USTA Junior Pathways

Final Presentation

Pramit Vyas, Dominick Robinson, Tyler Quinn

May 1, 2026

Table of contents

01

Introduction

About the team, advisor, and project

Meet Dominick Robinson

Academic Background
Undergraduate: CMU in Math (Statistics) and minor in Creative Writing
Graduate: CMU in Masters of Applied Data Science

Professional Path
Future Goals: Data Scientist/Software Engineer (in tennis, if possible)
Hobbies: Tennis, Pickleball, Game Development

Tennis
Favorite Players: Federer, Sincaraz, Medvedev, Isner
USTA Rating: 4.5
Current Setup: Pure Strike 98 (2014) with MSV Focus Hex @ 50lbs

Meet Pramit Vyas

Academic Background
Undergraduate: Iowa State with majors in Statistics and Data Science with minors in Computer Science and Political Science – Part of Sports Analytics Club, Tennis Club

Graduate: CMU in Masters of Applied Data Science – Carnegie Mellon Sports Analytics Center

Professional Path
Future Goals: Data Scientist, ML Engineer in Sports
Hobbies: Basketball, Tennis, Pickleball

Tennis
Favorite Players: Sinner, Tiafoe, Djokovic

Meet Tyler Quinn

Academic Background

University at Buffalo, double major in Statistics and Mathematics

Graduate: CMU, Masters of Applied Data Science – Carnegie Mellon Sports Analytics Center

Summer 2026: Hockey Research Intern for the Pittsburgh Penguins

Professional Path

Future goals: Data Scientist roles in sports

Hobbies: Weightlifting, Reading, Video Games

Tennis

Favorite Players: Djokovic, Serena Williams

Our "Coaches"

Will Townes

Research Interests:
Biomedical, Public Health
Statistics

Jamie McGovern

MADS Program Director

02 Executive Summary

Data sources, information

Executive Summary

Data Collection

Scraped over **3.7 million** match results from CoreTennis.

Integrated historical datasets to identify development pathways for junior tennis players.

Advanced Modeling

Normalized Elo: Quantified player skill & removed era-specific inflation.

Trajectory Similarity: Developed score to identify comparable historical players.

Predictive Insights

Used **GAM** with linear spline basis expansion to project career results.

Resulted in a data pipeline and Tableau dashboard for visualizations and uncertainty estimates.

Predict career trajectories of junior players and identify very successful players early

Primary Project Objectives

Deliverable Specifications

Deliverable

1

Predicts pathways

→ Idea: **Train predictive model**

2

Outlives collaboration

Automatically updates with new data

→ Idea: **Create ETL script**

3

Easy-to-use

For USTA Player Development

→ Idea: **Create dashboard**

Project Workflow

Scrape data

Obtain comprehensive match results across all ages

Ratings

Compute age-universal skill rating

Model

Predict future rating trajectories of current talent based on historical data

Dashboard

Interactively visualize results

Data Overview

03

What data we used

Data Sources and Rationale

	CoreTennis	Jeff Sackmann's ATP/WTA
Contains	Player and tournament info Match results across all ages starting from 2006	All professional match results and rankings in Open Era High-level match statistics
Chosen for	Comprehensive results across all ages required to compute accurate skill rating	Ranking data useful for rating computation and visualization

Scraping CoreTennis

CoreTennis has most comprehensive tennis match result data over all ages we found

But the data was not easily accessible!

So we developed a process to scrape it ourselves...

CoreTennis Scraping Steps

Find a Player

Last name

First name

Gender

Category

Country

GO

Player Search Results

(shiv) Patel	Shivam	Sweden (SWE)	» more info
-	Hridhay	India (IND)	» more info
- Oana Codrea	Nicola	Romania (ROU)	» more info
- Zoi Nikolakopoulou	Vasilia	Greece (GRE)	» more info
.	Abdullah	Pakistan (PAK)	» more info
A	Aakriti	India (IND)	» more info
A	Nikith Prasad	India (IND)	» more info
A	Scortt	South Africa (RSA)	» more info
A	Se	United States of America (USA)	» more info
A Benabe	Luis	Puerto Rico (PUR)	» more info
A M	Suraj	India (IND)	» more info
A R	Nur Siti Eissya	Malaysia (MAS)	» more info
A. F. Junior	Carlos	Brazil (BRA)	» more info
A.keshavan	Padmapriya	India (IND)	» more info
A.lobo Coliccio Ferreira	Armando	Brazil (BRA)	» more info
Aabbou	Benjamin	Belgium (BEL)	» more info
Aadamsoo	Villem	Estonia (EST)	» more info
Aadirai	K A	India (IND)	» more info
Aadith	Ka	India (IND)	» more info
Aadland	Noah Alexander	Norway (NOR)	» more info

< 1 -2-3-4-5--10670>

Roger Federer (SUI)

Country: **Switzerland**, Born in **1981** (44 years old), Category: **ATP**

Profile | Ranking | Results | Statistics | Pictures | More Info

Place of birth	Basel, Switzerland
Residence	Bottmingen, Switzerland
Plays	Right-handed
Backhand	One-handed backhand
Turned Pro	1998
Height	185 cm (6 ft 1 in)
Weight	85 Kg (187 Lbs)
Coaching	Ivan Ljubicic, Severin Luthi (part-time)
Management	Team8

About

Read more about Roger Federer

Arguably the Greatest Of All Time, Roger Federer owns a long list of tennis records, starting with 20 Grand Slam titles to his name: the Australian Open 6 times (2004, '06-07, '10 & '17-18), Roland Garros once (2009), Wimbledon 8 times (2003-07, '09, '12 & '17) and the US Open 5 times (2004-08). Other notable accomplishments include winning the Davis Cup (2014), the ATP Tour Finals 6 times (2003-04, '06-07, '10-11), 27 Masters 1000 and holding the ATP #1 spot for 310 weeks.

Roger Federer Results (SUI)

Country: **Switzerland**, Born **August 8, 1981** (44 years old), Category: **ATP**

Profile | Ranking | **Results** | Statistics | Pictures | More Info

See all tierlevels | See all surfaces | Hide qualif. | See all matches

2021 | 2020 | 2019 | 2018 | 2017 | 2016 | 2015 | 2014 | 2013 | 2012 | 2011 | 2010 | 2009 | 2008 | 2007 | 2006

Jun 28 Jul 10	**Wimbledon** - All England Lawn Tennis and Croquet Club Wimbledon (GBR) - Grand Slam (Men's) - Grass - £17,066,000			
1/64	W	vs	Adrian Mannarino (FRA)	64 67(3) 36 62 Ret'd
1/32	W	vs	Richard Gasquet (FRA)	76(1) 61 64
1/16	W	vs	Cameron Norrie (GBR)	64 64 57 64
1/8	W	vs	Lorenzo Sonego (ITA)	75 64 62
QF	L	vs	Hubert Hurkacz (POL)	63 76(4) 60

Jun 14 Jun 20	**Terra Wortmann Open** - Gerry Weber Stadion Halle (GER) - ATP 500 - Grass - €1,455,925			
1/16	W	vs	Ilya Ivashka (BLR)	76(4) 75
1/8	L	vs	Felix Auger-Aliassime (CAN)	46 63 62

Tournament Info

» Full results

Name	Wimbledon
Location	Wimbledon - Great Britain
Gender	Masc.
Venue	All England Lawn Tennis and Croquet Club
Date	2021-06-28 / 2021-07-10
Surface	Grass
Organisation	Grand Slam Committee
Tier Level	Grand Slam (Men's)

» more pictures and infos

Draw	7 rounds / 128 players
Prize Money	GBP 17,066,000
Sets	5 sets
Website	http://www.wimbledon.org

CoreTennis Data Cleaning Steps

1. Extract

1. Scrape player urls
2. Scrape players
3. Scrape matches from players
4. Scrape tournaments from matches
5. Convert jsonl to csv

2. Transform

1. Transform tournaments
2. Transform players
3. Transform matches
4. Add missing tournaments
5. Label wheelchair tournaments
6. Link qualifying with main draw
7. Clean columns
8. Estimate match dates
9. Add sex to players
10. Map Jeff Sackmann id's
11. Add rankings to matches
12. Add ages to matches
13. Classify tournament

3. Model

1. Compute ratings
2. Find nearest neighbors
3. Predict trajectories with GAM

4. Load

1. Reformat matches table
2. Standardize column names

As of (3/16/2026)

3,723,412
Matches

117,345
Tournaments

212,677
Players

2006-2026
Years

04 Computing Skill Ratings

How do we quantify player success?

We want to predict junior pathways...

But what exactly should we predict?

Well, we can’t use ATP/WTA/Junior ITF rankings

All age groups need to be on the same scale

What is Elo?

Zero-sum method for calculating relative skill levels

Elo recalculates after every match based on the result

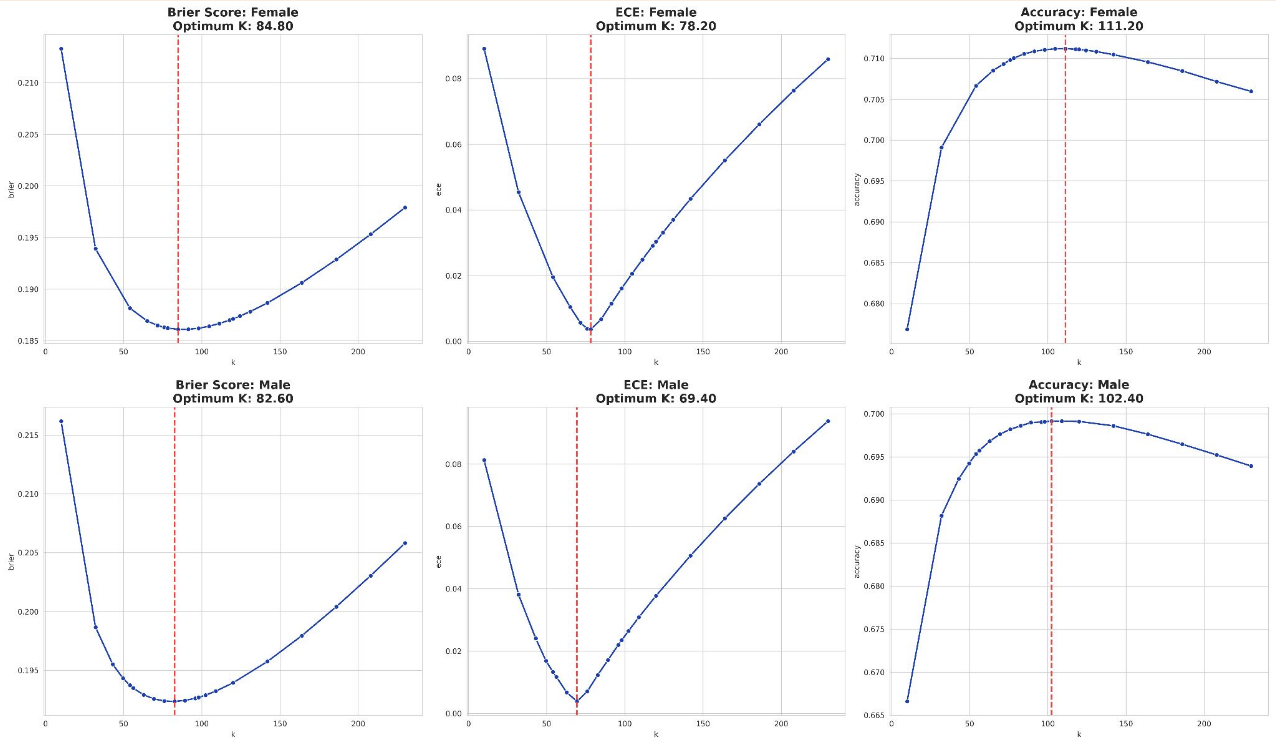
The amount of rating change depends on the opponent's rating (and parameter k)

Possible to calibrate win probability between two players based on ratings

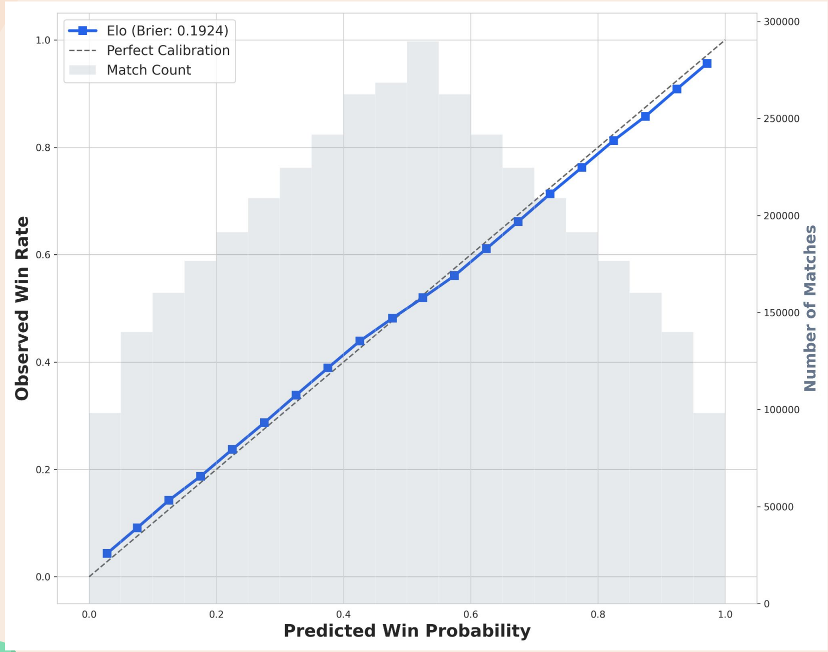
Limitations include possible inflation/deflation over time, and therefore inability to compare ratings across time

Image credit to: Michel Spingler/Associated Press

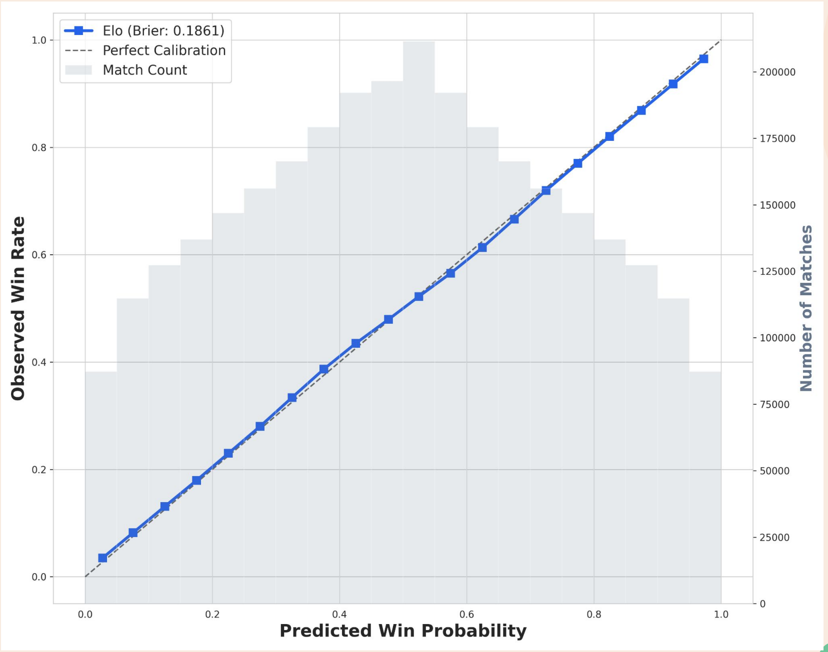
Optimized k parameter for ratings for men and women separately, found k ≈ 80 for each.



Estimated win probabilities are accurate, suggesting that Elo rating accurately reflects skill.

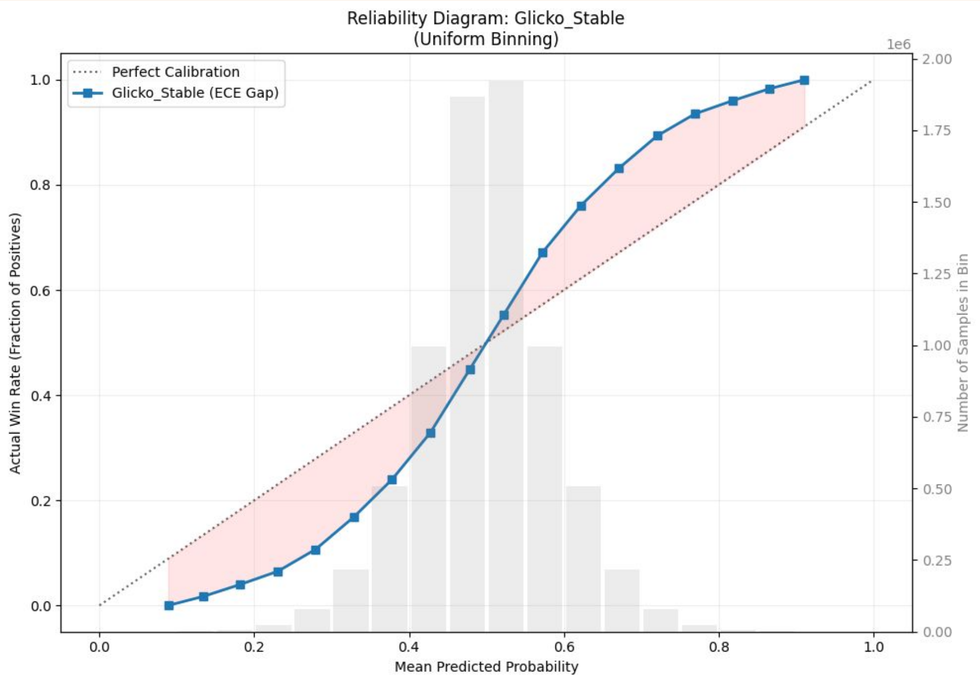


Male

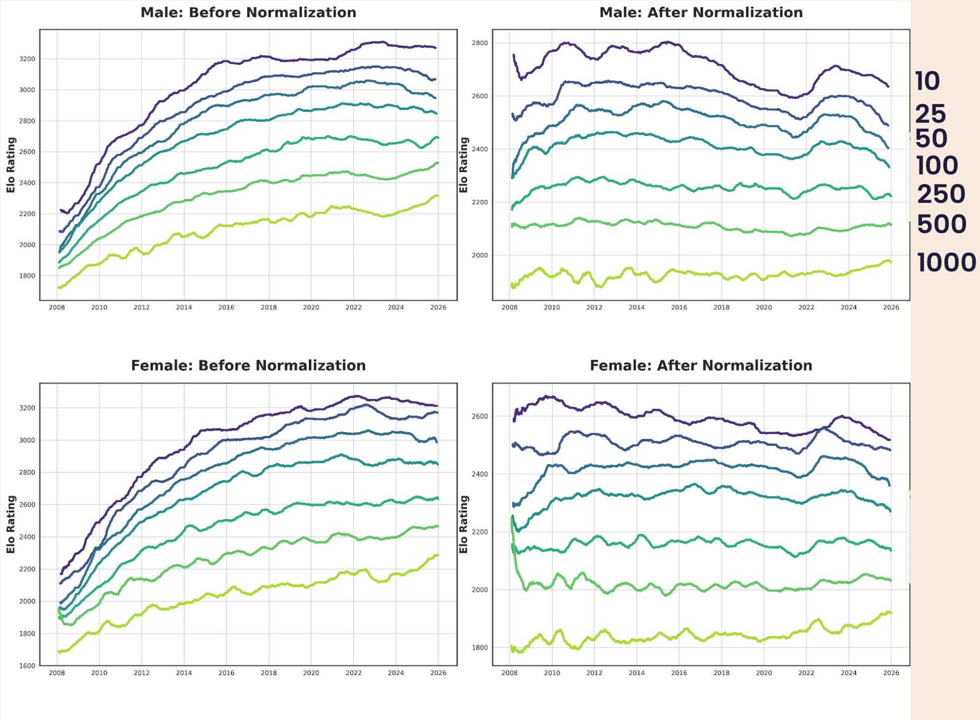


Female

This is opposed to our best Glicko Calibration



Plot of ranking benchmarks by rating over time before and after normalization reveals that rating inflation seems to be dealt with, enabling ratings comparisons across time (crucial for modeling!).



05 Player Similarity

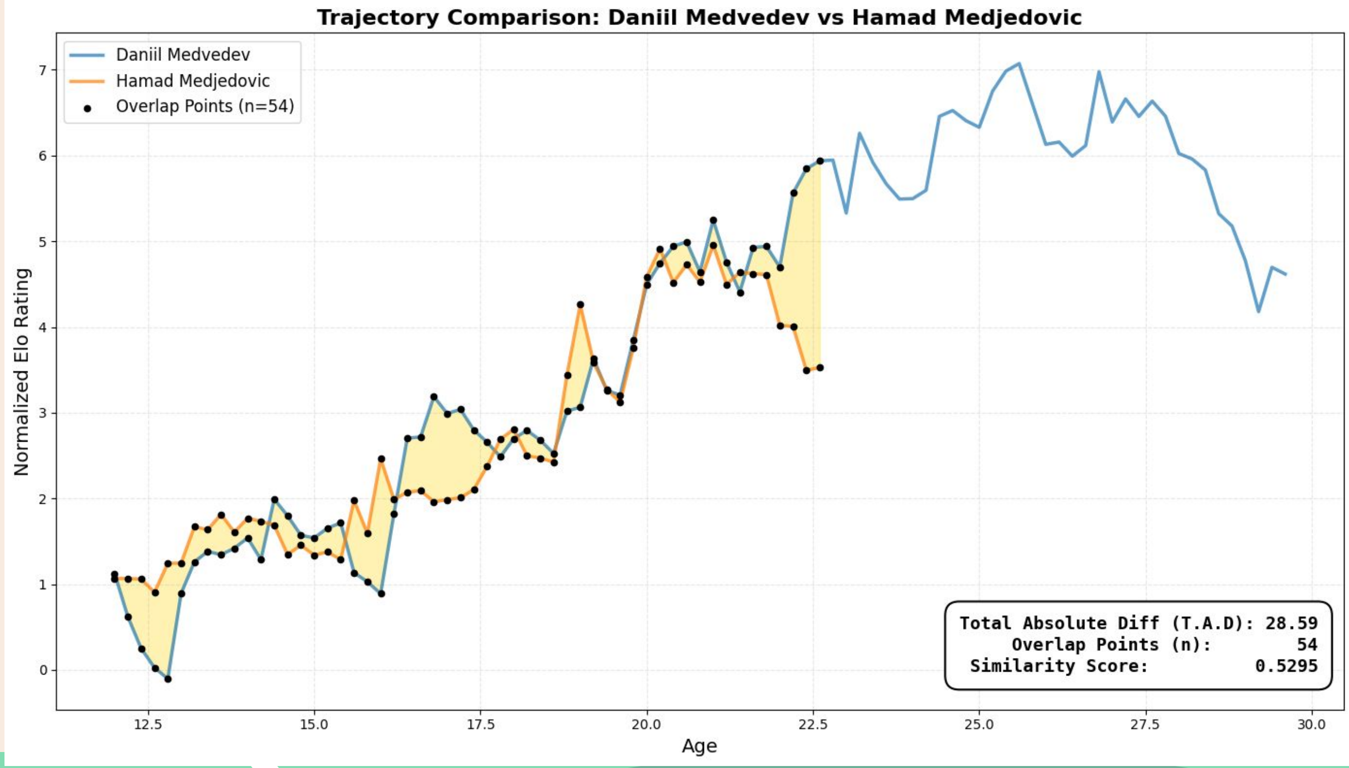
How we determined which players had similar trajectories

How did we measure similarity?

We took the average distance for all ages where the players overlap.

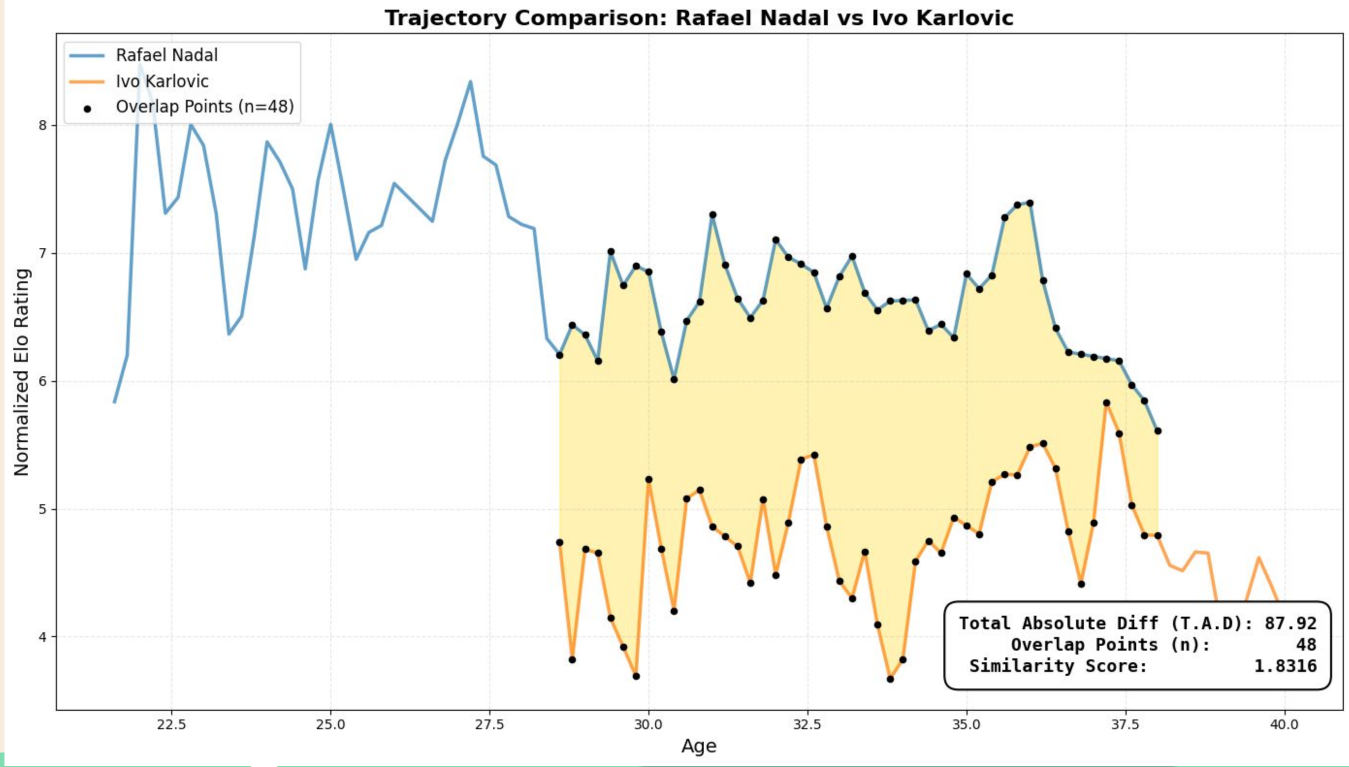
Example– Medvedev and Medjedovic

Their ratings are generally similar at the same ages, so their overall Similarity Score is low (more similar).



Example– Nadal and Karlovic

Nadal and Karlovic's curves are generally farther apart at the same ages, so their overall Similarity Score is higher (less similar).



Restrictions on Comparables for Modeling Inclusion

Players can only show up in a target player’s nearest neighbor list if they meet all of the following criteria:

- Must have at least 2 years of overlap
- Must have played at least 3 years after the target player

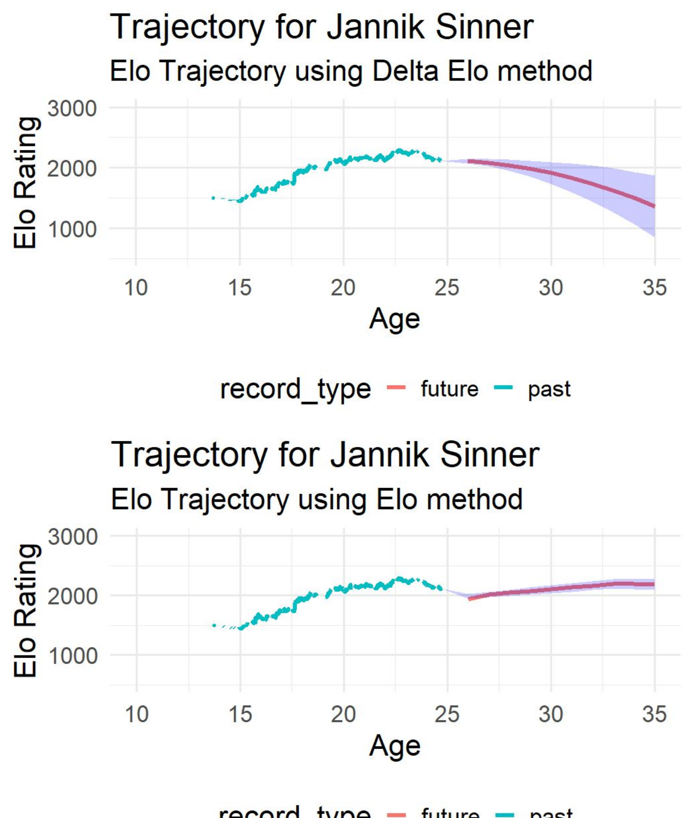
This ensures that we have a decent sample of similar performance, and that their observed trajectories are actually informative for what may happen with the target player

06 Modeling

Statistical Technique to Inform Trajectories

We are predicting ΔElo

Delta Elo makes better projections for Sinner, and the previous method was overconfident.



- We convert the data so every row represents a half year of play for a player
- We take the change in Elo between averages of half year stints
  - Helps in Continuity of results
  - Helps us model the bottom line of trajectory (series of changes in skill)

Variables Used

Results
Historical Individual Player Elos

Age
How old is the player we're projecting

Neighbors Trajectories
List of 30 closest comparable players

Global Trajectories
How the overall averages move

Final Modeling Framework

Hybrid Generalized Additive Model

Combines individual player history with population-level trends and peer-group insights.

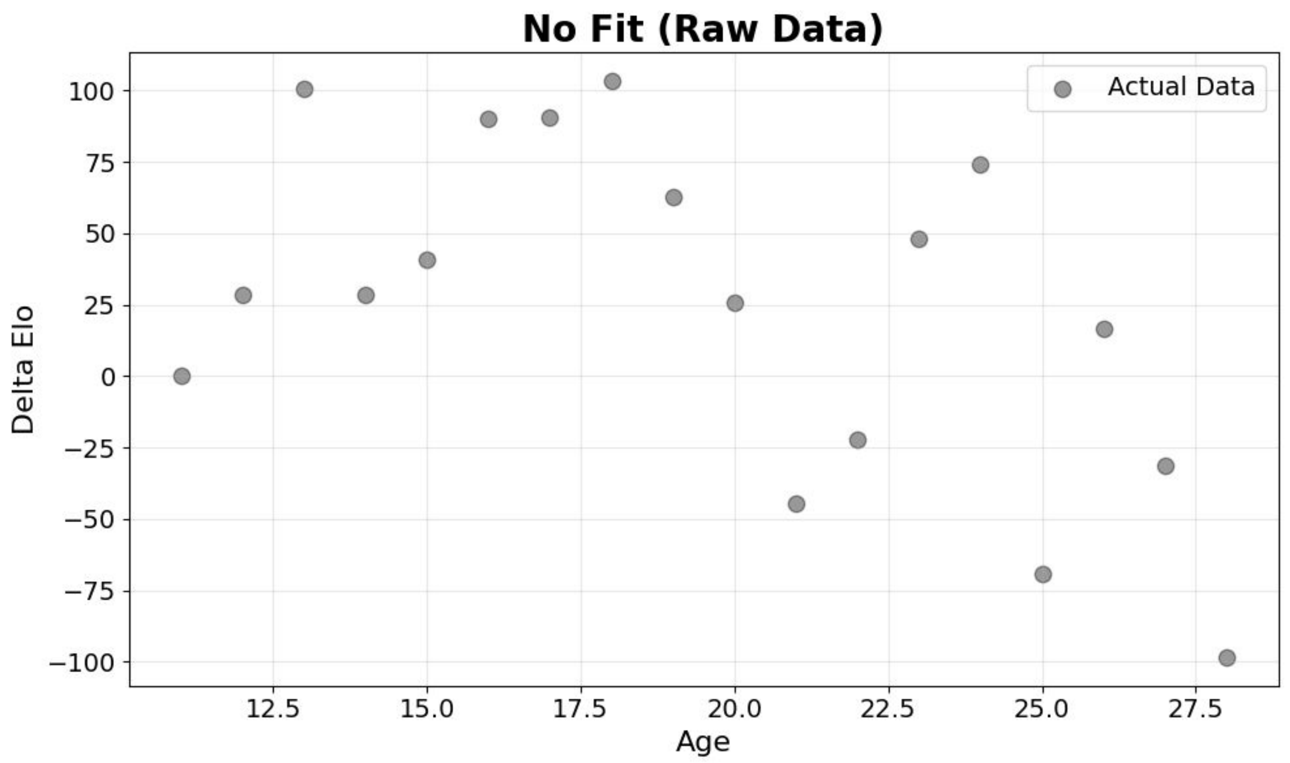
Individual Spline

Global Prior

Similarity Ensemble

3 Pillars of Prediction

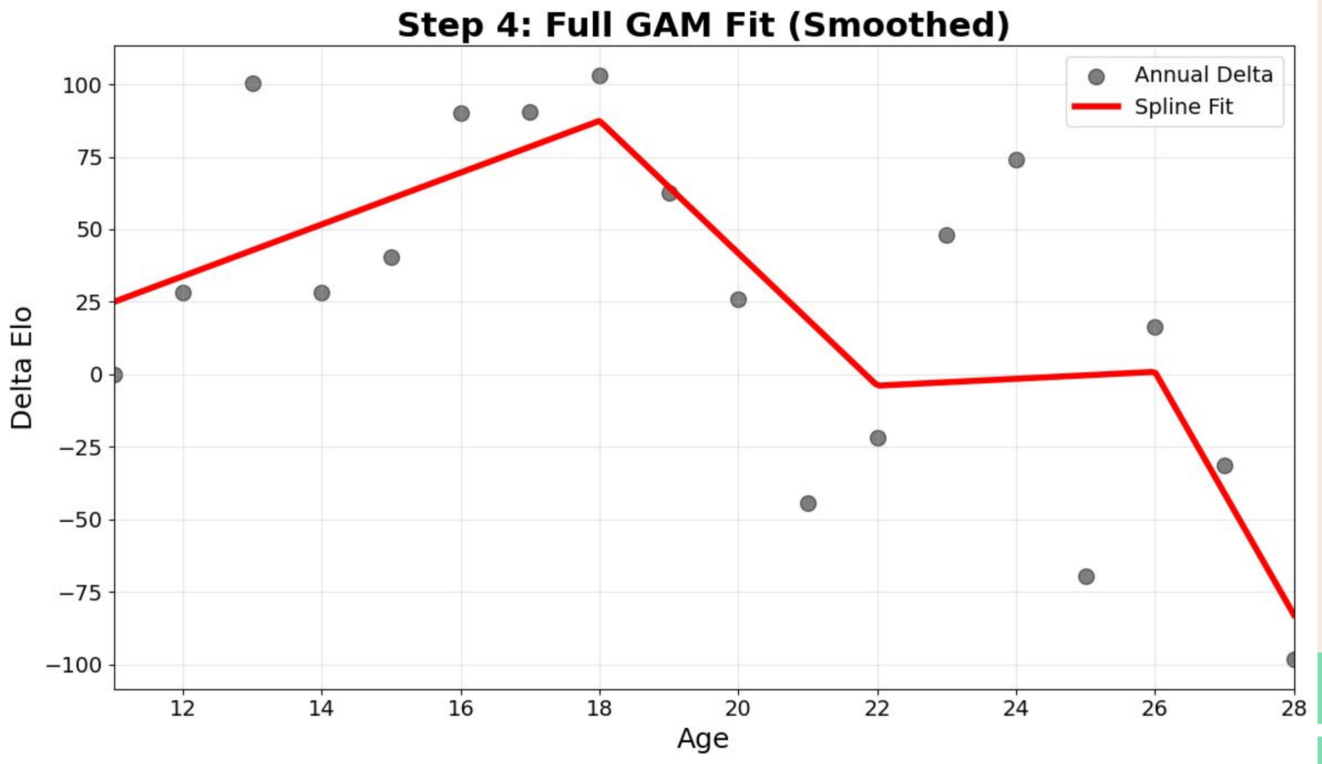
Individual Splines –Example
Zverev



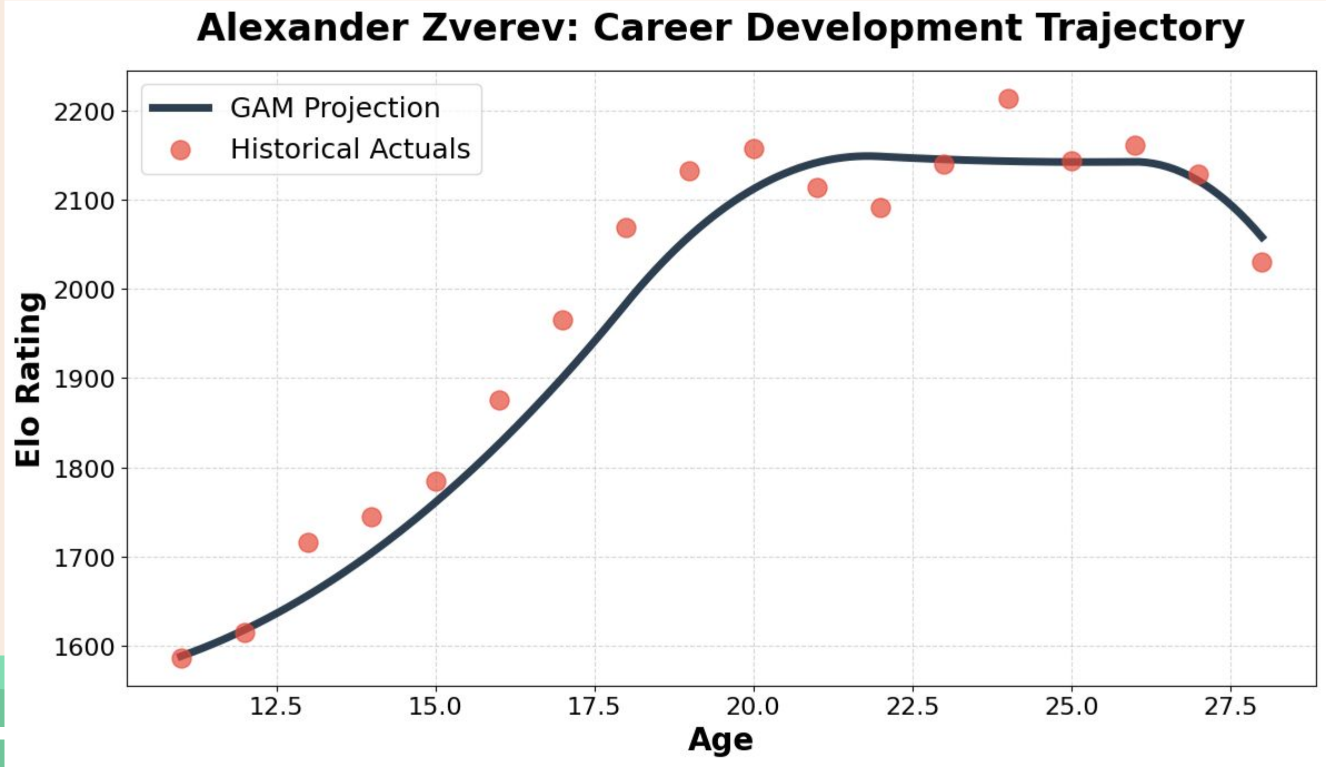
Individual Splines –Example Zverev

Individual Splines –Example
Zverev

Individual Splines –Example
Zverev



Individual Splines –Example Zverev

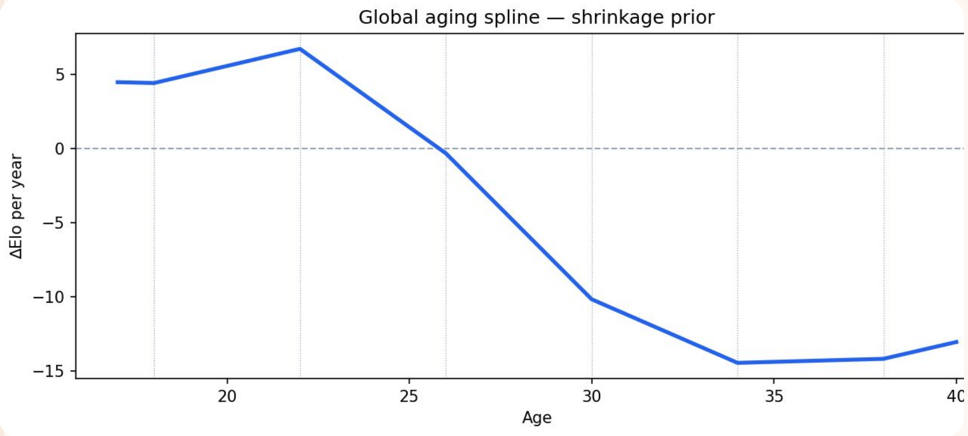


Global Spline

Key Insights

- If a player does not have enough data, we regularize with a global spline.
- This shows overall what ages players gain vs lose Elo.
- Implied peak age identified at 25.8.

Global aging spline — shrinkage prior



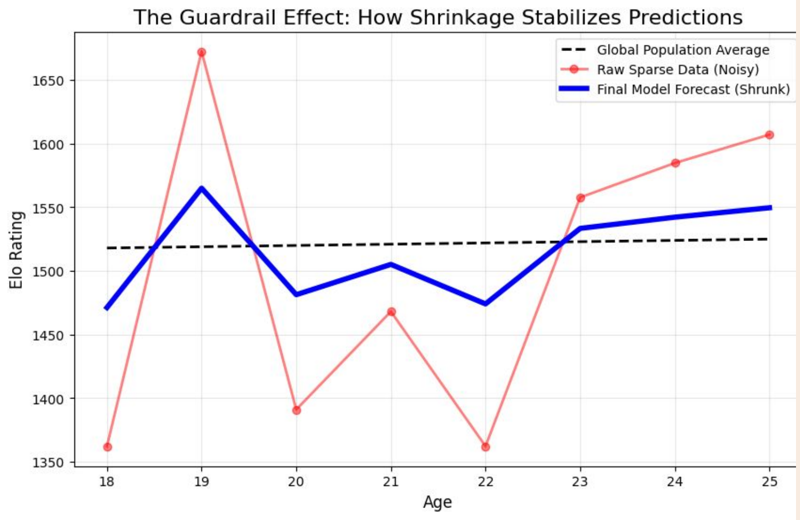
"Borrowing" Strength

The "Cold Start" Problem

What happens if a player has only a few matches?

Our Solution: Shrinkage

The less data we have, the more the model pulls the trajectory toward the **Global Average**.



Outcome: Prevents "hallucinated" peaks for young players with insufficient data.

What happens if neighbors don’t have data past a certain age

Global Spline Alignment

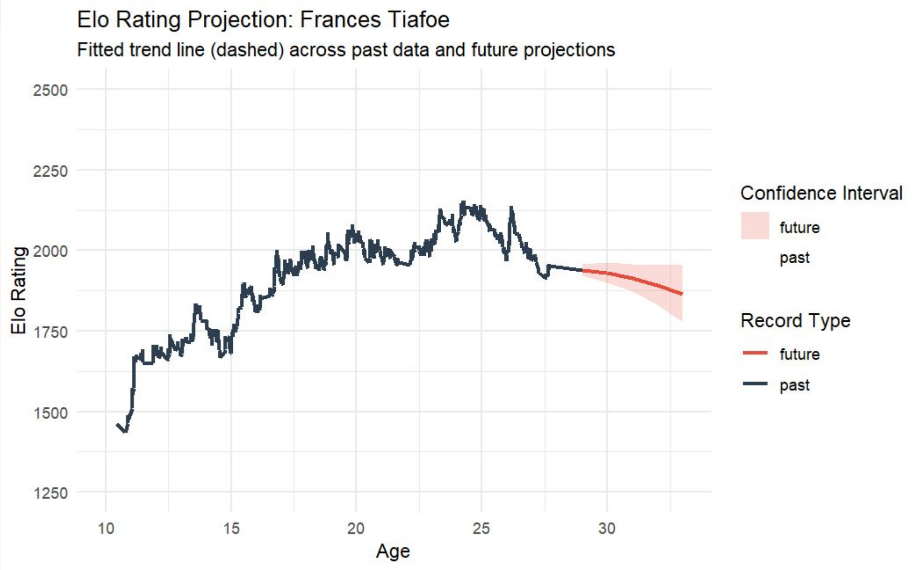
Regularized Ratings towards the global spline for consistent baseline comparison.

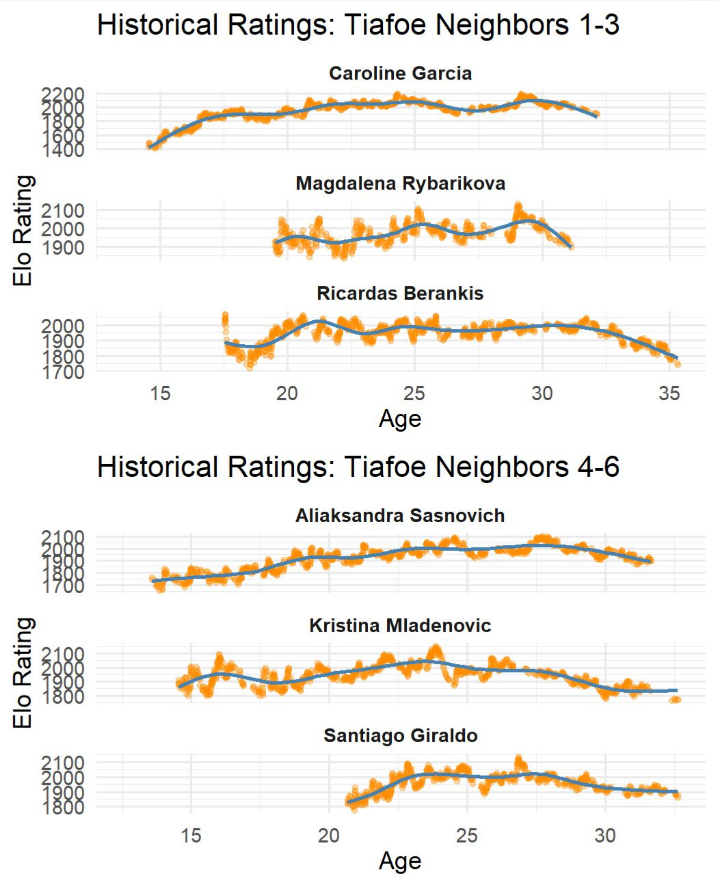
Predictive Trajectory

Their own trajectory predicted up to that point, based on available historical data.

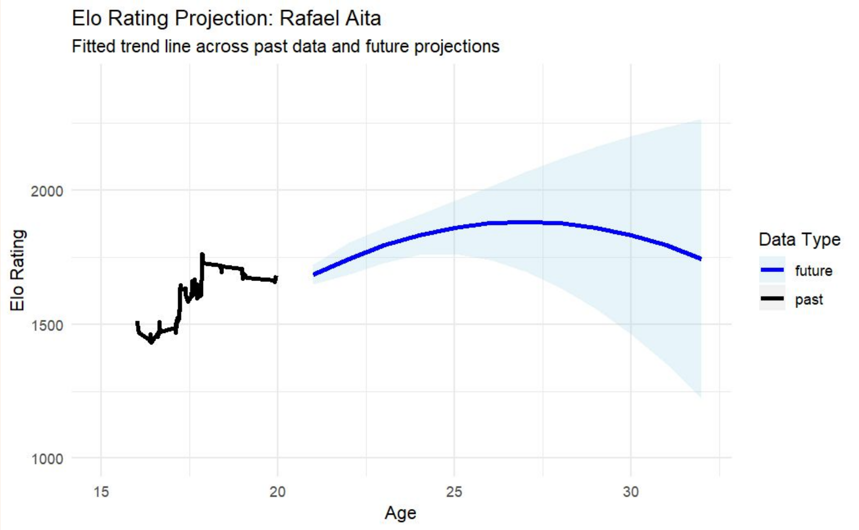
We consider these 2 factors to handle cases where neighbors lack data past a certain age.

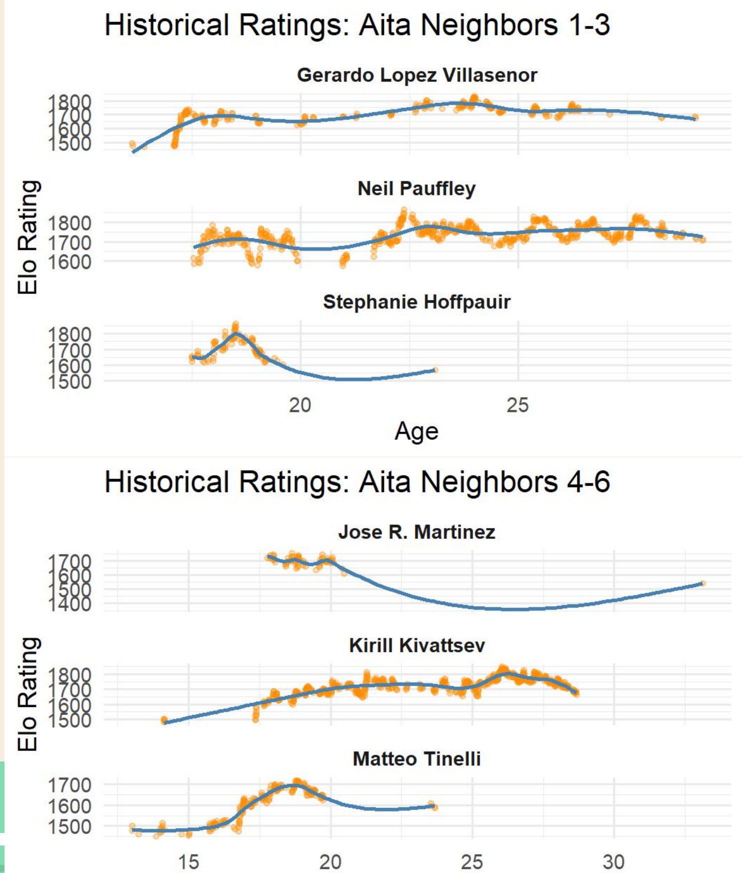
Model use examples



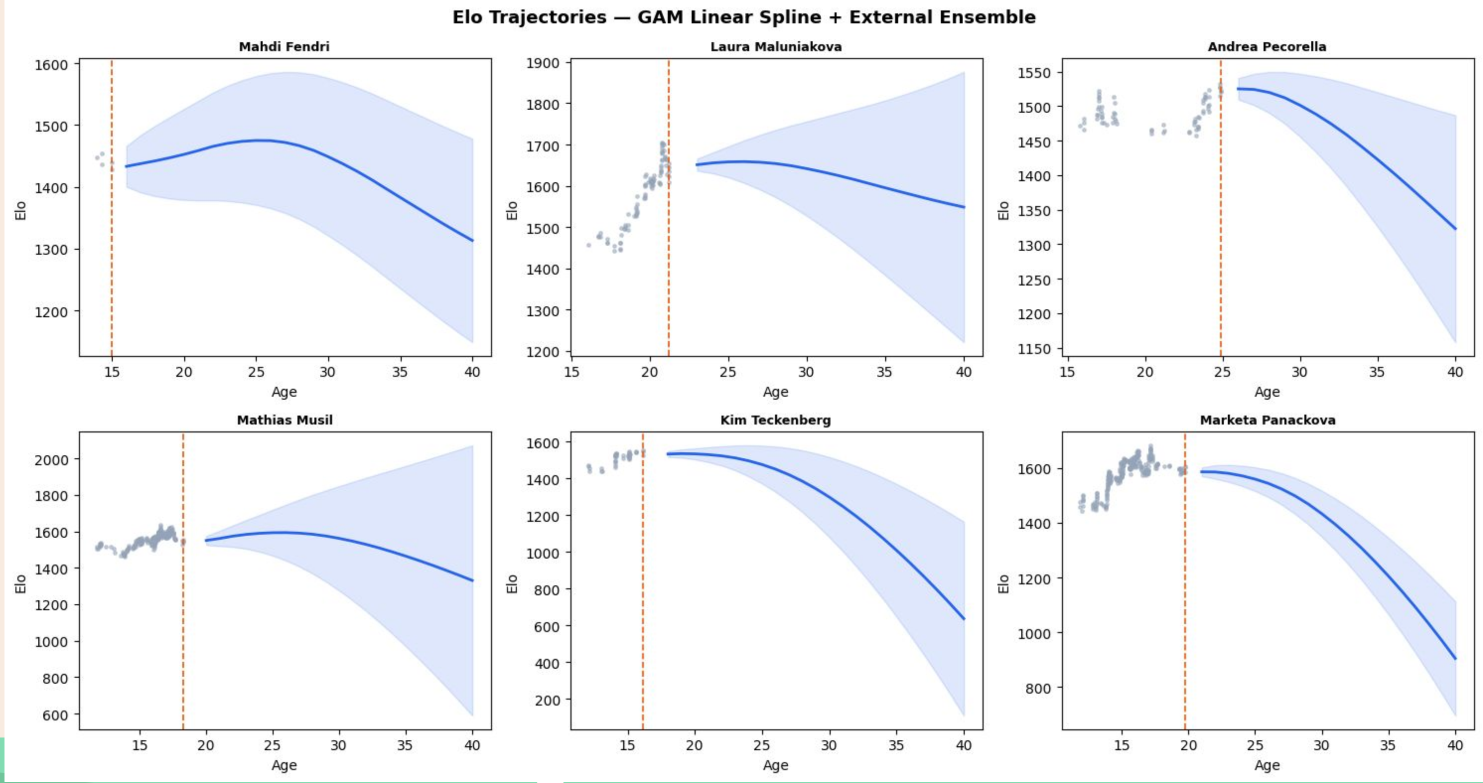


Model use examples





Model use examples



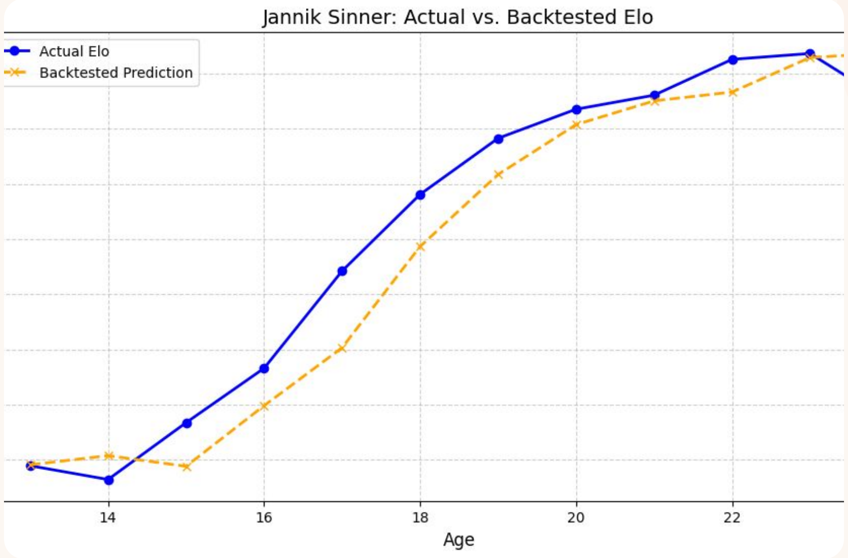
How we tested

Testing Methodology

For every select player we would like to test on, we follow these core steps:

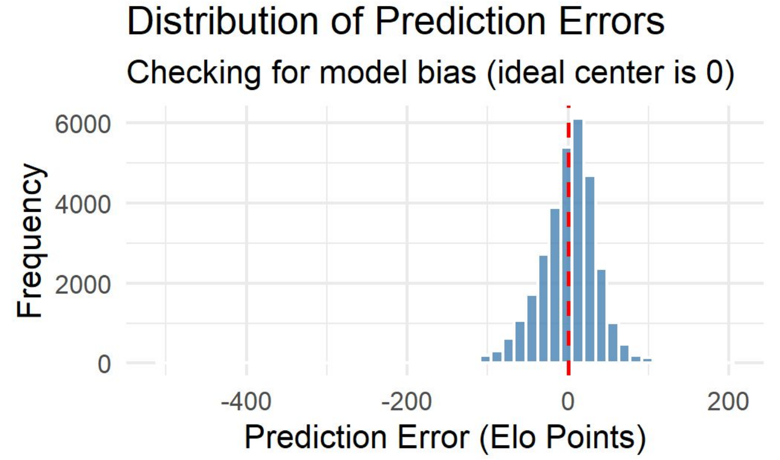
- Subset player data to events occurring past a previous age.
- Filter peer trajectories to those at or before the target calculation year.
- Fit the new trajectory incorporating the updated information.
- Calculate the delta between predicted and actual trajectory by age.

Sample Backtest: Actual vs. Prediction

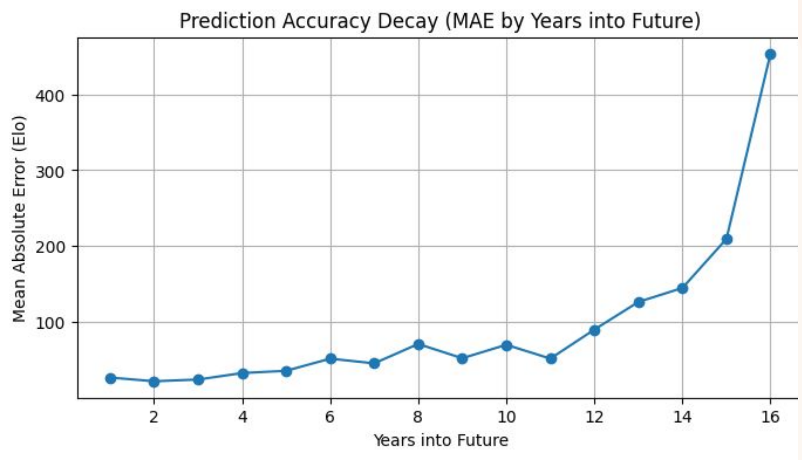


Model Diagnostics

Error Distribution

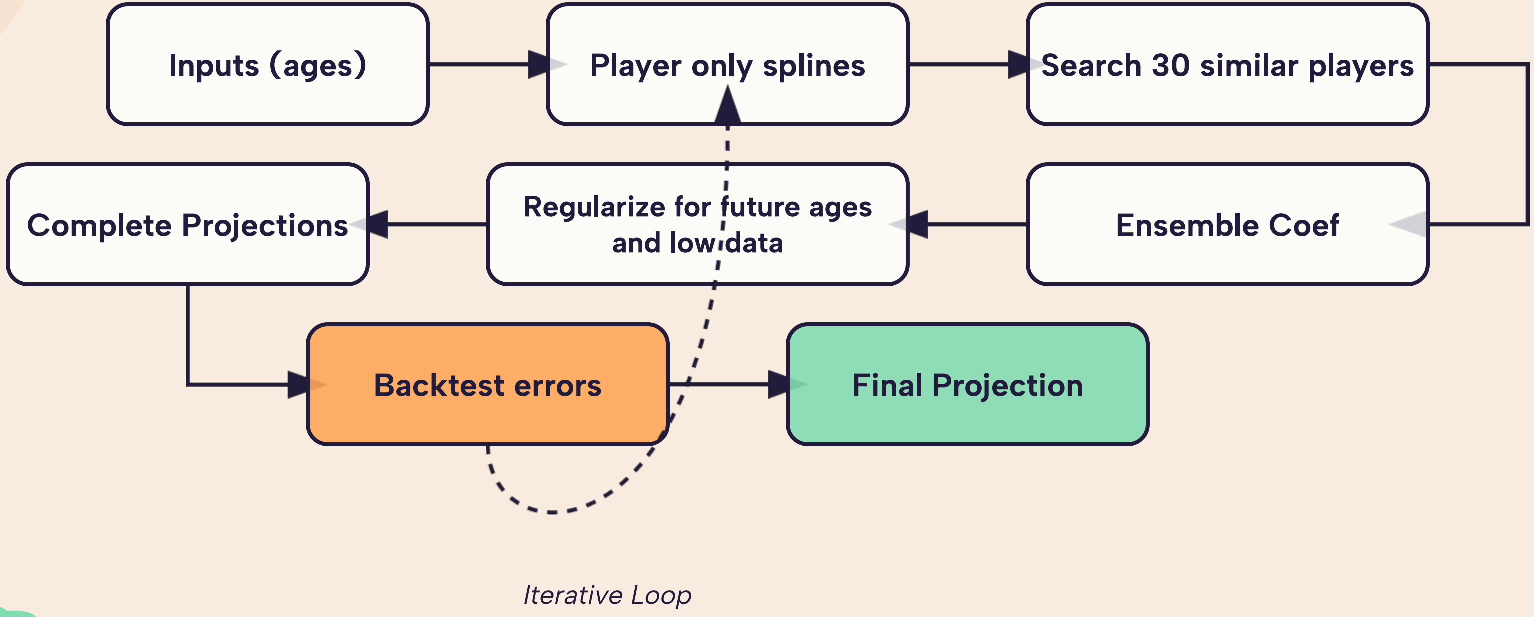


Accuracy Over Time



MAE: 27.2
Bias: Well Calibrated

Model Overall Flow



07

Dashboard

Live Demo of our Deliverable

Player Profile

USTA Junior Player Pathways (by CMU MADS)

Player Profile | Players with Similar Rating Trajectories | Rating Predictions with Ranking Thresholds

Player Name: Carlos Alcaraz

Match Date: 1/2/2006 – 3/16/2026

Tournament Level: (All)

Tournament Category: (All)

Tournament Round: (All)

Bio

Country: Spain
Sex: Male
Born: 2003

Height: 183
Weight: 74
Handedness: Right-handed
Backhand: Two-handed backhand

Highest Rated Wins

Name	Rating	Tournament	Date
Jannik Sinner	2944.21	BNP Paribas Open	March 15, 2024
Jannik Sinner	2934.49	Internazionali BNL d'Italia	May 18, 2025
Rafael Nadal	2921.77	Mutua Madrid Open	May 6, 2022
Novak Djokovic	2916.49	Wimbledon	July 16, 2023
Jannik Sinner	2914.66	French Open / Roland Garros	June 8, 2025

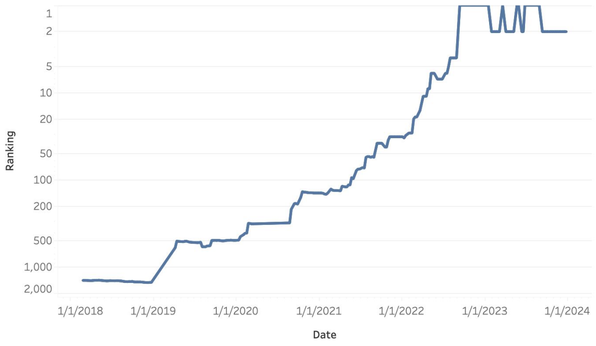
Career Stats

WIN/LOSS
Wins: 574
Losses: 109
Total Matches: 683

Tournaments Won: 42
Finals Lost: 13
Finals Played: 55

RANKING PERFORMANCE
Peak Ranking: 1 (Date: September 12, 2022)
Latest Ranking: 2 (Date: December 25, 2023)

Ranking Plot



Players with Similar Rating Trajectories

USTA Junior Player Pathways (by CMU MADS)

Player Profile | Players with Similar Rating Trajectories | Rating Predictions with Ranking Thresholds

Player

Carlos Alcaraz

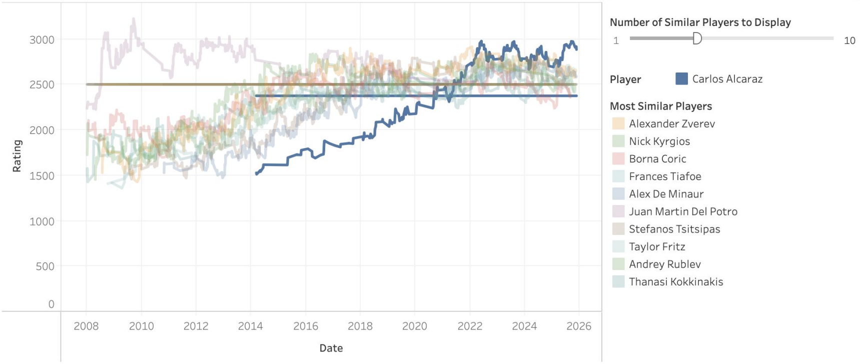
Similar Players Table

Index	Player	Similarity Score
1	Alexander Zverev	122.79
2	Nick Kyrgios	168.36
3	Borna Coric	207.13
4	Frances Tiafoe	209.58
5	Alex De Minaur	212.28
6	Juan Martin Del Potro	213.61

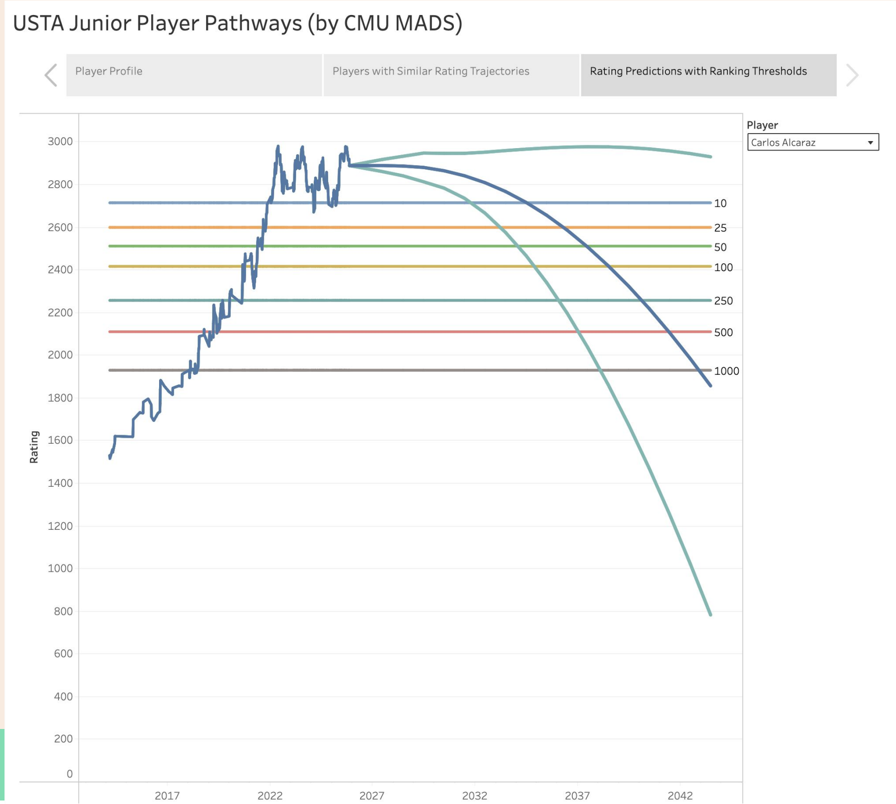
Similar Players Plot

Select X-Axis

Date



Rating Predictions with Ranking Thresholds



Dashboard Live Demo

08

Discussion

Summary, limitations, and possible future work

What We Delivered

CoreTennis Scraping + Cleaning Pipeline

Gather data/clean it

Modeling

Create player similarity scores, use them for projections

Dashboard

Visualize all in a digestible format

Limitations

Elo Normalization

We assume a Normal distribution for the Elo ratings at any given time, but in reality it likely has heavier tails (outliers are more common)

Age

We assume every player is born July 1st of their birth year because we don't have access to birthdays for all players

30 Nearest Neighbors

We didn't rigorously optimize the number of neighbors included in the model, this parameter can be tuned

Confidence Intervals

We calculate confidence intervals from the GAM coefficients, but they could be inaccurate if our assumptions are untrue

Future Work

Scrape ITF Website

They have matches from earlier than CoreTennis to provide more data, but they seem much harder to scrape

Alternative Rating Systems

Try further tuning Glicko 2 or using WTN (as long as data is still comprehensive)

Automate/Improve Pipeline

Change scrape order to not scrape entire CoreTennis website each run

Automate to run entire ETL-to-Tableau pipeline regularly in USTA server

Survivorship Analysis

Account for the probability of players dropping out of the dataset when making projections (e.g. injuries, retirement)

Learnings for Future Projects

Data Availability

CoreTennis API or other long-term data source

Project Planning

Picking a direction and strict schedule early in the process is helpful

Data Documentation

Clear information on what is contained in the data lake would be helpful so we know what is available

Dashboarding

We learned how to use Tableau!

Thanks!

Do you have any questions?

Special thanks to:

Dave Ramos

Peter Rios

Lauren Rothstein

Alex Torres

Appendix

Model Coefficients

Player Name	Years Data	Shrinkage (λ)	Intercept	Age Coef	Knot @ 22	Knot @ 26
Rafael Nadal	18.0	0.5488	1.25	-3.83	-33.32	9.67
Frances Tiafoe	18.0	0.5488	56.41	-2.10	-4.28	-22.23
Jannik Sinner	13.0	0.6484	-115.09	9.42	-9.74	-0.45
Carlos Alcaraz	13.0	0.6484	-45.53	5.14	-1.51	-0.45
Holger Rune	11.0	0.6930	-78.21	4.88	-1.22	-0.45

Formulas Explained

Shrinkage Coefficient

Balance between a single player and global trajectory

$$\lambda = \exp(-\text{n\_years} / \text{shrink\_threshold})$$

Bayesian Coefficient Blending

Weighted average of individual and global priors

$$\text{Coef\_blended} = (1 - \lambda) \cdot \text{Coef\_own} + \lambda \cdot \text{Coef\_global}$$

Linear Spline Basis

Accounts for the flexibility of a curve during individual spline fitting

$$\text{b\_k}(x) = \max(0, x - k)$$

* All formulas applied to individual spline fitting processes

Formulas Explained

Weighting of Coefficients

We want to weight neighbors coefficients by how similar they are

$$\text{Ensemble}(\beta) = \Sigma(\text{Sim\_Weight}_i \times \text{Coef}_i) \,/\, \Sigma\text{Weights}$$

Elo Predictions

Calculate predictions for a player using ages and coefficients

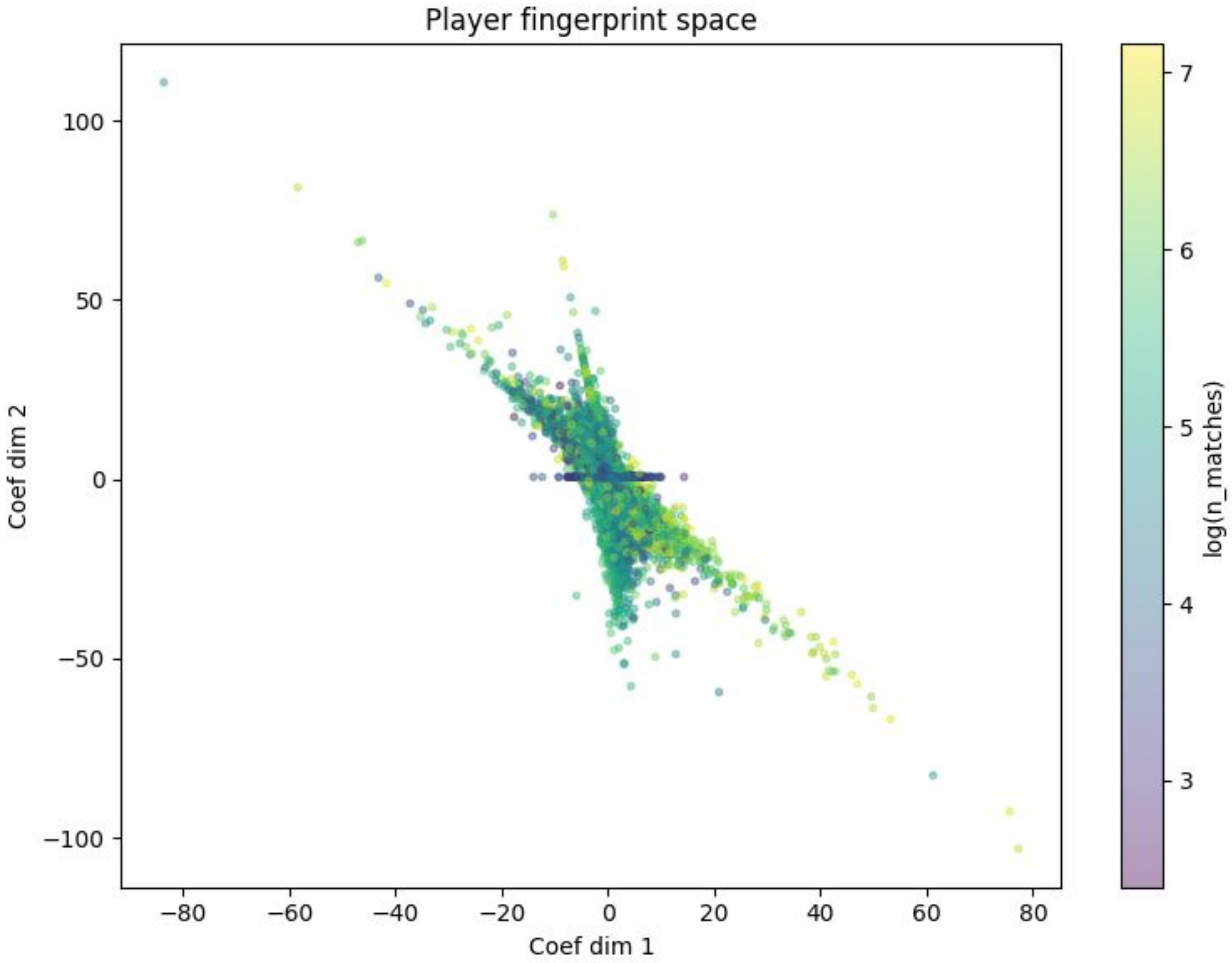
$$\Delta\text{Elo}\square_{re0} = \beta_0 + \beta_1(\text{age}) + \Sigma(\beta\square \cdot b\square(\text{age}))$$

Calculating Full Trajectory

Convert delta Elo predictions into actual Elo trajectory over time

$$\text{Elo\_\{age T\}} = \text{Elo\_\{last\}} + \Sigma\text{\_\{t=last\}\^{}\{T\}}\ (\Delta\text{Elo}\square \cdot \Delta t)$$

* All formulas applied to make predictions


Player fingerprint space
Coef dim 1
Coef dim 2
log(n_matches)
100
50
0
−50
−100
−80
−60
−40
−20
0
20
40
60
80
7
6
5
4
3