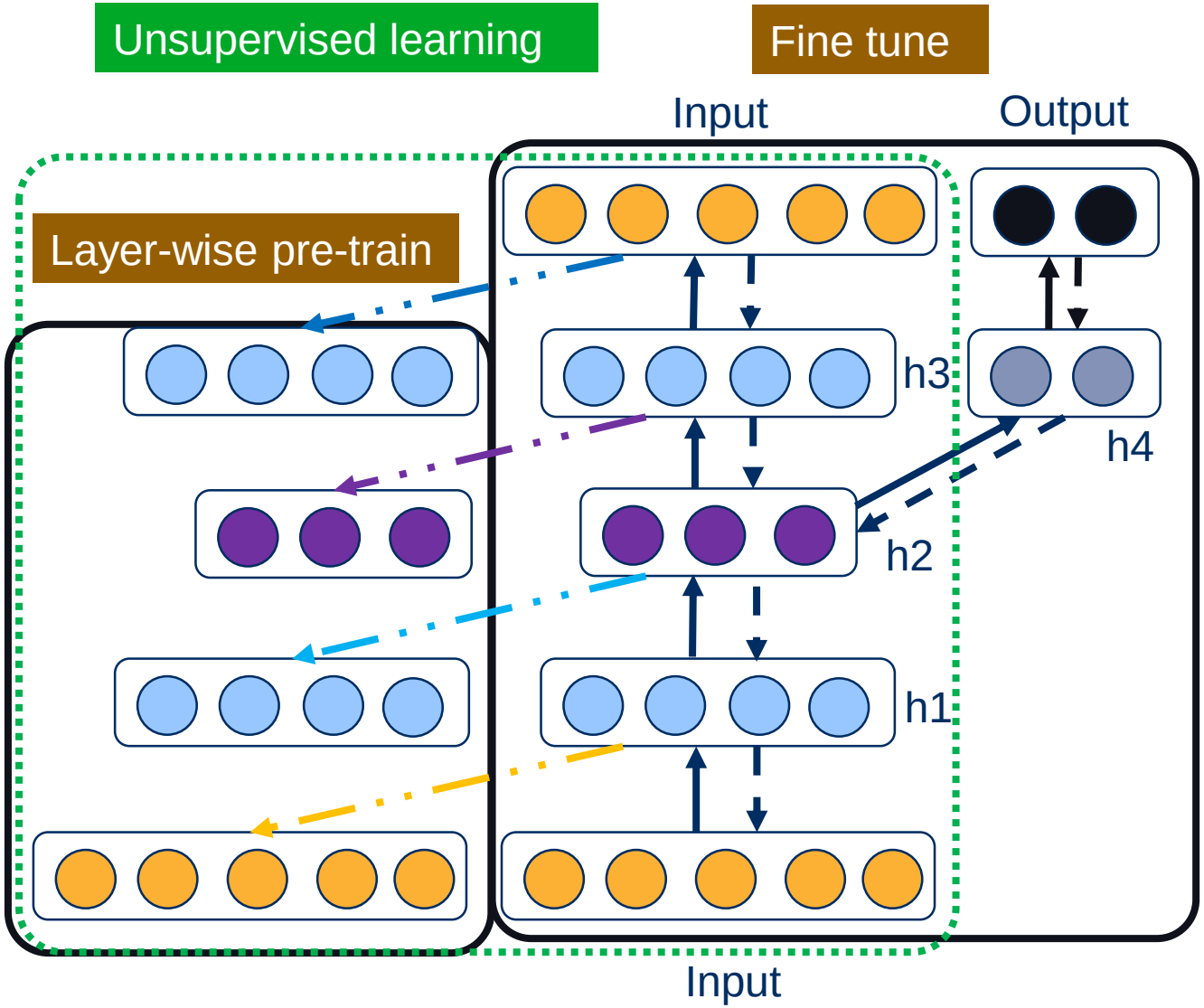
Classification with unsupervised learning





Multi-output deep neural network-train

Cope with imbalanced classes



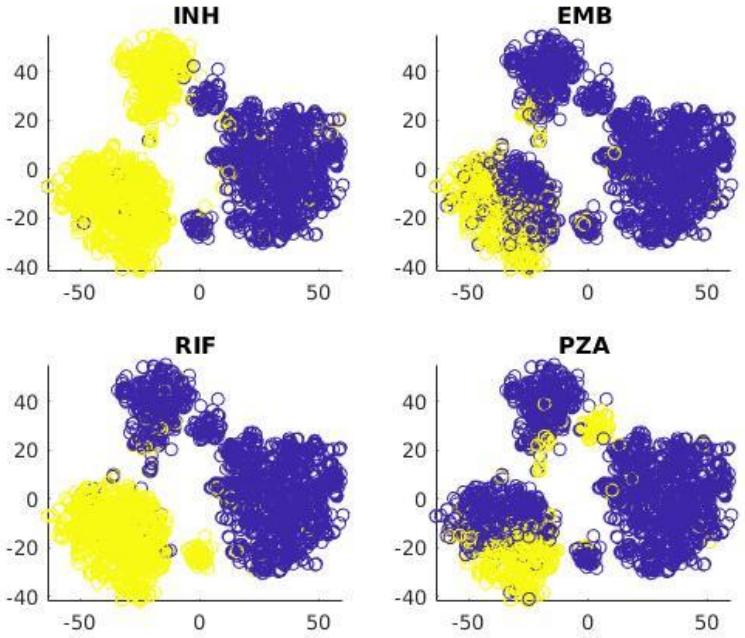
- **Cross-validation**
  - Construct balanced dataset, e.g., down sampling susceptible class
  - **Application of class weight**
  - **Application of iterative stratified cross-validation**

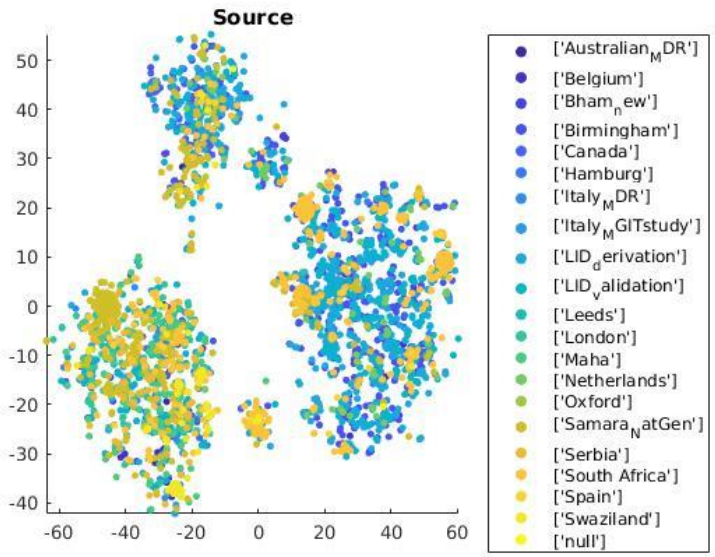
Classification performance

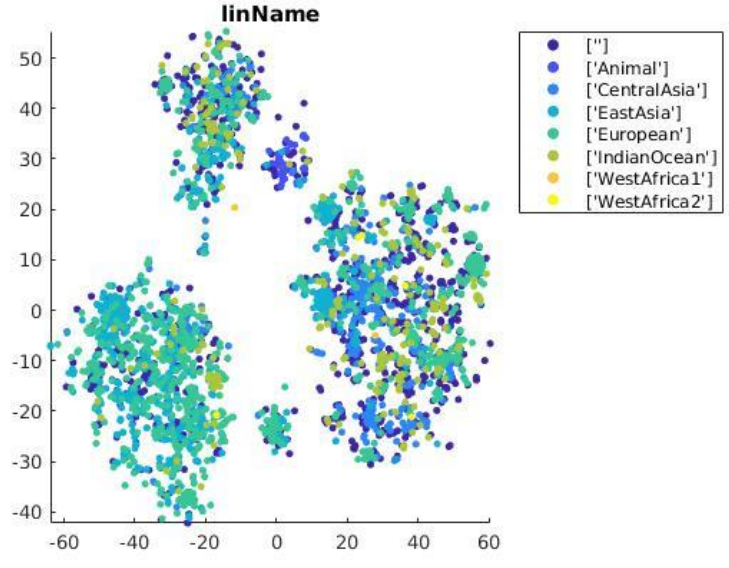

Drug	SVM-linear		SVM-RBF		RF		BSMLL		Deep AE+classification		Stacked DAE encoders+classficaiton	
	Sen (std)	Spec (std)	Sen (std)	Spec (std)	Sen (std)	Spec (std)	Sen (std)	Spec (std)	Sen (std)	Spec (std)	Sen (std)	Spec (std)
INH	94.2 (6.2)	96.9 (1.9)	93.9 (4.3)	97.4 (1.4)	90.3 (11.9)	96.7 (1.8)	91.9 (10.4)	96.5 (3.3)	93.7 (7.6)	90.1 (6.0)	94.7 (6.2)	90.2 (7.3)
EMB	89.3 (7.1)	92.1 (9.3)	79.3 (12.4)	95.4 (5.9)	74.0 (10.5)	95.0 (6.8)	76.8 (14.4)	95.6 (6.7)	86.9 (7.7)	92.4 (8.9)	89.4 (6.5)	92.2 (9.2)
RIF	95.5 (4.7)	95.6 (5.7)	92.6 (5.3)	96.4 (4.3)	89.0 (12.1)	96.5 (5.0)	92.4 (10.7)	98.6 (1.0)	95.6 (3.5)	92.7 (6.6)	95.9 (3.8)	92.5 (7.3)
PZA	75.4 (17.3)	89.6 (15.1)	60.9 (23.3)	91.2 (13.2)	60.7 (19.5)	92.5 (10.6)	66.7 (18.4)	92.2 (16.6)	76.7 (24.1)	88.5 (16.0)	78.4 (23.8)	89.3 (16.3)

Hidden space


- **BSMLL(t-distributed stochastic neighbor embedding-t-SNE)**



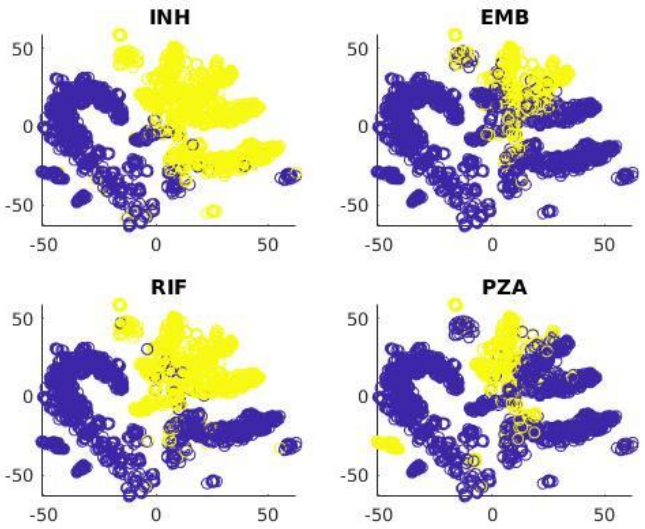


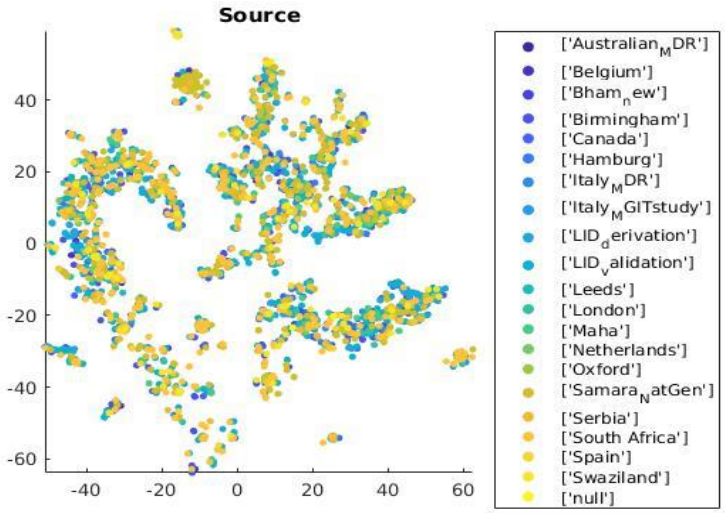


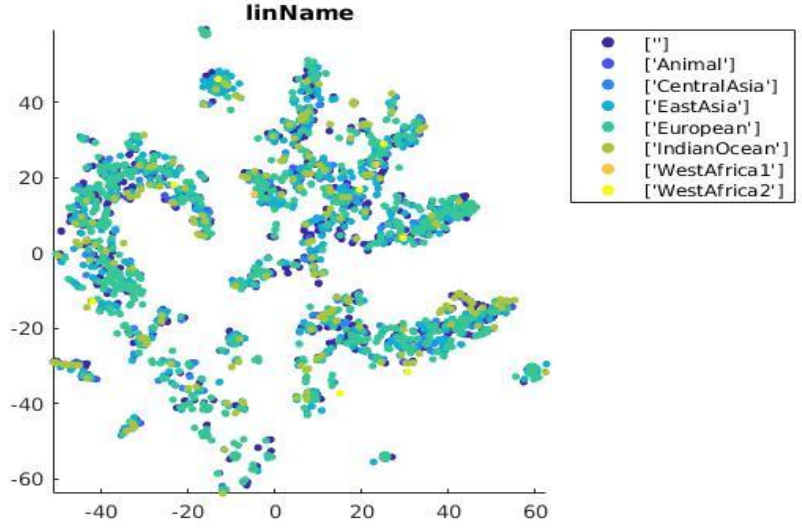
Hidden space



- **DNN(T-SNE)**

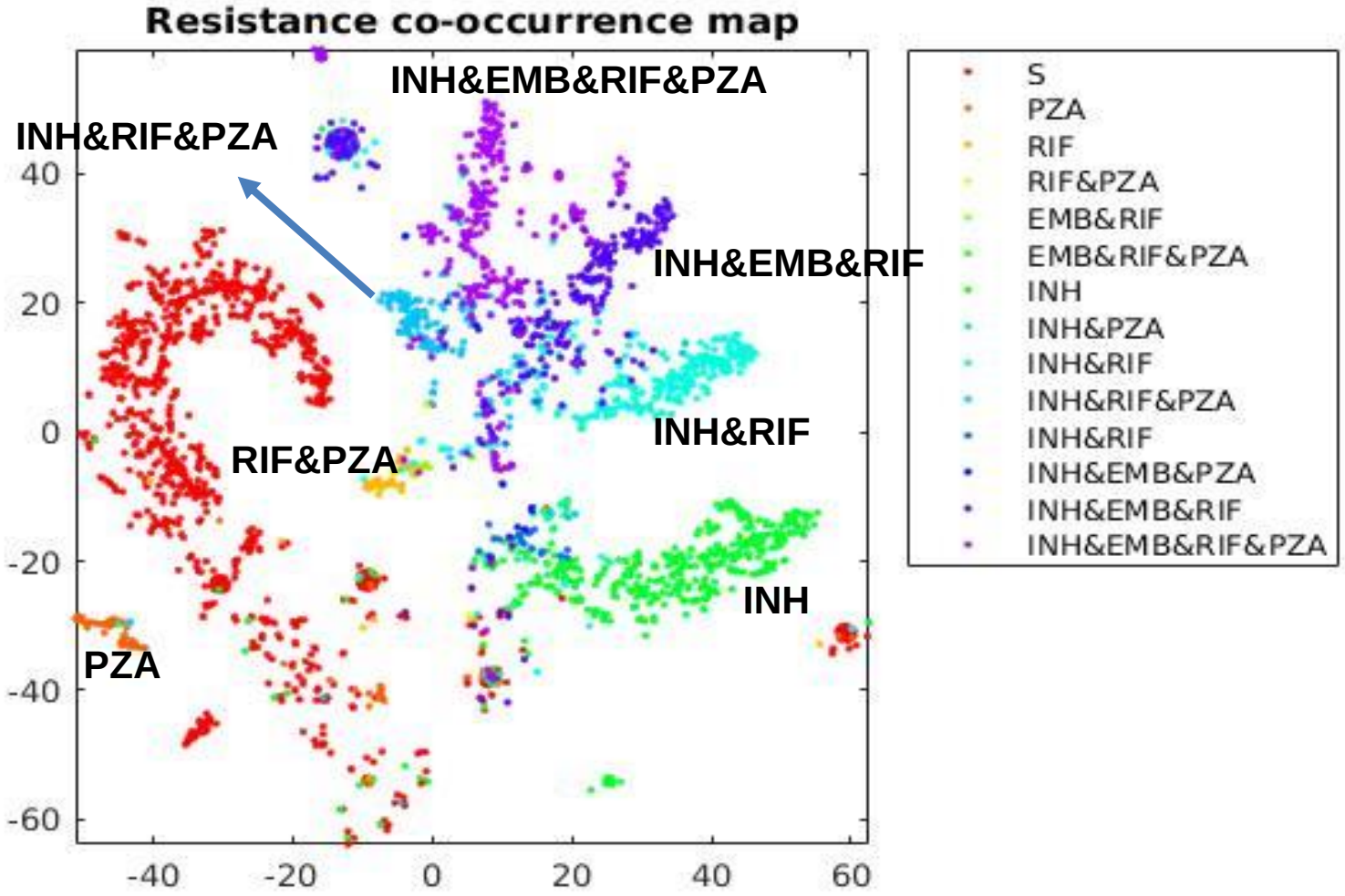






Hidden space


- DNN(T-SNE)



Discussion


- **High variance for EMB and PZA might result from stratified cross-validation or multi-label learning.**
- **Deep neural network captures nonlinear latent structure of genomic data and multiple drug resistance jointly.**
- **Clusters of resistance co-occurrence are well separated in hidden space of proposed deep neural network.**


THANKS