Key Findings Summary

ECMarker is a neural network-based model integrating semi- and discriminative-restricted Boltzmann machines (SRBM/DRBM) to predict disease phenotypes (e.g., early vs. late cancer stages) from gene expression data. It identifies biomarkers, reveals gene regulatory networks (GRNs), and links them to clinical outcomes like survival rates. Applied to non-small-cell lung cancer (NSCLC), ECMarker achieved 74% accuracy in stage prediction and identified 14 early-stage biomarkers significantly associated with survival ($p < 0.005$)

ECMarker Model Review

Assumptions

1. **Gene Regulatory Networks (GRNs) Drive Phenotypes:** ECMarker assumes that disease phenotypes (e.g., cancer stages) arise from disruptions in gene regulatory networks (GRNs), where transcription factors (TFs) and non-coding RNAs interact dynamically.

2. **Non-Linear Gene Interactions:** Unlike correlation-based models, ECMarker assumes cancer development involves non-linear gene interactions, which neural networks can capture.

3. **Implicit Feature Selection:** The model assumes that prioritizing genes via integrated gradients (without prior feature selection) retains system-level biological insights.

4. **Clinical Interpretability:** ECMarker presumes that biomarker genes linked to pathways can predict clinical outcomes like survival rates.

Methods

1. **Neural Network Architecture:**

    a. Integrates Semi-Restricted Boltzmann Machine (SRBM) for modeling gene-gene lateral connections (matrix L) and Discriminative RBM (DRBM) for phenotype classification.

b. Three layers:

i. Input (genes): Continuous gene expression values.
ii. Hidden: Binary units detecting patterns.
iii. Output: Phenotypes (e.g., early/late cancer).

2. **Training:**

a. Uses stochastic gradient descent (SGD) with L1 regularization to handle high-dimensional data.

b. Energy function:

$$E(v, y, h) = -\mathbf{h}^T\mathbf{W}\mathbf{v} - \mathbf{a}^T\mathbf{v} - \mathbf{b}^T\mathbf{h} - \mathbf{c}^T\mathbf{y} - \mathbf{h}^T\mathbf{U}\mathbf{y} - \mathbf{v}^T\mathbf{L}\mathbf{v}$$

c. Gibbs sampling generates synthetic data for training.

3. **Biomarker Prioritization:** Integrated gradients calculate gene importance scores based on prediction sensitivity.

4. **Pathway Analysis:** Gene Set Enrichment Analysis (GSEA) links biomarkers to pathways (e.g., HIPPO signaling) using MSigDB.

Key Features

1. **Interpretability:**

a. Lateral connections (L matrix) reveal gene networks.
b. Top biomarkers (e.g., KLF2 for apoptosis) are functionally enriched.

2. **Scalability:** Processes 10,102 genes without prior feature selection.

3. **Clinical Relevance:** Predicts survival rates (e.g., 14 early-stage biomarkers stratify patients with $P<0.005$).

4. **Drug Discovery:** Identifies drugs targeting biomarkers (e.g., Gefitinib for KRT13) via drug-gene z-scores.

Use Cases

1. **Cancer Stage Prediction:** Achieves 74% accuracy in NSCLC stage classification (vs. 50% for eRBMs).

2. **Biomarker Identification:** Prioritizes genes like CRTAP (early stage) and SETD2 (late stage).

3. **Mechanistic Insights:** Links WAC to Vorinostat (HDAC inhibitor) and TROAP to Trametinib (MEK inhibitor).

4. **Drug Repurposing:** Predicts FDA-approved drugs (e.g., Gefitinib) for early-stage biomarkers.

1. What Makes ECMarker Explainable?

ECMarker's explainability stems from three features:

1. **Lateral gene connections:** The model's SRBM component allows connections between input genes (visible layer), forming a gene network represented by matrix L. These connections highlight gene-gene interactions critical for phenotype prediction.

2. **Integrated gradients:** Gene importance scores are calculated using gradients of the model's predictions relative to input genes. Higher scores indicate stronger contributions to predicting a phenotype (e.g., early cancer).

3. **Functional enrichment:** Biomarker genes are analyzed via Gene Set Enrichment Analysis (GSEA) to identify pathways (e.g., "HIPPO signaling") and drug targets (e.g., Gefitinib for early-stage gene KRT13).

2. Scope of the Model

ECMarker is designed for:

- **Early-stage biomarker discovery:** Identifies genes like CRTAP (early NSCLC) and SETD2 (late NSCLC) with stage-specific expression patterns.

- **Mechanistic insights:** Links biomarkers to pathways (e.g., early-stage KLF2 to apoptosis regulation) and drug responses (e.g., WAC targeted by Vorinostat).

- **Clinical applications:** Predicts survival rates and suggests repurposed drugs (e.g., Trametinib for late-stage TROAP).

3. Basic Pipeline

1. **Input:** Gene expression data (e.g., 10,102 genes from 766 early/late NSCLC patients).

2. **Model Training:**

    a. **SRBM:** Models gene-gene interactions via lateral connections (L matrix).
    b. **DRBM:** Classifies phenotypes (early/late) using hidden layers.

3. **Biomarker prioritization:** Genes ranked by importance scores (e.g., HPSE score = 2.08 in early stage).

4. **Pathway/drug analysis:** GSEA and drug-gene z-scores (e.g., KRT13 targeted by Gefitinib).

4. Identifying Reaction Pathways

- **Gene network construction:** The L matrix from SRBM serves as an adjacency matrix to build gene networks.

- **Enrichment analysis:** Top-ranked genes (e.g., early-stage KLF2) are analyzed via GSEA against MSigDB pathways. For example, KLF2 enrichment in "negative regulation of cell proliferation".

- **TF-target inference:** Network modules are linked to transcription factors (TFs) like SOX2 using regulatory databases.

Numerical Analysis of Datasets

- **Training data:** 1,103 NSCLC patients (Gentles2015), balanced to 766 early (I/IA/IB) and 766 late (II/III/IV) cases.

- **Validation:**
    - **TCGA-LUAD:** 741 patients (409 early, 332 late).
    - **TCGA-LUSC:** 758 patients (380 early, 378 late).
- **Performance:** 74% balanced accuracy (10-fold cross-validation), outperforming baseline models like eRBMs (50% accuracy).
- **Survival prediction:** Early-stage patients stratified by 14 biomarkers showed significant survival differences ($p < 0.005$).