FRANCO CASPE
& ANDREA MARTELLONI

# DEEP LEARNING FOR DSP ENGINEERS:

*CHALLENGES AND TRICKS TO BE PRODUCTIVE WITH AI IN REAL-TIME AUDIO*

# Who we are

## Andrea Martelloni

PhD Candidate at Artificial Intelligence and Music CDT, Queen Mary University of London. DSP engineer, guitar player/songwriter.

## Franco Caspe

PhD Candidate at Artificial Intelligence and Music CDT, Queen Mary University of London. Electronic & Computer Vision Engineer, guitar player.

# What this talk is not about

Not about Deep Learning frameworks.

Not a hands-on DL tutorial.

Not about model optimisation for efficiency.

Not about (offline) Music Information Retrieval (MIR).

Not about LLMs.

Not about symbolic music.

# What this talk is about

We want to make DL for Audio **more approachable** from your point of view.

We will focus on the use of DL for **real-time musical audio**, e.g. a plugin in a DAW.

We will build upon **DSP engineering** knowledge.

We will present **our own experiences** in building DL-based audio tools.

# Outline

Uses of DL in audio

Basic working principles of DL

Steps of the design of a DL model through our experiences:

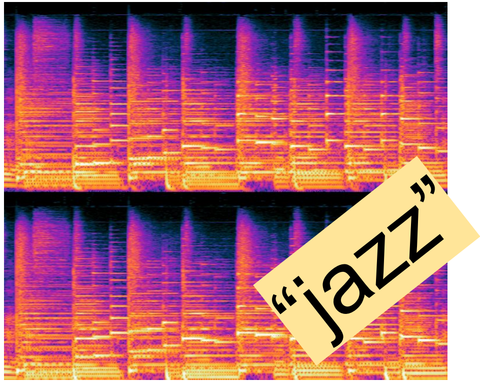
- The HITar
- Bessel’s Trick

Takeaways and lessons learned

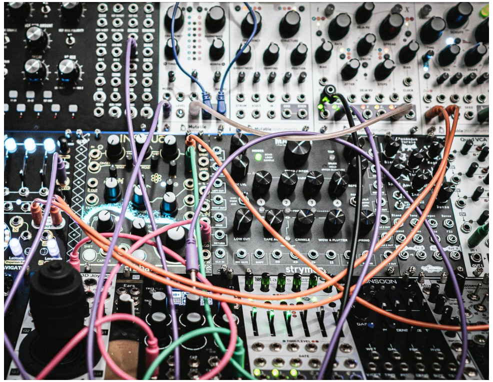
# Uses of Deep Learning in Audio

# What is Deep Learning good for?

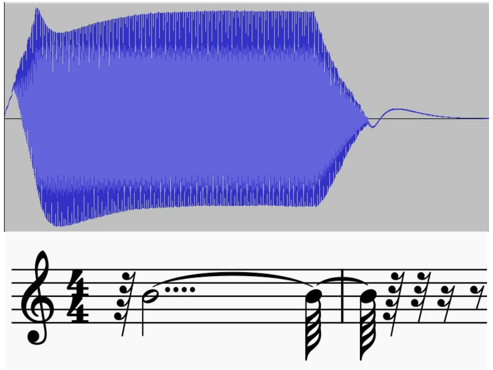
Complex Problem




Complex Solution

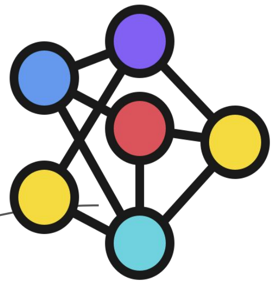


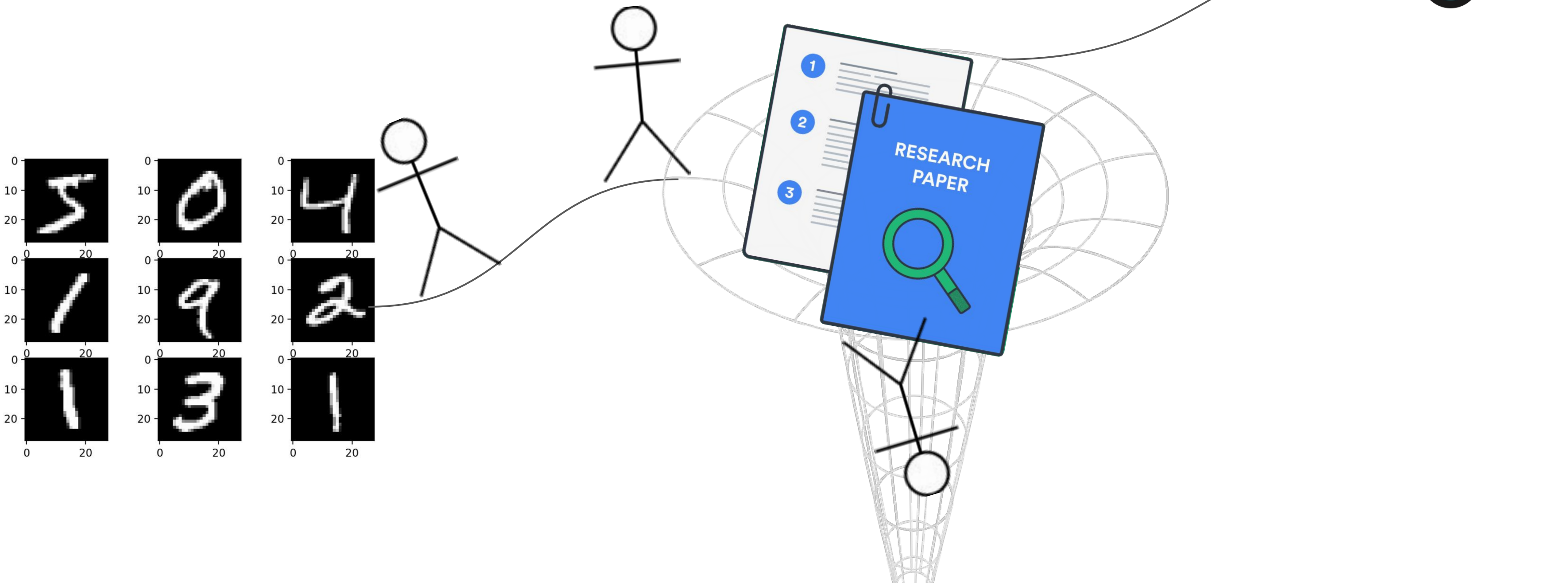
Brittle Solution



We optimize a system with **data,** without explicitly defining **how** to fit that data.

# **Kickstarting a DL 4 Audio Adventure:**
# Don’t work from first principles!



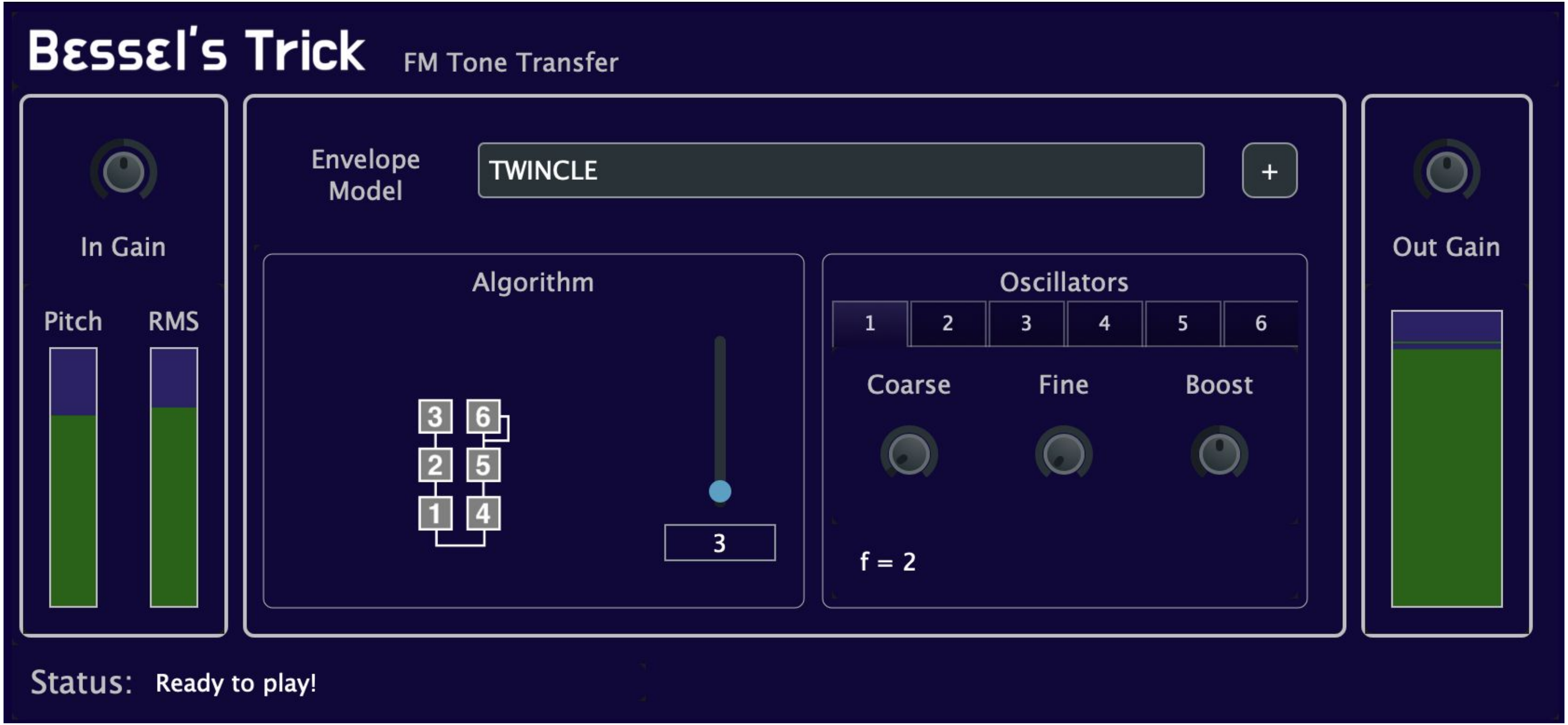


# How to start making sense of Deep Learning?

- We are going to use **informed intuition** from DSP engineering knowledge.

- We are going to walk you through our **projects** and **experiences**,

- Recalling how we used our **DSP background** to **navigate** the complexities of DL.


Bεssεl's Trick
FM Tone Transfer
In Gain
Pitch
RMS
Envelope Model
TWINCLE
+
Algorithm
3
6
2
5
1
4
3
Oscillators
1
2
3
4
5
6
Coarse
Fine
Boost
f = 2
Out Gain
Status:
Ready to play!

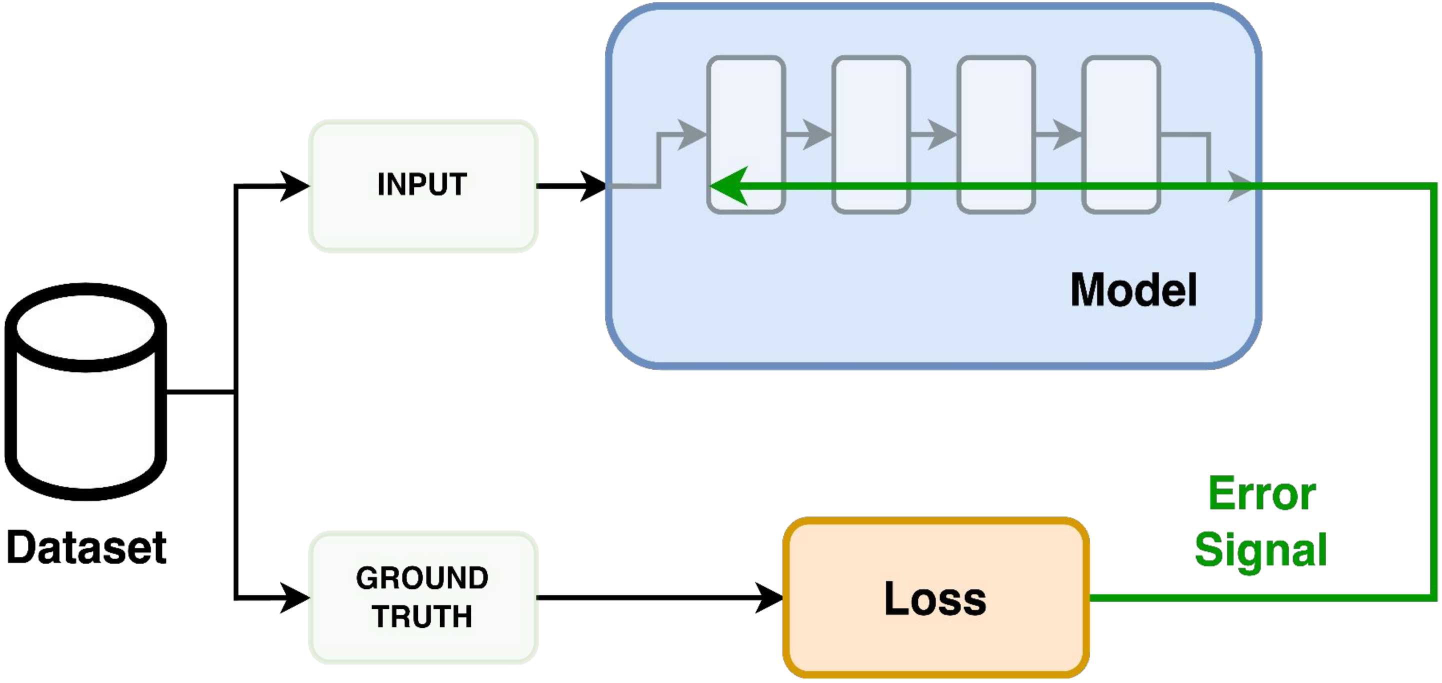
# HITAR

Augmented guitar capturing percussive techniques on the acoustic guitar.

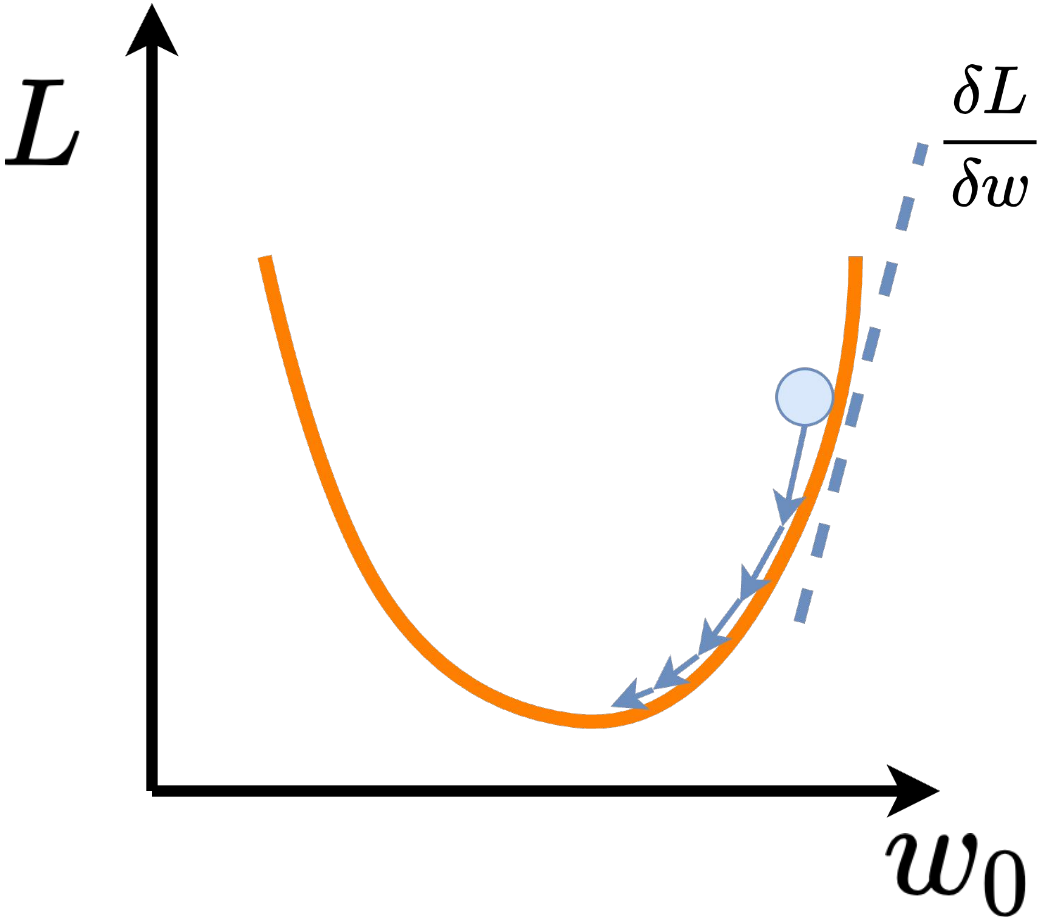
**Expressive gesture representation**: complex/brittle deterministic solutions.

# How Deep Learning works

# Overview

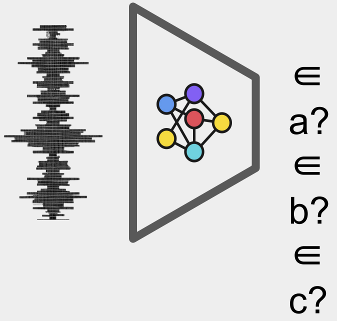


# Training process



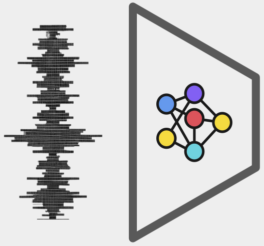
# Relevant ML Tasks for Real-Time Audio

Transcription, discrete gesture mapping.




**Classification**
Alignment to a particular taxonomy

Parameter estimation, waveform prediction



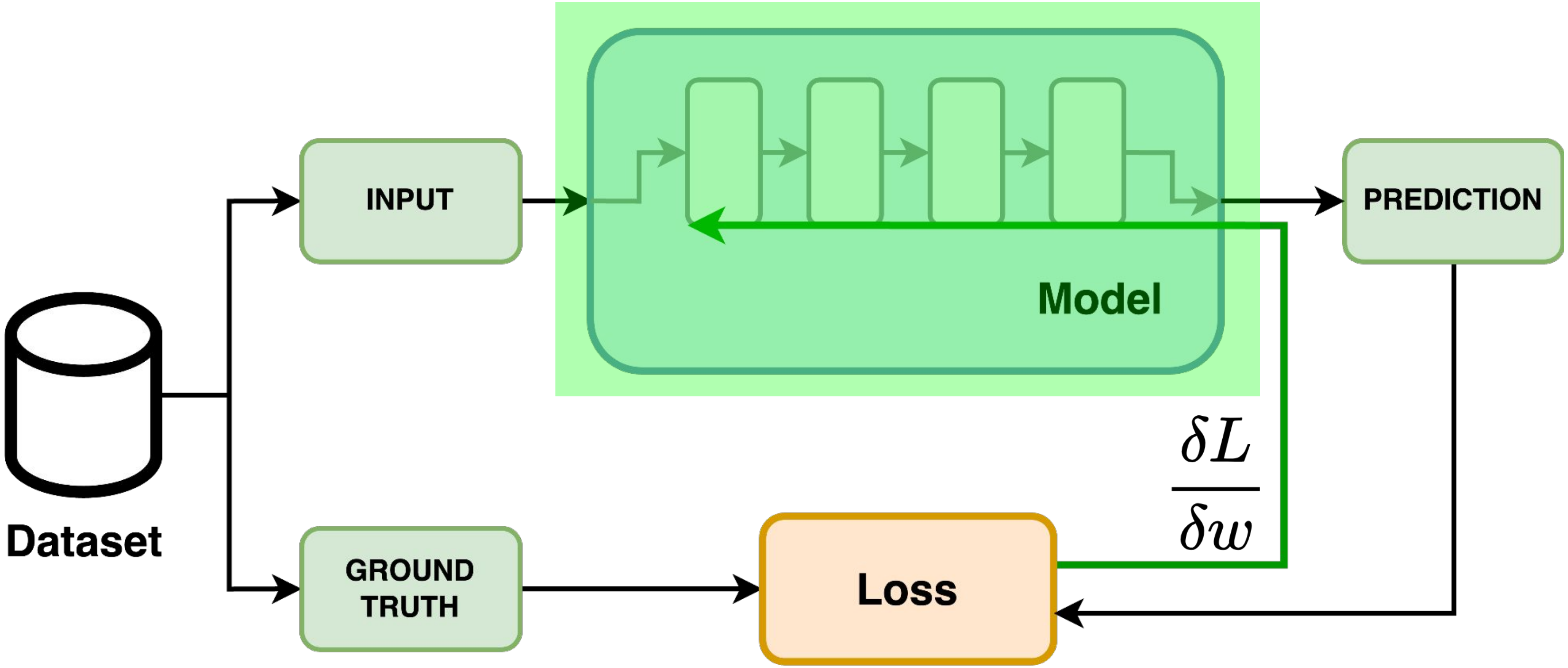
| | |
|---|---|
| a | 0.5 |
| b | 0.2 |
| c | 0.7 |

**Regression**
Estimation of quantities

Compression, Generation, Sampling

**Representation Learning**
Fit a distribution on a dataset

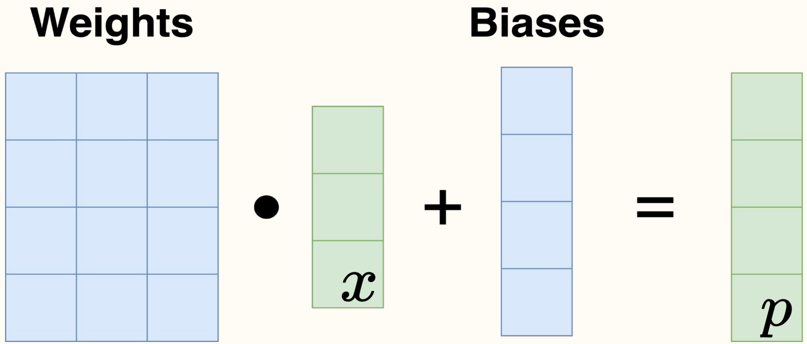
# Overview



INPUT
Model
PREDICTION
Dataset
$\frac{\delta L}{\delta w}$
GROUND TRUTH
Loss

# The Building Blocks

## Matrix Multiplication

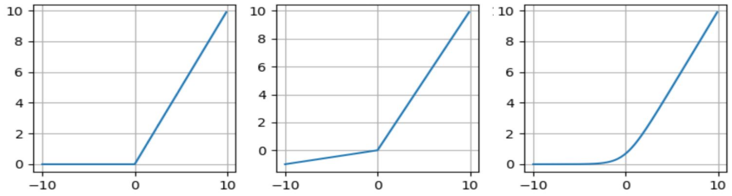


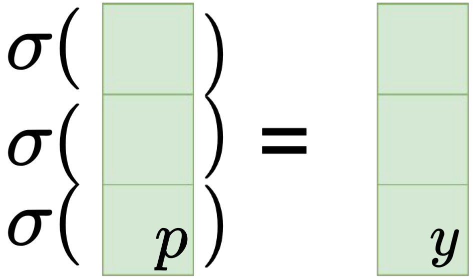

Learnable Parameters

Layer's I/O

$$p = Ax + b$$

## Activation Functions

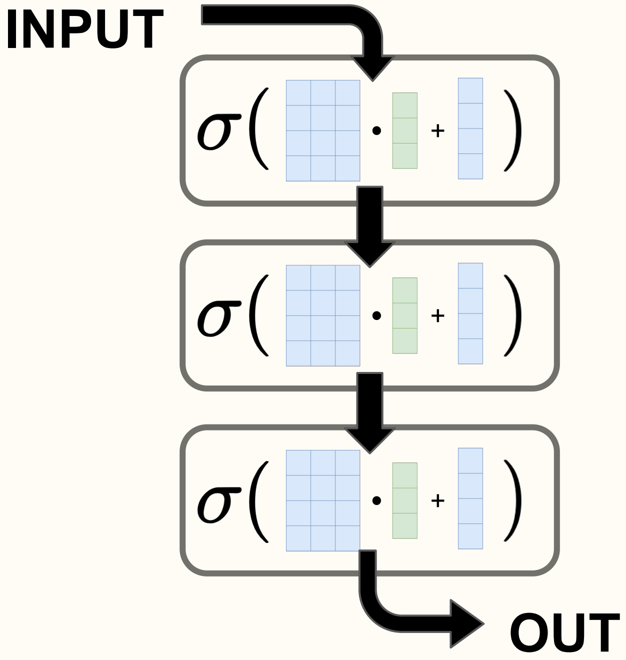





$$y = \sigma(Ax + b)$$

# Architectures

## Stacking Operations



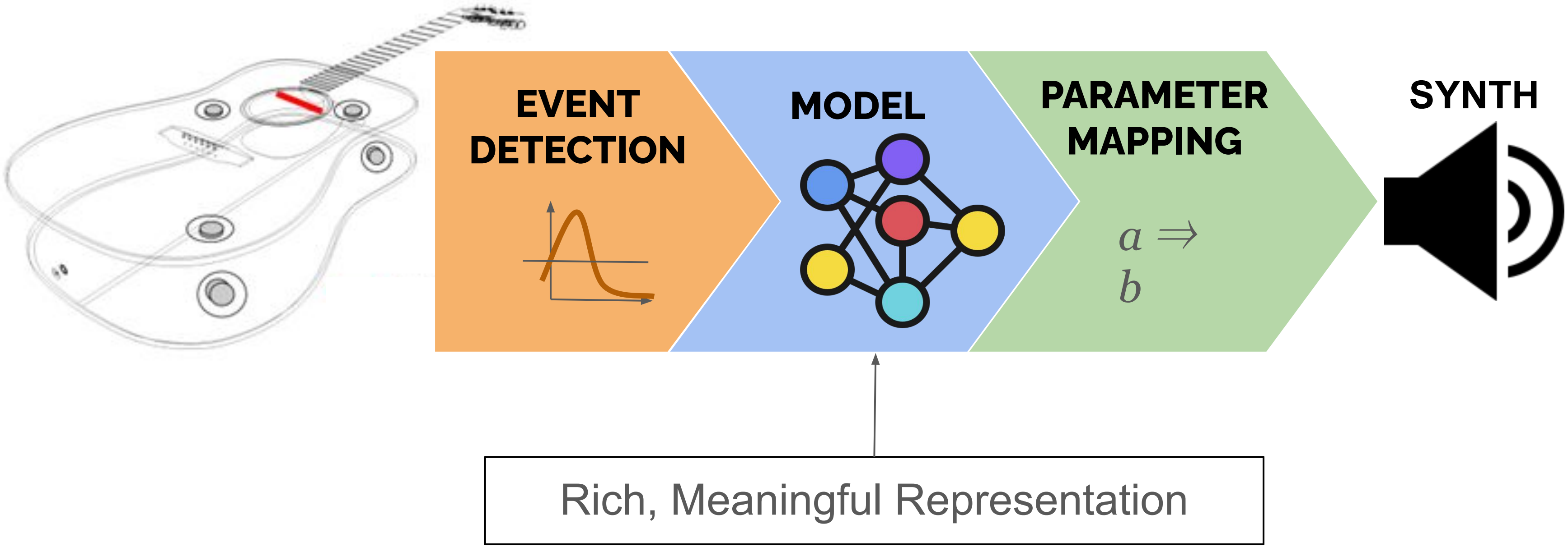

## Layer Design

**Clever ways of doing matrix multiplication to exploit data structures.**

- Dense Layer
- Convolution
- Recurrency

# The HITar

# The HITar's pipeline



# Where I started from

Representation of musical/artistic gestural language with a DL model.

- Real-time Automatic Drum Transcription. (Jacques 2018, MA 2021)
- Semantic description of musical gestures and dance. (Murray-Browne 2021)


Jacques and Roebel, ‘Automatic Drum Transcription with Convolutional Neural Networks’. DAFx 2018.

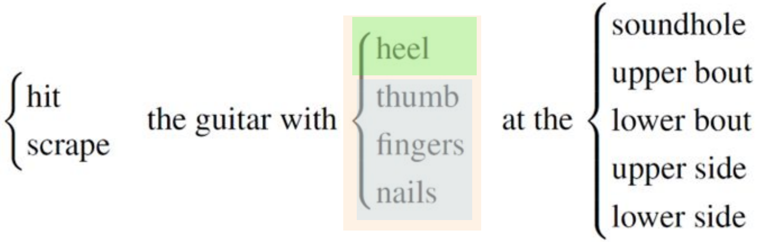
MA, Bhattacharjee, and Rao, ‘Four-Way Classification of Tabla Strokes with Models Adapted from Automatic Drum Transcription’. ISMIR 2021.

Murray-Browne and Tigas, ‘Latent Mappings’. NIME 2021.

# The task behind the HITar

## The “language”

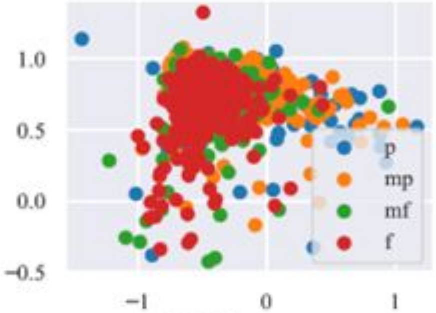
Taxonomy of percussive playing




**CLASSIFICATION**

## The “nuance”

Represent sound events uniquely based on what they mean




**REPRESENTATION LEARNING**

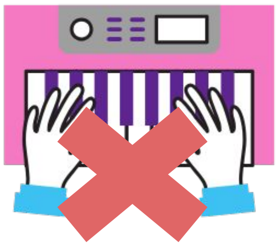
Both things at the same time!

# The Data for the HITar

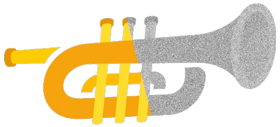
Making good data was the **first**, **longest** and most **expensive** part of my project.

**High Variance**

**Realistic Setting**



**No Confounding Factors**



Make **detailed labels**!

REAScript (and reapy) are your friends… ;)

# The Real-Time Constraint of the HITar

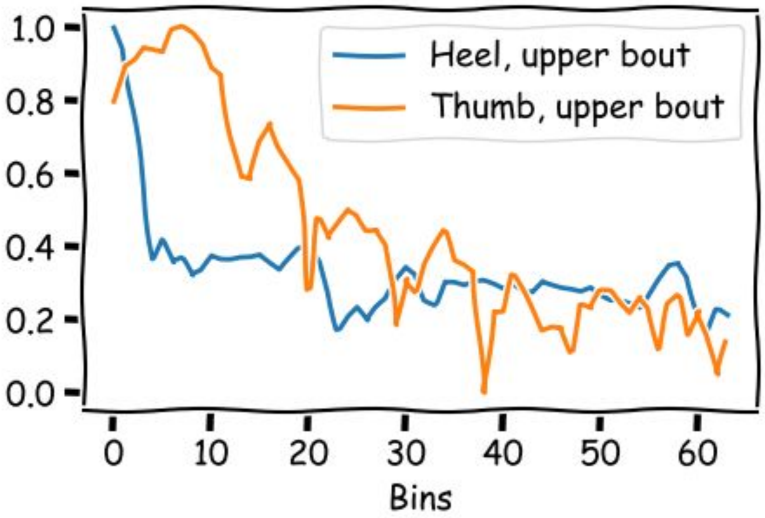

McPherson, Jack, and Moro, ‘Action-Sound Latency: Are Our Tools Fast Enough?’. NIME 2016.

# Data Representation

**Representation**

Input data:

- First 10 ms of the event
- Input representation:
  *Downsampled magnitude spectrum*




**Data augmentation**

Phase inversion, filtering, clipping…

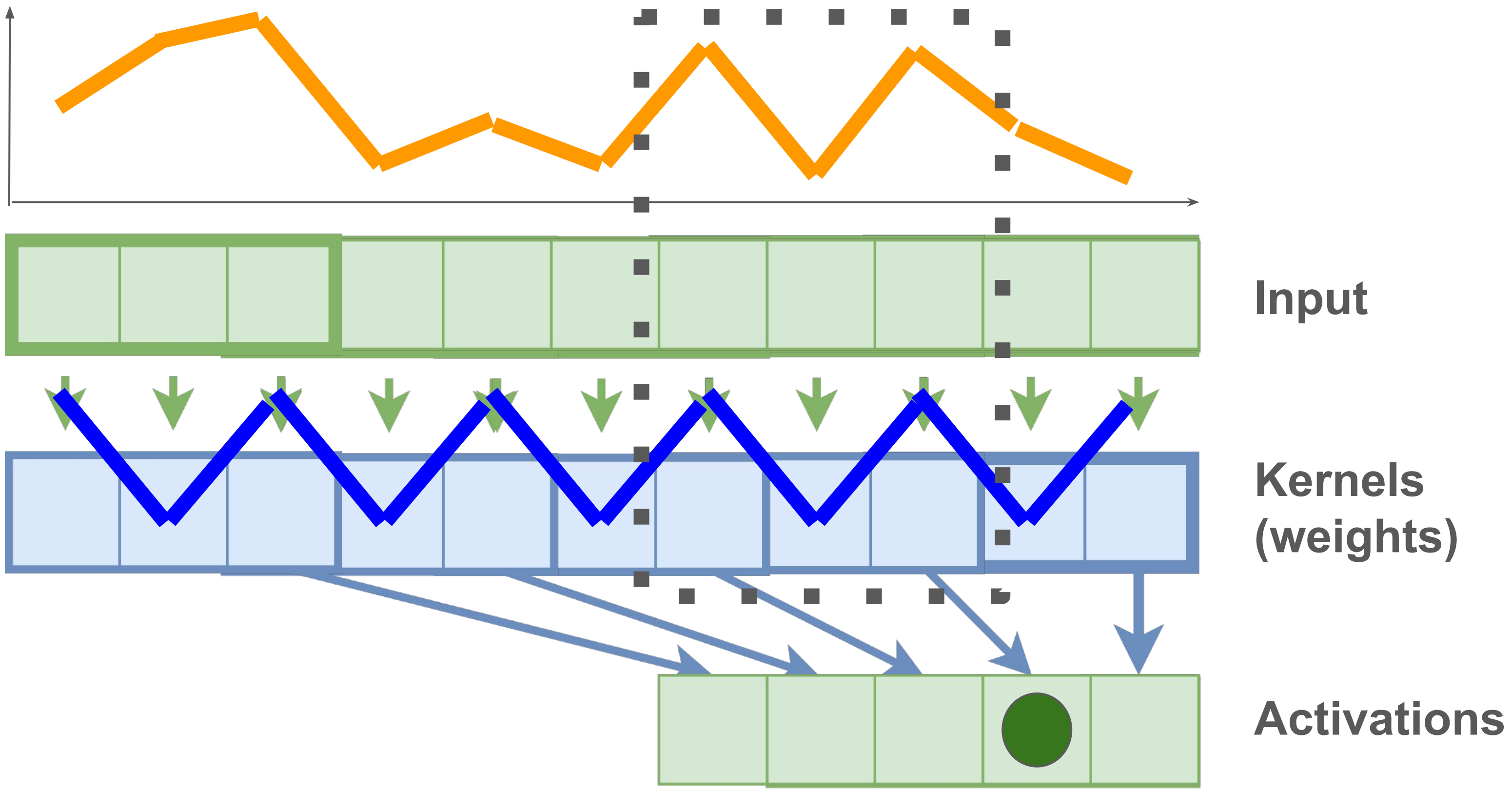
# Intuitions behind the HITar’s model

**Task:** learn a meaningful, readable representation of guitar body hits.

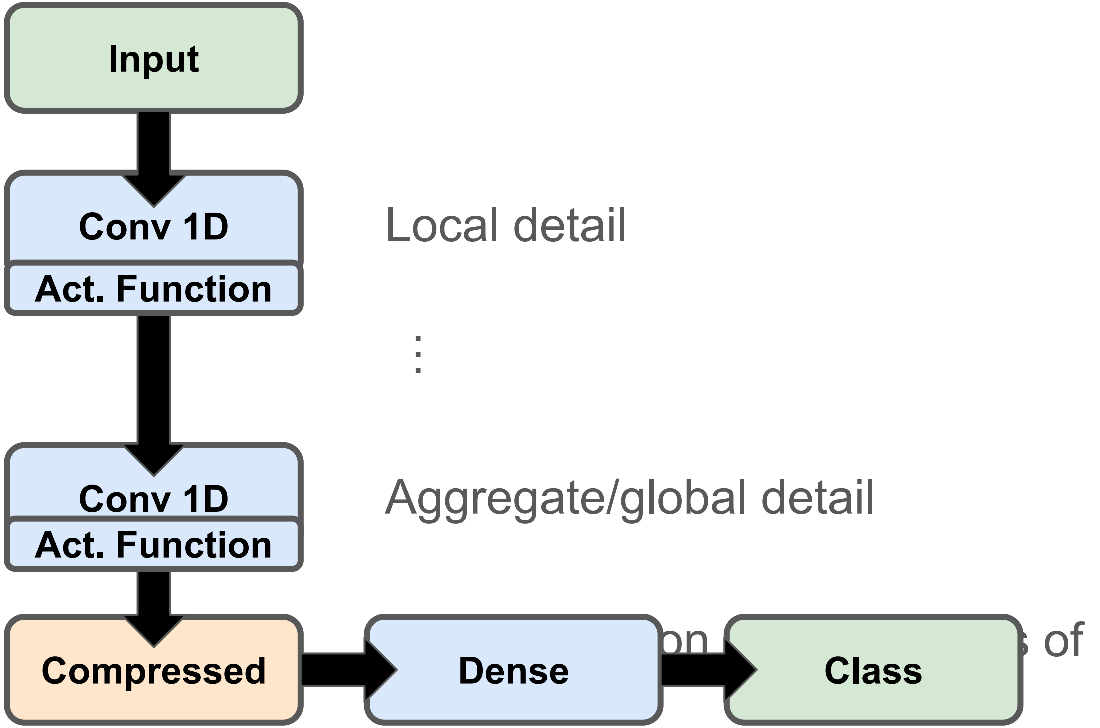
**Input:** magnitude spectrum.

**Intuition:** convolution to find patterns in spectra.

# Extracting features with Convolutional Layers



# Convolutional Encoding



# HITar - features of each <u>unique</u> hit?

In a musical instrument, there are **many ways to play the same thing**.

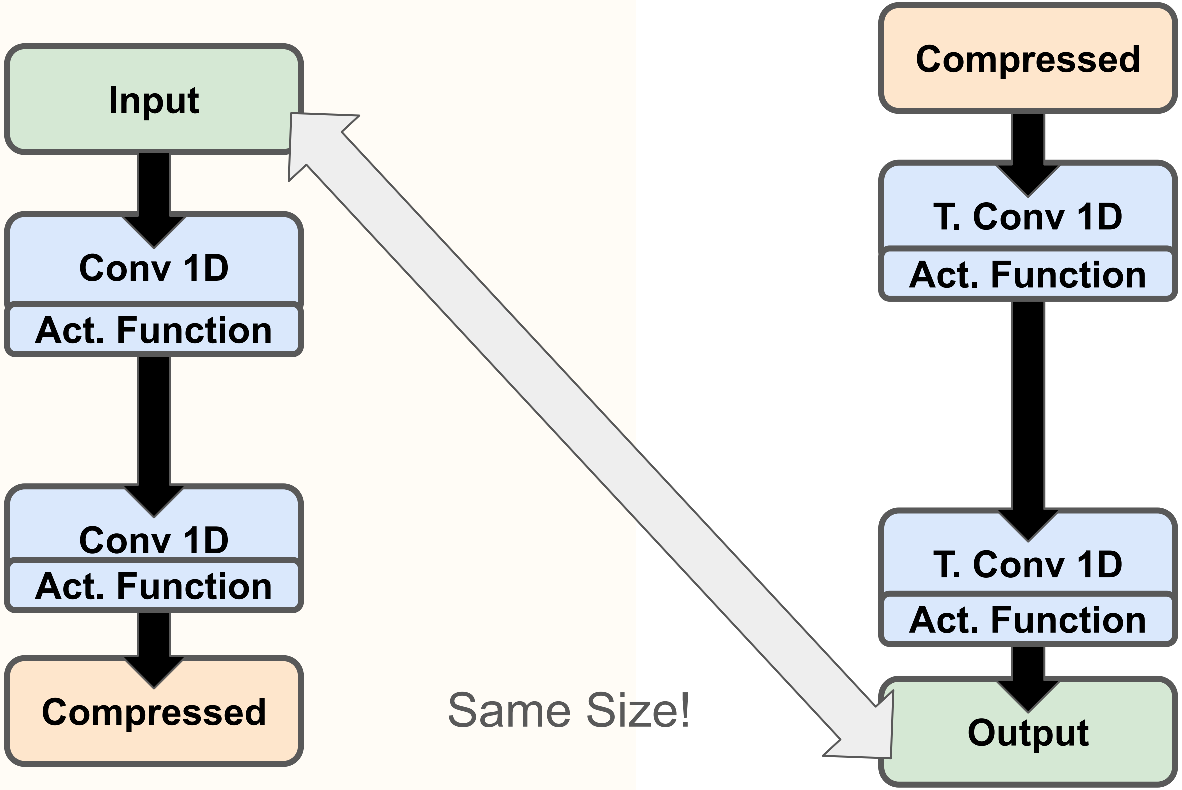
Can we capture that in a NN?

**Intuition:** expand the compressed representation back, to reproduce the input.

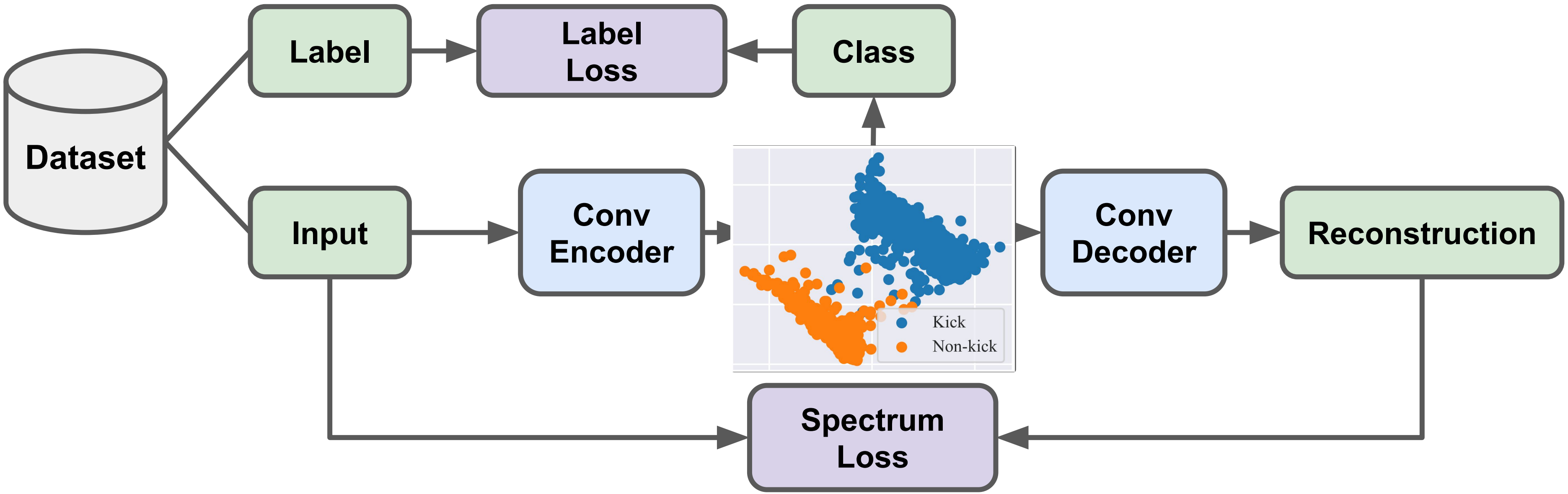
We can make a CNN de-compress data!

# Convolutional Encoding

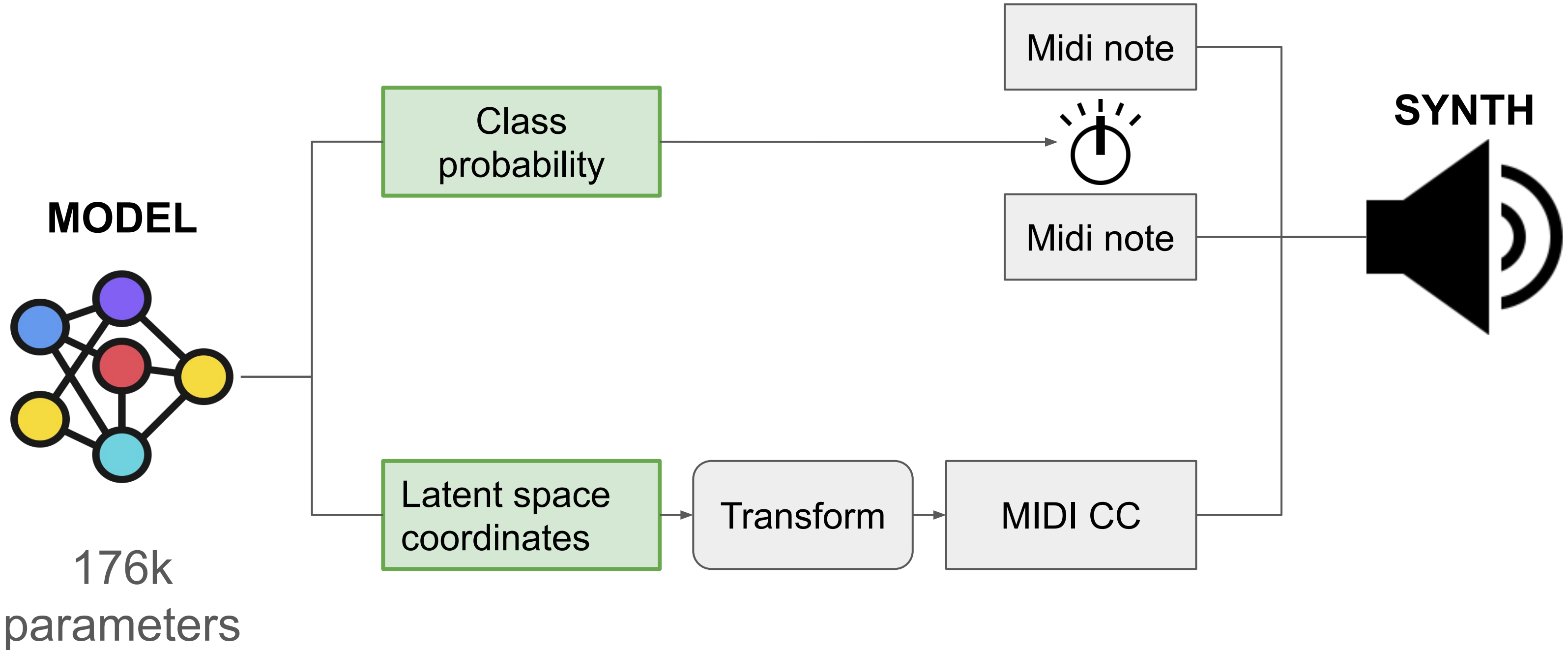
# Convolutional Decoding



# CNN for AutoEncoding



# The Mapping


MODEL
176k
parameters
Class
probability
Latent space
coordinates
Transform
MIDI CC
Midi note
Midi note
SYNTH

WINNER
MIDI®
Innovation
Awards
2023

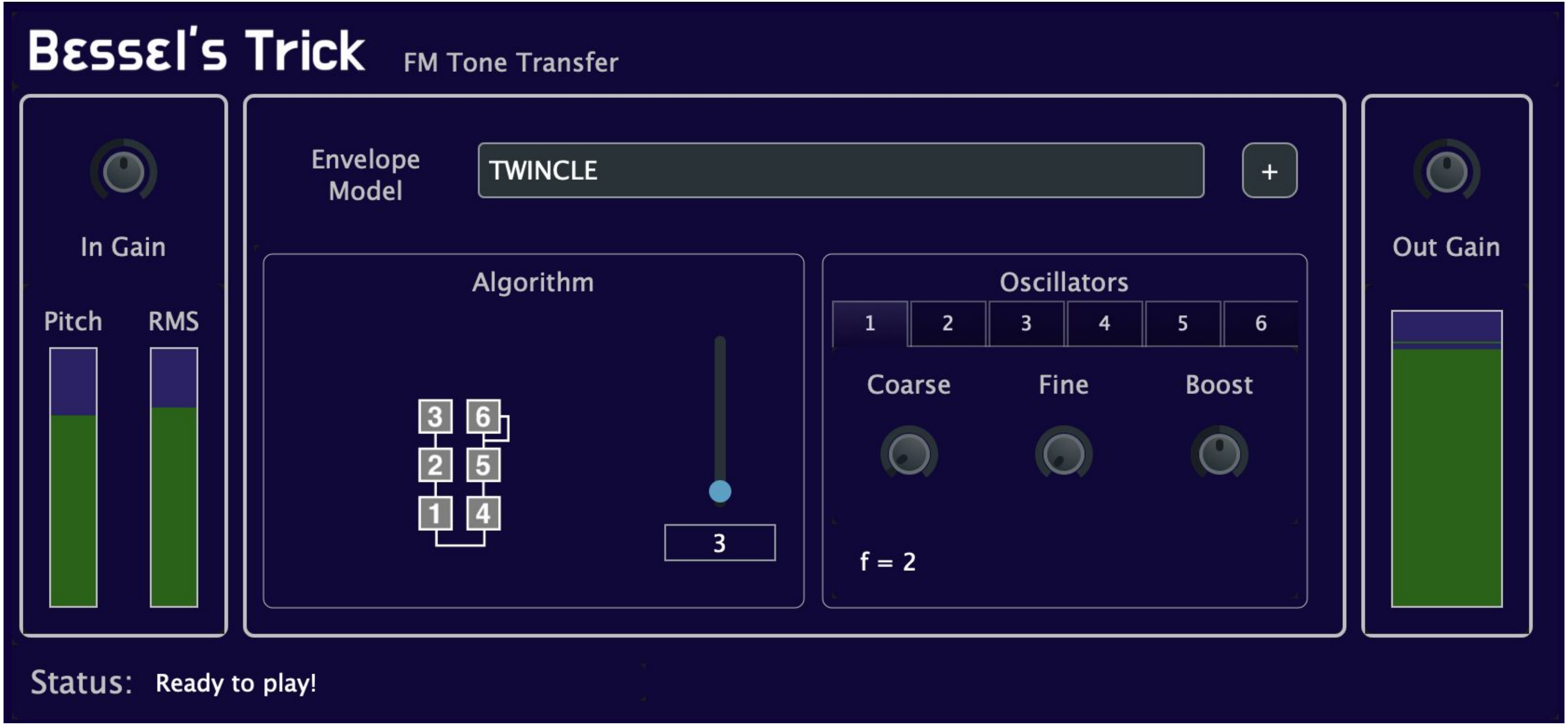
# …in action!

# My Takeaways

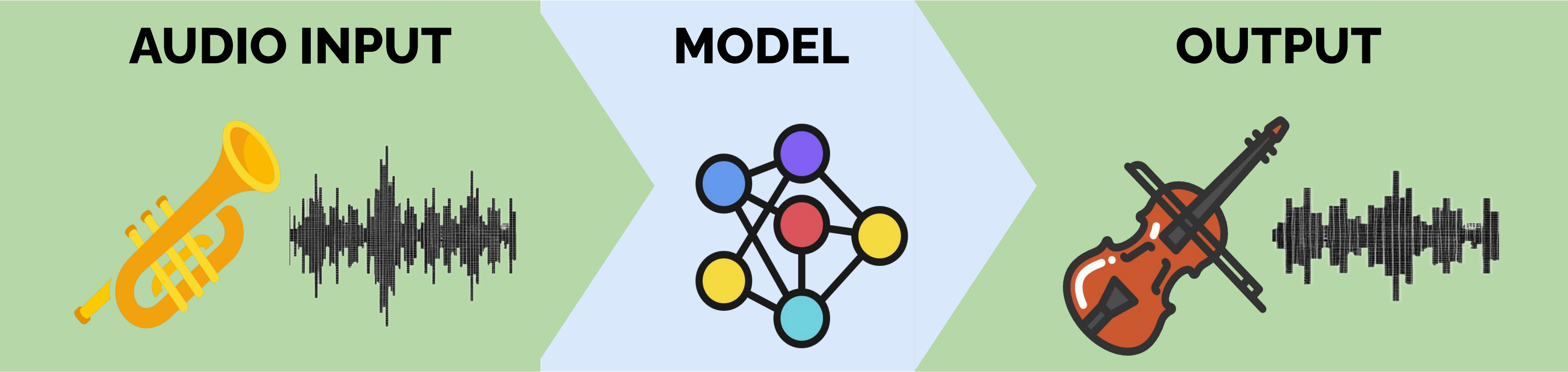
**Good data** is fundamental.

…but DSP engineering knowledge helps with **data augmentation**.

Real-time is a **constraint**, not an optimisation.


Bεssεl's Trick
FM Tone Transfer
In Gain
Pitch
RMS
Envelope Model
TWINCLE
+
Algorithm
3
6
2
5
1
4
3
Oscillators
1
2
3
4
5
6
Coarse
Fine
Boost
f = 2
Out Gain
Status:
Ready to play!

# Tone Transfer



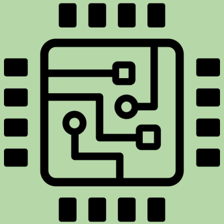
# Design Premises

**PERFORMANCE ORIENTED**

***- Focus on music phrasing and articulation***

**DIVERSE & INTERVENABLE**

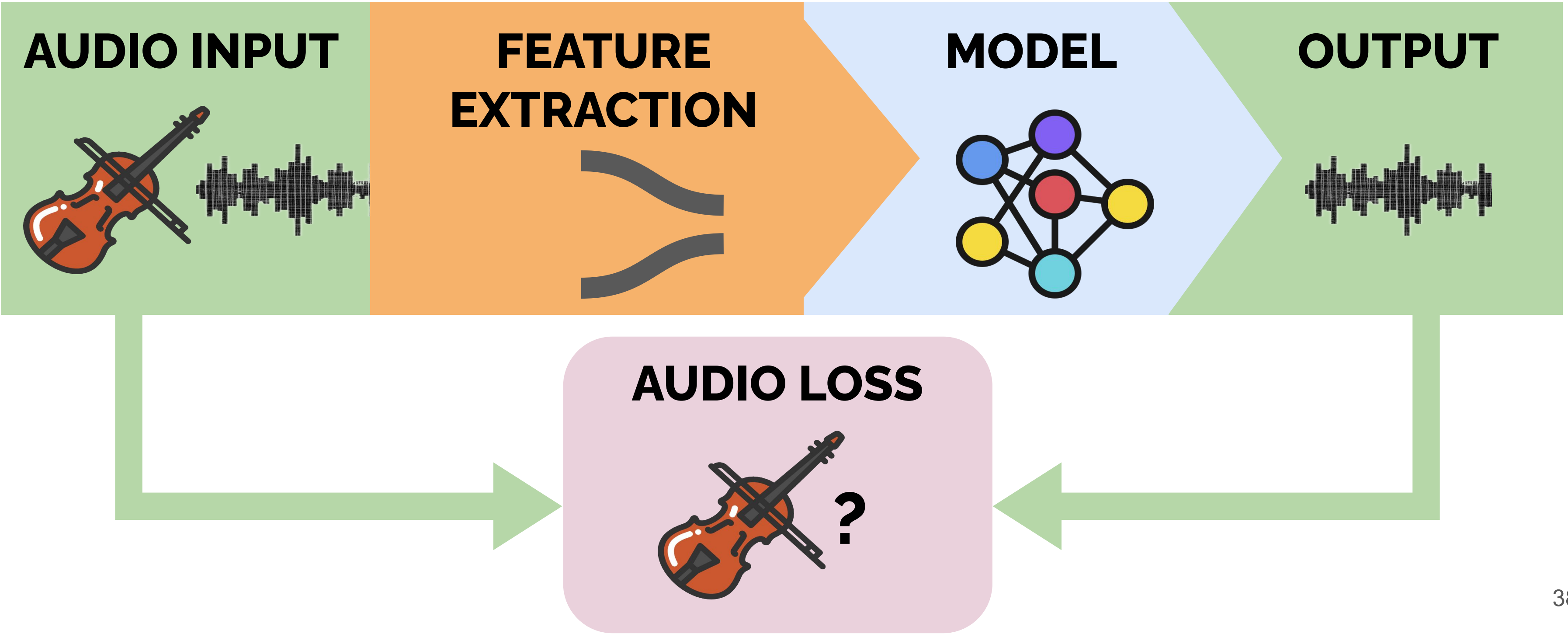
***- Variety of timbres***
***- Ability to edit sounds***



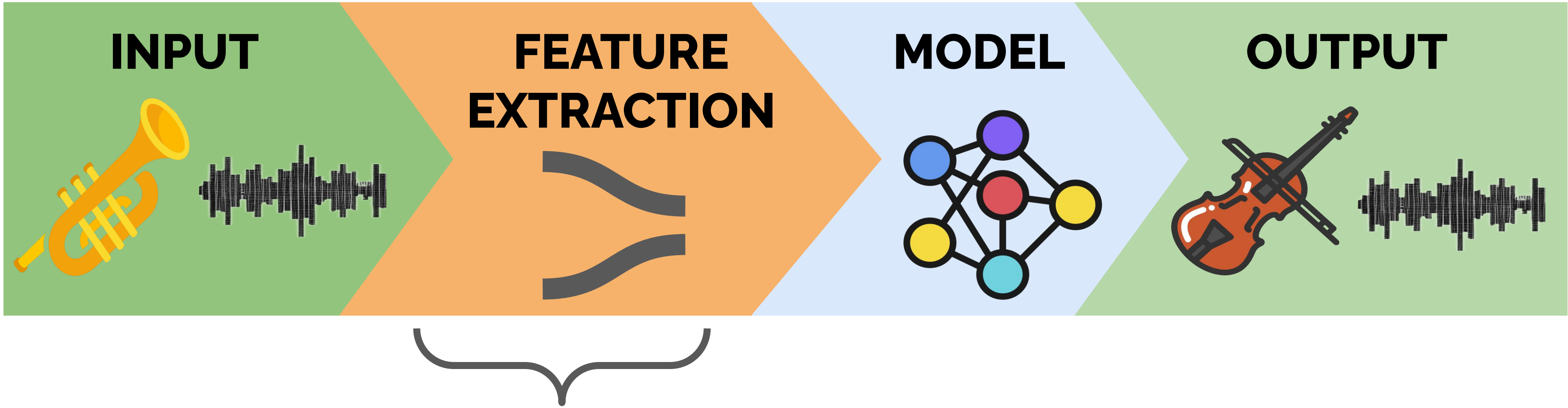
**REAL TIME !**

***- Low-latency inference***

# Tone Transfer Pipeline

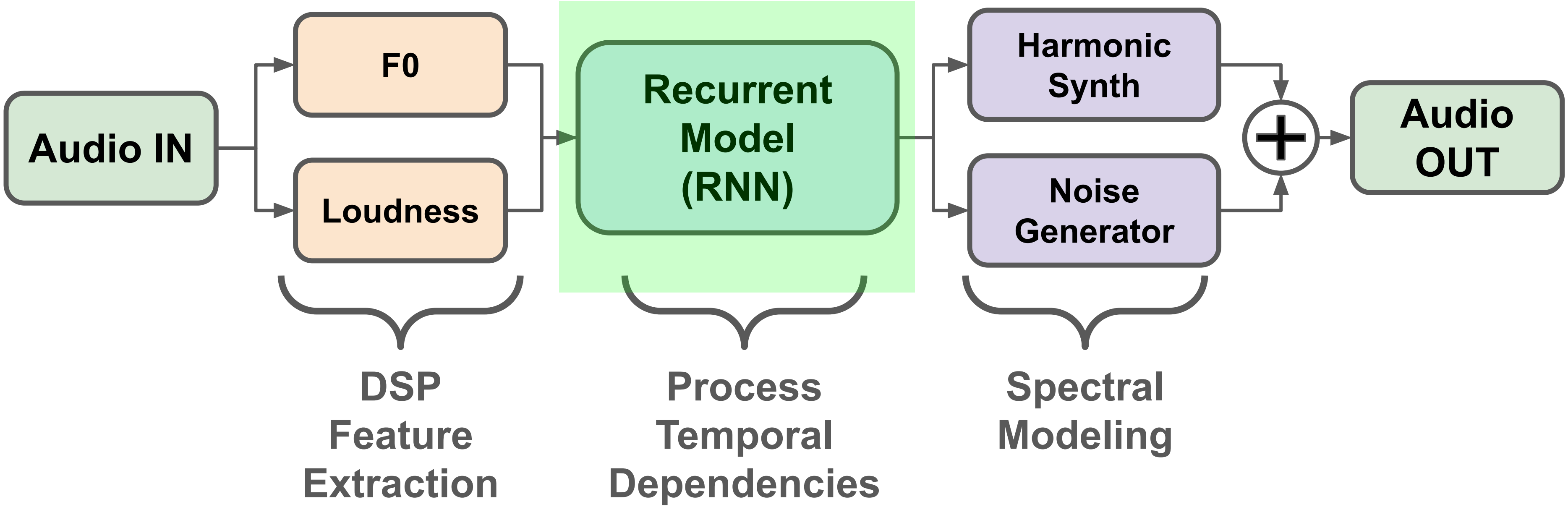


# The claim



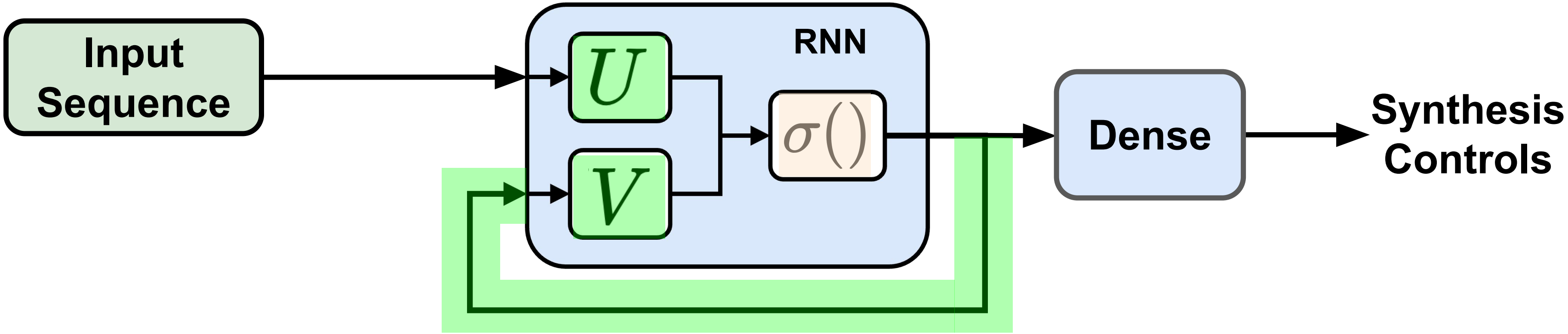

Instrument Independent

# Our Reference Model: The DDSP Decoder (simplified)



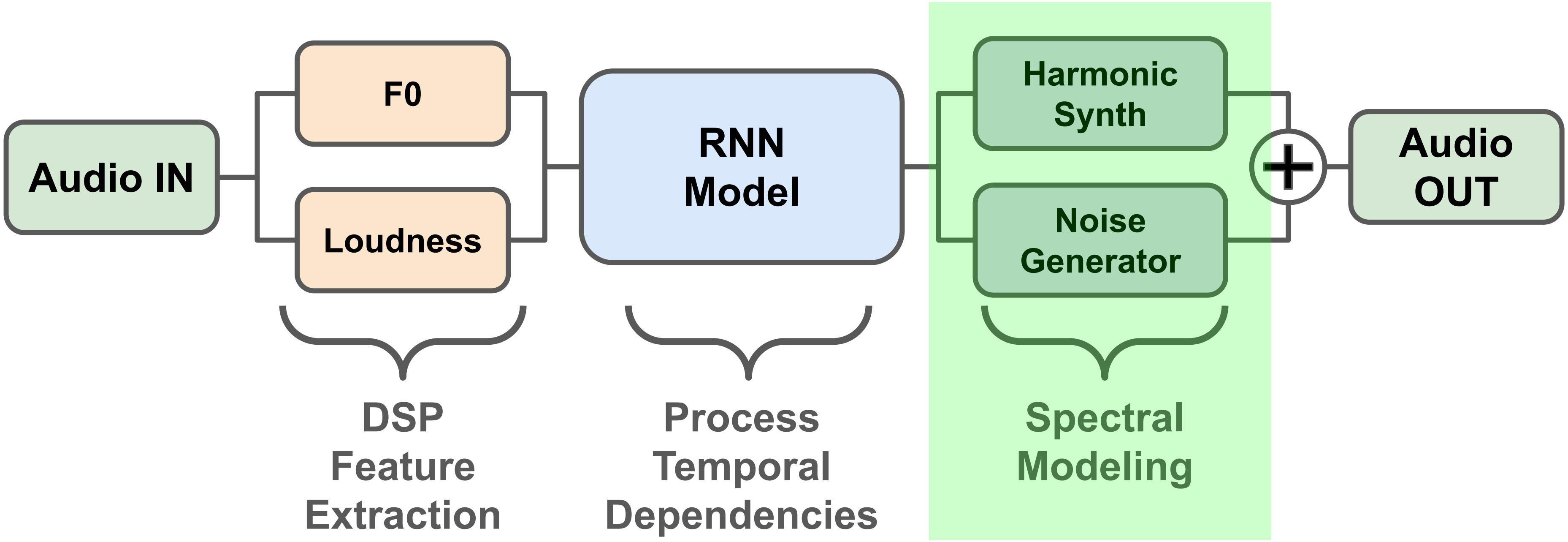

Engel et. al. DDSP: Differentiable Digital Signal Processing. ICLR 2020

# Temporal modelling with RNNs

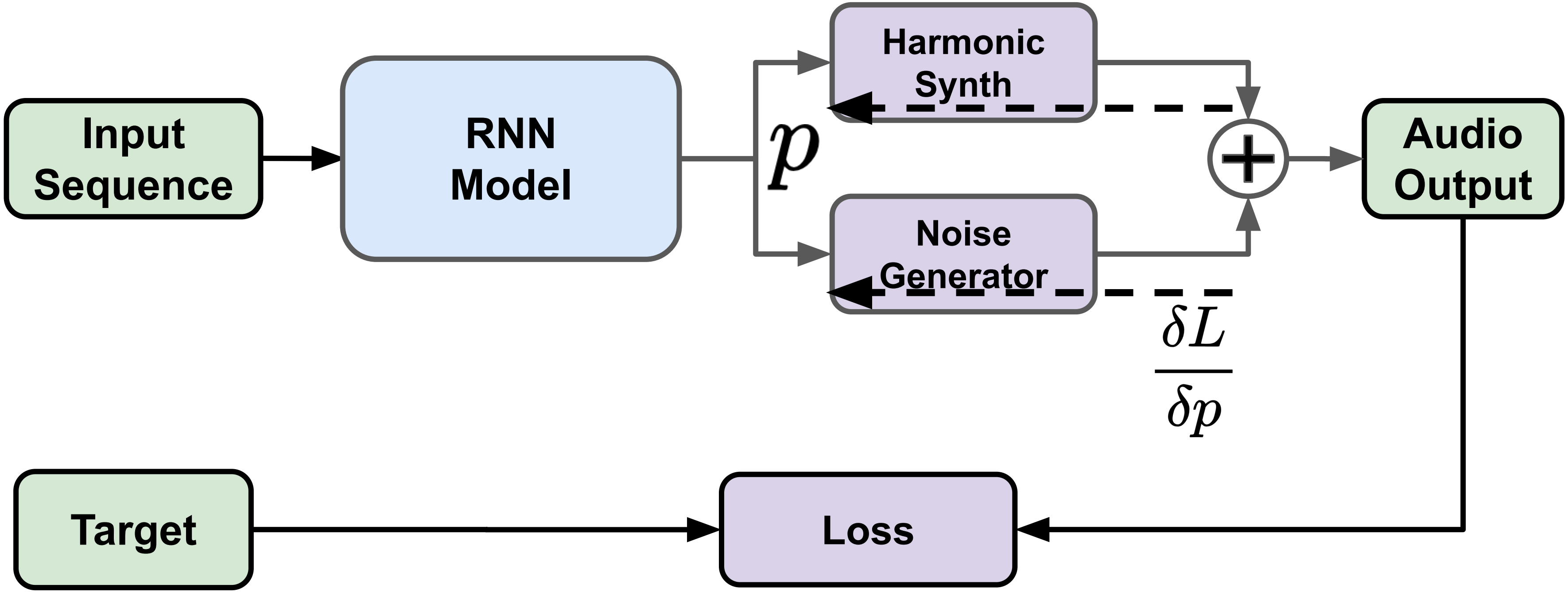



- Temporal aggregation through feedback, like an IIR filter

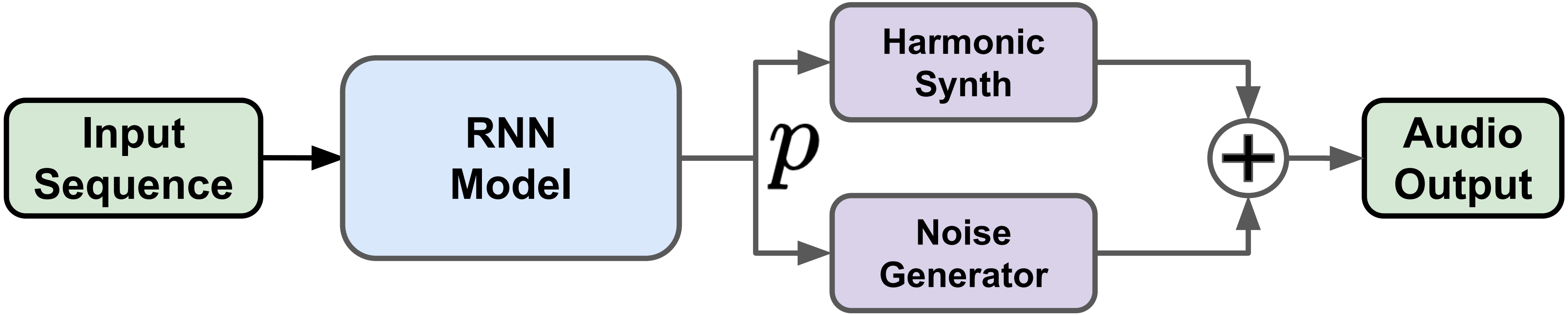
# Reference Tone Transfer Model



# Differentiable Signal Processing



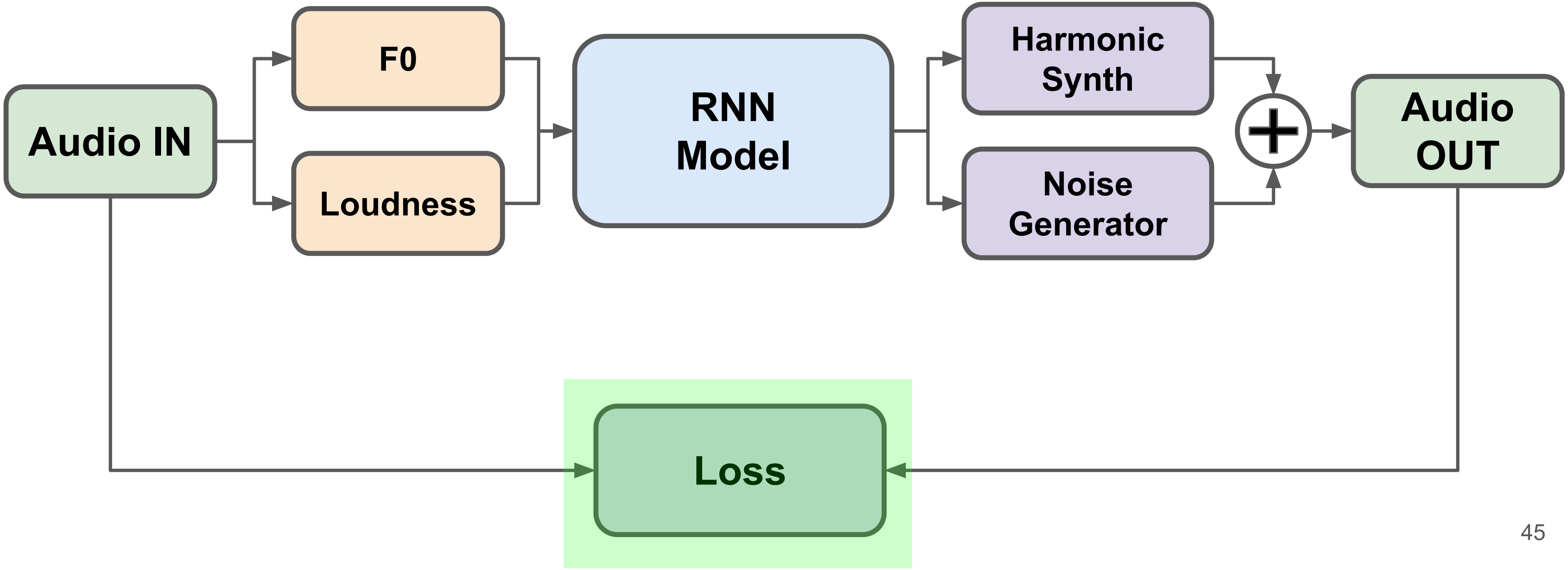
# Differentiable Signal Processing



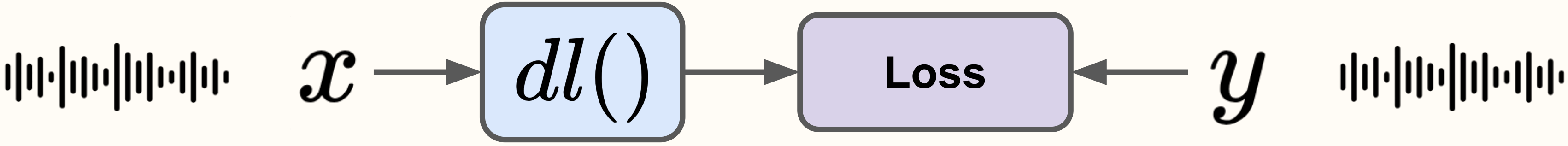

- **Inductive bias towards Signal Generation / Processing.**
- **Oscillators, Filter Banks, Compressors, Waveshapers, FM, Reverb, FFT, Denoisers . . .**

Hayes et. al. “A Review of Differentiable Digital Signal Processing for Music & Speech Synthesis”. Arxiv Preprint
Steinmetz et. al. “High-Fidelity Noise Reduction with Differentiable Signal Processing”. AES

# Reference Tone Transfer Model

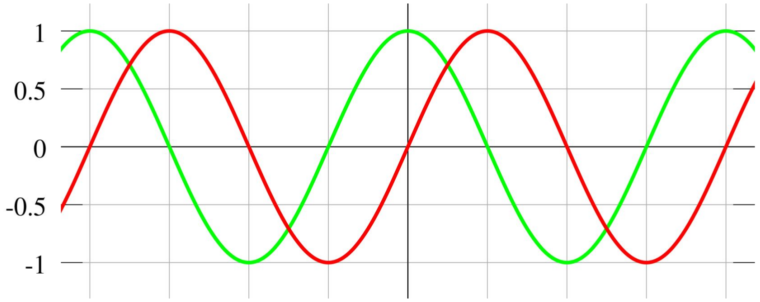


# Loss Functions for Audio



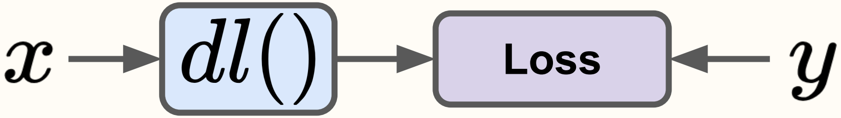

$$L = ||dl(x) - y||$$

- **Strong phase enforcement**




- **Suitable for sample-by-sample generation**

# Loss Functions for Audio




## Sample-by-sample Generation

- MAE / MSE

## Block-by-Block Generation

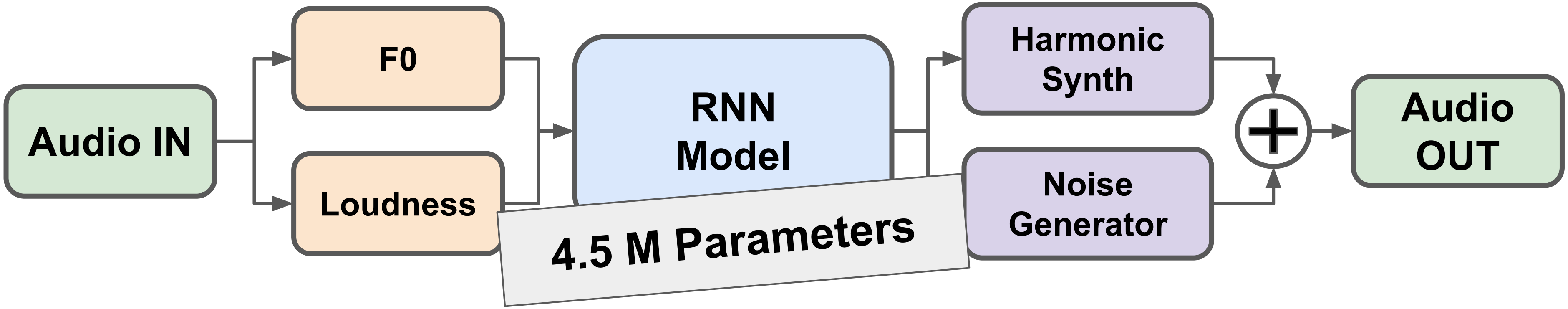
- Multi Scale Spectral Loss

$$L = ||S_{g(x)} - S_y||_{1024} + ||S_{g(x)} - S_y||_{256} + ||S_{g(x)} - S_y||_{64}$$

***Higher Frequency Resolution***

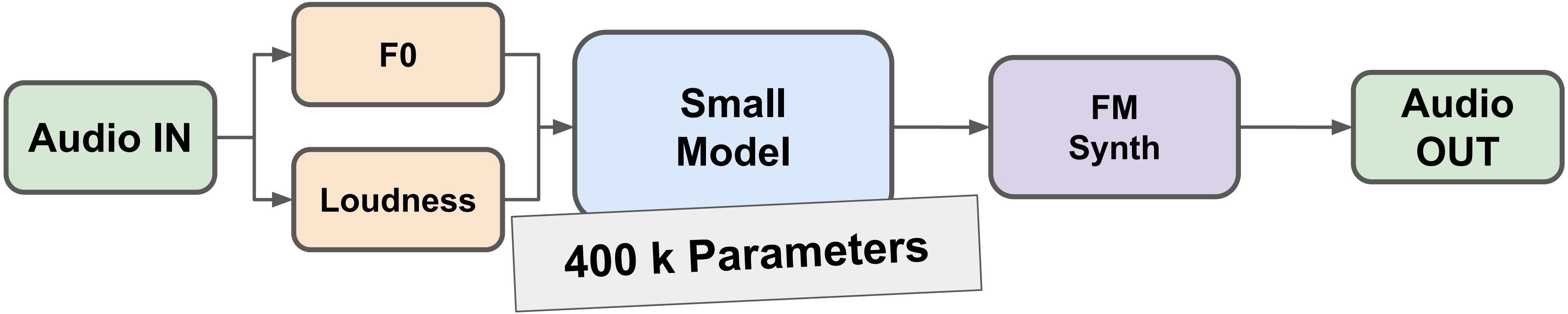
***Higher Temporal Resolution***

# Reference Tone Transfer Model




- **Block by Block Operation**
  - **Suitable for Real-Time.**

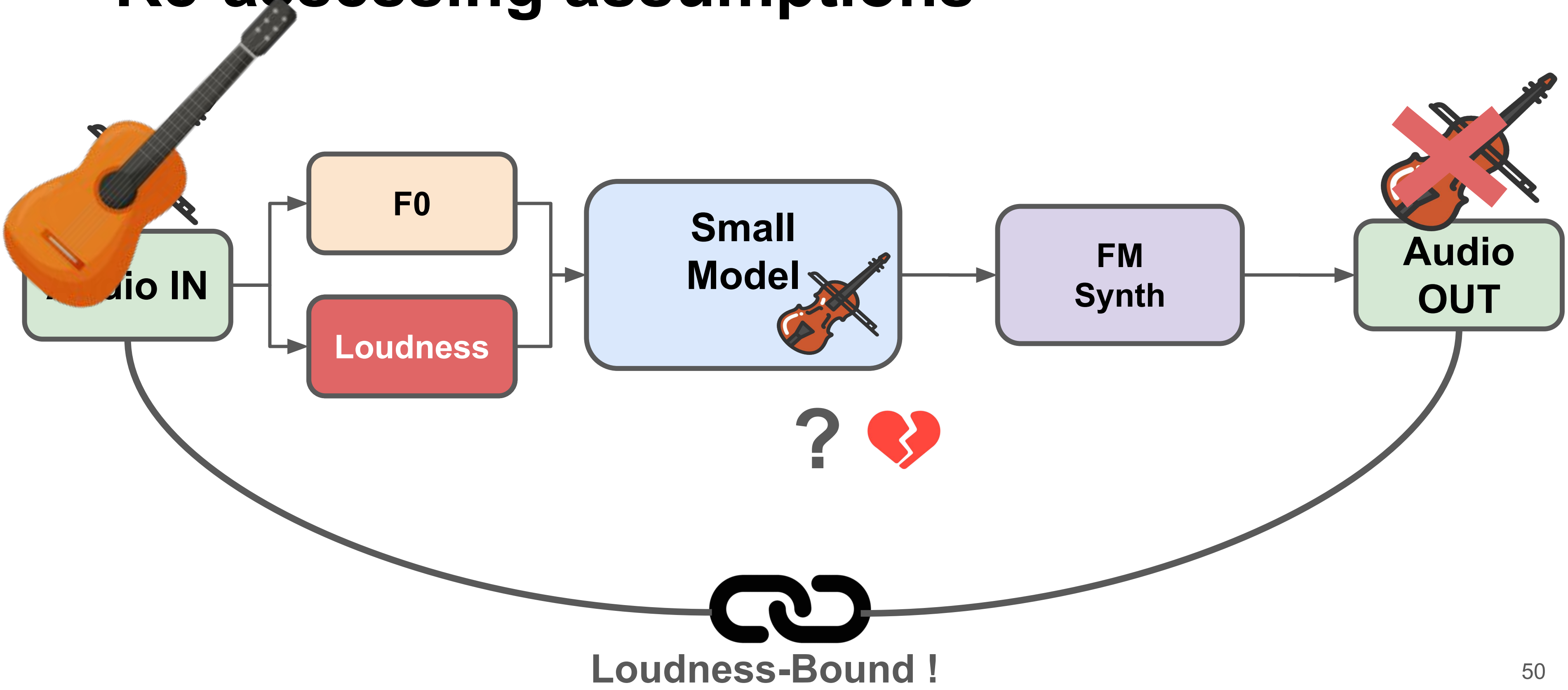
# Idea: use a compact yet versatile synth (FM)



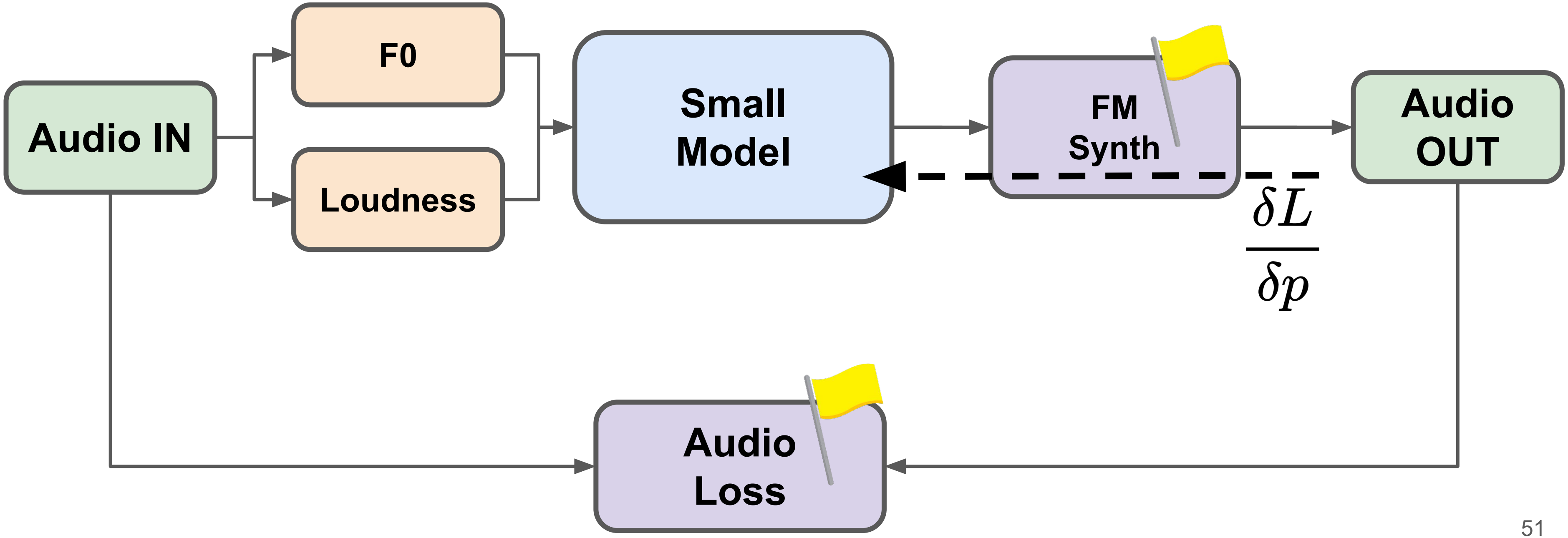

- Works great on the test set!
- 10x Smaller, works in real time!
- NOT Great during live use.

Caspe et. al. “DDX7: Differentiable FM Synthesis of Musical Instrument Sounds” ISMIR 2022

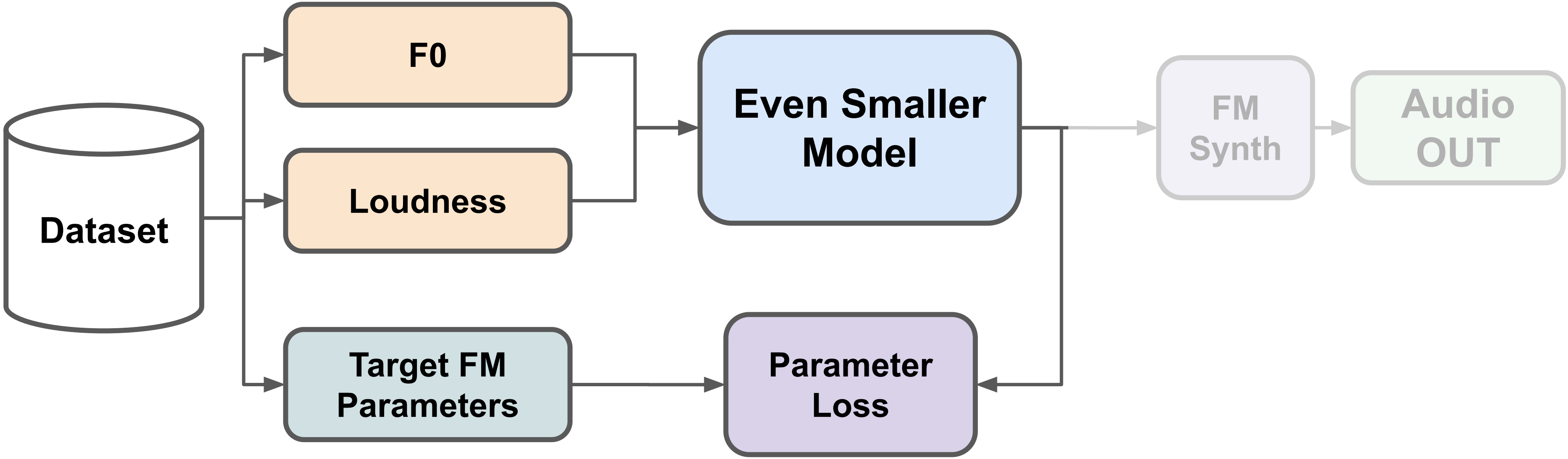
# Re-assessing assumptions



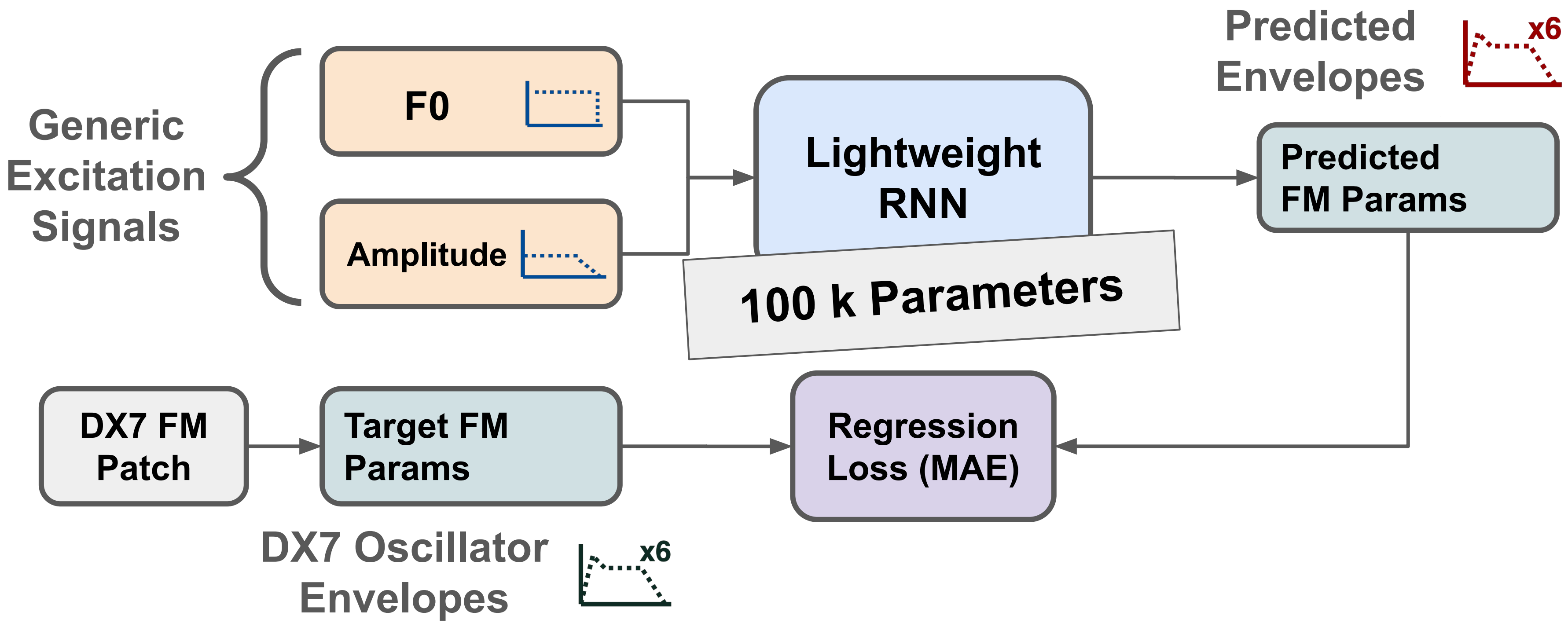
# Using FM Synthesis with Audio Losses



# Skipping the Synth during Training

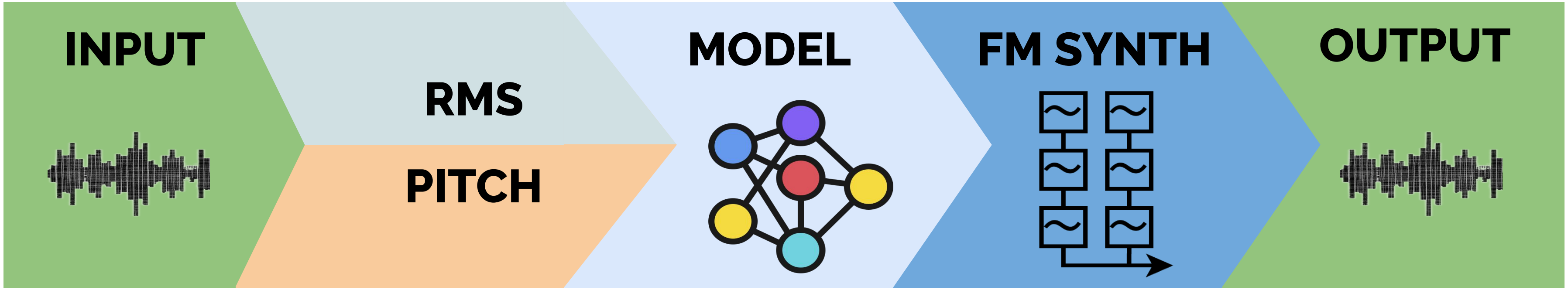


# Skipping the Synth during Training




Caspe et. al. “FM Tone Transfer with  Envelope Learning” Audio Mostly 2023

# Bessel's Trick
# FM Tone Transfer Pipeline



See it on Github!

# Takeaways

Have multiple loss functions at hand.

DSP Primitives can help simplify the network architecture.

Think of what is **explicitly** and **implicitly** learned in the model.

# Final Remarks

# Takeaways

- There’s a strong DSP intuition behind DL for Audio.
- Clarify your real-time constraints early.
- Don’t work from first principles.
- Be data-driven
    - Think about WHAT data.
    - Think about HOW the data is treated.

# The Augmented Instruments Lab is recruiting!

**1-2 PhD positions**

**1 postdoc position**

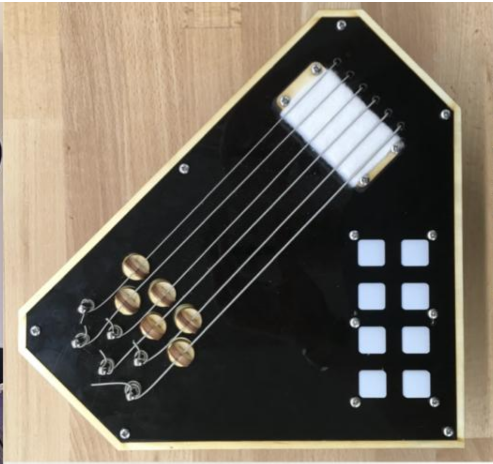
Areas of interest:

- Musical interaction
- Embedded hardware for audio and sensors
- Critical perspectives on technology and culture

Applications due December/January

See **instrumentslab.org**

Email: **andrew.mcpherson@imperial.ac.uk**



# Thank you

@the_hitar

@caspefranco