Let me describe to you a long-lasting frustration and a new hope.

The current channels of musical expression are like thin tubes that musical ideas have to squeeze through. A Digital Audio Workstation (DAW) is expressive but not real-time. A trumpet is real-time but not expressive. How can a machine extract humans' internal music imagination, so that a genius like Bach could compose the four parts of a four-part fugue simultaneously; so that a non-expert could improvise music without mastering any instrument or DAW?

Allow me to show you my answer: a dynamically scaffolded multi-modal *co-adaptation* between a human and her bespoke instrument through interactive machine learning. Picture an instrument that translates the human’s hand gesture, body motions, breathing, micro facial expressions, tongue movement details, muscle activations, and EEG signal into the music she is imagining. The instrument uses an ensemble of variational neural networks supervised by parallel data generated when she listens to or improvises music. Two training techniques foster better generalization from limited training data: 1) KL divergence loss ensures *low-frequency* *continuity* of the mapping; 2) Cycle consistency makes the entire output space *reachable*. Additionally, the human plays the instrument, marks undesirable behaviors, and scores the various ensemble learners in offline review sessions.

Furthermore, the human learns the instrument! Contrary to a static decoding task where the ground truth is passive, in co-adaptation the ground truth also moves towards the decoder. Concretely, when the human sees something the instrument does, she latches onto that and has some low-dimensional control over the instrument. She tries exposing different features for the instrument to learn. The human-instrument bi-directional learning never ends and the dimensionality of control increases. Moreover, the instrument applies haptic guidance to train the human. Specifically, the human selects a piece she wants to play, and the haptic ground truth is computed by the opposite encoder in charge of the cycle consistency. Using haptic guidance, the human can also review old training data and tell the instrument to “forget” outdated ones.

Here is the big picture in my eyes. Mind-reading the human would offer perfect expressivity, but unintrusive readings like EEG are unassailably noisy – an unreliable mapping. A piano translates finger motions to music, but is not nearly as expressive – a collapsed mapping. Co-adaptation gradually finds a mapping where the human can interface with the instrument at maximal information throughput. With many users, we can summarize several "pruned / principal instruments" for beginners to fork and jump start the training.

We will finally be free! Non-experts will be able to jam together in symphony. Musicians will be able to improvise multi-part novel-timbre music while hearing it in real time!

Viability-wise, I see various possible first steps. The initialization of the mapping can be copied from an existing instrument. I can go from a *generator* that decodes random noise to music, to a *controlled generator* that decodes some noise in junction with human input, and finally to an *instrument* that decodes the human.