# Ledidi: Designing genomic edits that induce functional activity


## Jacob Schreiber[1*], Yang Young Lu[1*], and William Noble[1,2]

1. Deparment of Genome Science, University of Washington
2. Paul G. Allen School of Computer Science and Engineering, University of Washington

* These authors contributed equally to this work.


## Genome editing is powerful but challenging

Driven by the development of the CRISPR-Cas9 system, there now exist several technologies capable of making precise edits to the genome. However, using these tools to their maximal effectiveness involves overcoming two challenges:

(1) Identifying the precise edits that must be made to achieve the desired effect.

(2) Ensuring that this set of edits does not have unintended consequences that interfere with the normal function of a cell.

## Ledidi is a method for designing edits

We propose Ledidi, an approach that can use any pre-trained machine learning model to design edits that achieve a particular output for that model. More formally, for some model *f*, Ledidi optimizes the following mixture of objectives

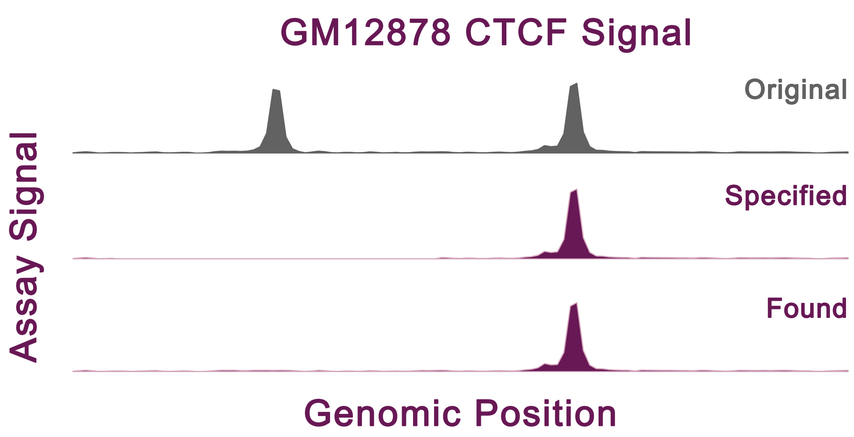
$$\min_{X} ||X - X_0||_1 + \lambda ||f(X) - \hat{y}||_2^2$$

where X is the edited sequence, $X_0$ is the original sequence, f(X) is the output of the pre-trained model on the edited sequence, and $\hat{y}$ is the desired output from the model.

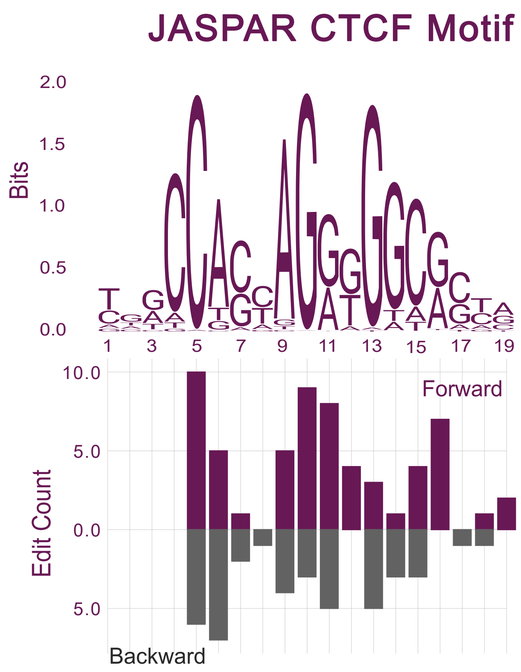
## Ledidi can knock-in and knock-out CTCF binding

We paired Ledidi with Basenji (Kelley et al, 2018), a model that takes in nucleotide sequence and outputs functional profiles in the form of signal from ChIP-seq, DNase-seq, and CAGE experiments.

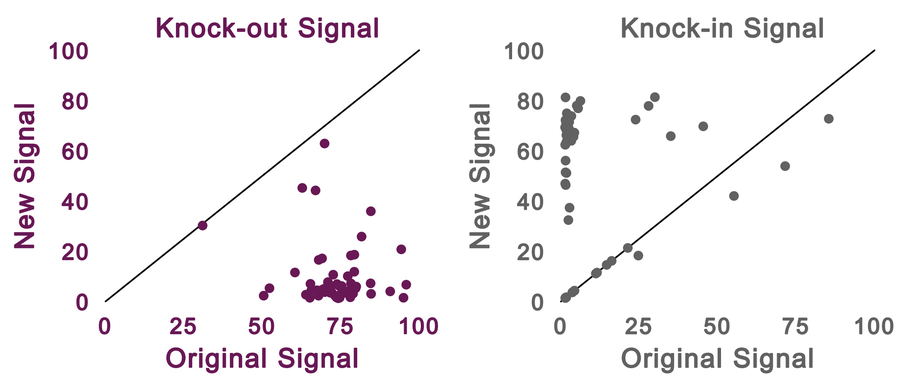
In an initial experiment (below), we found that Ledidi could easily design a sequence that disrupts binding of the protein CTCF, as seen by the removal of the left peak, while preserving binding at the location under the right peak.




When we applied Ledidi to knock out CTCF binding at 53 loci, we found that the proposed edits mostly fell at conserved sites within the CTCF motif.




These edits successfully knocked out binding and reduced the median predicted signal from a fold change of 74.2 to 5.2 (left, below).

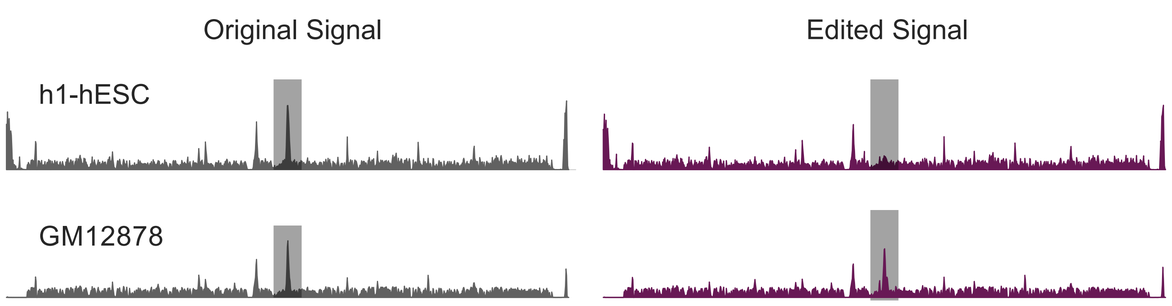


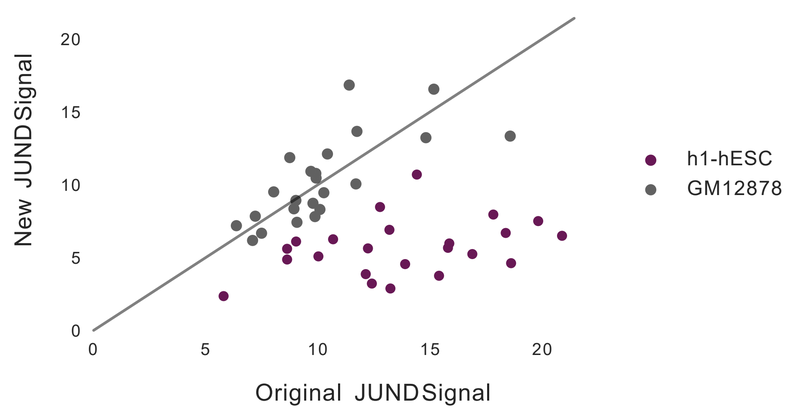

Ledidi can also successfully induce CTCF binding. When applied to 53 loci, we found that Ledidi increased the predicted fold change values for CTCF binding from 2.4 to 63.7 (right, below).

## Ledidi can design sequences that exhibit cell type-specific protein binding

We next applied Ledidi to the more challenging task of designing motifs that exhibit cell type-specific binding. We extracted 23 loci where JUND binds in both h1-hESC and GM12878 and tasked Ledidi with deleting the binding in h1-hESC while preserving it in GM12878.

Despite the complexity of this task, we found that Ledidi was able to perform quite well at each of the loci that it was applied to. Below is an example of one of these loci.




More comprehensively, we found that Ledidi was able to reduce the predicted fold change signal from a median of 13.2 to 5.6 in h1-hESC while keeping the predicted fold change the same in GM12878.

Ledidi requires changes in the hyperparameter settings to work well for this task. In general, the more complex the task, the larger the weight of the functional loss, the lower the learning rate, and the larger the maximum number of iterations should be.