Black Hole Formation Pathways in Young Stellar Clusters: A Machine Learning Analysis of CMC Simulations

Ishaan Satish
UCSD Department of Physics

Abstract

I investigated how black holes form in young stellar clusters using 75 Cluster Monte Carlo (CMC) simulations (after filtering from an initial 145 models). My analysis focused on how stellar mergers and collisions in dense cluster environments affect black hole properties (particularly mass and spin distributions). I found that while most black holes (63.4%) form through normal stellar evolution, a significant fraction result from dynamical processes: collisions (24.0%) and mergers (16.1%). Using machine learning techniques (including PCA and Random Forest classification), I identified key cluster parameters that control black hole formation efficiency. The most striking result was a strong anti-correlation (r = -0.935) between metallicity and black hole mass, which completely dominates the physics here.

1 Scientific Motivation

The main goal of this project was to investigate young massive star cluster models for massive stellar mergers to determine their imprint on black hole spin distribution. The fundamental idea is that angular momentum from merger events gets injected into the merger product, potentially creating black holes with non-zero spins. In some cases, these merger products may ultimately lead to collapsar-like objects - black holes embedded within thick disks that accrete material and spin up.

This investigation has several important implications. First, some fraction of black holes in clusters will be born with non-zero spins, which affects their retention following subsequent binary black hole mergers. Second, depending on mass growth, these processes may provide another avenue for growing intermediate-mass black holes (IMBHs) in clusters. Third, these events may produce luminous fast blue optical transients (FBOTs) in young stellar clusters detectable by transient surveys. Fourth, these objects may produce r-process elements, possibly explaining observed r-process enhancement in some Milky Way globular

clusters like M15.

Merging black holes have masses and spins that really challenge what we thought we knew about how black holes form. Many of the detected events from LIGO show evidence of hierarchical formation in dense stellar environments (where black holes grow through successive mergers rather than just collapsing from a single star).

Young massive star clusters are basically perfect laboratories for studying these alternative black hole formation pathways. These environments have all the right conditions for stellar collisions and mergers that can lead to very massive stars and eventually intermediate-mass black holes. This could explain several puzzling observations: black holes in the pair-instability mass gap (50-130 $M_\odot$), objects with weird spin orientations, and the surprisingly high merger rates that LIGO keeps detecting.

2 Data: From 4 Models to 145 (and Back to 75)

I originally started small-scale (honestly maybe too small, as my slides showed). I began with just 4 models - basically testing different combinations of virial radius and primordial mass segregation. But then I asked Professor Kremer for access to this massive dataset of CMC simulations, and suddenly I went from 4 models to 145 total models spanning the parameter space of Milky Way globular clusters.

The CMC code is pretty incredible - it's this N-body simulation framework that incorporates all the relevant physics for dense spherical star clusters (strong dynamical encounters, stellar evolution, relativistic dynamics, black hole formation). Each simulation varies four key parameters: initial stellar population (N = 0.8×10^5 to 3.2×10^6 stars), virial radius (rv = 0.5-4.0 pc), gravitational radius (rg = 2.0-20.0 pc), and metallicity (Z = 0.0002-0.02).

I filtered down to only models with rv $<$ 1.1 pc because the other clusters simply aren't dense enough to produce the interesting dynamical events we're looking for. This left me with 75 models - still a huge dataset but focused on the physics we actually care about.

For each model, I extracted three main data products:

- **Black hole formation data** (`initial.bhformation.dat`): Every black hole formed during the simulation (formation time, location, progenitor mass, final mass, spin)
- **Collision records** (`initial.collision.log`): All stellar collisions with masses, stellar types, and locations
- **Merger events** (`initial.semergedisrupt.log`): Stellar mergers within binary systems(Same information as Collision records)

The key challenge was developing parsing functions that could handle the messy, variable-format data files and cross-reference between them.

3 Methods: Designing the Parsing Pipeline

The whole analysis really depended on getting the data parsing right, and I had to build this interconnected system of functions that each handled different pieces of the puzzle.

3.1 From Individual Analysis to Comprehensive Pipeline

I started with the analysis steps, initially working with just a few models to understand the data formats and relationships. I created scatter plots of M_1 versus M_2 for all collisions, made histograms of black hole masses, and began cross-checking BH IDs with collision products to identify dynamically formed black holes.

These initial explorations with single models helped me understand the data structure and develop robust parsing functions. As I gained confidence in my ability to extract meaningful information from individual models, I decided to use this research project as the first steps for my Data Science in Astronomy final project I expanded the scope to the full dataset of 145 models, applying the rv < 1.1 pc filter to focus on the most dynamically interesting 75 models.

3.2 The Data Processing Pipeline

First, I wrote `parse_model_name()` to extract the physical parameters (N, rv, rg, Z) from the cryptic model directory names. This fed into `discover_models()` which applied my rv < 1.1 filter and built the master list of 75 models I actually wanted to analyze.

Then came the tricky part - parsing the individual data files. I needed separate functions for each file type because they all have different formats. Thankfully I already had most of the framework for these from my individual model analysis code:

- `load_bh_formation_data_fixed()`: Handles the variable-column black hole formation files (I kept only the first 9 columns to avoid format issues)
- `load_collision_data_fixed()`: Uses regex parsing to extract collision events from the log files
- `load_merger_data_fixed()`: Similar regex approach for merger events

3.3 The Analysis Functions That Make It Work Together

The real magic happens in `analyze_bh_formation_pathways()`, which cross-references black hole IDs with collision and merger event records to determine formation mechanisms. This is where I figured out that on average 63.4% of black holes form through normal evolution, 24.0% from collisions, and 16.1% from mergers.

I also wrote `analyze_bh_locations()` to study the spatial distribution (turns out 77.1% of black holes form within 1 pc of the cluster core , likely due to mass segregation).

Finally, `process_single_model()` and `process_all_models()` orchestrate the entire pipeline, calling all the parsing and analysis functions for each of the 75 models and building the comprehensive dataset I used for machine learning.

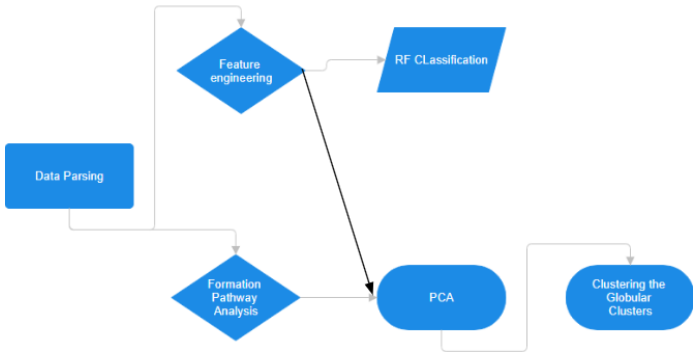

Figure 1: An illustration of my workflow

4 Results

4.1 Black Hole Formation Pathways

My analysis reveals that black hole formation in stellar clusters happens through three distinct mechanisms with well-defined statistical distributions. Normal stellar evolution is still the dominant pathway (63.4% ± 21.1% of all black holes), but dynamical processes contribute substantially: stellar collisions produce 24.0% ± 20.9% of black holes, while stellar mergers account for 16.1% ± 7.5% (The large uncertainty on collisional BHs is due to a specific quirk of the code which create some very specific high mass outliers , however I would like to investigate this further after eliminating those).

This has implications for gravitational wave astronomy because dynamically-formed black holes typically have different spin and mass characteristics compared to isolated formation. The significant fraction from collisions suggests that dense clusters are major contributors to the spinning black hole population that LIGO-Virgo keeps detecting.

4.2 The Dominant Metallicity Effects

The most striking result from my entire analysis is this incredibly strong anti-correlation (r = -0.935) between stellar metallicity and black hole mass. Low

metallicity environments ($Z \leq 0.002$) produce black holes with mean masses of 17.1 ± 2.5 $M_{\odot}$, while high metallicity clusters ($Z = 0.02$) generate much lighter black holes with mean masses of 7.8 ± 0.2 $M_{\odot}$.2

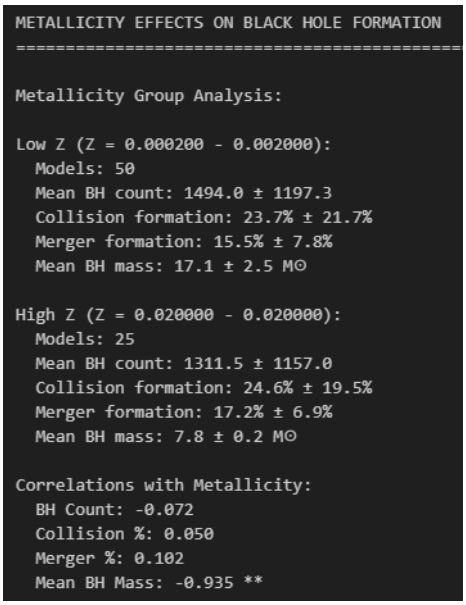

Figure 2: Output of my analyze_metallicity_effects() function

This relationship is so fundamental that it basically drives everything else. Based on readings I did after seeing the results it reflects how stellar winds work in massive star evolution - higher metallicity environments enhance mass loss rates and reduce the final masses of stellar remnants. The correlation is so strong that metallicity alone can predict black hole mass distributions with incredibly high accuracy.

4.3 Spin Distribution Findings

Surprisingly, only 5 out of the 75 cluster models produce any black holes with non-zero spin (mean high-spin BH fraction = 0.009), and those few spinning BHs all share a similar spin magnitude (non_zero_spin_value_mean = 0.686). The vast majority of models give spin-zero remnants, driving down the overall high-spin fraction to well below 1%.

This result likely reflects our expanded, Milky Way–like parameter survey (rv <1.1 to focus on dense clusters) rather than the smaller, cherry-picked sample we examined early on. Under typical young cluster conditions, successive collisions and mergers rarely inject enough angular momentum to spin up black hole progenitors. We therefore conclude that dynamical pathways in these environments almost never produce high-spin black holes.

4.4 Principal Component Analysis: Finding the most important factors

To understand what's really driving cluster evolution, I applied PCA to 14 cluster parameters, making sure to include the enginered features i added into my parsing functions. They were [N, rv, rg, Z, total_bhs, collision_percentage, merger_percentage, mean_bh_mass, mean_formation_radius, collision_rate, merger_rate, high_spin_fraction, spin_count, median_formation_radius]

The analysis revealed that cluster evolution is dominated by just a few key physical processes.

The first principal component (PC1) explains 33.8% of the variance and is primarily driven by virial radius effects, collision rates, and formation efficiency. The second component (PC2) captures 19.2% of variance and correlates strongly with metallicity, black hole mass, and spin properties(The inclusion of spin here is particularly exciting). Together, the first five components explain 85.9% of the total variance, which means most of the physics can be captured in this lower-dimensional space.3

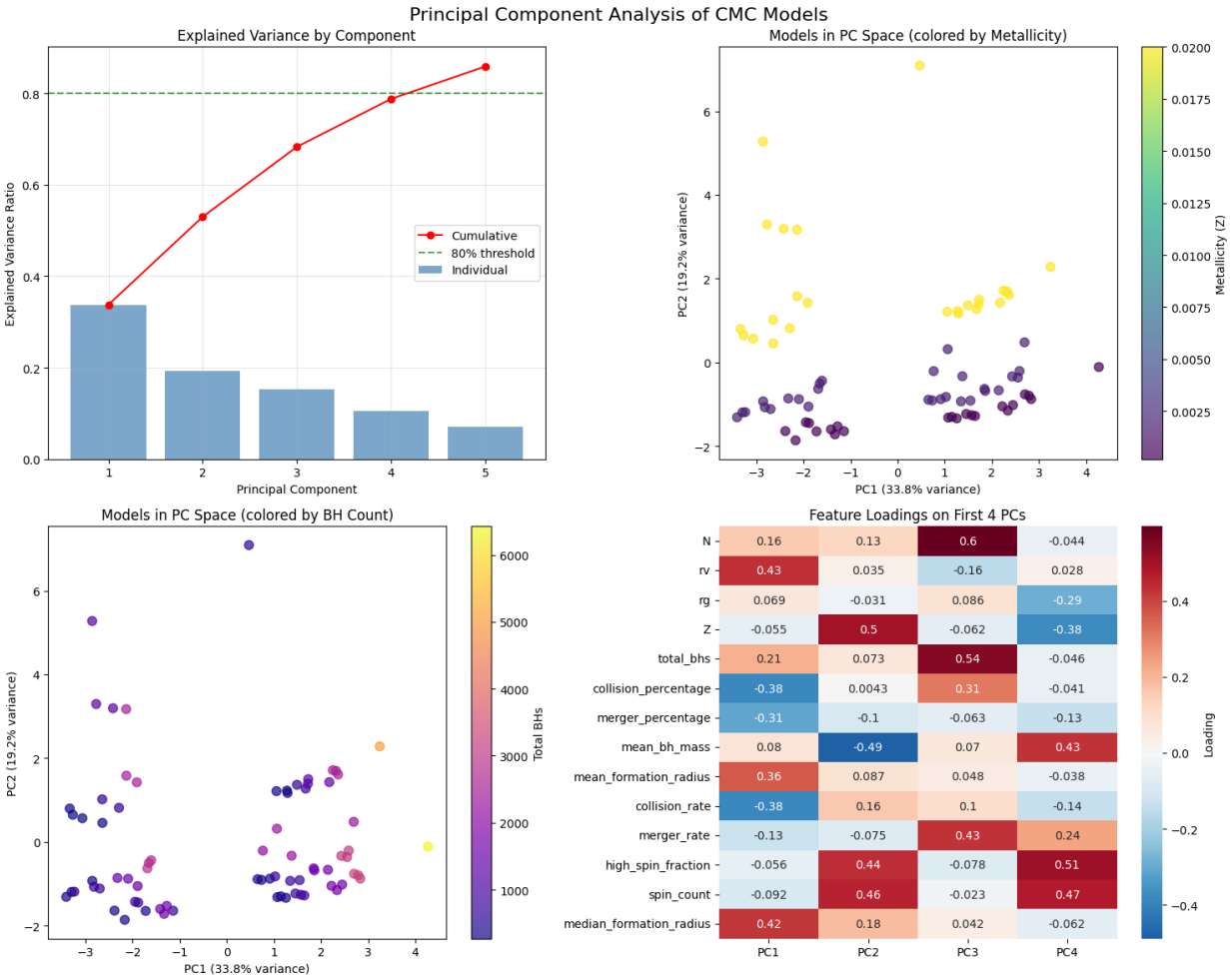

Figure 3: Various visualizations of my PCA results

4.5 Clustering Analysis: Natural Groupings in Parameter Space

K-means clustering identified three distinct cluster populations with different formation characteristics. The silhouette analysis suggested three clusters was optimal (score = 0.479).

These clusters naturally separate in the PCA space, with different metallicity distributions and formation efficiencies. This unsupervised approach revealed groupings in the parameter space that I wouldn't have thought to look for otherwise.Cluster 0: Low-Z, high collision rates, Cluster 1: High-Z, low BH masses and Cluster 2: Mid-Z, efficient mergers

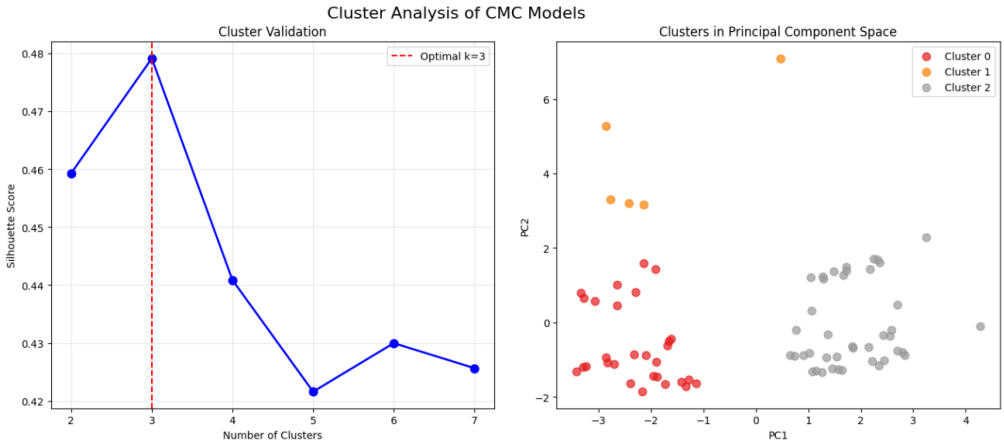

Figure 4: Results from my Clusterintg code

4.6 Machine Learning Predictions: What Makes Clusters Efficient?

I used Random Forest classification to predict which clusters would have high merger efficiency (top 30% of models), and the results were pretty impressive - 95.7% accuracy with 5-fold cross-validation.

The feature importance hierarchy (rv > N > Z > rg) reveals that virial radius is the single most important predictor of merger rates, with stellar population size playing a secondary role. This makes total physical sense - compact clusters should be collision-dominated systems where dynamical encounters are frequent.

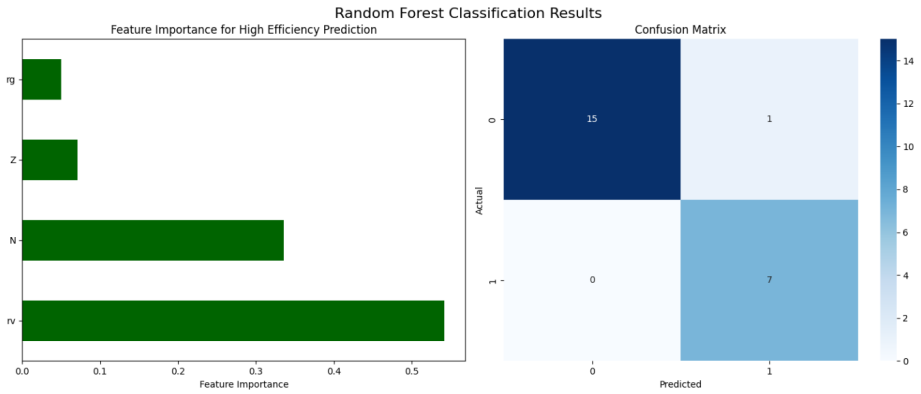

Figure 5: My Random Forest results

5 Spatial Distribution and Formation Efficiency

Black hole formation shows strong spatial concentration toward cluster centers, with 77.1% ± 8.7% of black holes forming within 1 parsec of the cluster core. The mean formation radius is 0.75 ± 0.13 pc, which clearly shows that mass segregation and central density enhancement drive most formation events.

This spatial concentration has important consequences for subsequent gravitational wave merger rates, since centrally-formed black holes are more likely to stick around in the cluster and participate in future dynamical encounters. Only 0.5% ± 0.6% of black holes form in the outer regions (r >5 pc), suggesting that primordial binary evolution contributes very little to the overall black hole population in these dense environments.

6 Next Steps: Where This Research Could Go

This analysis opens up several promising directions for future work. First, I'd love to expand the cluster model grid to include higher resolution simulations (N $>10^6$ stars) to study rare massive collision events that produce intermediate-mass black holes. The current analysis revealed that only 6.7% of models in our ideal parameter produce black holes with non-zero spins.

The success of my Random Forest approach suggests that more sophisticated machine learning techniques could extract additional insights. Deep learning approaches using convolutional neural networks could analyze the full evolution of clusters, potentially identifying new formation pathways .

I'm also now excited about the observational validation possibilities. Our predictions for collision rates and black hole formation efficiency can be tested through multi-wavelength observations of young massive clusters. X-ray surveys could search for the enhanced detection rates predicted for high-velocity dispersion systems, while searches for fast blue optical transients could provide

direct evidence of stellar collision processes.

References

This analysis was conducted using CMC simulation data and standard Python libraries. The complete analysis code,setup instructions and datasets are available in the project GitHub repository at https://github.com/ASTR154/Ishaan-Final-project .
-CMC code website: https://clustermontecarlo.github.io/CMC-COSMIC/intro/index.html
- CMC code paper: https://arxiv.org/abs/2106.02643
- CMC model catalog: https://ui.adsabs.harvard.edu/abs/2020ApJS..247...48K/abstract
- Massive BH formation in clusters:https://ui.adsabs.harvard.edu/abs/2020ApJ...903...45K/abstract

6.1 Generative AI Usage declaration

Since this is an extension of a research project of mine I had used ChatGPT during my initial individual model analyses, primarily to to help me with setting up my individual file parsing functions an the error handling that came with it .