New Tools for Old Questions: Applying Feature Extraction and Machine Learning to Rodin's "The Josquin Canon at 500"

Cory McKay
Marianopolis College and CIRMMT, Canada

Julie Cumming
McGill University and CIRMMT, Canada

July 6, 2024. Medieval and Renaissance Music Conference (MedRen) Granada, Spain

Topics

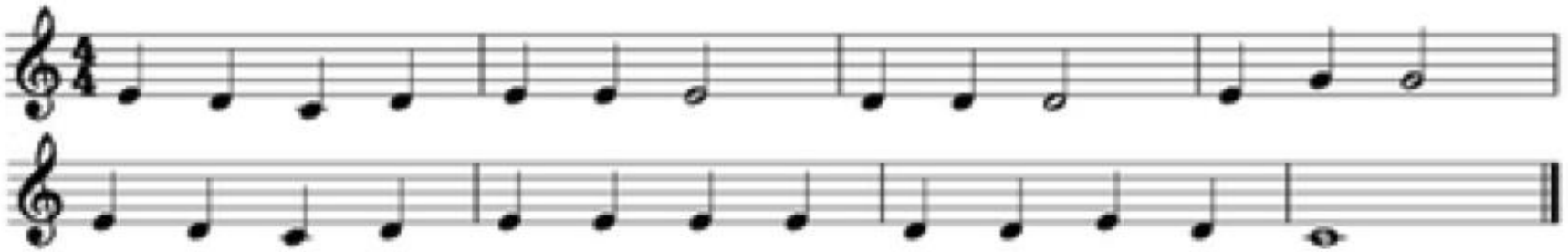
- Introduction to statistical musical features
- Rodin and Rifkin's "The Josquin Canon at 500"
  - And applying a feature-based approach to it
- Data / corpus used
- Experiments, results and discussion

What do we mean by 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 values)
- Provides a summary description of the characteristic being measured
  - Usually provides a macro rather than local view
- Usually extracted from complete pieces or distinct sections of music (e.g., mass movements) in their entirety
  - But can also be extracted from smaller segments of music if wanted

Example: A simple feature

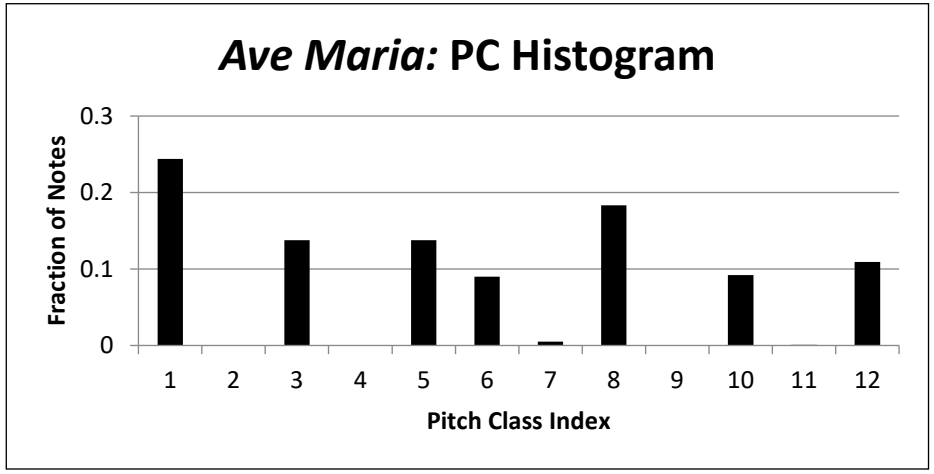
- Range: Difference in semitones between the lowest and highest pitches present

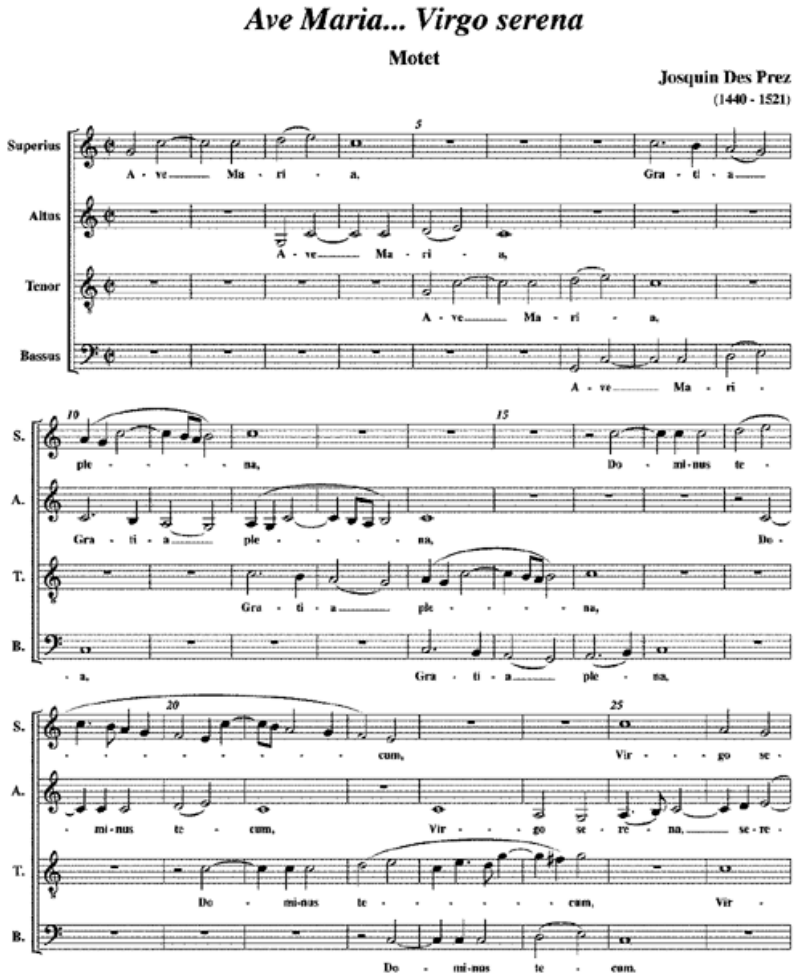


- Value of this feature for this music: 7
  - G - C = 7 semitones

Josquin's *Ave Maria . . . Virgo serena*

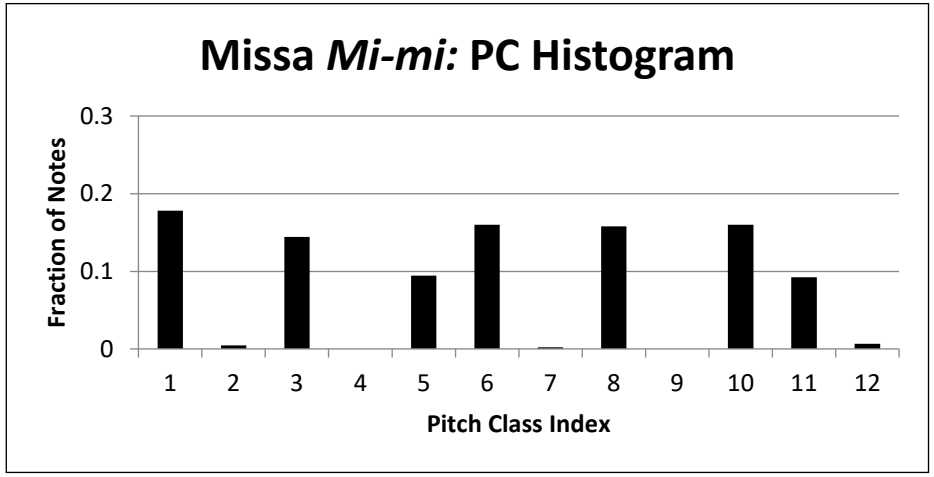
- Range: 34 (semitones)
- Repeated notes: 0.181 (18.1%)
- Vertical perfect 4^{ths}: 0.070 (7.0%)
- Rhythmic variability: 0.032
- Parallel motion: 0.039 (3.9%)

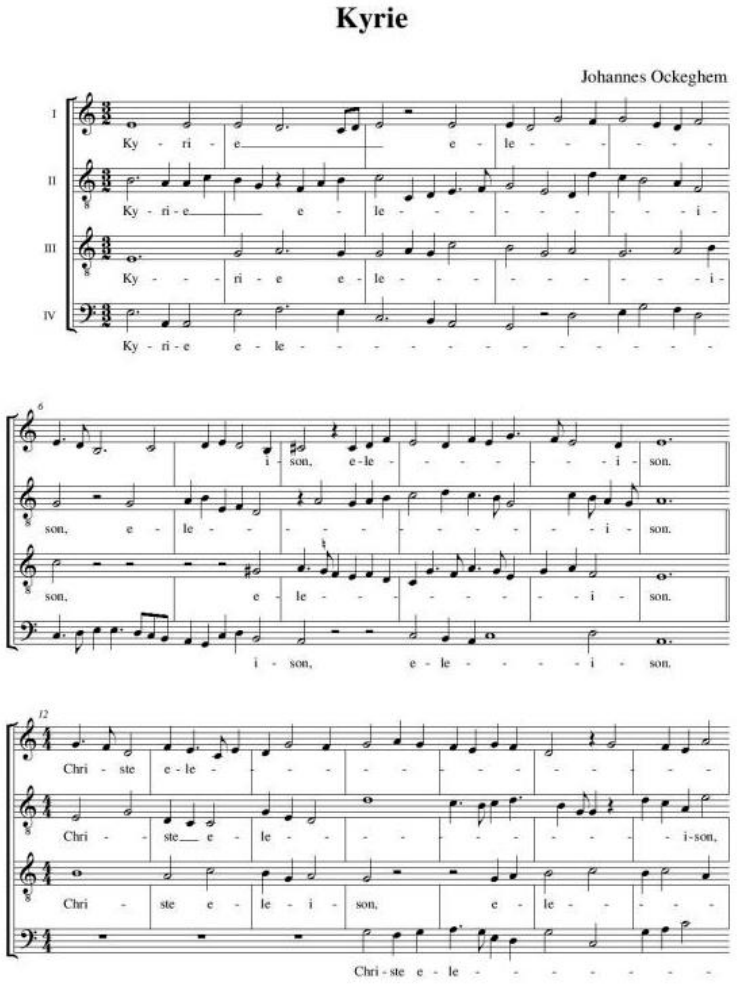




Ockeghem's Missa *Mi-mi* (Kyrie)

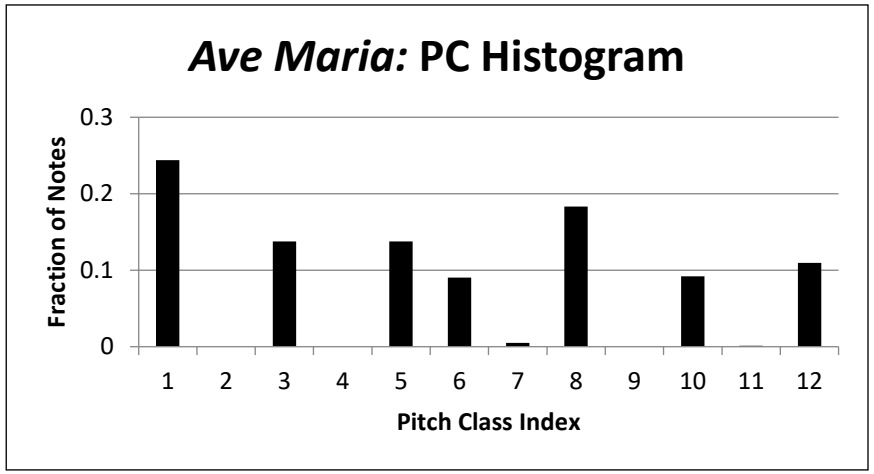
- Range: 26 (semitones)
- Repeated notes: 0.084 (8.4%)
- Vertical perfect 4^{ths}: 0.109 (10.9%)
- Rhythmic variability: 0.042
- Parallel motion: 0.076 (7.6%)

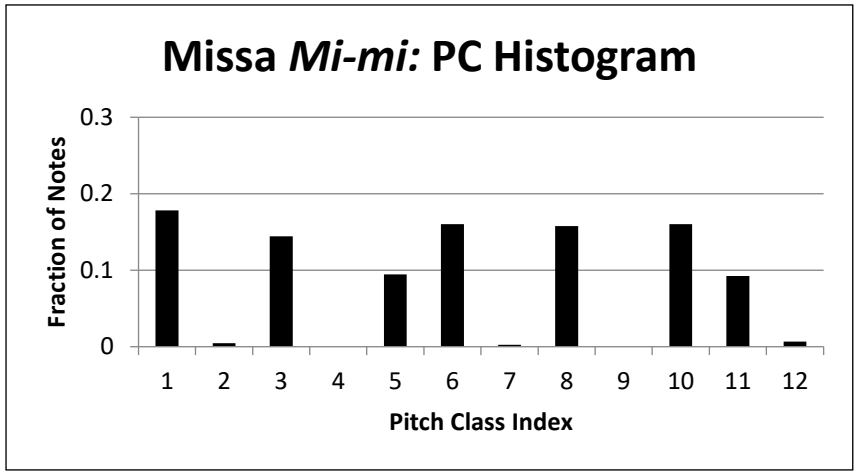




Feature value comparison

Feature	*Ave Maria*	Missa *Mi-mi*
Range	34	26
Repeated notes	0.181	0.084
Vertical perfect 4^{ths}	0.070	0.109
Rhythmic variability	0.032	0.042
Parallel motion	0.039	0.076



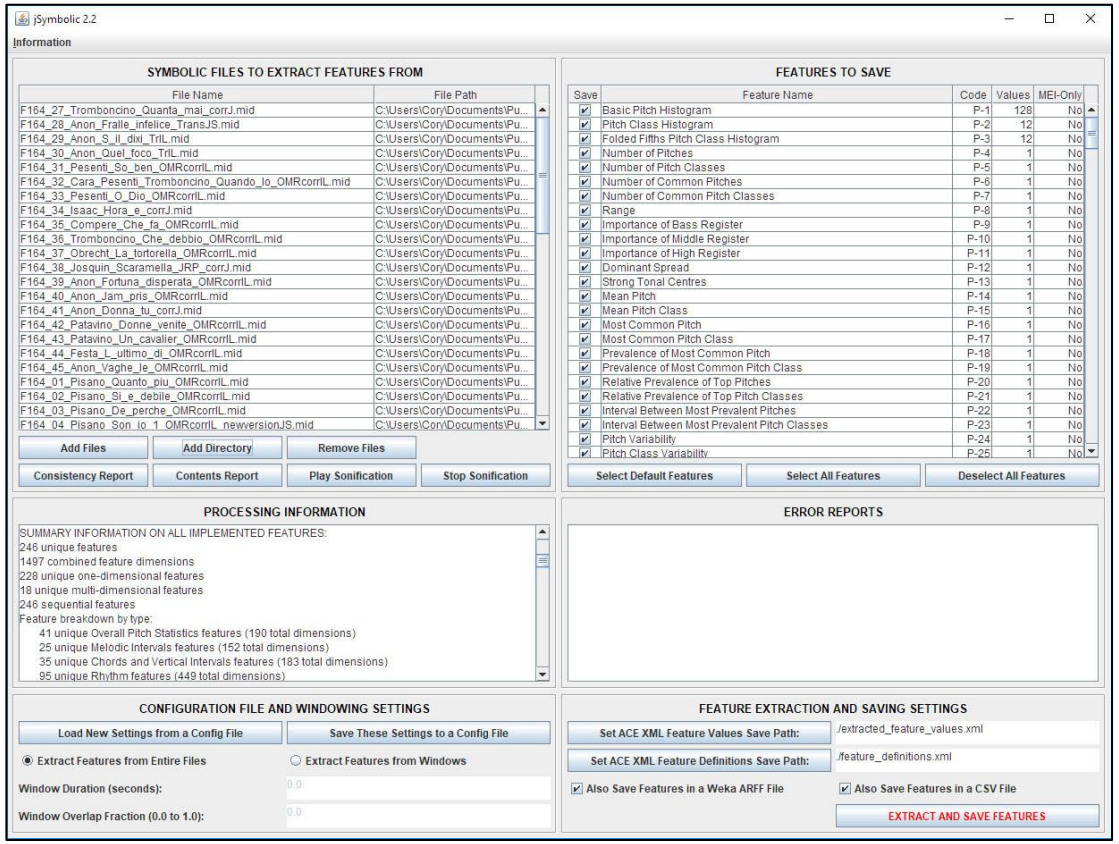


Comparing features

- Comparing pieces in terms of features can be particularly revealing when hundreds or thousands of features are involved, not just six
- Things get even more interesting when comparisons are made between hundreds or thousands of pieces, not just two
  - Especially when the music is divided into groups of interest, whose features can then be collectively contrasted with one another
  - e.g. comparing the styles of composers, genres, regions, time periods, etc.
- Comparing features manually can certainly be useful, but:
  - Statistical analysis and machine learning can reveal complex patterns that might be difficult to discover manually

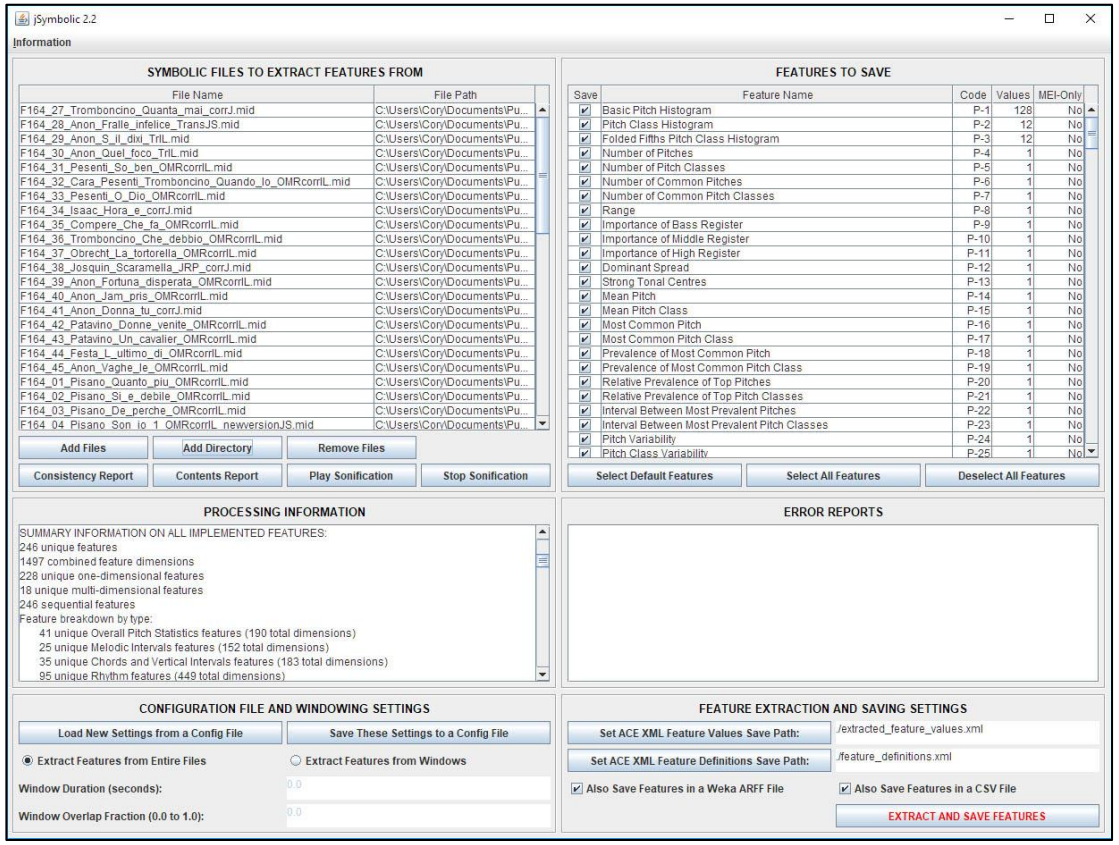
How might one calculate feature values?

- The open-source jSymbolic research software (McKay et al. 2018) can be used to automatically extract features from symbolic digital scores (e.g., MIDI)
- Version 2.2 extracts 246 unique features
  - 1497 separate feature values, since many features a multi-dimensional (e.g. histogram vectors)
- The upcoming Version 3 extracts 533 unique features
  - 2040 feature values, including n-gram features



jSymbolic 2.2'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



Which works did Josquin compose?

- Jesse Rodin (Stanford) and Joshua Rifkin (Boston University): “The Josquin Canon at 500” (*Early Music*, 2022)
- Includes a list of the works they believe are by Josquin
  - “Guilty unless proven innocent” – assume it is not by Josquin unless there is good evidence

RR1: The core group (54 works)

- Source created at a time and place close to the composer
- Attributions to pieces by other composers in the same source are convincing
- Musical variants in source are convincing (e.g., no contrapuntal errors)
- External evidence – piece can be connected to a time and place close to Josquin

Additional levels

RR2: Provisional Acceptance, based on source and style (49 works)

RR3: Problematic (35 works)

RR4: The rest (205 works)

- a: no convincing argument (135 works)
- b: almost certainly by another named composer (70 works)

Can we test this attribution taxonomy with musical features? Yes!

- Features are based only on musical content
  - Complements the Rodin-Rifkin Josquin taxonomy, which is based primarily on historical evidence
- A feature-based approach therefore provides an independent way of providing confirming (or contrasting!) evidence for the Rodin-Rifkin taxonomy
  - All hail the scientific method!

- But, to do this, we need both music securely by Josquin and stylistically relevant music securely not by Josquin, which we can use to train statistical models ("classifiers") that can distinguish Josquin's style
  - And we ideally want all this music to be encoded using the same methodology, to avoid bias (Cumming, McKay, Stuchbery, and Fujinaga 2019)

Source of our digital scores

- The Josquin Research Project (JRP) contains not only the digital scores for all the music in the Rodin-Rifkin taxonomy, but also works by many other relevant composers
  - And they are all encoded using a uniform methodology
  - All hail the JRP!
- See https://josquin.stanford.edu for more

Data partitioning

- We downloaded all the music we needed from the JRP and broke it into two groups:
  - RR: all the music referred to in the Rodin-Rifkin Josquin groups
    - Sub-divided into RR1, RR2, RR3, RR4a and RR4b, according to the Rodin-Rifkin taxonomy
    - With some minor modifications for consistency and quality, such as excluding fragments
  - NonRR: stylistically relevant music securely not by Josquin
- Further broke down each of these groups based on genre:
  - Masses
    - Treating each Mass movement as a separate piece
  - Motets
  - Secular Music
    - French- and Italian-texted works, and pieces without text

Details of the NonRR (not by Josquin) data

- 384 Compositions by 8 composers from the previous generation
  - Busnoys, Du Fay, Frye, Japart, Martini, Ockeghem, Regis, Tinctoris
- 330 Compositions by 10 contemporaries of Josquin
  - Agricola, Brumel, Compere, de Orto, Fevin, Isaac, La Rue, Mouton, Obrecht, Pipelare
- 85 Anonymous pieces (mostly 3-voice chansons from the 15th c.)

- 27% of the NonRR dataset are for fewer than 4 voices:
  - Chansons and (mostly Martini) hymn settings

Methodology

- We used jSymbolic to extract features from all of this music
  - Only extracted that subset of jSymbolic features that is both relevant to Renaissance music and not susceptible to potential biases in our corpus
  - 183 unique features and 801 feature values
- Used machine learning to train a series of models that could differentiate between various groupings in our corpus
  - The only input to the resulting classifiers was extracted jSymbolic features
  - Used the Weka (https://ml.cms.waikato.ac.nz/weka) data mining software
    - Specifically, used support vector machines (SVMs) with a linear kernel

Methodology

- Classifiers were trained separately on different genre groupings:
  - All three genres (mass movements, motets, secular music), combined
  - Just mass movements and motets, combined
  - Just mass movements
  - Just motets
  - Just secular music
- Separate experiments were performed where the "ground truth" (data assumed to be perfectly attributed) Josquin consisted of either:
  - RR1 alone
    - More secure (according to Rodin-Rifkin), but less data
  - RR1 and RR 2 combined
    - Adding RR2 makes it less secure than just having RR1 (according Rodin-Rifkin), but this provides more training data, which can improve performance

Experiment Set 1: Cross-validation

- We began by performing "cross-validation" experiments to see:
  - How statistically distinguishable the ground truth Josquin (RR1 or RR1+RR2) is from the ground truth music not by Josquin (NonRR) for each genre grouping
  - How reliable we can expect our various classifiers to be when later applied to the less secure RR3, RR4a and RR4b groups

Experiment Set 1: Cross-validation *(Ground truth Josquin: RR1)*

	Josquin Accuracy	NonRR Accuracy	Overall Accuracy
Masses + Motets + Secular	63%	98%	**94%**
Masses + Motets	75%	98%	**94%**
Masses	79%	97%	**95%**
Motets	65%	93%	**89%**
Secular Music	20%	96%	**92%**

- Overall, the classifiers appear to be quite reliable (89% or higher)
- But they are much better at correctly identifying music NOT by Josquin than they are at correctly identifying music by Josquin!
  - i.e. as a Josquin classifier, false negatives are a greater problem than false positives
  - This is especially true for the secular group (only 20% Josquin accuracy!)
- They are also better at identifying Josquin in mass movements alone (79%) than motets alone (65%)

Experiment Set 1: Cross-validation *(Ground truth Josquin: RR1+RR2 combined)*

	Josquin Accuracy	NonRR Accuracy	Overall Accuracy
Masses + Motets + Secular	68%	95%	**91%**
Masses + Motets	74%	96%	**92%**
Masses	74%	96%	**92%**
Motets	69%	91%	**86%**
Secular Music	52%	96%	**90%**

- The results are quite similar when the ground truth Josquin is taken to be RR1 and RR2 combined
  - In terms of both in classification accuracies and overall trends
- Two notable differences:
  - The differences in correctly identifying Josquin in masses only vs. motets only is less extreme (5% vs. 14%), although still present
  - In the secular group, Josquin is correctly identified as Josquin 52% of the time (as opposed to 20% of the time with just RR1), which is much better but still very poor

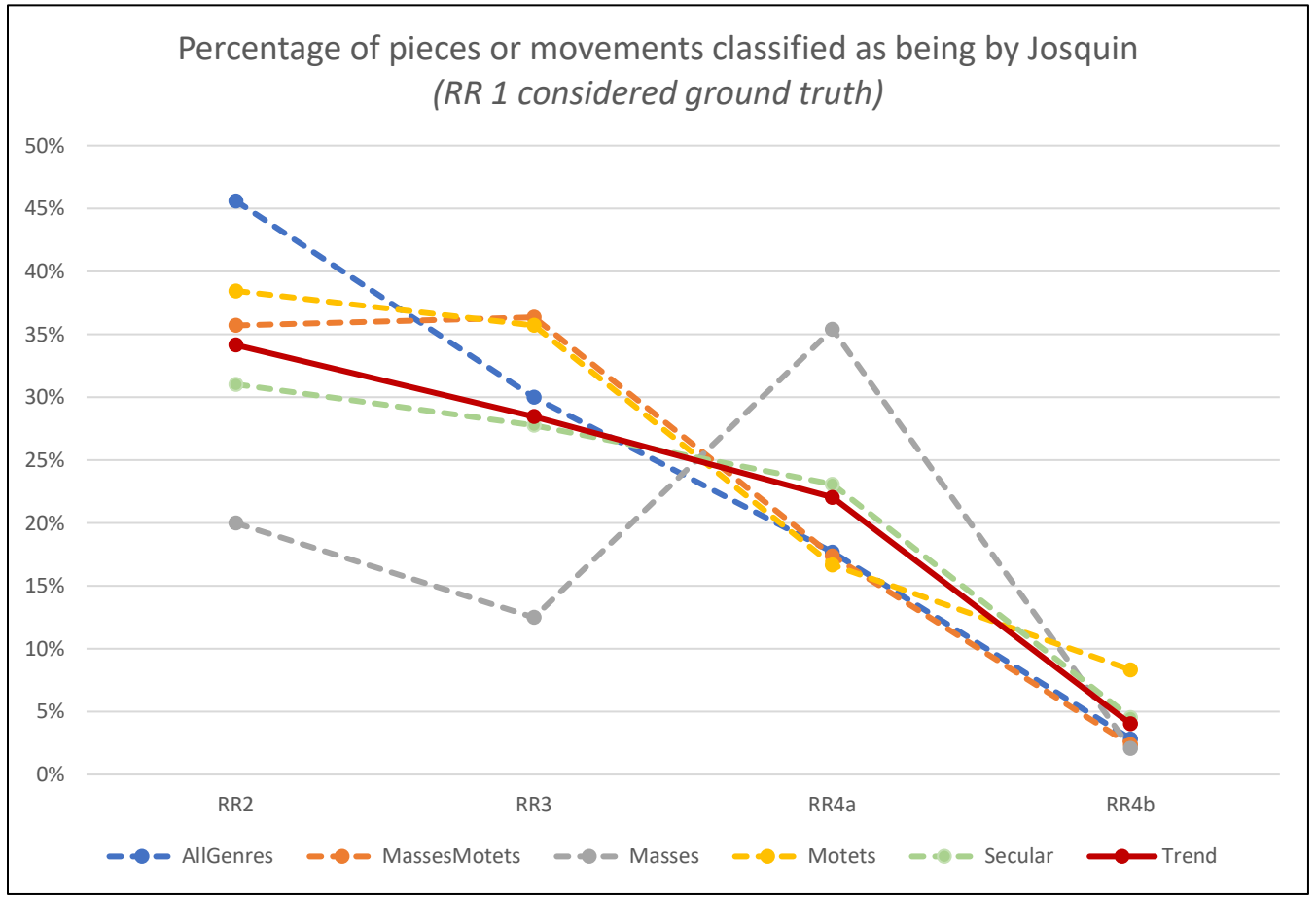
Experiment Set 1: Discussion

- Overall conclusions from Experiment Set 1:
  - As a whole, the classifiers were quite effective
  - When the classifiers say a piece is by Josquin, they are usually right
  - However, if they say a piece is not by Josquin, this is somewhat less reliable
    - But results are still reasonably good for mass movements and okay for motets
    - Terrible at secular music, however
  - This is likely because we had much less secure Josquin music in RR1 and RR2 than we did NonRR music to train the classifiers with
- This context is important to keep in mind for all the experiments to come

Experiment Set 2: Classifying (RR2), RR3, RR4a and RR4b

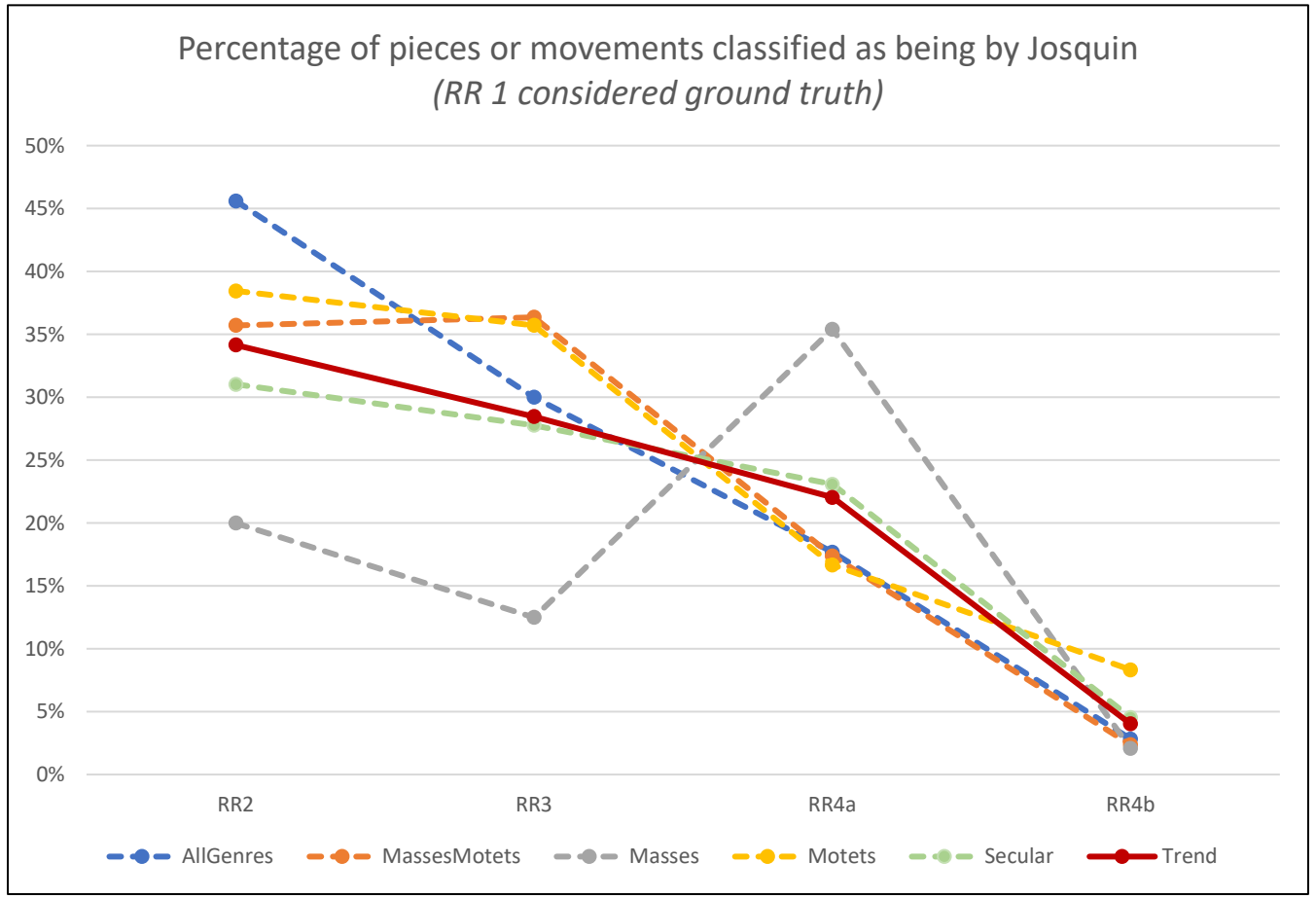
- Next, we trained classifiers on all the secure Josquin and secure NonRR music
  - And use them to classify all of the music in each of the less secure RR groups
- Hypotheses:
  - A greater fraction of the music in a more secure group should be classified as being by Josquin than in a less secure group
  - If this is consistently the case, then this empirically supports the Rodin-Rifkin taxonomy as a whole
    - Although not necessarily with respect to individual pieces
  - If this is not the case, then this suggests that there may be certain problems with the Rodin-Rifkin taxonomy
- Important note:
  - Even if the classifiers make mistakes with a few individual pieces (which they almost certainly will), in aggregate the trends that appear will very likely be meaningful overall
  - This expectation is supported by the results from Experiment Set 1

Experiment Set 2: Classifying RR2, RR3, RR4a and RR4b *(Ground truth Josquin: RR1)*



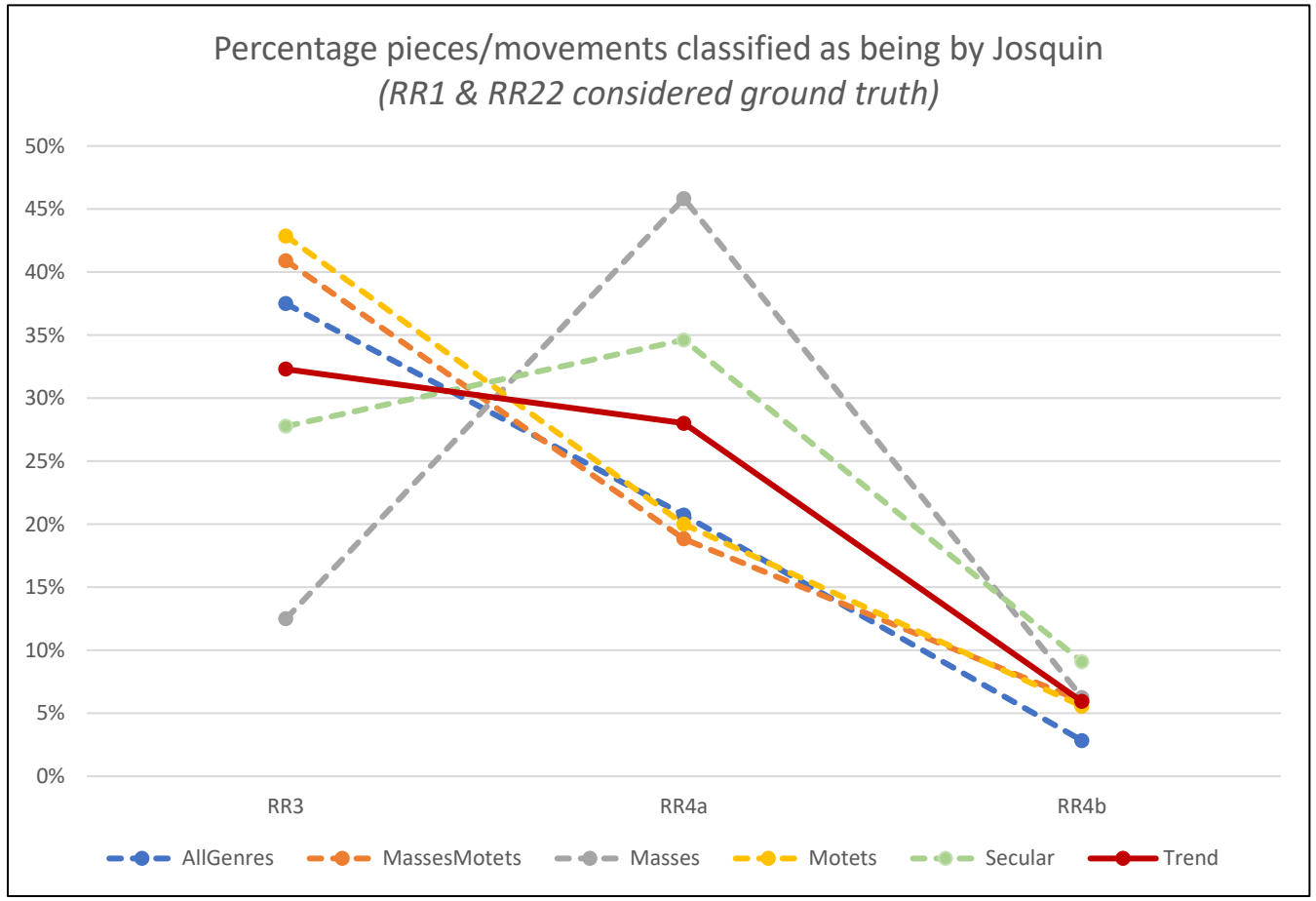
- Each curve on the graph represents the percentage of the music identified as being by Josquin for a particular genre grouping, as a function of RR grouping
- The solid red curve shows the average over all five of the genre groupings

Experiment Set 2: Classifying RR2, RR3, RR4a and RR4b *(Ground truth Josquin: RR1)*



- Overall, the trend shows that the less secure an RR group is, the lower the probability is that its music will be classified as being by Josquin
  - Note the particularly precipitous drop in 4b
- This provides general empirical support for the Rodin-Rifkin taxonomy!
- BUT a surprisingly large proportion of mass movements are classified as being by Josquin in the RR4a group
  - Perhaps the masses in this group should be revisited?

Experiment Set 2: Classifying RR3, RR4a and RR4b *(Ground truth Josquin: RR1+RR2 combined)*



- The results when RR1 and RR2 are combined to train the classifier (instead of just RR1, as in the previous graph)
- Exhibits the same trends

Experiment Set 2: Classifying (RR2), RR3, RR4a and RR4b

- Overall conclusions from Experiment Set 2:
  - The results provide good overall supporting evidence for the Rodin-Rifkin taxonomy
    - Overall, but not necessarily with respect to specific pieces (yet)
  - There is some evidence for revisiting at least some of the mass cycles and movements in RR4a ("no convincing argument"), as an unexpectedly high number of mass movements in this group were classified as being by Josquin
    - And recall that our classifiers were in fact relatively good at correctly identifying Josquin's mass movements in particular in Experiment Set 1

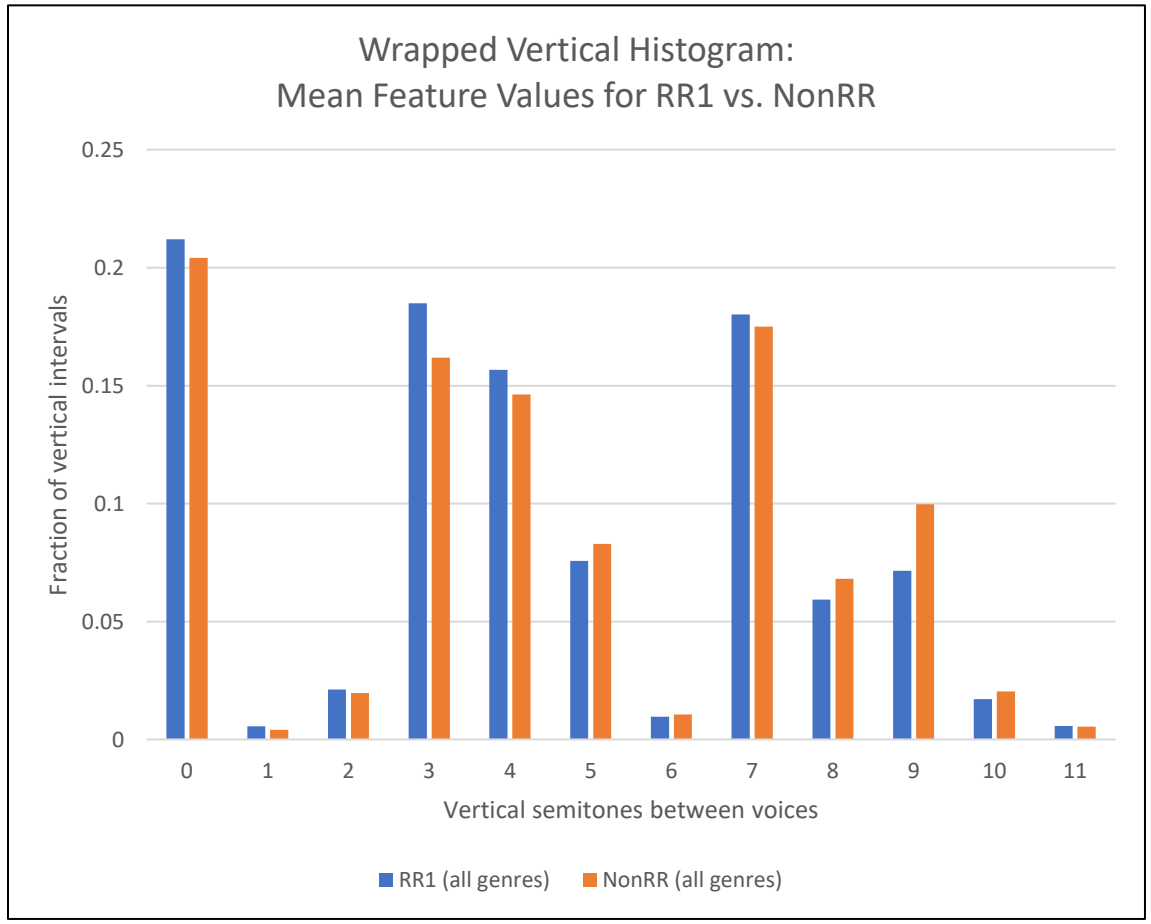
Experiment Set 3: Identifying statistically relevant features

- We also (separately) used statistical methods to identify specific features that were (individually) particularly effective in separating secure Josquin from NonRR across all of the data groups from Experiment Set 1
  - i.e. various genre groupings of RR1 vs. NonRR or RR1+RR2 vs. NonRR
  - Used Weka's implementation of information gain and Pearson correlation to do this
- Goal:
  - Identify stylistic elements especially statistically characteristic of Josquin's personal style

Experiment Set 3: Identifying statistically relevant features

- It turns out that most of the features statistically highlighted as most discriminative of Josquin were associated with vertical intervals between voices
  - *Note:* vertical intervals were measured here in terms of number of semitones separating notes, not diatonically, and are weighted by note duration
    - e.g., four semitones corresponds to a major third
- Some of the differences were proportionally quite large:
  - e.g., 4.5% of vertical intervals are unisons for RR1, on average, compared to 3.0% in nonRR
    - i.e. proportionally speaking, Josquin used vertical unisons 50% more often
- Other differences highlighted were smaller, but still statistically meaningful:
  - e.g., 7.6% of vertical intervals are perfect fourths for RR1, compared to 8.3% for nonRR
  - e.g., 34.2% of vertical intervals are thirds (major or minor) for RR1, compared to 30.8% for nonRR

Experiment Set 3: Identifying statistically relevant features



- This graph shows the wrapped vertical interval histogram feature values, averaged across all pieces in RR1 vs. NonRR
  - "Wrapped" means that intervals separated by an octave are counted together
  - e.g., wrapped "7" = perfect 5ths & 12ths & . . .
- This represents a kind of signature harmonic profile for Josquin (in blue) relative to the NonRR composers (in orange)
  - e.g. note how Josquin uses intervals of major sixths & octaves (bin 9) much less often than other composers

Experiment Set 3: Identifying statistically relevant features

- Certain melodic features were also especially statistically discriminative:
  - Melodic sixths (major & minor combined): 0.46% for RR1 vs. 0.26% for NonRR
    - Melodic minor sixths: 0.31% for RR1 vs. 0.16 for NonRR
  - Melodic octaves: for 1.4% for RR1 vs. 1.0% for NonRR
- As were certain rest-related features:
  - Average rest fraction across voices: 25% for RR1 vs. 18% for NonRR
    - The fraction of a voice's duration during which no note is sounding, averaged across all voices
  - Partial rests fraction: 62% for RR1 vs. 47% for NonRR
    - Measures the fraction of piece's duration during which at least one voice is resting

Experiment Set 4: Considering individual pieces

- In our final set of experiments, we used our trained models to classify selected pieces individually
- Recall from Experiment Set 1:
  - The models can make mistakes when looking at individual pieces
  - In particular, they are substantially more likely to incorrectly say that a piece is not by Josquin than they are to incorrectly say that it is by Josquin
- The results from this section are therefore not definitive
  - But they can be used as supporting evidence, or reason to take a second, more detailed look at certain pieces

Experiment Set 4: Classifying selected RR motets individually

	RR1 vs. NonRR All Genres	RR1 vs. NonRR Masses & Motets	RR1 vs. NonRR Motets
RR4a: Jos1401: Absalon fili mi	Not Josquin	Not Josquin	Not Josquin
RR4a: Jos1409: Planxit autem David	Not Josquin	Not Josquin	Not Josquin
RR4a: Jos2811: Dulces exuviae	Not Josquin	Not Josquin	Not Josquin
RR3: Jos2815: Fama malum	Josquin	Josquin	Not Josquin

- Perhaps Jos2815: Fama malum actually is by Josquin?
  - Despite being in the RR3 "questionable" group
- These results are not definitive proof, but they merit further attention
  - Recall from Experiment Set 1 that classifiers rarely incorrectly identify music as being by Josquin

Experiment Set 4:
Classifying the RR4a mass movements individually

- Back in Experiment Set 2 we found that a surprisingly high fraction of the RR4a (“no convincing argument) mass movements were classified as being by Josquin
- But which ones?
- To answer this, we noted all the masses in RR4a that had more than half of their movements classified as being by Josquin in all three tests
  - i.e. all genres, masses and motets combined and masses only
  - We discounted two incomplete masses:
    - Missa Mon seul plaisir has only two surviving voices
    - Missa Rosina has only one surviving voice (except for the Credo)
  - We discounted masses with only one movement
- Three masses met these conditions . . .

Experiment Set 4:
Classifying the RR4a mass movements individually

- *Jos0903:* Missa Di dadi
- *Jos1001:* Missa Mater patris
- *Jos1201:* Missa Ad fugam
  - *Caveat:* this is a canonic mass, which differentiates it from much of the rest of our corpus
- All three of these masses warrant further investigation as potentially being by Josquin

- *Interesting side note:* in the case of Missa Allez Regretz I (Jos0701), which Steib suggests could be Josquin, not a single one of the mass movements was classified as being by Josquin in any of the three tests

Thanks for your attention

cory.mckay@mail.mcgill.ca

julie.cumming@mcgill.ca