HaydnModels: Symbolic Music Composition with Ensemble Models

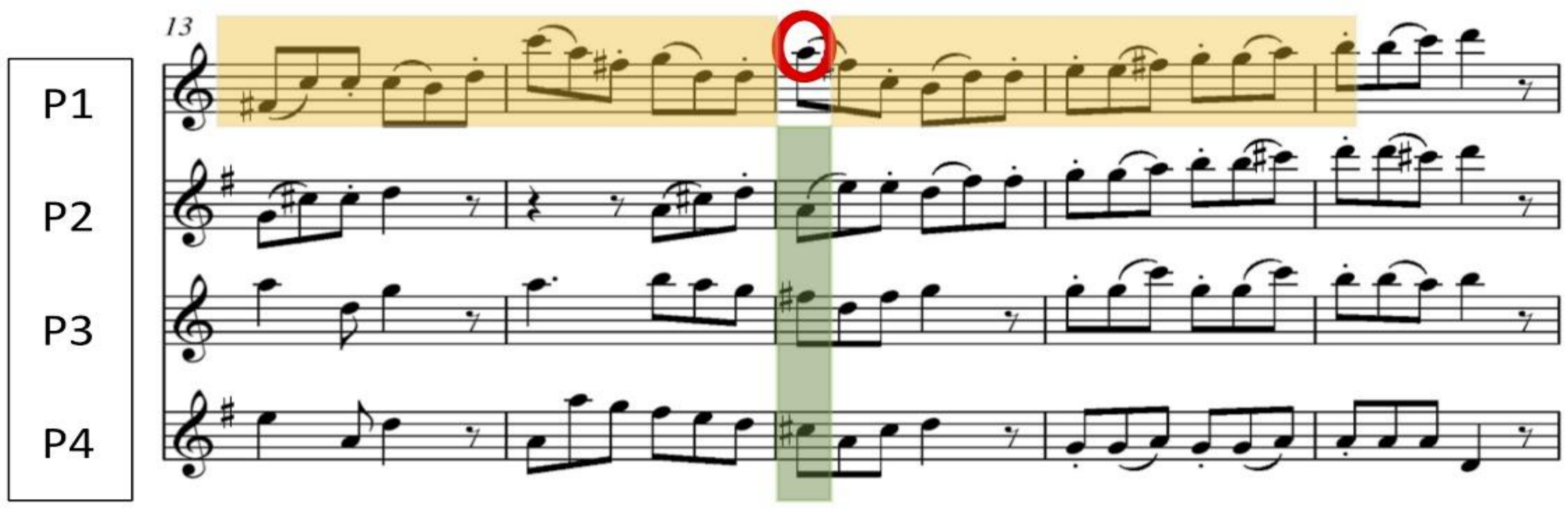
Ziqiang Guan, Xiaohan Ding

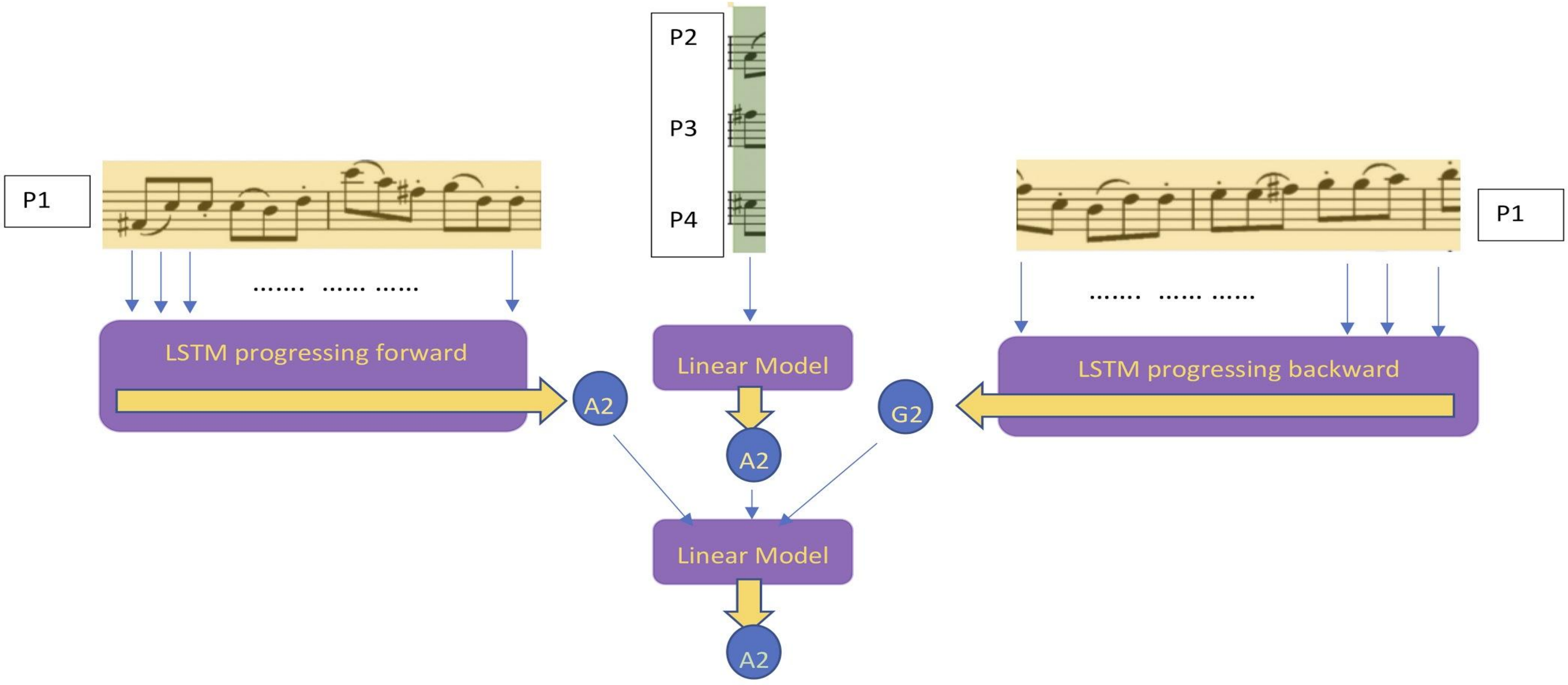
University of Massachusetts, Amherst

Introduction:

In this project, we explore new ways to let computer learn how to generate music. HaydnModel is an ensemble of models that can generate music sounds like composed by Haydn. In HaydnModel, we experimented with a new way to represent music notes, and a new way to teach computer to learn music composition.

Soprano Saxophone
Alto Saxophone
Tenor Saxophone
Baritone Saxophone
7
13





Data

- 81 MIDI files of string quartets by Haydn,
- 12 transpositions per piece to augment dataset,
- ~80,000 sequences of length 64 as training data,
- ~20,000 sequences of length 64 as validation data,
- embedding layer to learn note representation.

For a given chunk of four-part music X, parts $\{X_1, X_2, X_3, X_4\}$ and the notes $x_i^t \in X_i$ at various time step t can be visualized as follows,

$$\begin{aligned} X_1 &= \{..., x_1^{t-2}, x_1^{t-1,}, x_1^t, x_1^{t+1}, x_1^{t+2}, ...\} \\ X_2 &= \{..., x_2^{t-2}, x_2^{t-1,}, x_2^t, x_2^{t+1}, x_2^{t+2}, ...\} \\ X_3 &= \{..., x_3^{t-2}, x_3^{t-1,}, x_3^t, x_3^{t+1}, x_3^{t+2}, ...\} \\ X_4 &= \{..., x_4^{t-2}, x_4^{t-1,}, x_4^t, x_4^{t+1}, x_4^{t+2}, ...\} \end{aligned} \tag{1}$$

Relationships exists both across the time steps of the same part, as well as the across parts at the same time step. The first can be intuitively understood as melody, and the latter harmony.

To model these relationships, two LSTMs networks (the Note models, F for the forward model and B for the backward model) and one partially connected network (the Harmony model H) is used to model the conditional probability distribution of each x_i^t,

$$\begin{aligned} F(\{..., x_m^{t-3}, x_m^{t-2}, x_m^{t-1}\}) =& P(x_m^t = n \mid ..., x_1^{t-3}, x_1^{t-2}, x_1^{t-1,}) \\ B(\{x_m^{t+1}, x_m^{t+2}, x_m^{t+3}, ...\}) =& P(x_m^t = n \mid x_m^{t+1}, x_m^{t+2}, x_m^{t+3}, ...) \\ H(\{x_j^t, x_k^t, x_l^t\}) =& P(x_m^t = n \mid x_j^t, x_k^t, x_l^t) \\ \{j, k, l, m\} \in& \{1, 2, 3, 4\} \end{aligned} \tag{2}$$

Lastly, a fully-connected judge network J takes in three suggestions and comes up with a final decision,

$$J(F, B, H) = P(x_m^t = n \mid F, B, H) \tag{3}$$

Ensemble methods

Each note generated by the model is a combined learned results from four models, two LSTM and two linear models. Given a sequence, the note needs to be predicted is at the center of the sequence.

One LSTM will progress from the beginning of the sequence towards the center, and generate a predicted result of the center note. This models the forward music composition.

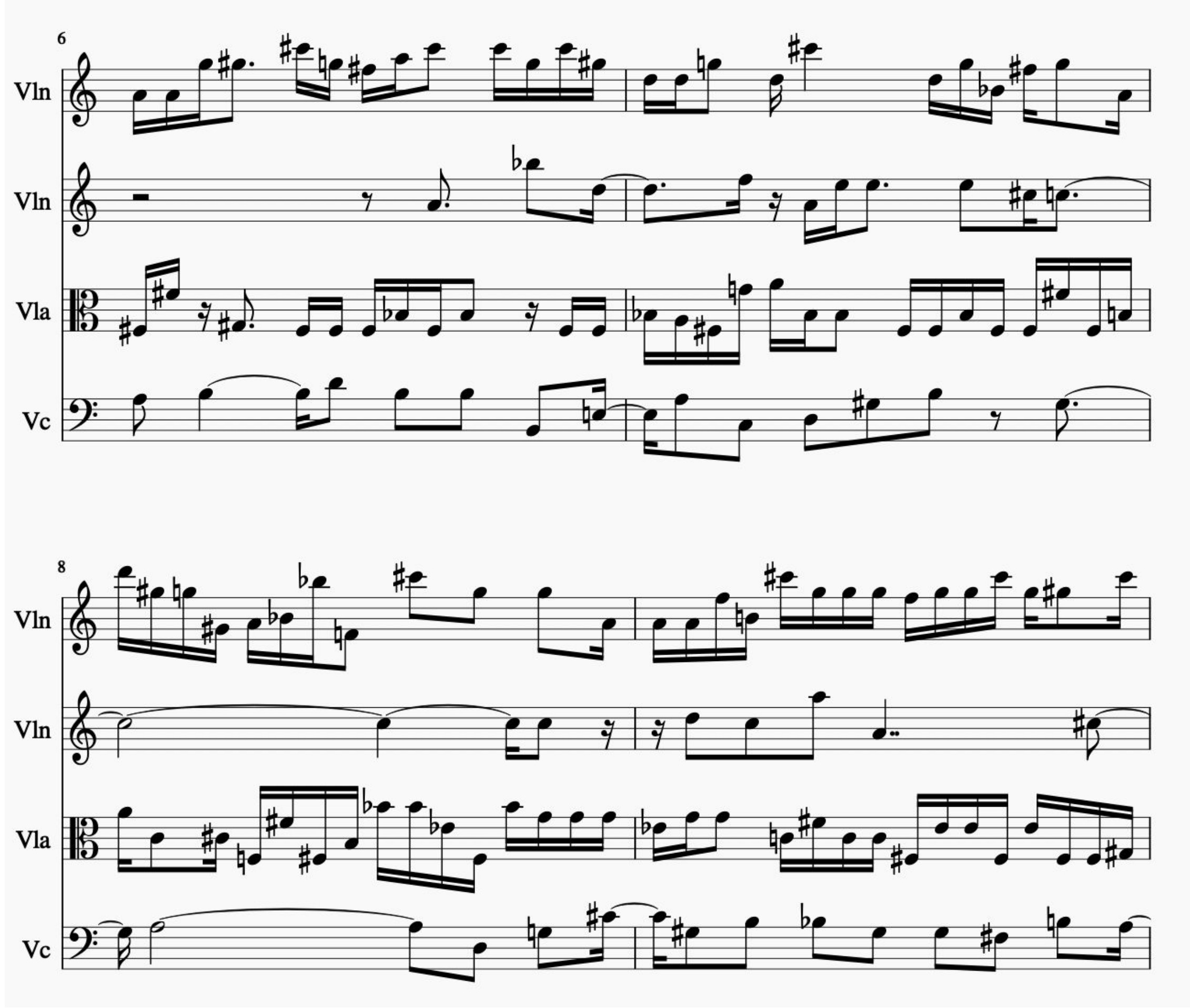
The second LSTM will progress from the end of the sequence towards the center, thus, models the human-like composition process that composers often work backwards to keep music structure.

The first linear model models the harmony. It takes in the notes from the other three parts and predicts the note for the remaining one part.

The last linear model combines the predictions from the above three models, and generate the final prediction.

Training

Result


6
Vln
Vln
Vla
Vc
8
Vln
Vln
Vla
Vc