

# SIMSSA DB: An Introduction

**Cory McKay**

Marianopolis College, Canada

CIRMMT, Canada

# An initial meander: What is a “feature?”

- Information that **measures a characteristic** of a segment of music in a **simple, consistent** and **precisely-defined** way
- Represented using **numbers**
  - Can be a single value, or can be a set of related values (e.g., a vector of histogram bin values)
- Provides a **summary description** of the characteristic being measured
  - Usually provides a **macro** rather than local view
- Usually extracted from pieces or distinct sections (e.g., mass movements) **in their entirety**
  - But can also be extracted from smaller segments of music

# Example: A simple feature

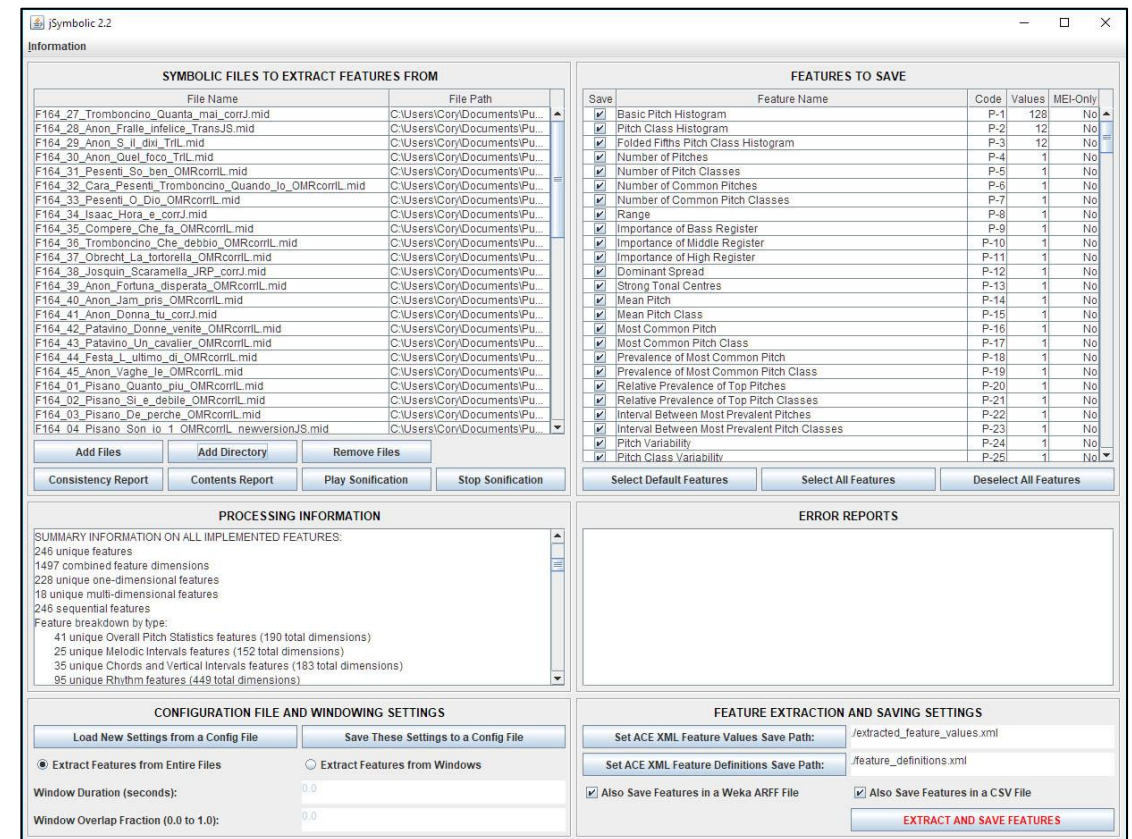
- **Range:** Difference in semitones between the lowest and highest pitches present
  - A 1-dimensional feature



- **Value of this feature** for this music: 7
  - $G - C = 7$  semitones

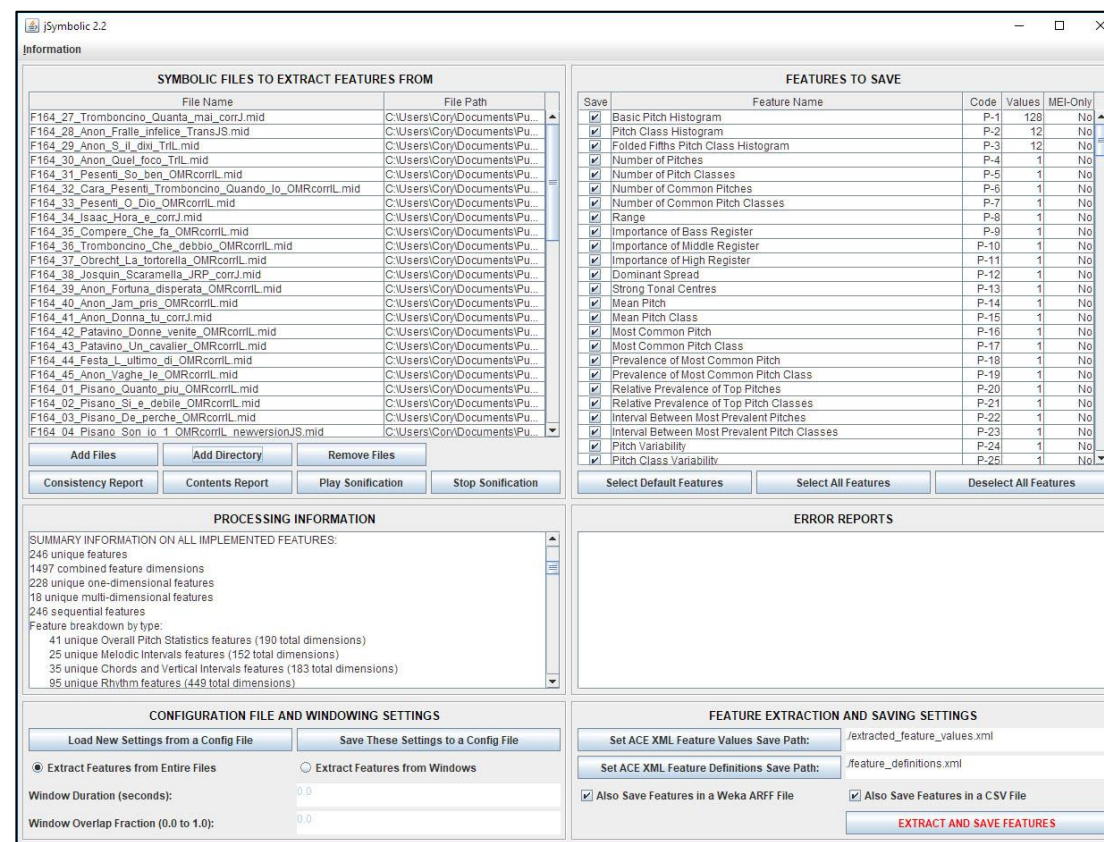
# How can one calculate features?

- The **jSymbolic** research software (McKay et al. 2018) can be used to automatically extract features from **symbolic digital scores**
  - Open source
  - General purpose
- Version 2.2 extracts **246 unique features**
  - 1497 separate feature values, since many features a multi-dimensional (e.g. histogram vectors)



# jSymbolic's feature types

- Pitch statistics
  - e.g. Range
- Melody / horizontal intervals
  - e.g. Most Common Melodic Interval
- Chords / vertical intervals
  - e.g. Vertical Minor Third Prevalence
- Texture
  - e.g. Parallel Motion
- Rhythm
  - e.g. Note Density per Quarter Note
- Instrumentation
  - e.g. Note Prevalence of Unpitched Instruments
- Dynamics
  - e.g. Variation of Dynamics



# Features, what are they good for?

- Provide an empirical basis for musicological **comparison** and **classification**
  - Using automated **machine learning** and **statistical analysis**
  - Manually by experts
- Can study very large large quantities of music
- Can explore a very broad range of musical characteristics and their interrelationships
  - Including aspects one may not have thought to consider
- No need to specify specific queries or heuristics before beginning analyses if one does not wish to
  - Facilitates **exploratory research**
- Help to bypass potentially incorrect ingrained assumptions and biases

# Samples of previous early music research with jSymbolic features

- Composer attribution
  - McKay et al. 2017
- Coimbra manuscripts
  - Cuenca & McKay 2019; Cuenca & McKay 2021
- *Ave verum corpus* and *O decus virgineum*
  - Rodriguez-Garcia & McKay 2021
- *Ave festiva ferculis*
  - Rodriguez-Garcia & McKay 2021
- Morales and Guerrero
  - McKay & Cuenca 2021
- Origins of the madrigal
  - Cumming & McKay 2018; Cumming & McKay 2021
- Buch
  - Cuenca & McKay 2022

# Cory's status circa 2017

- I could extract lots of great features
- I was collaborating with wonderfully insightful musicologists
- How could things possibly get **even better**?
  - Wanted to find new ways to use features
  - Wanted to collaborate more directly with wonderful co-grantees working on OMR
  - Wanted more high-quality symbolic scores
  - Wanted to be able to share extracted features
    - Both generally and related to specific studies
  - Wanted to search for scores based on content, not just metadata



# And so SIMSSA DB was born!

- **Collaborative** database **prototype infrastructure** for holding and accessing **symbolic music files** and **associated features** (and **more**)
  - Web **browser interface**
- Populated by:
  - **Now:** Samples from datasets we have constructed
  - **Medium-term:** Import existing open symbolic datasets that musicologists, libraries and others have already constructed
  - **Long-term:** Auto-population via (verified) OMR
- Focused (for now) on **early music**

# An infrastructure, not a corpus

- The SIMSSA DB is **not** intended just as a repository of music we have transcribed ourselves
  - Although it is seeded with datasets we have made, such as JLSDD (Cumming et al. 2018), Florence 164 (Cumming & McKay 2018), etc.
- Rather, it is a **general unified infrastructure** to which **other** scholars can **contribute** and share symbolic music files (and more) that they have used in their own work

# SIMSSA DB prototype contribution form

**Title**

What is the title of the work? Click the green button to add variant titles or nicknames. Please include opus number or catalogue numbers if applicable (e.g., Op. 55, D960, BWV 202)

Title \*:

Variant Titles:

+ -

Sections:

+ -

**Contributions**

Who created the work? Use the drop-down menu to choose between different kinds of contributions. Add more contributors with the green button.

**Genre(s)**

What type of piece is this? (e.g., song, symphony, motet)

What style is this piece? (e.g., classical, jazz)

Sacred Or Secular:

Not Applicable ▾

**Medium of Performance**

Please enter the instruments or voices below.

Instruments:

SUBMIT

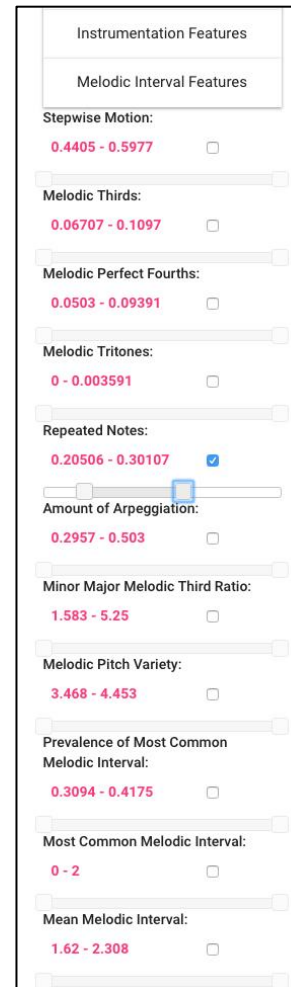
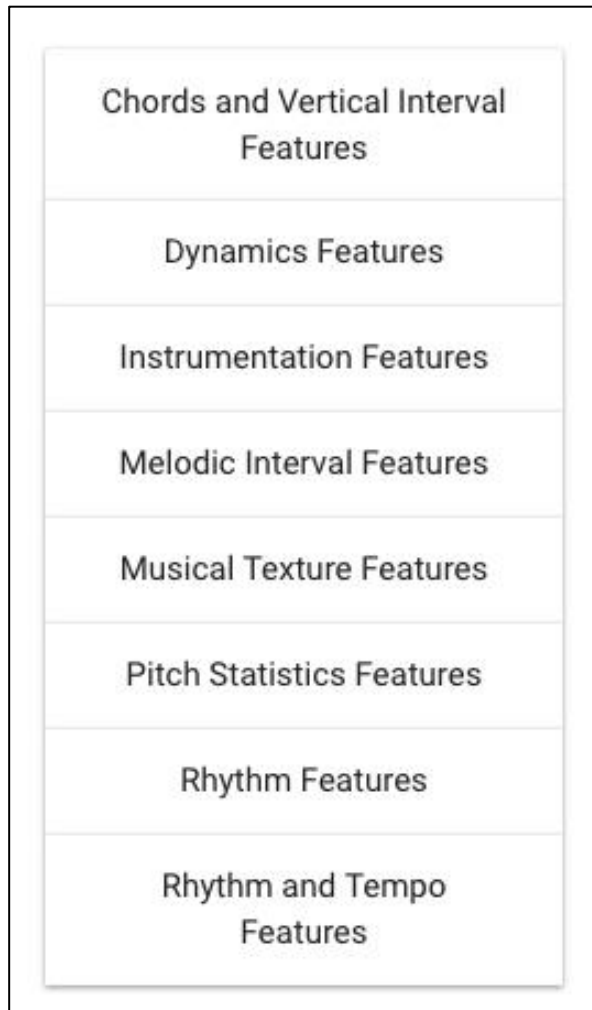
# Metadata searches

- SIMSSA DB may be searched using traditional metadata queries:
  - **Free-text** search
  - **Faceted** metadata filters, such as:
    - Contributor
      - Composer, arranger, author of text, transcriber, etc.
    - Sacred, secular, etc.
    - Instruments / voices
    - Genre / type of work
      - e.g. madrigal, motet, etc.
    - Etc.

# SIMSSA DB and features (1/3)

- SIMSSA DB may also be searched based on **musical content via features**
  - **jSymbolic 2.2** features, specifically
- jSymbolic has been **integrated into the SIMSSA DB**
  - Whenever a file is uploaded to the SIMSSA DB, features are automatically pre-extracted and stored
- Users can **search the database based on musical content** using these features

# SIMSSA DB and features (2/3)



- Users can specify **feature-range queries** via a **slider** for each feature they are interested in

# SIMSSA DB and features (3/3)

- Can also **download complete feature sets** directly and use them as input to statistical analysis and machine learning tools (or analyze them manually)
  - As in the jSybmolic studies referred to earlier
- Feature searches can also be **combined with metadata searches**
  - e.g. retrieve all sacred pieces attributed to Josquin that contain parallel fifths

# Sample query

The screenshot displays a music search interface with the following components:

- Search:** A search bar containing the text "amor".
- Genre (Type of Work):** A list of filters including Frottola(1) and Madrigal(8).
- Genre (Style):** A list of filters including Renaissance(9).
- Composer:** A list of filters including Festa, Sebastiano(4), Pisano, Bernardo(4), and Tromboncino, Bartolomeo(1).
- Instrument/Voice:** A list of filters including Voice(9).
- Sacred or Secular:** A list of filters including Sacred(9).
- File Format:** A list of filters including midi(8), sibelius(8), and xml(8).
- Filter Button:** A purple button labeled "FILTER".
- Results:** A list of 9 results for "amor". The first result is "Amore amor quando io speravo" by Pisano, Bernardo (1490-1548). It is categorized as Madrigal (Type of Work) and Renaissance (Style). The collection of sources is "Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167". The files holding the complete musical work are sibelius, midi, and xml. The files holding an individual section are also listed.
- Feature Filters:** A list of feature filters including Chords and Vertical Interval Features, Dynamics Features, Instrumentation Features, Melodic Interval Features, and Musical Texture Features (highlighted in blue). Below these are sliders for "Average Number of Independent Voices" (1 - 3.804), "Contrary Motion" (0.1578 - 0.1924), "Importance of Loudest Voice" (0 - 1), "Maximum Number of Independent Voices" (1 - 4), and "Oblique Motion" (0.5066 - 0.6374).
- Disclaimer:** A note stating that features only apply to valid MIDI, Music XML and MEI files, and will exclude file formats from Sibelius, Finale, etc. For an explanation of all features, please consult the jSymbolic Manual.



# Highlights of the SIMSSA DB

- Designed to meet the specific needs of scholars wishing to engage in **large-scale computational musicological research**
  - Emphasis on usability
- **Feature-based search** combined with free-text and faceted **metadata search**
  - Full sets of auto-extracted feature values can also be downloaded
- Emphasis on research-relevant data structuring (**more on this tomorrow**)
  - Modeling of **complex abstract musical relationships**
    - e.g., relationships between sources and (abstract) works, sections and parts
    - e.g., linking different kinds of musical documents
  - **Provenance** chains
  - **Authority control** and cataloguing standards
  - **Archiving** of specific **corpora** and associated features from specific **studies**

# Credit to the deserving

- I designed the original data model and provided high-level guidance to the project
  - Along with **Julie Cumming**
- **Emily Hopkins** supervised most of the actual development work
- **Gustavo Polins Pedro** and **Yaolong Ju** did all of the actual implementation work
- We also received important insight and suggestions from a variety of generous contributors
  - Especially **Ichiro Fujinaga**

# Current status

- We are coming out of a two-year development hiatus that followed the loss of our entire development team during the pandemic lockdowns
  - Prior to the pandemic they had completed a prototype interface and were beginning user testing and consultation with domain experts
- The hope is to recruit a new development team in the near future and take up where we left off
- **Tim de Reuse** recently kindly resurrected a limited partial version of the DB test instance running on Compute Canada for demo and test purposes
  - <https://db.simssa.ca>
  - Will briefly demo it at tomorrow's session

# Thanks for your attention!

cory.mckay@mail.mcgill.ca



Social Sciences and Humanities  
Research Council of Canada

Conseil de recherches en  
sciences humaines du Canada

Canada



Centre for Interdisciplinary Research  
in Music Media and Technology

