ECHO: A Bayesian approximation model for *in silico* drug screening in genomic medicines

Abstract

In silico exploration of drug design space in genomic medicines can expedite drug development, serving as a cost-effective replacement for large experimental screens. However, building trustworthy *in silico* screening models often requires high data volumes *a priori*.

Here, we demonstrate ECHO, a lightweight Bayesian approximation machine learning paradigm for *in silico* drug screening, designed for sparse datasets. We benchmark this method against state-of-the-art methods used for target-primed reverse transcription (TPRT)-based genome engineering, and we show that ECHO can be used to identify highly potent hits across multiple diverse, therapeutically relevant targets.

Our modeling paradigm presents an agile framework for *in silico* screening and optimization using small contextually-relevant datasets, thereby enabling researchers to cost-effectively explore the vast design space in genomic medicines.

AUTHORS

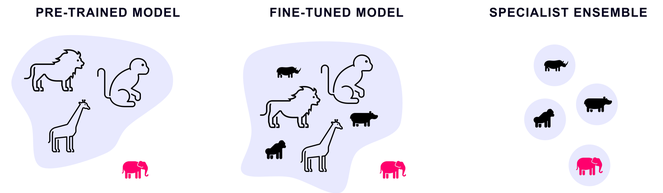
Divya Ramamoorthy*, Giulia I. Corsi*, Selina Sun*, Robin Chan, Tom Noonan, Athanasios Dousis, Pradeep Ramesh, Aamir Mir, Anne Bothmer, Gregory McAllister, Cecilia Cotta-Ramusino

*Indicates equal contribution

TESSERA

Background

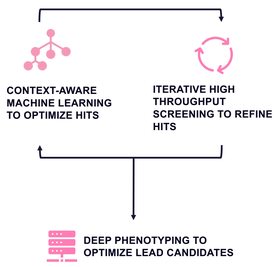
- Optimization of genomic medicines presents a large **combinatorial problem**: many components drive efficacy and only a fraction of the design space can be assayed
- Generating new large-scale datasets in therapeutic contexts is cost-prohibitive, and existing datasets are conducted in settings (cell types, delivery mechanisms, etc.) that **translate poorly** to real-application contexts
- There is a need for tools that enable the **cost-effective optimization** of genomic medicines while working with **sparse datasets**



***Strategies for modeling sparse datasets**: TPRT-optimization is heavily context-dependent, resulting in pre-training and fine-tuning strategies translating poorly*

Methodology

ECHO uses a lab-in-the-loop modeling workflow. We iteratively assay a small panel of in-context data, generate context-aware predictions, then assay predicted hits



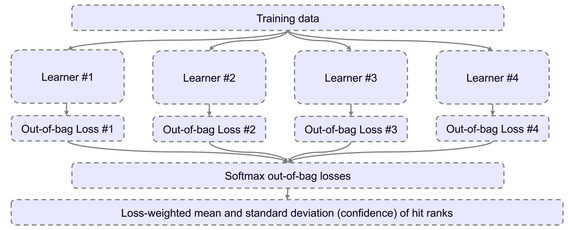
***Workflow**: Iterative screening and modeling to optimize hits from sparse datasets*

ECHO leverages a weighted ensemble of gradient-boosted decision trees, with loss-weighted ranks to predict top screen hits and a measure of epistemic uncertainty for each hit. The model trains on 76 features, which include position-independent nucleotide composition features and nucleic acid binding energies.

Algorithm 1: Pseudocode of the loss-weighted ensemble algorithm

Input: Training dataset $D = \{(s_1, e_1), (s_2, e_2), (s_3, e_3)\}$ where s are sequences, e are measured efficacies
Complete dataset $U = \{s \mid s \text{ in design space}\}$
Number of learners m
Out-of-bag (OOB) subsample fraction f
Output: Weighted μ and σ of sample ranks for U

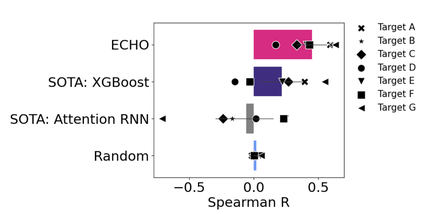
START
Step 1: Train base learners T for dataset D, and get predicted ranks for complete dataset
FOR i = 1…m, **DO**
$T_i, b_i = L_i(D, f)$ where L_i is gradient boosted regressor initialized with rand seed set to i, b_i is OOB score for a learner with subsampling defined by f, and T_i is the trained learner
$\boldsymbol{r}_i = T_i(U)$ where $\boldsymbol{r}_i$ is predicted ranks of U for learner i
END FOR
Step 2: Calculate ensemble weights from softmax-normalized OOB losses
$w = softmax(b)$
Step 3: Combine models with weighted μ and σ
$\mu = \sum_i^m w_i r_i \qquad \sigma = \sqrt{\sum_i^m w_i (r_i - \mu_i)^2}$
RETURN μ , σ
END



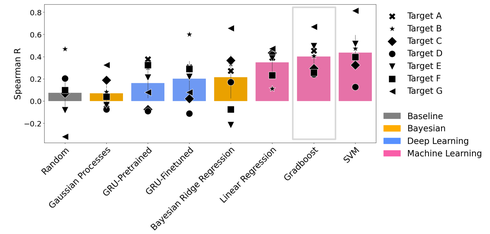
***Model steps**: Learners are each initialized with separate random seed and different out-of-bag subsamples of 10%, enabling stochasticity*

Results

ECHO outperforms state-of-art methods for *in silico* benchmarks

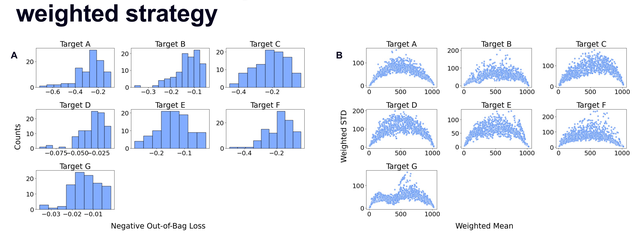


***Benchmark against state of art methods for template engineering task.** ECHO performance evaluated on 10 withheld test samples, averaged over 100 random draws of samples. State of art models (SOTA) are publicly available benchmarks. Targets are anonymized T-cell and Hematopoietic Stem Cell (HSC) endogenous loci.*



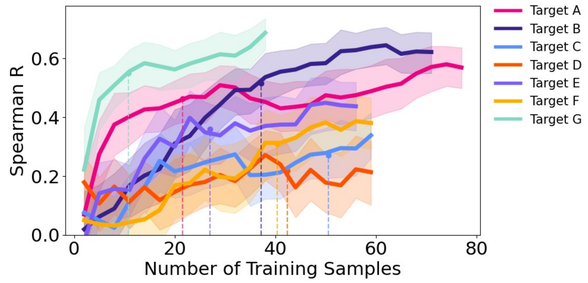
***Selection of learner.** Algorithms evaluated across machine learning, deep learning, and Bayesian domains. Final gradient boosted regression learner selected due to improved feature interpretability relative to SVM learner.*

Loss and uncertainty distributions show ECHO-weighted strategy



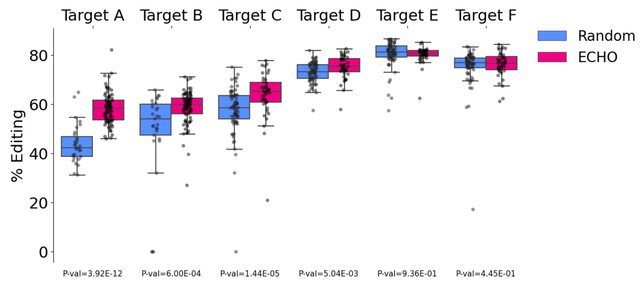
***A) Distribution of out-of-box loss.** Sparse learners show spread of out-of-box loss, enabling weighted ranks. **B) Correlation of weighted mean and std of hit ranks.** Targets show distinct uncertainty patterns, with middle ranks having high uncertainty.*

Model demonstrates high data efficiency across diverse targets



***Learning curve for training sample volume.** Vertical lines indicate data volume required for 80% of maximum Spearman R score per target.*

Experimental *in vitro* validation demonstrates that ECHO enhances hit potency for diverse therapeutic targets



***Percent editing for predicted hits.** Targets tested in T-cells, with editing at endogenous loci. Percent editing evaluated by amplicon sequencing. P-values for one-sided Mann-Whitney-U test of ECHO vs. benchmarks.*

	Top ECHO hit (absolute editing efficiency)	Difference from Benchmark (percent points)
Target A	82.2%	17.3%
Target B	71.1%	5.2%
Target C	77.8%	2.7%
Target D	82.6%	0.8%
Target E	85.2%	-1.4%
Target F	84.3%	0.9%

***Editing efficiency of Top ECHO Hit.** Absolute editing efficiency for targets in T-cells, and the difference in efficiency between the top hit from ECHO and the top hit from benchmarks*

Conclusions

Increased potency

AI predictions generated candidates with high efficiencies, with max rewriting efficiencies between 71-85%

Reduced cost

Identified candidates at 10% the cost of a full screen

Reduced TAT

Accelerated design-build-test cycle, reducing time to lead candidates by up to 50%

Agility

Modular ML strategy designed to enable rapid pivots to new cell types

Data efficiency

Model requires < 55 data points to train for new contexts

Future Work & Limitations

- Model only captures epistemic uncertainty; method may be extended to approximate aleatoric uncertainty when biological or technical replicates are available
- Model is limited to sparse datasets where out-of-bag weighted distributions show meaningful diversity among learners