Supplement for Target Predictions using LINCS Data
Yan Xia\textsuperscript{1}, Nick Pabon\textsuperscript{2}, Carlos Camacho\textsuperscript{2}, Ziv Bar-Joseph\textsuperscript{1,*}
\textsuperscript{1}Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15217, USA.
\textsuperscript{2}Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15261, USA.
* Email: zivbj@cs.cmu.edu

1 Supplementary Methods

1.1 Extracting Experiments from LINCS

After determining the subsets of small molecules and cell lines, we obtained the associated experiment identifiers known as “distil ID” from LINCS meta-information. We included only the reproducible distil IDs known as “Gold” IDs.

We then extracted the corresponding signature values from LINCS using the L1000 Analysis Tools (l1ktools)\textsuperscript{[1]} We chose only to extract the signature values of 978 “landmark” genes, because their expression were directly measured, and the values of other genes were imputed from the data of these landmark genes.

**Drug-response experiments** There exists multiple experiments (distil IDs) corresponding to a combination of drug \(d\) and cell line \(c\) (applying drug \(d\) to cell line \(c\)). Denote the \(N_{dc}\) as the number of experiments for the combination \(d,c\). We extracted a matrix of signature values of size \(978 \times N_{dc}\) (number of landmark genes \(\times\) number of experiments) per combination. We next take the median of signature values across different experiments, and obtain a \(987 \times 1\) signature vector per combination. The overall drug-response data \(\Delta\), therefore, is implemented as a MATLAB structure with \(D = 152\) entries, each containing the following fields.

\[
\begin{align*}
\text{name: } & \text{PertID}_d \text{ (string)} \\
\text{cells: } & \text{Cells}_{C_d} \text{ (} |C_d| \times 1 \text{ string array)} \\
\text{signature: } & \Delta_{d..} \text{ (} 978 \times |C_d| \text{)}
\end{align*}
\]

where \(\text{PertID}_d\) is the unique internal identifier of a small molecule in LINCS. \(\Delta_{d..}\) contains the expression values of drug \(d\) across \(C_d\) different cell lines. The \(\text{Cells}_{C_d}\) field contains cell line names corresponding to the column of \(\Delta_{d..}\).

**Gene knockdown experiments** We follow the similar protocol to extract the signature values of gene knockdown experiments. Denote \(N_{gc}\)
as the number of experiments for the combination of gene $g$ and cell line $c$ (knocking down gene $g$ in cell line $c$). Then, for each combination of $g$ and $c$ we extracted signature values of size $978 \times \text{N}_{gc}$. After taking the medians across different experiments, we obtain a $978 \times 1$ vector per combination. The overall gene knockdown data $\Gamma$ has $C = 7$ entries and each entry contains the following fields:

- **name**: $\text{Cells}_c$ (string)
- **genes**: $\text{Symbols}_{G_c} (G_c \times 1 \text{ string array})$
- **signature**: $\Gamma_{c.} (978 \times |G_c|)$

where $\text{Cells}_c$ is the name of a cell indexed by $c$. $\Gamma_{c.}$ contains the signature values of the knockdown of genes in cell line $c$. The $\text{Symbols}_{G_c}$ field is a subset of gene symbols corresponding to the column identifiers of $\Gamma_{c.}$ under the HGNC naming scheme.

**Control experiments** We also extracted the signatures of control experiments. The signature values for each cell line were extracted and we obtained a $978 \times 1$ vector after taking the medians. We denote the overall control experiment data as $\Psi$. $\Psi$ is of size $978 \times C$ and implemented with the following format:

- **name**: $\text{Cells}_c$ (string)
- **control**: $\Psi_{c.} (978 \times 1)$

where $\Psi_{c.}$ is the signature column vector for a cell line $c$.

### 1.2 Extracting and Integrating Features from Different Data Sources

#### 1.2.1 Correlation feature

The correlation feature, denoted as $f_{cor}$ is constructed as follows:

For each drug $d$ in $\Delta$

- Denote $T_d$ as the intersection of gene symbol indices for cells in $C_d$.
  \[ T_d = \bigcap_{c \in C_d} G_c \]

- Obtain the knock-down signature values of $T_d$ from $\Gamma$. Denote this data matrix as $\Gamma_{C_d \cdot T_d}$ which is of size $|C_d| \times 978 \times |T_d|$, where for each cell line in $C_d$ there is a signature matrix of size $978 \times |T_d|$.

- Compute the Pearson’s correlation between $\Delta_{d.} \cdot (978 \times |C_d|)$ and $\Gamma_{C_d \cdot T_d} (|C_d| \times 978 \times |T_d|)$. Specifically, for each cell line $c \in C_d$, we compute the correlation between $\Delta_{d.}$ and $\Gamma_{c \cdot T_d}$, and obtain a correlation vector of size $|T_d|$. This is the correlation between the responses of the cells to the drug
treatment and their response to the gene KD. Each entry in this vector is
the correlation of 978 landmark genes of the drug \( d \) in one cell line \((\Delta_{d,c})\)
and a knockdown of gene \( g \) in the same cell line \((\Gamma_{c,g})\). In other words, if
we collect these correlation vectors for all cell lines in \( C_d \) and denote the
overall correlation feature as \( f_{\text{cor}} \) has the following definition:

\[
f_{\text{cor}}(d, g, c) = \text{corr} (\Delta_{d,c}, \Gamma_{c,g}) \quad \forall g \in T_d
\]

The correlation feature for one drug \( d \), \( i.e. f_{\text{cor}}(d, \cdot, \cdot) \), has
a dimension of \( |T_d| \times |C_d| \).

**Cell selection feature**

The cell selection feature, denoted as \( f_{\text{CS}} \), is computed as follows.

- For each drug \( d \) in \( \Delta \) (\( \Delta_\cdot \)):
  - For each cell line \( c \) in \( C_d \):
    - Compute the correlation between \( \Delta_{d,c} \) and \( \Psi_c \)

\[
f_{\text{CS}}(d, c) = \text{corr} (\Delta_{d,c}, \Psi_c)
\]

In other words, \( f_{\text{CS}}(d, \cdot) \) produces a \( |C_d| \times 1 \) vector, each entry corresponds
to the correlation between the drug-response and control experiments for one
cell line in \( C_d \). This feature is used to determine the relevance of the drug to
the cell type being studied.

**1.2.2 PPI correlation score**

The “PPI correlation Score”, denoted as \( f_{\text{PC}} \) is constructed as follows:

For each drug \( d \), we first obtain \( T_d \), the intersection of gene symbol indices,
as before. Then for each cell line \( c \) in \( C_d \), we sort \( T_d \) in descending order using
the correlation values \( f_{\text{cor}}(d, \cdot, c) \), and we denote the sorted gene symbol indices
for cell line \( c \) as \( \sigma_c(T_d) \).

We then construct the PPI correlation feature \( f_{\text{PC}} \) as follows:

- For each knockdown gene \( g \) in \( T_d \): Obtain the set of neighbor gene symbol
  indices from PPI adjacency list, and denote it as \( B_g \).

\[
f_{\text{PC}}(d, g, c) = \frac{|B_g \cap \sigma_c(T_d)|_{1:100}}{|B_g \cap \sigma_c(T_d)| + 50}
\]

- \( f_{\text{PC}} \) feature is of the same dimension as \( f_{\text{cor}} \), which is \( |T_d| \times |C_d| \).

In other words, \( f_{\text{PC}}(d, g, c) \) reflects the fraction of gene \( g \)’s binding partners
that is more correlated with drug \( d \) in the context of cell line \( c \). We use 50 as the
pseudo-count to penalize hub proteins which have substantially more neighbors
than others.
PPI expression score

We compute two types of PPI expression scores, denoted as $f_{PE_{\text{max}}}$ and $f_{PE_{\text{avg}}}$, as follows:

- For a drug $d$:
  - For each knockdown gene $g$ in $T_d$:
    - Obtain $N_g$ as above.
    * For each cell line $c$, find the set of signature values for the neighbors: $\Delta_{d,N_g,c}$ (size: $|N_g| \times 1$). Then, the two PPI expression scores computed as
      
      
      \[
      f_{PE_{\text{max}}}(d,g,c) = \max(\Delta_{d,N_g,c})
      \]
      
      \[
      f_{PE_{\text{avg}}}(d,g,c) = \text{avg}(\Delta_{d,N_g,c})
      \]

1.2.3 Feature data structure

We combined the features for all drugs in a MATLAB structure $\Omega$. $\Omega$ has $D$ entries, and each entry $\Omega^{(d)}$ has the following fields:

- Name: PertID$_d$ (string)
- Targets: $P_d$ (targets for $d$)
- Cells: $C_d$ (|$C_d| \times 1$ string array)
- Genes: $T_d$ (common genes across $G_c$)
- Correlation: $f_{\text{cor}}(d,\cdot,\cdot)$ (|$T_d| \times |C_d|$)
- PPI Correlation: $f_{PC}(d,\cdot,\cdot)$ (|$T_d| \times |C_d|$)
- Max PPI Expression: $f_{PE_{\text{max}}}(d,\cdot,\cdot)$ (|$T_d| \times |C_d|$)
- Avg PPI Expression: $f_{PE_{\text{avg}}}(d,\cdot,\cdot)$ (|$T_d| \times |C_d|$)
- cell selection: $f_{CS}(d,\cdot)$ (|$C_d| \times 1$)

There are a total of $D = 152$ drugs in $\Omega$, and the number of drugs with different $|C_d|$ are summarized in Table ??.

1.3 Subcellular Localization Assignment

We obtained the cellular localization of genes from the Gene Ontology Consortium. The GO database provides web services to query genes in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner\footnote{http://geneontology.org/page/go-enrichment-analysis}. We further assign the locations as either “intracellular” (inside of cell) and “extracellular” (outside of cell). The detailed assignments are shown in Table ??.
<table>
<thead>
<tr>
<th>Localization</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Membrane</td>
<td>External</td>
</tr>
<tr>
<td>Endosome</td>
<td>Internal</td>
</tr>
<tr>
<td>Secreted</td>
<td>External</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>Internal</td>
</tr>
<tr>
<td>Nucleus</td>
<td>External</td>
</tr>
<tr>
<td>Chromosomes</td>
<td>Internal</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>Internal</td>
</tr>
<tr>
<td>ER</td>
<td>Internal</td>
</tr>
<tr>
<td>Lysosome</td>
<td>Internal</td>
</tr>
<tr>
<td>Golgi</td>
<td>Internal</td>
</tr>
<tr>
<td>Peroxisome</td>
<td>Internal</td>
</tr>
<tr>
<td>Ribosome</td>
<td>Internal</td>
</tr>
<tr>
<td>Microsome</td>
<td>Internal</td>
</tr>
<tr>
<td>Endomembrane</td>
<td>Internal</td>
</tr>
<tr>
<td>Cytoskeleton</td>
<td>Internal</td>
</tr>
<tr>
<td>Centrosome</td>
<td>Internal</td>
</tr>
<tr>
<td>Vesicle</td>
<td>External</td>
</tr>
<tr>
<td>Vacuole</td>
<td>External</td>
</tr>
<tr>
<td>Membrane</td>
<td>External</td>
</tr>
<tr>
<td>Cell</td>
<td>Internal</td>
</tr>
</tbody>
</table>

Table 1: Possible cellular localizations retrieved from GO and their assignment.

1.4 Classification Procedure

Criterion of successful classification

Due to the intrinsic noise from the data, we define a successful classification for a drug if any of its correct targets is enriched into the top $K$ ranked genes, where $K$ can be either 50 or 100.

1.4.1 Single Feature

The evaluation of single features was performed using the drugs that have been applied on all 7 cell lines. There are 29 of these drugs from $\Omega$. We sort the common genes $T_d$ descendingly for a drug $d$ and a cell line $c$ using an individual feature $f(d,\cdot,c)$, where $f$ is either $f_{cor}$ or $f_{PC}$. Denote $\sigma_d(g,c)$ as the ranking of a gene $g \in T_d$ in the context of cell line $c$. Then, we define the overall ranking of a gene $\sigma_d(g)$ to be the best ranking across all seven cell lines, i.e., $\min(\sigma_d(g,c))$ for $c \in C_d$.

1.4.2 Constructing training dataset

Next, we wish to learn and evaluate classifiers that predict drug target using all features from the feature dataset $\Omega$. Therefore, we first construct training data (design matrix $X$ and its associated labels $y$) from the feature dataset $\Omega$. 

5
For each drug $d$ in $\Omega$, we select the rows corresponding to the targets in $P_d$ from the other feature matrices and concatenate them into a row vector. The same cell selection vector is appended to every row of targets. These rows are assigned with a positive label 1. We then randomly sampled 100 non-target genes (denoted as $\nu_d$) and construct the row vectors the same way as the target genes, and these rows are assigned with a negative label 0. In other words, the training matrix and label vector constructed from a drug $d$ are of the following format.

\[
\begin{bmatrix}
  f_{cor}(d, P_{d1}, \cdot) & f_{PC}(d, P_{d1}, \cdot) & f_{PE_{max}}(d, P_{d1}, \cdot) & f_{PE_{avg}}(d, P_{d1}, \cdot) & f_{CS}(d, \cdot) \\
  f_{cor}(d, P_{d2}, \cdot) & f_{PC}(d, P_{d2}, \cdot) & f_{PE_{max}}(d, P_{d2}, \cdot) & f_{PE_{avg}}(d, P_{d2}, \cdot) & f_{CS}(d, \cdot) \\
    \vdots & \vdots & \vdots & \vdots & \vdots \\
  f_{cor}(d, \nu_{d1}, \cdot) & f_{PC}(d, \nu_{d1}, \cdot) & f_{PE_{max}}(d, \nu_{d1}, \cdot) & f_{PE_{avg}}(d, \nu_{d1}, \cdot) & f_{CS}(d, \cdot) \\
  f_{cor}(d, \nu_{d2}, \cdot) & f_{PC}(d, \nu_{d2}, \cdot) & f_{PE_{max}}(d, \nu_{d2}, \cdot) & f_{PE_{avg}}(d, \nu_{d2}, \cdot) & f_{CS}(d, \cdot) \\
    \vdots & \vdots & \vdots & \vdots & \vdots \\
  f_{cor}(d, \nu_{d100}, \cdot) & f_{PC}(d, \nu_{d100}, \cdot) & f_{PE_{max}}(d, \nu_{d100}, \cdot) & f_{PE_{avg}}(d, \nu_{d100}, \cdot) & f_{CS}(d, \cdot)
\end{bmatrix}
\begin{bmatrix}
  1 \\
  1 \\
  \vdots \\
  0 \\
  0 \\
  \vdots \\
  0
\end{bmatrix}
\]

where $m = |P_d|$, which is the total number of targets for drug $d$. Therefore, the training matrix $X_d$ for drug $d$ is of size $(m + 100) \times 5 |C_d|$, and label vector $y_d$ has length $m + 100$.

1.4.3 Logistic Regression and Random Forest

We used MATLAB’s `lassoglm` package to train a lasso-regularized logistic regression model. We applied MATLAB’s `TreeBagger` package for random forest based models. Leave-One-Out Cross Validation (LOOCV) was performed to evaluate the performance of both methods.

1.5 Structural Screening Protocol