# SIMPLE SIAMESE NETWORKS ON FOUNDATION MODEL EMBEDDINGS OUTPERFORM COMPLEX MODELS ON UNSEEN SYNTHETIC LETHALITY PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Synthetic lethality (SL) is a biological phenomenon whereby co-deletion of two genes results in cell death, whereas deletion of either has minimal effect, and is widely considered a promising route for cancer therapeutic target discovery (O'Neil et al., 2017). Over the past decade, various machine learning methods have been proposed to predict synthetic lethal interactions from gene-level features. However, recent advances in AI for biology have produced a rich set of pretrained embeddings, each capturing distinct multi-modal representations defining genes and proteins, which have as-yet not been integrated into SL prediction models. Here, we propose a lightweight deep learning framework that uses multiple pretrained embeddings for the prediction of SL. On the largest database of SL interactions, we show such lightweight models achieve comparable performance to existing methods despite having orders of magnitude fewer trainable parameters, while the advantage of graph-based methods on previously seen genes disappears under strict gene holdout, likely due to existing SL methods leveraging the SL graph itself. We conduct a comprehensive ablation analysis, finding embeddings of curated knowledge generalize best, and several individual embeddings outperform naive multi-modal fusion. We further find integrating cellular context improves generalization capacity. Finally, we evaluate SL prediction on small focussed screens unseen during training, to our knowledge first evaluation to test whether such models could help actively guide future SL screens. We find poor zero-shot performance across all models, and evaluate fine-tuning strategies possible with lightweight models to improve predictive performance.

## 1 INTRODUCTION

Synthetic lethality (SL) is a genetic interaction between two genes whereby disrupting either gene alone leads to minimal cellular effects but disrupting both leads to cell death (Dobzhansky, 1946; Bridges, 1922). SL is a promising route within cancer therapeutic target discovery (Hartwell et al., 1997), leading to the clinical success of PARP inhibitors in BRCA1/2-mutant cancers(Farmer et al., 2005; Bryant et al., 2005; Lord & Ashworth, 2017). Given this success, there has been a push within the community to develop machine learning approaches to discover additional SL interactions across additional indications (see *related work* for an overview). Such models are trained on data sources such as SynLethDB, which stores known SL pairs derived from several sources (Guo et al., 2015; Wang et al., 2022a).

However, prediction of synthetic lethality for unobserved gene pairs is a challenging problem for several reasons. First, known SL pairs are non-uniform among genes, highly concentrated on a small number of well-studied genes (Horlbeck et al., 2018). Second, while positives themselves are rare (e.g. $0.0567\%$ of all possible pairs over $8{,}844$ genes that appeared in screened pairs in SynlethDB 3.0), confirmed negatives are also scarce ($0.273\%$). Third, as gene pairs are considered, observations are not independent nor identically distributed, making evaluation challenging. Finally, labels are aggregated at the screen (population) level (Shalem et al., 2014), so within-screen variation is typically unobserved (Dixit et al., 2016; Norman et al., 2019).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recent work by Feng et al. (2024) conducted a comprehensive benchmarking study across many models for predicting SL interactions (see *Related Work*). They introduced three tiers of test data splitting depending on the difficulty of the task[1]:

- **Tier I** randomly selects pairs, allowing both genes in a test pair to be included in the training set (though not together).
- **Tier II** selects pairs where at least one gene in each test pair is unseen during training.
- **Tier III** selects pairs where all genes in any test pair are unseen during training.

Since genetic screening typically measures all possible combinations across a given geneset, Tier II and III correspond to the scenario that motivates SL prediction enabling nomination of interactions that have not been screened for validation.

Concurrent with the rise of SL-prediction algorithms, multiple language models have been pretrained to capture gene-specific embeddings from various data types. Such models and resulting representations vary depending on the data modality on which they are pretrained:

- **Amino acid sequence:** protein language models such as ESM-2, are trained on hundreds of millions of protein sequences and are predictive of structural and functional properties
- **Expression:** the single-cell foundation models like Geneformer and scGPT are pretrained on tens of millions of transcriptomes and encode how a gene behaves across cellular contexts (Theodoris et al., 2023; Cui et al., 2024),
- **Curated function:** Methods such as GO2Vec (Zhong et al., 2019), Onto2Vec (Smaili et al., 2018) and Anc2Vec (Edera et al., 2022) all embed Gene Ontology, a highly curated measure of gene function and localization.
- **Text:** Methods such as BioConceptVec (Chen et al., 2020) are trained on PubMed abstracts and captures co-occurrence and context for gene concepts as they appear in the literature, while methods like GenePT(Chen & Zou, 2024) capture text embeddings that encode curated gene descriptions through an LLM embedding model .
- **Network topology:** node2vec on the protein–protein interaction graph, the raw adjacency spectrum, and Mashup describe a gene by its position among interaction partners rather than by any intrinsic property (Grover & Leskovec, 2016; Szklarczyk et al., 2022; Cho et al., 2016); we keep the raw adjacency deliberately, as an untrained control against the learned network embeddings.
- **Knowledge graph:** ComplEx(Trouillon et al., 2016) treats genes as entities in a heterogeneous graph of biomedical facts and learns relational structure rather than similarity.

However, despite the rise of such multi-modal pretrained embeddings, their utility to predict synthetic lethal interactions has never been tested. In this paper, our main contributions are threefold:

1. We propose SLeuth (Synthetic Lethality Estimation Using Twin Heads), a lightweight Siamese network for SL prediction using multi-modal pretrained embeddings as input, with orders of magnitude fewer trainable parameters than the current state-of-the-art graph model baseline.
2. Demonstrate this model achieves comparable performance compared to existing models across multiple levels of generalization.
3. Perform comprehensive ablation studies, to characterise the contribution of different data sources to the model's performance.
4. Investigate the zero-shot and fine-tuned performance of such models on smaller focussed SL screening datasets.

## 2 Related Work

Feng et al. (2024) evaluated 12 SL prediction methods, finding a set of three based on neural networks learning fixed gene representations to be the strongest performing (SLMGAE(Hao et al., 2021),

[1]Note these were originally refered to as CV1/2/3. We have renamed for clarity.

SLGNN(Zhu et al., 2023), and NSF4SL (Wang et al., 2022b)) SLMGAE (Hao et al., 2021) casts SL prediction as multi-view graph reconstruction: the SL graph forms a main view, and PPI and GO similarity serve as support views, and a graph auto-encoder reconstructs each view separately. An element-wise attention module then weights the support views and fuses their reconstructions, with the objective function formulated jointly by reconstruction and prediction losses. SLGNN Zhu et al. (2023) instead adds a knowledge graph, SynLethKG, with $11$ entity and $24$ relation types. It embeds genes through relation-aware message passing and decomposes each gene into latent factors, where a factor is a learned weighting over relation types processed by its own graph attention network. Pairs are then scored by inner product under an independence penalty that keeps the factors distinct. NSF4SL (Wang et al., 2022b) departs from the classification framing altogether and ranks candidate partners for a query gene. Starting from TransE embeddings of SynLethKG (Bordes et al., 2013), it trains a contrastive objective over positive pairs only, sidestepping the negative-sampling bias that pair-classification models inherit. All three learn gene representations from the SL graph (or SynLethKG), so a gene's representation depends on its observed interactions.

# 3 METHOD

## 3.1 OVERVIEW

Our proposed model predicts the probability that two genes are synthetically lethal in a given cell-line context given multi-modal pretrained embeddings. First, input representations for each of the two genes are retrieved from pretrained embeddings from six different biological categories via lookup of fixed vectors. These are concatenated and undergo dimensionality reduction. An input learnable representation of the cell line being predicted is then concatenated to each of the two genes' preprocessed input representations. The gene pair representations are fed through an identical lightweight encoder that maps them onto a latent Hilbert space. A positive-definite kernel scores the pair in that space, and a shared scale, a per-cell-line bias and a sigmoid turn the score into the probability of an SL pair. A masked cross entropy loss then removes every unseen pair-and-cell-line from both the numerator and the denominator.

## 3.2 MULTI-MODAL GENE EMBEDDING CONSTRUCTION AND PREPROCESSING

Pre-trained models were leveraged to compute fixed embeddings for each category as outlined in Appendix A Then, as some genes lack corresponding embeddings in specific categories, we employ mean imputation to get complete gene embeddings. Genes appearing only in held-out pairs from both test or validation data are excluded from imputation (Ambroise & McLachlan, 2002) to maintain strict train-test splitting. Each embedding matrix is then rescaled by a single mean and standard deviation taken over the whole matrix, and six matrices are concatenated. The principal component projection of the concatenated matrix using PCA is performed to reduce redundancy across dimensions, with components retaining $> 80\%$ of the variance being retained. The preprocessing pipeline is shown in Figure 1.

## 3.3 MODEL ARCHITECTURE

Our model is based on the Siamese network framework Bromley et al. (1993b), a symmetric architecture to map two inputs to a shared latent space. The overall architecture is shown in Figure **??**. Siamese neural networks map the input feature space into a Hilbert space (Bromley et al., 1993a; Chopra et al., 2005), wherein the inner product can be used as a measure of similarity. Hence, the approximation power depends largely on the depth of this map in our piecewise linear setting (Telgarsky, 2016; Raghu et al., 2016; Montúfar et al., 2014). This encoder network is made up of an entry projection from $\mathbb{R}^d \mapsto \mathbb{R}^{16}$, followed by $5$ residual blocks $\in \mathbb{R}^{16} \mapsto \mathbb{R}^{16}$, and finally a linear map $W_{-1} \in \mathbb{R}^{16\times16}$. Skip connections are used for each block as it keeps gradients flowing and stops the representation losing rank block by block (He et al., 2016; Pennington et al., 2017; Hanin & Rolnick, 2019). Skips fire only between equal-width blocks, so the entry block does not have skip connection. Each block applies an unbiased linear map, then layer normalization (Ba et al., 2016), then a Leaky ReLU with slope $0.2$ for the negative input (Maas et al., 2013), then dropout with ratio $0.2$ (Srivastava et al., 2014).

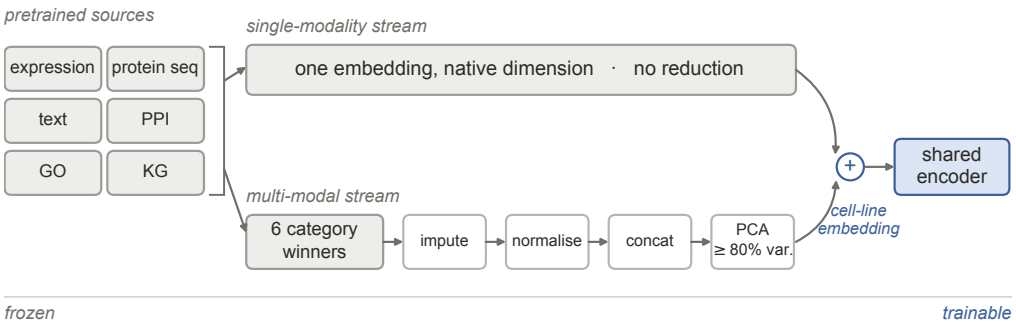


Figure 1: **Preprocessing and the two training streams.** Pretrained embeddings spanning six biological categories enter frozen; no gradient reaches the pretrained models. The single-modality stream trains one model per embedding at its native dimensionality, with no reduction of any kind. The multi-modal stream takes the six category winners, imputes genes missing from a source, normalises each matrix, concatenates them, and applies a single principal component projection at a variance target of 80%. Both streams concatenate a trainable cell-line embedding to each gene's features before the shared encoder.

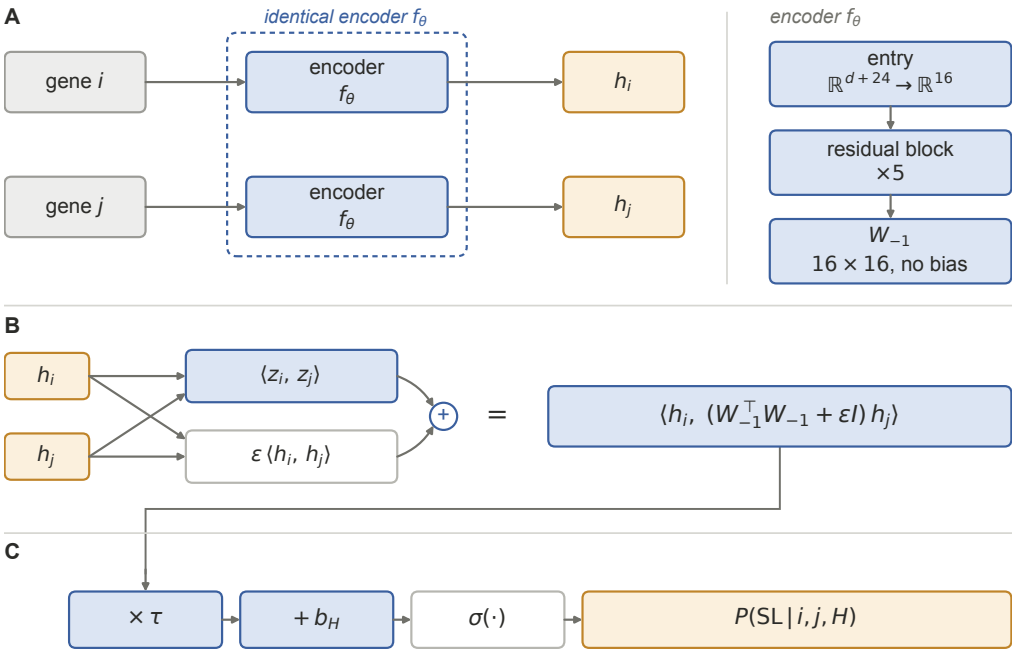


Figure 2: **Model Architecture: A.** Encoder **B.** Latent Inner Product Space **C.** Calculating Score per Cell Line Head (H)

The eventual SL score is calculated by

$$\left\langle z_i^{(H)}, z_j^{(H)} \right\rangle + \varepsilon \left\langle h_i^{(H)}, h_j^{(H)} \right\rangle = \left\langle h_i^{(H)}, \left(W_{-1}^T W_{-1} + \varepsilon I\right) h_j^{(H)} \right\rangle = \left\langle h_i^{(H)}, h_j^{(H)} \right\rangle_{W_{-1}^T W_{-1} + \varepsilon I},$$

where $z_\cdot^{(H)}$ denotes the output of the last linear layer and $h_\cdot^{(H)}$ the input to it, as summarized in Figure 2. The superscript records that the shared encoder consumes the gene features concatenated with the embedding of cell line type $H$, so each gene carries one hidden state per head, and $\varepsilon = 10^{-3}$ is the weight assigned to identity matrix to ensure positive definiteness. The last linear layer carries no bias, so $z_\cdot^{(H)} = W_{-1} h_\cdot^{(H)}$ and the identity above is exact. It follows that the inner product is

well-defined and that the output SL score per cell line type is calculated by

$$\text{logit}\left(\text{similarity score}_H\right) = \tau \cdot \left\langle h_i^{(H)}, h_j^{(H)} \right\rangle_{W_{-1}^T W_{-1} + \varepsilon I} + b_H,$$

where $\log \tau$ is trainable and shared across different cell line types, and $b_H$ is a per-head bias to absorb the base rate for each cell line type. The predicted probability that the pair is synthetically lethal in the context of head $H$ is then similarity score$_H$, that is, the logistic transform of the expression above.

In our model, the gene representation entering the encoder is augmented with cell-line context. We learn one trainable embedding $\in \mathbb{R}^{24}$ per cell-line category and concatenate it to each gene's representation, so that a given gene pair is encoded once per category. As SynLethDB 3.0 features cell line types for both positive and confirmed negative SL pairs, such information is reflected using these cell line embeddings in the model. Based on the cell line types, we designed $8$ different cell line corresponding to K562, JURKAT, A549, HELA, A375, 293T, PC9, plus OTHER.

### 3.4 Masked Objective for Partially Screened Pairs

Screening is virtually always incomplete in SL measurements, distinct experiments may measure subsets of gene (pairs) in different cell lines. Therefore, for a gene pair $(i, j)$ and head $H \in \{1, \ldots, 8\}$, we define a label $y_{ij,H} \in \{0, 1\}$ and a mask $m_{ij,H} \in \{0, 1\}$, where $m_{ij,H} = 1$ denotes that the pair was screened in at least one cell line mapping to head $H$. The label follows a disjunctive rule: $y_{ij,H} = 1$ if any screen assigned to head $H$ reported synthetic lethality, and $y_{ij,H} = 0$ if every such screen reported its absence.

Three cases then arise. First, a pair reported as SL in cell line $A$ and non-SL in cell line $B$, where the two cell lines map to different heads, sets those two heads independently. There is no contradiction to resolve as the pair is positive in one cellular context and negative in another, which is the phenomenon the multi-head formulation exists to represent. Second, pair reported as both SL and non-SL with in the same head. This may happen for two reasons: (i) the two screening methods give conflicting results; we still treat this as positive SL. (ii) For the "OTHER" head, which aggregates data from many small heads, it is possible that one head reports SL while another reports non-SL; in this case, we also classify it as SL. Third, a pair with no usable cell line recorded is folded into the "OTHER" head. If "MISSING" is set up as a separate head, it would produce many positives and zero negatives, making it impossible to train.

The training objective is the masked mean binary cross-entropy over observed scenarios. For a minibatch $\mathcal{B}$ of pairs with logits $\ell_{ij,H}$, the loss is defined by

$$\mathcal{L}(\mathcal{B}) = \frac{\displaystyle\sum_{(i,j)\in\mathcal{B}} \sum_{H=1}^{8} m_{ij,H} \cdot \text{BCE}\left(\ell_{ij,H}, y_{ij,H}\right)}{\max\left(1, \displaystyle\sum_{(i,j)\in\mathcal{B}} \sum_{H=1}^{8} m_{ij,H}\right)}.$$

### 3.5 Training

We employ AdamW with cosine warm restarts for optimization (Loshchilov & Hutter, 2017; 2016). Beyond dropout, the encoder is regularized two further ways. First, a decoupled weight decay of $10^{-5}$ applied by AdamW, and a proximal $\ell_1$ soft-thresholding step applied to each of the weight matrices after every optimizer step (Tibshirani, 1996; Beck & Teboulle, 2009; Parikh & Boyd, 2014), outside autograd, with a per-layer schedule decreasing from $0.01$ at the entry projection to $0.001$ at $W_{-1}$. The cell-line embedding is deliberately exempt from the $\ell_1$ penalty.

When evaluating SLeuth , modality selection uses a validation split generated separately within each of the $5$ training splits. It is impossible to create a universal validation split, as the part for training is unique to the particular split and does not intersect with the test pairs belonging to the split. At Tier I it is a random $10\%$ of the training pairs. For each of the six biological categories, the model takes an embedding with the highest AUPRG averaged from $5$ Tier I validation splits, concatenates

them all and performs principal component analysis of the concatenation, keeping at least $80\%$ of the variance. The selection is done only once, not per split and not per tier, and the selected embeddings are used unchanged in tiers II and III. Beyond controlling the number of parameters, this multi-modal PCA step also reduces repeated information across multi-modal data. Because the target is a variance fraction rather than a fixed width, and the projection is re-fit on the genes each fold leaves eligible, the retained dimension and therefore the model's parameter count varies slightly between folds and tiers.

## 4 Experimental Setup

We use SynLethDB 3.0 (Wang et al., 2022a) for our experiments, which is the largest and most up-to-date database of SL pairs. We restricted the data experimentally screened pairs only, containing $22{,}166$ positives and $106{,}704$ negatives. The folds for training and evaluation were produced only once using a set seed $42$ and shared among all models for fair comparison. For SLeuth, the top performing embedding in each category was selected from the Tier I validation splits, and reused unchanged at every tier. These are `geneformer` (expression), `prot_t5` (amino acid sequence), `text_embed` (text), `ppi_raw` (network topology), `go2vec` (curated function) and `kg_complex` (knowledge graph). Every SLeuth row in Tables 1 is this six-modality model. SLMGAE, SLGNN and NSF4SL were trained as per (Feng et al., 2024). We give SLGNN and NSF4SL the same ComplEx knowledge graph embeddings used by SLeuth in place of SynLethKG.

While AUROC is invariant to prevalence (Fawcett, 2006), it weights every positive-negative comparison equally, meaning the score is dominated by the bulk of the ranking and is largely insensitive to the high-ranked interactions, which would be prioritized for screening. AUPR is more sensitive to the rare high-ranked interactions in comparison (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). However, AUPR is heavily affected by the baseline prevalence. Thus, we adopt AUPRG (Flach & Kull, 2015) as it rescales precision and recall against the always-positive baseline,

$$\mathrm{precG} = 1 - \frac{\pi}{1-\pi} \cdot \frac{1-P}{P};\ \mathrm{recG} = 1 - \frac{\pi}{1-\pi} \cdot \frac{1-R}{R},$$

and integrates $\mathrm{precG}$ over $\mathrm{recG} \in [0, 1]$. Random prediction scores $0$ and perfect prediction scores $1$ at any prevalence, so unlike AUPR the value is not effected by the baseline.

## 5 Results

### 5.1 Comparison to State-of-the-Art Models on SynlethDB 3.0

| Tier | Model | Trainable parameters | AUROC | AUPR | AUPRG |
|---|---|---|---|---|---|
| I | SLeuth | 29,248 (22) | 0.900 (0.007) | 0.763 (0.011) | 0.955 (0.004) |
| | SLMGAE | $\approx 250M$ | **0.921** (0.002) | **0.856** (0.004) | **0.978** (0.001) |
| | SLGNN | $\approx 900k$ | 0.664 (0.005) | 0.336 (0.007) | 0.648 (0.012) |
| | NSF4SL | $\approx 500k$ | 0.495 (0.040) | 0.173 (0.012) | −0.228 (0.169) |
| II | SLeuth | 28,177 (170) | **0.796** (0.047) | **0.597** (0.061) | **0.871** (0.076) |
| | SLMGAE | $\approx 250M$ | 0.765 (0.013) | 0.573 (0.026) | 0.864 (0.056) |
| | NSF4SL | $\approx 500k$ | 0.517 (0.065) | 0.202 (0.044) | 0.131 (0.293) |
| | SLGNN | $\approx 900k$ | 0.512 (0.011) | 0.188 (0.031) | 0.174 (0.059) |
| III | SLeuth | 29,173 (69) | 0.677 (0.031) | **0.419** (0.089) | **0.732** (0.136) |
| | SLMGAE | $\approx 250M$ | **0.705** (0.038) | 0.371 (0.113) | 0.713 (0.080) |
| | SLGNN | $\approx 900k$ | 0.499 (0.014) | 0.167 (0.064) | −0.004 (0.075) |
| | NSF4SL | $\approx 500k$ | 0.455 (0.051) | 0.150 (0.060) | −0.262 (0.239) |

Table 1: **Performance across three evaluation tiers of increasing difficulty.** AUROC, AUPR, and AUPRG on SynlethDB 3.0 evaluated for SLeuth and three existing models across Tiers I-III.

Table 1 reports the results for SLeuth as well as three other methods identified as top performing in a recent benchmarking (Feng et al., 2024) at each generalization tier, with every value reporting a mean over $5$ training/testing data splits as well as its standard deviation. Model size is reported as trainable parameters. For SLeuth this excludes the frozen pretrained encoders, which are used for feature extraction but never fine-tuned. We find at Tier I, where a random split of pairs leaves

both genes of a test pair present in training (but not observed together), SLMGAE leads across all metrics. However, interestingly for Tier II (where only one gene is observed during training) and Tier III (where neither gene is observed), SLeuth has highest performance on all metrics with the exception of Tier III AUROC where SLMGAE has highest performance. That said, the smaller data size available at Tiers II and III mean certainty around the advantage is small. However, we interpret this to mean at increasing levels of generalization, small efficient models trained on fixed pretrained embeddings can at least match large models with orders of magnitude more parameters.

### 5.2 Ablation Study

We ablated Sleuth in two ways: (1) We fitted a model to each embedding on its own removing all others, and (2) We removed one category at a time from the six-category multi-modal model. In doing so we aimed to separate what information each embedding category carries on its own and what each adds beyond the other five in a multi-modal context.

#### 5.2.1 Single-Modality Models

| Category | Embedding | Tier I | Tier II | Tier III |
|---|---|---|---|---|
| expression | geneformer | **0.945** (0.008) | 0.786 (0.109) | 0.430 (0.167) |
| | scgpt | 0.943 (0.010) | **0.788** (0.096) | **0.556** (0.189) |
| | gene2vec | 0.939 (0.021) | 0.740 (0.117) | 0.535 (0.218) |
| protein_seq | esm2 | 0.950 (0.004) | **0.892** (0.035) | **0.790** (0.025) |
| | esm1b | **0.951** (0.005) | 0.864 (0.054) | 0.682 (0.187) |
| | prot_t5 | 0.948 (0.006) | 0.886 (0.039) | 0.785 (0.055) |
| text | text_embed | **0.959** (0.003) | 0.887 (0.046) | **0.841** (0.039) |
| | bioconceptvec | 0.945 (0.014) | **0.889** (0.036) | 0.816 (0.028) |
| ppi | node2vec_ppi | 0.952 (0.006) | 0.892 (0.042) | **0.815** (0.059) |
| | ppi_svd | 0.954 (0.002) | **0.892** (0.042) | 0.748 (0.143) |
| | ppi_raw | **0.971** (0.003) | 0.886 (0.055) | 0.726 (0.100) |
| | mashup | 0.931 (0.003) | 0.873 (0.038) | 0.727 (0.210) |
| go | go | 0.946 (0.005) | 0.838 (0.079) | 0.792 (0.062) |
| | go2vec | **0.958** (0.004) | **0.872** (0.054) | **0.802** (0.066) |
| | onto2vec | 0.909 (0.011) | 0.834 (0.062) | 0.799 (0.035) |
| kg | kg_complex | 0.951 (0.007) | 0.835 (0.063) | 0.411 (0.226) |

Table 2: **Single-modality models, one per embedding, grouped by biological category.** Collapsed AUPRG, mean (standard deviation) across the five training/testing splits. Bold font marks the best embedding within a category *for that tier*.

Table 2 reports the results of using each modality individually as a feature of the encoder. Unlike the multi-modal model, no principal component analysis was applied. From the collapsed AUPRG, expression and knowledge graph embeddings seem to underperform relative to the other embedding types. Averaging the collapsed AUPRG of each category's members, PPI is the strongest category at Tier I, while text embeddings are the strongest category at Tier II and III.

Interestingly, these findings suggest models trained solely with `go2vec` or `text_embed` embeddings perform better than the multi-modal model at every tier, and at Tiers II and III 9/16 single-modality models do so. We reason that this may be because the dimensionality reduction step may remove some informative variation, and because the six-modality combination was fixed once on Tier I validation evidence and never re-selected for the harder tiers.

#### 5.2.2 Leave-One-Category-Out

From Table 3, in our specific model setup, removing expression or knowledge graph embeddings does not measurably reduce Tier III performance — both ablated models in fact score marginally above the full model ($0.738$ and $0.736$ against $0.732$). This matches what we observe in the single modality models, where models built on expression and knowledge graph embeddings do not perform as well as other single-modal models. Dropping the Gene Ontology category is the largest single effect in the table, reducing AUPRG by $0.132$ at Tier III.

| Category removed | Embedding removed | Tier I | Tier II | Tier III |
|---|---|---|---|---|
| expression | geneformer | 0.956 (0.005) | 0.872 (0.073) | 0.738 (0.185) |
| go | go2vec | 0.953 (0.003) | 0.890 (0.057) | 0.600 (0.202) |
| kg | kg_complex | 0.955 (0.008) | 0.884 (0.061) | 0.736 (0.139) |
| ppi | ppi_raw | 0.949 (0.006) | 0.863 (0.055) | 0.710 (0.143) |
| protein_seq | prot_t5 | 0.958 (0.003) | 0.879 (0.063) | 0.715 (0.137) |
| text | text_embed | 0.946 (0.006) | 0.859 (0.090) | 0.689 (0.148) |
| none (full six-category model) | | 0.955 (0.004) | 0.871 (0.076) | 0.732 (0.136) |

Table 3: **Multi-modal model with one biological category removed.** Collapsed AUPRG, mean (standard deviation) across the 5 training/testing splits; each row is a model retrained from scratch on the remaining five categories, under the splits, architecture and preprocessing used for the full model. The second column names the embedding that category contributes to the

#### 5.2.3 Cell Line Context Abalation

We next investigated the contribution of cell line embedding to model performance. We ablated cell context information and re-computed test predictive performance across all three tiers (Table 4). Interestingly, masking cell line information results in virtually no change at Tier I, but results in larger performance drops in Tiers II&III, implying cell line context may be important for generalization in SL-prediction.

| Tier | Cell-line embedding | AUROC | AUPR | AUPRG |
|---|---|---|---|---|
| I | included | 0.900 (0.007) | **0.763** (0.011) | **0.955** (0.004) |
| | removed | **0.905** (0.008) | 0.747 (0.019) | 0.952 (0.007) |
| II | included | **0.796** (0.047) | **0.597** (0.061) | **0.871** (0.076) |
| | removed | 0.741 (0.086) | 0.522 (0.086) | 0.783 (0.154) |
| III | included | **0.677** (0.031) | **0.419** (0.089) | **0.732** (0.136) |
| | removed | 0.662 (0.029) | 0.392 (0.072) | 0.649 (0.196) |

Table 4: **Multi-modal model with the cell-line embedding zeroed at inference.** Mean (population standard deviation) across the 5 splits, on the same test edges and evaluator as Table 1.

### 5.3 Generalization to External Dataset

While large-scale benchmarking studies (Feng et al., 2024) typically rely on SynlethDB, there are many smaller studies that assess synthetic lethality in specific contexts or under specific pharmacological perturbation. Therefore, we assessed whether SLeuth could transfer predictions to smaller experiments not used in training or selection. For this we selected three recent CRISPR screening datasets: Adamson (Simpson et al., 2023) and Gilbert (Horlbeck et al., 2018), both in K562 cells, and Corn (Fielden et al., 2025) in RPE-1 cells. Each reports a continuous phenotype, so we used Spearman's correlation to measure agreement between the continuous phenotype and predicted SL score. No calibration or fine-tuning is applied, so the comparison is strictly zero-shot and identical for all four methods. We assess these models on the $54\%$, $56\%$ and $94\%$ of each screen's gene pairs whose genes overlap with those in SynLethDB.

The results in Table 5 show the results for all models across the three datasets. In general, zero-shot performance is low, with several methods even anti-correlated with ground-truth scores. Interestingly, NSF4SL performed best on the Adamson dataset by a clear margin, while SLeuth performed best on Gilbert, and there was no consensus for Corn.

Given this poor performance and the relatively low parameter count of SLeuth, we asked whether fine-tuning could improve predictions. To do so, we trained only the final layer freezing all other weights of SLeuth, or fully retrained the model, and re-evaluted across the three datasets. Fine-tuning the whole encoder improved Adamson from $-0.203$ to $0.974$ on the random-pair split, but the gain falls to $0.655$ when one gene of each test pair is held out and to $0.427$ when both are (Table 6. However, this pattern was not consistant across all datasets, though in general last-layer fine tuning and full retraining did improve test-set predictive performance across generalization layers.

| Screen | Split | SLeuth | SLMGAE | SLGNN | NSF4SL |
|---|---|---|---|---|---|
| Adamson | random pairs | −0.203 (0.060) | −0.035 (0.027) | −0.012 (0.011) | **0.179** (0.039) |
| | one gene unseen | −0.195 (0.075) | −0.017 (0.035) | −0.013 (0.006) | **0.237** (0.069) |
| | both unseen | −0.196 (0.096) | −0.007 (0.048) | −0.027 (0.012) | **0.301** (0.126) |
| Corn | random pairs | 0.031 (0.009) | 0.031 (0.006) | 0.007 (0.007) | **0.036** (0.016) |
| | one gene unseen | **0.045** (0.019) | 0.031 (0.011) | 0.002 (0.007) | 0.040 (0.019) |
| | both unseen | 0.049 (0.039) | **0.060** (0.020) | −0.028 (0.005) | 0.044 (0.023) |
| Gilbert | random pairs | **0.058** (0.018) | 0.029 (0.012) | −0.019 (0.003) | 0.018 (0.020) |
| | one gene unseen | **0.062** (0.016) | 0.022 (0.014) | −0.006 (0.010) | 0.020 (0.034) |
| | both unseen | **0.067** (0.018) | 0.016 (0.020) | −0.007 (0.022) | 0.028 (0.064) |

Table 5: **Zero-shot transfer to three external screens.** Spearman correlation between each model's score outputs and the screen phenotype; mean (standard deviation) across the five Tier III fold models.

| Screen | Split | Zero-shot | `LP_last` | `Full` |
|---|---|---|---|---|
| Adamson | random pairs | −0.203 (0.060) | 0.290 (0.052) | 0.974 (0.001) |
| | one gene unseen | −0.195 (0.075) | 0.258 (0.063) | 0.655 (0.031) |
| | both unseen | −0.196 (0.096) | 0.231 (0.081) | 0.427 (0.047) |
| Corn | random pairs | 0.031 (0.009) | 0.061 (0.015) | 0.442 (0.002) |
| | one gene unseen | 0.045 (0.019) | 0.051 (0.008) | 0.210 (0.011) |
| | both unseen | 0.049 (0.039) | 0.041 (0.023) | −0.026 (0.026) |
| Gilbert | random pairs | 0.058 (0.018) | 0.087 (0.007) | 0.303 (0.002) |
| | one gene unseen | 0.062 (0.016) | 0.084 (0.014) | 0.195 (0.007) |
| | both unseen | 0.067 (0.018) | 0.084 (0.028) | 0.105 (0.006) |

Table 6: **Fine-tuning strategies on small datasets.** Spearman correlation with the screen's phenotype, mean (standard deviation) over the five fold models, on SynLethDB-filtered test pairs, scored at the cell-line head matching each screen, for zero-shot, last layer training (`LP_last`), and full retraining.

## 6 Discussion and Conclusion

In this paper, we introduced lightweight Siamese networks on top of multi-modal pretrained embeddings for synthetic lethality prediction. We showed that this lightweight method with frozen, pretrained embeddings is competitive to leading graph models with orders of magnitude more parameters, whose large advantage on already seen gene pairs (Tier I) does not transfer to increasingly difficult levels of generalization. Interestingly, through comprehensive ablation analysis, we find representations derived from curated knowledge (including text descriptions and GO annotation) help the most with generalization, while some representations that are essential for Tier 1 prediction provide virtually no advantage at more difficult levels of generalization. However, we caution that this interpretation may be confounded by the fact that many SL relationships (including those in SynlethDB) come from genes that are well studied, which is precisely those that will be well annotated with GO or text annotations. Furthermore, interestingly we find that multi-modal fusion of embeddings frequently under-performs relative to single-mode embeddings, especially on Tiers II-III. This may be due to optimal modalities being selected on Tier I validation splits, which was chosen in part due to decreased data sizes in Tiers II-III. As synthetic lethality databases become larger and more comprehensive, future studies may examine the impact of Tier-specific modality selection. Finally, we found poor zero-shot generalization across all models to unseen focused screening datasets, which should arguably be the major use-case in order to guide future screening. Due to the lightweight nature of our model, we were able to evaluate fine-tuning strategies, and showed strong performance with last layer and full training.

Our work has several limitations. Due to the small number of genes left in Tier II-III analysis, the large standard deviations make it difficult to definitively say which models worked best. Second, as noted above, modality selection for Tiers II-III was performed on Tier I, which may be sub-optimal overall. Finally, due to computational limitations, we only evaluated against the top three existing models, but from the myriad other SL prediction models proposed, others may outperform these for unseen screen generalization despite not scoring top in benchmarking.

## AI USE STATEMENT

Generative AI tools were used to: (1) Reword specific paragraphs of the paper, (2) Get overall feedback on the paper to strengthen the work, (3) Clean specific datasets, (4) Initial code generation to execute experiments. (5) Anonymizing the repository and structuring the code and results. Generative AI was not used for research ideation, helping develop theoretical models or conceptual frameworks, or design the architecture for the Siamese network models. Additionally, we have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work.

## ETHICS STATEMENT

This work uses only publicly available data from de-identified human cell lines. No patient data was incorporated at any point, and findings were not used for further therapeutic recommendation or development.

## REPRODUCIBILITY STATEMENT

Every input is public. Positives and screened negatives come from SynLethDB, restricted to experimentally confirmed pairs. The five folds at each tier were generated once from seed 42 and shared byte-identically across all four models; the fold metadata stores a hash of every partition, so a rerun can be checked against ours rather than merely resembling it. Section 3.2 states the preprocessing pipeline and the scope on which each statistic is fit, Section 3.3 gives the architecture and the regularization settings, and Section 4 gives the selection protocol and how validation pairs are carved. Code and configuration are available at

`https://anonymous.4open.science/r/SLeuth-679B/`.

## REFERENCES

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. Pykeen 1.0: A python library for training and evaluating knowledge graph embeddings. July 2020. doi: 10.48550/ARXIV.2007.14175.

Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10): 6562–6566, April 2002. ISSN 1091-6490. doi: 10.1073/pnas.102102699.

Suzi A and Aleksander, James P Balhoff, Seth Carbon, J Michael Cherry, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Patrick Kalita, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganujan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Edith D Wong, Valerie Wood, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Robert Dodson, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Ximena Aixa Castro Naser, Silvio C E Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Nicholas H Brown, Gil dos Santos, Steven Marygold, Katja Röper, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Rossana Zaru, Ruth C Lovering, Colin Logie, Daiqing Chen, Alexandra Naba, Karen Christie, Lori Corbani, Li Ni, Dmitry Sitnikov, Cynthia Smith, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Sushma Naithani, Pascal Carme, Kim Rutherford, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Gene Goldbold, Stacia R Engel, Stuart R Miyasato, Robert S Nash, Gavin Sherlock, Shuai Weng, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Aduragbemi Adesina, Emily Bowler-Barnett, David

Carpentier, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Nadya Urakova, Kate Warner, Conny Wing-Hen Yu, Juancarlos Chan, Stavros Diamantakis, Mark Quinton-Tulloch, Daniela Raciti, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Doug Howe, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The gene ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, December 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf1292.

Alex and Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Emily H Bowler-Barnett, Hema Bye-A-Jee, David Carpentier, Paul Denny, Jun Fan, Penelope Garmiri, Leonardo Jose da Costa Gonzales, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Jie Luo, Yvonne Lussi, Juan Sebastian Martinez Marin, Pedro Raposo, Daniel L Rice, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Nidhi Tyagi, Nadya Urakova, Preethi Vasudev, Kate Warner, Supun Wijerathne, Conny Wing-Heng Yu, Rossana Zaru, Alan J Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H Auchincloss, Kristian B Axelsen, Parit Bansal, Delphine Baratin, Teresa M Batista Neto, Marie-Claude Blatter, Jerven T Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard de Castro, Anne Estreicher, Maria L Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J A Sigrist, Karin Sonesson, Shyamala Sundaram, Anastasia Sveshnikova, Cathy H Wu, Cecilia N Arighi, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Minna Lehvaslaiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Yuqi Wang, and Jian Zhang. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, November 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1010.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, July 2016.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, January 2009. ISSN 1936-4954. doi: 10.1137/080716542.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, volume 26, pp. 2787–2795. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.

Calvin B. Bridges. The origin of variations in sexual and sex-limited characters. *The American Naturalist*, 56(642):51–63, January 1922. ISSN 1537-5323. doi: 10.1086/279847.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector (eds.), *Advances in Neural Information Processing Systems 6 (NIPS 1993)*, volume 6, pp. 737–744. Morgan Kaufmann, 1993a. URL https://proceedings.neurips.cc/paper/1993/hash/288cc0ff022877bd3df94bc9360b9c5d-Abstract.html.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'93, pp. 737–744, San Francisco, CA, USA, 1993b. Morgan Kaufmann Publishers Inc.

Helen E. Bryant, Niklas Schultz, Huw D. Thomas, Kayan M. Parker, Dan Flower, Elena Lopez, Suzanne Kyle, Mark Meuth, Nicola J. Curtin, and Thomas Helleday. Specific killing of brca2-deficient tumours with inhibitors of poly(adp-ribose) polymerase. *Nature*, 434(7035):913–917, April 2005. ISSN 1476-4687. doi: 10.1038/nature03443.

Qingyu Chen, Kyubum Lee, Shankai Yan, Sun Kim, Chih-Hsuan Wei, and Zhiyong Lu. Biocon-ceptvec: Creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLOS Computational Biology*, 16(4):e1007617, April 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007617.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature Biomedical Engineering*, 9(4):483–493, December 2024. ISSN 2157-846X. doi: 10.1038/s41551-024-01284-6.

Hyunghoon Cho, Bonnie Berger, and Jian Peng. Compact integration of multi-network topology for functional analysis of genes. *Cell Systems*, 3(6):540–548.e5, December 2016. ISSN 2405-4712. doi: 10.1016/j.cels.2016.10.017.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005. doi: 10.1109/cvpr.2005.202.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, February 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, ICML '06, pp. 233–240. ACM Press, 2006. doi: 10.1145/1143844.1143874.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, December 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.11.038.

Th Dobzhansky. Genetics of natural populations. xiii. recombination and variability in populations of drosophila pseudoobscura. *Genetics*, 31(3):269–290, May 1946. ISSN 1943-2631. doi: 10.1093/genetics/31.3.269.

Alejandro A Edera, Diego H Milone, and Georgina Stegmayer. Anc2vec: embedding gene ontology terms by preserving ancestors relationships. *Briefings in Bioinformatics*, 23(2), February 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac003.

Hannah Farmer, Nuala McCabe, Christopher J. Lord, Andrew N. J. Tutt, Damian A. Johnson, Tobias B. Richardson, Manuela Santarosa, Krystyna J. Dillon, Ian Hickson, Charlotte Knights, Niall M. B. Martin, Stephen P. Jackson, Graeme C. M. Smith, and Alan Ashworth. Targeting the dna repair defect in brca mutant cells as a therapeutic strategy. *Nature*, 434(7035):917–921, April 2005. ISSN 1476-4687. doi: 10.1038/nature03445.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655. doi: 10.1016/j.patrec.2005.10.010.

Yimiao Feng, Yahui Long, He Wang, Yang Ouyang, Quan Li, Min Wu, and Jie Zheng. Benchmarking machine learning methods for synthetic lethality prediction in cancer. *Nature Communications*, 15 (1), October 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-52900-7.

John Fielden, Sebastian M. Siegner, Danielle N. Gallagher, Markus S. Schröder, Maria Rosaria Dello Stritto, Simon Lam, Lena Kobel, Moritz F. Schlapansky, Stephen P. Jackson, Petr Cejka, Marco Jost, and Jacob E. Corn. Comprehensive interrogation of synthetic lethality in the dna damage response. *Nature*, 640(8060):1093–1102, April 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08815-4.

Peter A. Flach and Meelis Kull. Precision-recall-gain curves: PR analysis done right. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, volume 28, pp. 838–846. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/hash/33e8075e9970de0cfea955afd4644bb2-Abstract.html.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 855–864. ACM, August 2016. doi: 10.1145/2939672.2939754.

Jing Guo, Hui Liu, and Jie Zheng. Synlethdb: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Research*, 44(D1):D1011–D1017, October 2015. ISSN 1362-4962. doi: 10.1093/nar/gkv1108.

Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns. *NeurIPS 2019*, June 2019. doi: 10.48550/ARXIV.1906.00904.

Zhifeng Hao, Di Wu, Yuan Fang, Min Wu, Ruichu Cai, and Xiaoli Li. Prediction of synthetic lethal interactions in human cancers using multi-view graph auto-encoder. *IEEE Journal of Biomedical and Health Informatics*, 25(10):4041–4051, October 2021. ISSN 2168-2208. doi: 10.1109/jbhi.2021.3079302.

Leland H. Hartwell, Philippe Szankasi, Christopher J. Roberts, Andrew W. Murray, and Stephen H. Friend. Integrating genetic approaches into the discovery of anticancer drugs. *Science*, 278(5340): 1064–1068, November 1997. ISSN 1095-9203. doi: 10.1126/science.278.5340.1064.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016. doi: 10.1109/cvpr.2016.90.

Max A. Horlbeck, Albert Xu, Min Wang, Neal K. Bennett, Chong Y. Park, Derek Bogdanoff, Britt Adamson, Eric D. Chow, Martin Kampmann, Tim R. Peterson, Ken Nakamura, Michael A. Fischbach, Jonathan S. Weissman, and Luke A. Gilbert. Mapping the genetic landscape of human cells. *Cell*, 174(4):953–967.e22, August 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.06.010.

Rachael P. Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J. Martin, and Claire O'Donovan. The goa database: Gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063, November 2014. ISSN 0305-1048. doi: 10.1093/nar/gku1113.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 1095-9203. doi: 10.1126/science.ade2574.

Christopher J. Lord and Alan Ashworth. Parp inhibitors: Synthetic lethality in the clinic. *Science*, 355(6330):1152–1158, March 2017. ISSN 1095-9203. doi: 10.1126/science.aam7344.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. August 2016. doi: 10.48550/ARXIV.1608.03983.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. November 2017. doi: 10.48550/ARXIV.1711.05101.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, Atlanta, Georgia, USA, 2013. URL https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. February 2014. doi: 10.48550/ARXIV.1402.1869.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, August 2019. ISSN 1095-9203. doi: 10.1126/science.aax4438.

Nigel J. O'Neil, Melanie L. Bailey, and Philip Hieter. Synthetic lethality and cancer. *Nature Reviews Genetics*, 18(10):613–623, 2017. ISSN 1471-0064. doi: 10.1038/nrg.2017.47.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1 (3):127–239, January 2014. ISSN 2167-3918. doi: 10.1561/2400000003.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. November 2017. doi: 10.48550/ARXIV.1711.04735.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. June 2016. doi: 10.48550/ARXIV.1606.05336.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, March 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0118432.

Eric W Sayers, Evan E Bolton, Anna M Fine, Christopher Kelly, Sunghwan Kim, Melissa Landrum, Stacy Lathrop, Adriana Malheiro, Terence D Murphy, Lon Phan, Shashikant Pujar, Barton W Trawick, Valerie A Schneider, and Kim D Pruitt. Database resources of the national center for biotechnology information in 2026. *Nucleic Acids Research*, 54(D1):D20–D27, December 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf1060.

Ophir Shalem, Neville E. Sanjana, Ella Hartenian, Xi Shi, David A. Scott, Tarjei S. Mikkelsen, Dirk Heckl, Benjamin L. Ebert, David E. Root, John G. Doench, and Feng Zhang. Genome-scale crispr-cas9 knockout screening in human cells. *Science*, 343(6166):84–87, January 2014. ISSN 1095-9203. doi: 10.1126/science.1247005.

Danny Simpson, Jia Ling, Yangwode Jing, and Britt Adamson. Mapping the genetic interaction network of parp inhibitor response. August 2023. doi: 10.1101/2023.08.19.553986.

Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. Onto2vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, 34(13):i52–i60, 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty259.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL `https://jmlr.org/papers/v15/srivastava14a.html`.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, November 2022. ISSN 1362-4962. doi: 10.1093/nar/gkac1000.

Alp Tartici, Gowri Nayar, and Russ B Altman. Pool parti: a pagerank-based pooling method for identifying critical residues and enhancing protein sequence representations. *Bioinformatics*, 41 (6), 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf330.

Matus Telgarsky. Benefits of depth in neural networks. February 2016. doi: 10.48550/ARXIV.1602.04485.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, January 1996. ISSN 1467-9868. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. June 2016. doi: 10.48550/ARXIV.1606.06357.

Jie Wang, Min Wu, Xuhui Huang, Li Wang, Sophia Zhang, Hui Liu, and Jie Zheng. Synlethdb 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. *Database*, 2022, January 2022a. ISSN 1758-0463. doi: 10.1093/database/baac030.

Shike Wang, Yimiao Feng, Xin Liu, Yong Liu, Min Wu, and Jie Zheng. Nsf4sl: negative-sample-free contrastive learning for ranking synthetic lethal partner genes in human cancers. *Bioinformatics*, 38(Supplement_2):ii13–ii19, 2022b. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac462.

Xiaoshi Zhong, Rama Kaalia, and Jagath C. Rajapakse. Go2vec: transforming go terms and proteins to vector representations via graph embeddings. *BMC Genomics*, 20(S9), December 2019. ISSN 1471-2164. doi: 10.1186/s12864-019-6272-2.

Yan Zhu, Yuhuan Zhou, Yang Liu, Xuan Wang, and Junyi Li. Slgnn: synthetic lethality prediction in human cancers based on factor-aware knowledge graph neural network. *Bioinformatics*, 39(2), January 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad015.

## A Input representation processing

- **ESM-2.** The protein produced by a gene is a string of amino acids (AA). We retrieve the sequence for each gene from NCBI (Sayers et al., 2025) and run it through ESM-2 (Lin et al., 2023), which is a language model that was trained on hundreds of millions of protein sequences in the same fashion as a text model is trained on sentences. This generates a $\mathbb{R}^{1280}$ vector for every AA position. We then get a single vector for every gene using PoolPaRTI, where we take the attention scores of the model and use them to calculate the importance of each position and then average the vectors accordingly, which ensures the position of each AA is considered in the construction of this $\mathbb{R}^{1280}$ embedding for the gene (Tartici et al., 2025).
- **ProtT5.** Same sequences as before, but the pooling here is a mean across residues normalized by the sequence length.
- **Geneformer.** No forward pass is required since the gene vector is the token input to the model itself.
- **scGPT.** The vocabulary uses direct symbol indexing, and thus only one embedding matrix lookup for the gene encoder.
- **GO2Vec** and **Onto2Vec.** These models embed GO terms instead of genes, which requires aggregation. Gene-term annotations from the human GOA file were used with NOT-qualifiers stripped (Ashburner et al., 2000; and Aleksander et al., 2025; Huntley et al., 2014; and Bateman et al., 2024), and genes represented by an unweighted average of the vectors of the terms annotated to them.
- **BioConceptVec.** Gene concepts already use our gene identifiers. Multi-gene concepts were discarded, while the limit of coverage is genes not mentioned enough times in PubMed to be considered for extraction.
- **Text embedding.** We construct a string *symbol: description. Also known as: synonyms* for each gene from NCBI and then embed it. Virtually all genes have such a representation, but even one with a brief description gets the full vector length.
- **ComplEx.** Trained locally, not downloaded. The graph is built using the STRING v12.0 Human Interactome (Szklarczyk et al., 2022), where only edges with a score $\geq 700$ are retained as `interacts_with` triples, and the GO annotations are added as `annotated_with` triples. ComplEx is trained via PyKEEN at 256 complex dimensions (Ali et al., 2020).
- **PPI features (node2vec, raw adjacency spectrum).** STRING protein identifiers map to symbols via the STRING info file.