L1000 Ligand Screening Methodology

DOC ID: SZ-L1000-M1 | CLASSIFICATION: CONFIDENTIAL
PIPELINE: TRANSCRIPTOMIC ALIGNMENT | VERSION: 2.1.4

1. Abstract & Objective

This document outlines the computational pipeline utilized by StateZero Labs to extract, transform, and evaluate high-dimensional transcriptomic signatures from the NIH LINCS L1000 dataset (GSE92742). The core objective is to systematically identify novel biological ligand interventions capable of reversing disease-state transcriptomic profiles toward a predetermined healthy regenerative baseline. By explicitly restricting our search space to biologics, endogenous signaling proteins, and engineered peptides, we bypass the target promiscuity and severe toxicity limitations inherently associated with small-molecule perturbations.

The methodology relies on projecting both the post-perturbation transcriptomic signature and the target regenerative blueprint into a shared high-dimensional vector space, allowing for precise mathematical alignment and affinity scoring.

2. Data Ingestion and Preprocessing

The StateZero pipeline directly ingests Level 5 data (moderated Z-scores representing robust differential expression) from the LINCS Phase I matrix. The raw feature space comprises 978 landmark genes, computationally inferred to represent the broader ~22,000 gene transcriptome.

To ensure strict biological relevance to our proprietary regenerative blueprint, the ingestion protocol applies a rigid metadata filtering layer. We isolate perturbation types explicitly designated as `trt_lig` (ligand treatments). All `trt_cp` (compound/small molecule) and `trt_sh` (shRNA) profiles are systematically excluded from the primary screening matrix.

PREPROCESSING PARAMETERS

- **Dataset:** GSE92742 (Phase I)
- **Input Dimension:** 978 Landmark Genes (L1000 standard)
- **Filtering Constraint:** Perturbation == trt_lig
- **Normalization:** Level 5 Moderated Z-Scores (Replicate consensus)
- **Thresholding:** Absolute Z-score ≥ 1.5 for inclusion in DEG sets

3. High-Dimensional Distance Metrics

To quantify the efficacy of a candidate ligand, we project the perturbation vector (P) and the target disease-reversal vector (R) into a unified Cartesian space. The primary evaluation metric for determining transcriptomic trajectory is Cosine Similarity.

$$S_C(P, R) = (P \cdot R) / (||P|| \times ||R||)$$

A candidate ligand is prioritized if the cosine similarity approaches 1.0, indicating that the directional transcriptomic shift induced by the ligand perfectly parallels the desired regenerative trajectory. Negative scores indicate disease exacerbation.

To account for absolute expression magnitude differences that Cosine Similarity might obscure (e.g., matching direction but insufficient amplitude), the algorithm subsequently evaluates the Mean Squared Error (MSE) across the k most critical differentially expressed genes (DEGs), which are dynamically weighted by our upstream CRISPR knockout network simulations.

$$MSE = (1/k) \Sigma_{(i=1 \text{ to } k)} (P_i - R_i)^2$$

4. Earth Mover's Distance (Wasserstein Metric)

Beyond pointwise comparisons, StateZero employs the Earth Mover's Distance (EMD) to evaluate the global distribution shift of the transcriptome. This provides a macroscopic view of transcriptomic

topology, ensuring that a perturbation does not introduce chaotic off-target cascades while correcting the target genes.

$$EMD(P, R) = \inf_{\gamma \in \Gamma(P,R)} \int \|x - y\| \, d\gamma(x,y)$$

By minimizing the EMD, we guarantee that the overall cellular state requires the minimal possible "transcriptional energy" to transition from the perturbed state to the healthy blueprint, optimizing for thermodynamic cellular stability.

5. Hit Triage and Validation

Ligands that successfully pass the similarity thresholds (Cosine Similarity ≥ 0.85) and MSE minimization checks enter the hit triage phase. These candidates are cross-referenced against the StateZero proprietary target network using a simulated propagation model.

During this phase, we explicitly screen for network stabilization. Candidates that trigger pro-inflammatory cascade markers—specifically up-regulation of TNF-α, IL-6, or NF-κB pathways above a Z-score of 1.5—are automatically deprioritized regardless of their regenerative alignment scores. Surviving assets are flagged as Tier 1 candidates and advance to in vitro validation and advanced pharmacokinetic (PBPK) modeling.

OUTPUT ASSET FORMATTING

The final output of the TFE (Transcriptomic-to-Fingerprint Encoder) yields a synthesized Annotation Dossier for each Tier 1 asset, containing:

- Predicted Morgan Fingerprints and 3D conformations.
- Stratum Corneum Permeability coefficient ($K_{sc/v}$).
- Molecular Weight, Isoelectric Point (pI), and binding affinity predictions (ΔG).

6. Conclusion

By combining stringent metadata isolation with advanced topological distance metrics (Cosine Similarity, MSE, and EMD), StateZero's L1000 processing pipeline reliably identifies high-fidelity

biological ligands. This mathematical infrastructure effectively bridges the gap between massive transcriptomic datasets and synthesizable, actionable therapeutics.