Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess

Reid McIlroy-Young
University of Toronto
reidmcy@cs.toronto.edu

Russell Wang
University of Toronto
russell@cs.toronto.edu

Siddhartha Sen
Microsoft Research
sidsen@microsoft.com

Jon Kleinberg
Cornell University
kleinberg@cornell.edu

Ashton Anderson
University of Toronto
ashton@cs.toronto.edu

Abstract

The advent of machine learning models that surpass human decision-making ability in complex domains has initiated a movement towards building AI systems that interact with humans. Many building blocks are essential for this activity, with a central one being the algorithmic characterization of human behavior. While much of the existing work focuses on aggregate human behavior, an important long-range goal is to develop behavioral models that specialize to individual people and can differentiate among them.

To formalize this process, we study the problem of behavioral stylometry, in which the task is to identify a decision-maker from their decisions alone. We present a transformer-based approach to behavioral stylometry in the context of chess, where one attempts to identify the player who played a set of games. Our method operates in a few-shot classification framework, and can correctly identify a player from among thousands of candidate players with 98% accuracy given only 100 labeled games. Even when trained on amateur play, our method generalises to out-of-distribution samples of Grandmaster players, despite the dramatic differences between amateur and world-class players. Finally, we consider more broadly what our resulting embeddings reveal about human style in chess, as well as the potential ethical implications of powerful methods for identifying individuals from behavioral data.

1 Introduction

The advent of machine learning models that surpass human decision-making ability in complex domains raises the prospect of building AI systems that can collaborate and interact productively with people. While many advances will be essential in order to achieve this far-reaching goal, algorithmically characterizing aspects of human behavior will be a significant step. Whereas existing work focuses on characterizing human behavior in an aggregate or generic sense, genuine human-AI interaction will require specializing these characterizations to individual people.

Developing algorithms that can detect signatures of individual decision-making style would constitute an important advance for many reasons. First, being able to recognize an individual from their behavioural characteristics would greatly benefit an AI system’s ability to coordinate with that individual. In general, decision-making styles can either be complementary or clashing, and a predictive characterisation of an individual’s decisions would be useful for an AI system trying to adapt to a particular person’s style. Second, such fine-grained individual-level characterizations would help isolate the distinguishing features of individual styles. In a domain where ML systems

are able to match or surpass human performance, it becomes natural to try extending these systems to
be able to pinpoint types of decisions that a person is particularly good at, and types of decisions they
could potentially improve. Finally, applying these algorithms to many different individuals would
shed light on the structure of decision-making in the domain itself. For example, certain instances
could prove particularly discriminative in detecting individual style. In addition to these benefits, the
ability to identify individuals from behavioral traces also raises ethical concerns centered around
privacy and generalization to high-stakes domains. We discuss the ethical consequences of our work
in Section 5.

We formalize this by studying behavioral stylometry, in which the goal is to identify a decision-maker
from their decisions alone. We pursue this task in the domain of chess, an ideal model system for
developing human-compatible algorithms for many reasons. AI systems surpassed all human-level
ability in chess at least 15 years ago. Despite this, millions of people at all skill levels still actively
play. Through their activity, people generate detailed traces of their decisions, which are digitally
recorded and publicly available for processing. In chess, behavioral stylometry reduces to predicting
the identity of a player from the moves they played in a given set of games.

We present a transformer-based approach to behavioral stylometry in chess. We show how to adapt
recent work in speech verification to the problem of detecting individual decision-making style.
Working in the domain of chess requires that we address the challenge of extending these methods to
the complex types of data found in the game-playing structure of chess. We apply our model in a
few-shot classification setup, where only a small number of games are provided for the model to learn
a representation of a player’s decisions, and the model must identify the player from thousands of
possible candidates based on another small sample of their games. Compared to a previous approach,
our method achieves high accuracy on this task, and is much more scalable and data-efficient.

Our resulting embeddings of players and their games reveal a structure of human style in chess, with
players who are stylistically similar clustering together in the space. We consider the implications
of these findings, including the potential ethical implications of powerful methods for identifying
individuals from behavioral data.

2 Related Work

Speaker verification. Our work adapts recent work in speaker verification [46, 52], where given a
speaker’s known utterances, the task is to verify whether an utterance belongs to that speaker. Recent
approaches embed utterances into a vector space such that different speakers are distinguishable [28,
42]. These embedding approaches can also be used for downstream tasks such as voice cloning and
multi-speaker speech synthesis [21, 27, 34].

A key decision for our task is the choice of loss function, of which two families are most relevant [13].
Classification-based objectives are designed to classify a speaker out of all possible speakers. Cross-
entropy loss is a simple approach, with many variations including angular softmax loss [29], cosine
loss [50], and additive angular margin loss [14]. Contrast or metric-learning based objectives instead
directly optimize the contrastiveness for samples in different classes and similarity for those within
the same class. Contrastive loss [11], triplet loss [38] and generalized end-to-end loss (GE2E) [49]
are used in face verification, face clustering and speaker verification respectively. In this work, we
adapt GE2E to the problem of behavioral stylometry.

Stylometry. Both speaker verification and handwriting recognition are stylometry tasks where
the goal is to detect and analyze authorial style. Stylometry has historically been associated with
attribute studies [45] and has been applied to de-anonymizing programmers from their code [8],
identifying age and gender from blog posts [22] and distinguishing machine-generated text from
human-generated text [5]. It is often framed as a task to determine the similarity between samples or
features extracted from the samples [32]. Handwriting recognition has employed neural methods to
identify individuals for decades [7] and development is ongoing [23, 24]. We introduce the task of
behavioral stylometry, in which the goal is to identify decision-makers from their decisions.

Chess. Chess has a long history in academic research as both a test bed for artificial intelligence [39]
and for understanding human behaviour [1, 6, 9]. Recently these two goals have begun to merge, as
We begin by describing our model design for performing behavioral stylometry in chess. The model with the closest training samples and their associated classes. Several architectures have been developed, such as matching networks [4, 10, 48] and prototypical networks [36, 41]. Finally, meta-learning [2] [44] is another popular approach, which uses a two-level learning process to train a model across similar tasks and adapts it to solve new tasks with a small number of training samples [20, 35].

**Few-shot learning.** We approach behavioral stylometry as a few-shot learning task with tens to hundreds of samples per class, which is a well-studied problem formulation [18, 19, 51]. Transfer learning is a main approach to few-shot learning, where models are first trained on a large dataset and then specialized to several smaller datasets [16, 53, 43, 53]. Embedding learning [26] is another approach where an embedding space is constructed so that similar training samples are close together and different samples are further apart. Test samples can then be mapped to this space and compared with the closest training samples and their associated classes. Several architectures have been proposed, such as matching networks [4, 10, 48] and prototypical networks [36, 41]. Finally, meta-learning [2] [44] is another popular approach, which uses a two-level learning process to train a model across similar tasks and adapts it to solve new tasks with a small number of training samples [20, 35].

### 3 Methodology

We begin by describing our model design for performing behavioral stylometry in chess. The model has three main components: the first extracts features from individual moves, the second aggregates information from the move level to the game level, and the third aggregates information from the game level to the player level. Given a chess position as input, the first step uses residual blocks to extract move-level features. Then in the second step, we aggregate move-level features into a game representation with a transformer-based method. Finally, the third step aggregates a player’s game vectors by calculating their centroid to produce a player vector that captures the target player’s individual playing style.

**Move-level feature extraction.** To extract features from chess boards and moves, we draw upon previous work (AlphaZero and Maia) and use a series of residual blocks. The input to the model is encoded as a 34-channel $8 \times 8$ chess move representation, and we represent a move with two board positions, corresponding to the boards before and after the player’s move. The first 24 channels each encode a specific type of chess piece (6 for each side in one position, and we encode two positions), and the next 10 channels encode position metadata such as repetition, castling rights, the active player’s side, 50-move-rule count, and border information. The output is a 320-dimensional feature vector that characterizes each move. Each residual block consists of convolutional layers, batch normalization, ReLU activations, and squeeze-and-excitation (SE) blocks [25]. On top of the residual blocks we add global average pooling, which summarizes the output from residual blocks into a $1 \times 1 \times 320$ feature map, which we flatten into a 320-dimensional vector. The hyperparameters of
the model were chosen to match the Maia architecture: 6 residual blocks with channel size of 64 and a reduction ratio for SE blocks of 8.

**Aggregating information to the game level.** The next component of the model takes in a sequence of move features extracted from the first component and constructs a game representation. The input features correspond to the moves played by the target player in a particular game. The output is a game vector that aggregates information about the move features. Although previous work used an LSTM to perform the analogous task in speaker verification, our experimentation revealed that a transformer \[47\] worked better for our task. Hence, in our approach we use a modified version of the Vision Transformer (ViT) \[17\], which was originally proposed to replace CNNs with a pure transformer in vision tasks and achieved comparable results to state-of-the-art CNN architectures with fewer computational resources required. As we consider chess moves as 8 × 8 images with 34 channels, we can adapt ViT to fit in our setting.

To do this, we first modified the input to ViT to take in a sequence of move features instead of fixed-size patches split from an image. Both BERT \[15\] and ViT append a classification token to the input sequence, and the final hidden state of this token is considered as a representation of the entire sequence. However, in our task we achieve better performance if we directly average the output vectors of transformer blocks, so we remove the classification token in the input. As all move features are fed into the transformer simultaneously, we retain the ordering of the chess moves by adding a positional encoding to each move feature. The original ViT architecture includes a learnable 1D position embedding in input sequence, but it cannot easily accommodate variable length inputs during training. Therefore, we use sine and cosine functions to encode positions, a method proposed in the standard transformer, and we include support for up to 500 chess moves.

As an overview of our model, each 320-dimensional input move feature vector is first projected into 1024 dimensions and a positional encoding is appended to the resulting vectors before they are fed into the transformer encoder. The encoder has 12 transformer blocks, and each is composed of an attention layer followed by a feed-forward layer, and layer normalization \[3\] is applied before each layer. There are 8 attention heads, each with 64 dimensions in each attention layer. The feed-forward layer uses MLP with two layers to project the 1024-dimensional vector into 2048 dimensions, where GELU activation is applied, and then another layer projects it back to 1024 dimensions. Each output vector of the encoder corresponds to an input move feature, and we average these vectors to get a single vector of dimension 1024, which is finally projected with MLP into our final embedding space. This 512-dimensional embedding vector encodes the decisions the player made in the game. At inference time, the player vector is obtained by simply averaging all the game vectors (i.e. the centroid) that belong to the player.

**Loss Function.** In order to capture individual playing style, we adapt Generalized End-to-End Loss for Speaker Verification (GE2E) \[49\] to our setting, which allows us to train the game vectors to maximize both intra-player compactness and inter-player discrepancy in the resulting embedding space. GE2E builds a similarity matrix on a batch of N × M games, where N is number of players and M is number of games per player. For each game from each player, we extract the current 512-dimensional embedding vector \(y_{ji}\) (i-th game from the j-th player) output by the residual blocks and transformer network. We also find the centroid vector \(c_j\) of a specific target player by averaging all of their game vectors in the batch. However, we remove \(y_{ji}\) when computing the centroid of the true player as stated in \[49\] to help avoid trivial solutions.

We then construct a similarity matrix \(S_{ji,k}\) containing the cosine similarities between each game embedding vector and each player centroid (for all players in the batch), which we further scale by two learned parameters \(w\) and \(b\). For the target player we calculate the centroid \(c_j^{(-i)} = \frac{1}{M-1} \sum_{m \neq i} y_{jm}\), and otherwise simply use the average of game vectors for all other players \((k \neq j)\).

\[
S_{ji,k} = \begin{cases} 
  w \cdot \cos(y_{ji}, c_{j}^{(-i)}) + b & \text{if } k = j; \\
  w \cdot \cos(y_{ji}, c_k) + b & \text{otherwise.}
\end{cases}
\]  

(1)

The final loss function is designed to maximize the similarity for the correct entries while penalizing all other entries in the similarity matrix. As a result, the loss function \(L(y_{ji}) =\)
Figure 2: A schematic depiction of how players are organized in a behavioral stylometry task; how each player’s games are organized into training, query, and reference sets; and how each game’s moves are in a sequential order.

\[-S_{ji,j} + \log \sum_{k=1}^{N} \exp(S_{ji,k})\] pushes each embedding vector closer towards the player’s centroid and further from all other players’ centroids simultaneously.

**Training procedure.** We randomly sample $N \times M$ games as a batch, where $N = 40$ is number of players and $M = 20$ is number of games per player. In order to speed up training and perform batch computing, we randomly sample a 32-move sequence from each game and pad 0s to games with length less than 32 moves. The entire model was trained with 4 Tesla K80 GPUs and SGD optimizer with an initial learning rate of 0.01 and momentum of 0.9. The learning rate was reduced by half every 40K steps. Initial values of $w$ and $b$ for the similarity matrix were chosen to be $(w, b) = (10, -5)$, and we used a smaller gradient scale of 0.01 to match the original GE2E model.

4 Evaluation

We now train a model using our methodology and evaluate it against competitive baselines on several behavioral stylometry tasks.

4.1 Data

**Lichess.** We use chess games downloaded from Lichess, a popular open-source chess platform. Their game database contains over two billion games played by players ranging from rank beginners to the current world champion, Magnus Carlsen. From this database, we collect rated Blitz games (3–7 minutes per player per game) played on Lichess between January 2013 and December 2020 by players who match all of the following criteria: active as of December 2020, mean rating between 1000–2000, low variance in Blitz rating, and at least 1,000 games played during the time interval. We also restrict our attention to games with at least 10 moves (20 ply). To investigate whether behavioral stylometry performance varies with number of games played, we bucket the players according to how many games they have played: 1K–5K, 5K–10K, 10K–20K, 20K–30K, 30K–40K, and 40K+ (buckets are hereafter referred to by the minimum of their respective range, e.g. “20K” for the 20K–30K bucket). The union of these sets constitutes our universe of players $U$.

Players are further divided into seen and unseen sets, using the same players as our previous work for those with more than 10K games played \[31\]. Players are seen if their games are used in model training and unseen if not. Finally, each player’s games are randomly split into training games (80% of their games), reference games (10%), and query games (10%). Training games are only used for model training (and thus only seen players’ training games are used), reference games are what is known about each candidate player, and query games are what is given about an unknown target player we are attempting to identify. In all of our evaluations, we only use 100 games per player for both the reference and query sets, although in principle they could vary, and for each player their
reference and query sets are non-overlapping: \( x_r \cap x_q = \emptyset \). The numbers of players and games in each player set is shown in Table 1.

**High-ranking players.** We also assembled a dataset of high-ranking online players from both Lichess and Chess.com, the two most popular online chess platforms. To do so, we took the lists of the top 1,500 players on both sites’ leaderboards, downloaded all of their games, and filtered out games with fewer than 10 moves and players with fewer than 950 games. We split the remaining high-ranking players into seen (80%) and unseen (20%) sets, and their counts are shown in Table 1 (again, we use only 100 reference and 100 query games per player in our evaluations). Note that these games include all time controls and tend towards quicker games than the general population. These sets include many of the best players in the world, including the world champion.

### 4.2 Experimental Setup

**Task formulation.** Our central evaluation task is performing behavioral stylometry in chess, in which we are given a set of query games played by a particular player, and the task is to identify them from amongst a pool of candidate players \( C \), each of which is associated with a set of reference games.

Formally, we set this up as a few-shot classification task. Within our universe of players \( U \), we have an evaluation pool \( E \subseteq U \) of players we are trying to identify, and a candidate pool \( C \subseteq U \) of players which they are guaranteed to belong to \( \big(E \subseteq C\big) \). Every target player in \( E \) is associated with a small set of query games \( x_q \), and every candidate player in \( C \) is associated with a small set of reference games \( x_r \) (the conceptual organization of players, games, and moves is shown in Figure 2). In general the candidate pool \( C \) can be much larger than the evaluation pool \( E \) — for example, one possible task would be to try to identify some set of \( |E| = 10 \) unknown players, but we are only promised that they are members of some much larger pool of \( |C| = 10,000 \) players. Finally, we also specify a task parameter \( k \) where all games, both query and reference games, only contain moves starting at move \( k \). As we will see, the opening moves in chess can be highly discriminative because many players repeatedly play the same openings. For \( k = 0 \) we are given entire games, but for e.g. \( k = 15 \), we are only given moves 15 and beyond for each game. The behavioral stylometry task is thus fully specified by \( C, E, |x_r|, |x_q| \), and \( k \): for each player in \( E \) we are given \( |x_q| \) query games and for each player in \( C \) we are given \( |x_r| \) reference games, all starting at move \( k \). For each set of query games \( x_q \) we seek to identify which player in \( C \) played them.

**Openings baseline.** Chess games always start from the same initial position, and the opening moves are heavily studied and well-understood. Opening choice shapes the resulting game, and players often study a small set of openings in depth and stick to them. As a result, the opening moves are likely to contain a lot of information about who is playing the game, albeit for the somewhat unsatisfying reason that they are memorized patterns of play rather than new decisions being made. Indeed, when strong anonymous accounts play on online platforms, informed observers often rely on their opening moves to guess who they are. As a strong baseline for comparison, we create an embedding model that uses the opening moves. The model creates game vectors by sampling 5 sequential moves by the target player, converts each into a one-hot 4096-dimensional encoding vector (moves that are never used will be 0 for all vectors and do not affect our distance measurements), and adds them together to yield a 5-hot vector encoding of the 5 opening moves. Doing this process for each of the \( x_q \) games in the target player’s set of unlabeled games yields \( x_q \) 5-hot vectors in a 4096-dimensional space, which we then convert into a player vector by simply taking their centroid. These player vectors can then be used in the same manner as those in our main model to perform behavioral stylometry from the opening moves.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>1K</th>
<th>5K</th>
<th>10K</th>
<th>20K</th>
<th>30K</th>
<th>40K+</th>
<th>Ranked</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Seen Players</strong></td>
<td>8,617</td>
<td>5,298</td>
<td>1,866</td>
<td>295</td>
<td>86</td>
<td>19</td>
<td>1.813</td>
</tr>
<tr>
<td><strong>Seen Players’ Games</strong></td>
<td>14M</td>
<td>24M</td>
<td>17M</td>
<td>4.9M</td>
<td>2.0M</td>
<td>0.7M</td>
<td>9.5M</td>
</tr>
<tr>
<td><strong>Unseen Players</strong></td>
<td>23,101</td>
<td>1,324</td>
<td>466</td>
<td>72</td>
<td>20</td>
<td>20</td>
<td>451</td>
</tr>
<tr>
<td><strong>Unseen Players’ Games</strong></td>
<td>4.6M</td>
<td>264.8K</td>
<td>93.2K</td>
<td>14.4K</td>
<td>4K</td>
<td>4K</td>
<td>90.2K</td>
</tr>
</tbody>
</table>

Table 1: The numbers of players and games used in training (seen) and evaluation (unseen).
Our model. We apply our methodology to the behavioral stylometry task. To construct our main model, we trained on all training games by all players in the seen universe—63.7M games by 16,181 players in total. Our model learns player vectors in an embedding space that captures chess style into which new, unseen players can be encoded. To perform few-shot stylometry on an unseen player $p$ for some candidate pool $C$, we first infer player representations from the reference set of each player in $C$, then infer $p$'s representation by running the model on $p$’s query set. We then compute the cosine distance between $p$’s query representation and all candidate players’ reference representations and return the closest candidate player as $p$’s identity. In all of our evaluations, the reference sets are drawn from players’ 10% validation sets and the query sets are drawn from players’ 10% testing sets. We report precision at 1 (P@1) unless otherwise noted. We also trained two additional models, one on high-ranking players and one on the set of players used in prior work [31], which are discussed in the Supplement.

4.3 Behavioral Stylometry

Stylometry performance. Our main evaluation is the behavioral stylometry task where $C = 10K \cup 20K \cup 30K \cup 40K+$ (all 2,844 Lichess players with at least 10,000 games played, both seen and unseen, are possible identities), $E = 10K+$ (the 578 unseen Lichess players with 10K–40K+ games are the target players to be identified), $|x_q| = |x_s| = 100$ (reference and query sets are 100 games each), and $k = 15$ (we only get games starting from move 15, so that memorized openings do not factor into our evaluation of capturing human decision-making style). On this task, our model achieves a Precision@1 of 0.86, exactly identifying a player from 100 of their games out of a possible 2,844 players 86% of the time (see Table 2 (top)). Furthermore, this is without access to the opening, the most regimented, repeated, and identifiable part of chess. In comparison, the baseline with access to the first 5 given moves of each game (i.e. moves 16–20) only achieves a Precision@1 score of 0.234—far better than random guessing, indicating that exact moves indeed capture a significant amount of information, but far worse than our method. We also tested an alternative baseline that uses all moves of the game, which performs slightly better than the opening baseline but still much worse than our method (27.7% P@1, see Supplement for details). In the Supplement, we also show how performance varies for different choices of evaluation sets. Both our model and the baseline are robust to the choice of minimum total number of games per player, indicating that strong performance on this task is not contingent on players being very active on the platform.

A powerful property of our method is that it adapts to players it has never seen before. We compare performance on unseen players versus seen players (i.e. those whose training games were used during model training). The difference in performance between players the model explicitly trained on and players it has never seen before is surprisingly small. For example, the model achieves P@1 = 0.86 on unseen players and P@1 = 0.853 on seen players. This suggests that the model has learned a comprehensive space of decision-making style.

In Table 2 (bottom), we also show the version of the task where the opening moves are given to the models ($k = 0$). Our model performance increases to near-perfect accuracy, but the openings baseline also increases substantially. This indicates that using the opening moves to perform stylometry—identical moves that players repeat out of memorization—is less interesting than using moves later on during games, when players are making decisions in instances they (and our models) haven’t seen before. Although our model is still significantly better than the openings baseline even in this formulation of the task, we use $k = 15$ for the rest of our analyses.

To gauge whether our performance relies on discriminating between player skill levels, we show a variation of behavioral stylometry limited to subsets of players with similar ratings. Restricting our attention to players with ratings within a narrow range around 1900 yields P@1 = 92.6% (See Supplement Table 7 for full details).

Comparison with personalized move-matching models. The previous state of the art performance for chess player stylometry was achieved by training individual move-matching models for every player in $C$ [31]. This requires thousands of labeled games per player, and is only computationally feasible for small $|C|$, hence we are unable to include this method in our main analyses due to its lack of scalability. However, we directly compare our method on the small, 400-player candidate pool tested in [31]. For this comparison, the evaluation and candidate pools are both 400 players,
Table 2: Behavioral stylometry performance on Lichess players with all games starting at move $k=15$ (top) or $k=0$ (bottom). The candidate pool is the entire set of Lichess players with over 10K games, and the evaluation pool consists of all players that were seen, unseen, or both. Each players’ reference and query sets consisted of 100 games each.

| Evaluation Pool ($E$) | $|E|$ | $|C|$ | $k$ | Random Baseline | Our Model |
|-----------------------|------|------|-----|----------------|-----------|
| 10k+ Lichess (unseen) | 578  | 2844 | 15  | 0.0004         | 0.244     | 0.860     |
| 10k+ Lichess (seen)  | 2266 | 2844 | 15  | 0.0004         | 0.274     | 0.853     |
| 10k+ Lichess (all)   | 2844 | 2844 | 15  | 0.0004         | 0.268     | 0.854     |

| 10k+ Lichess (unseen) | 578  | 2844 | 0   | 0.0004         | 0.929     | 0.979     |
| 10k+ Lichess (seen)  | 2266 | 2844 | 0   | 0.0004         | 0.930     | 0.982     |
| 10k+ Lichess (all)   | 2844 | 2844 | 0   | 0.0004         | 0.929     | 0.982     |

Table 3: (Top) Comparison with personalized move-prediction models in [31] with $k=15$. (Bottom) Performance on high-ranked players, either with high-ranked players only in the candidate pool or high-ranked players and Lichess amateurs together, with $k=15$ and $|x_q| = |x_r| = 100$.

| Task | $|E|$ | $|C|$ | Baseline | Personalized | Our Model |
|------|------|------|----------|--------------|-----------|
| McIlroy-Young et al. | 400  | 400  | 0.478    | 0.552        | 0.953     |
| High-ranked players | 408  | 2157 | 0.039    | —            | 0.308     |
| " ∪ 10K–40K+ in $C$ | 408  | 5001 | 0.027    | —            | 0.301     |

$k = 15$, and $|x_q| = |x_r| = 100$. The results are shown in Table 3 (top), where personalized models achieve Precision@1 of 0.552. However, our model achieves Precision@1 of 0.953, far outstripping the previous method’s performance.

High-ranking players. To examine how our model performs on out-of-distribution examples, we evaluated on our set of high-ranking players (Section 4.1). These players are not only unseen by the model, but they are of much higher skill than any of the players it was trained on, and half of them are drawn from a separate online platform. We begin with the task of identifying the 408 unseen high-ranked players out of the complete set of both seen and unseen high-ranked players (Table 3 (bottom)). This task is significantly more difficult, as both the baseline and our model have lower P@1 scores. However, our model still outperforms the baseline by almost a factor of 10. Interestingly, when the set of Lichess amateurs is added to the candidate pool $C$, the model performance hardly changes at all, indicating that high-ranked players occupy a different part of the embedding space.

Reference and query set sizes. Here we show how model performance varies with $|x_r|$ and $|x_q|$. As Figure 3a shows, model performance rises sharply for early values and starts to saturate around $|x_r| = |x_q| = 50$ games. Even when given very small numbers of labeled examples, i.e. $|x_r| = 10$, our model performs well. In comparison, personalized move-prediction models require thousands of games to begin seeing improvements over unpersonalized models [31].

Expanded dataset. We examined performance on our complete dataset (1K to 40K+), comprising 41,184 players for both the candidate and evaluation pools. Even in this extensive regime, we achieve 54.0% P@1 while the baseline model achieves 8.49% P@1 (see Supplement for details).

4.4 Qualitative Analysis

Attention. Our use of attention allows us to examine which sections of the game are most useful for identifying individual style. Figure 3b shows the average per-move attention for the main model when applied to 100 games for each player in the 40K+ set. Four different starting moves were examined ($k = 0, 5, 10, 15$) to show how the models handle being given incomplete games. The earlier parts of the game are most informative, with a monotonic decay in attention as games progress.
5 Ethical Considerations

As this work pursues the identification of individuals from behavioral traces, it raises important ethical questions. In principle, although our main goal is to specialize behavioral models in the service of more human-compatible AI, the methodology we present and others like it could be abused by bad actors. The most salient concerns revolve around user privacy, as sufficiently powerful stylometry methods could be used to deanonymize individuals. We now discuss these issues in detail, as well as comment on our choice of domain and possible generalization to other domains.

Privacy concerns. Our methods can be used to accurately identify people from limited behavioral traces. This raises significant privacy concerns, as well as concerns about power differentials and inequality, since those who are marginalized in society often benefit most from anonymized communication and behavior. The relatively small amount of data required to identify an individual with a high degree of accuracy, at least in the context of chess, would likely surprise those who wish to remain anonymous. Engaged players often have several accounts (either on the same online server or across multiple platforms), and our embedding methodology could be used to link them together. This could be problematic if a player intended to keep an anonymous account separate from a main account, potentially concealing ideas they are experimenting with or details about when they are active online. Similar concerns are also present in the contexts of speech recognition, the domain we adapted our methodology from, and author identification in text, a subfield with a long history.

The history of author identification also makes clear that people have an interest in developing countermeasures to being identified (and again, especially those with limited power). Our work contains insights that can offer help to those seeking such countermeasures in our chosen domain of chess. Consider again a player with multiple accounts. We find that if their accounts were to differ non-trivially in their opening repertoire, this would reduce our model’s accuracy in identifying them as the same individual, since we find that the opening phase of the game is the most revealing part. To illustrate, if we divide the games played as White by a single player into two sets, depending on whether they started with the opening move e4 or d4, the resulting set centroids are on average as distant from each other (cosine distance 0.27) as they are from other players’ centroids. Of course, any such solution is far from ideal since it requires people to put effort into avoiding identification. But the framework we develop may be able to continue providing further insight into the design of countermeasures to identification requiring limited effort or adaptation from individuals.
Our methods could possibly reveal other information about players through the embedding representations we learn. We do not examine if characteristics such as gender or nationality are correlated with the embedding representations, since the public dataset we use contains no such personally identifiable information, but other machine learning representations have been found to unwittingly incorporate private characteristics [12].

**Domain and generalization.** We chose chess as a first domain to study since it is relatively benign compared to other possible contexts such as human speech, driving behavior, and cellphone mobility data. Having one’s chess behavior connected with one’s identity, while not always harmless, is arguably less serious than in other domains. It is not a priori clear, in any domain, how uniquely identifying particular forms of behavioral data will be; but it is crucial to understand this since it is easy to build individual-level classifiers from public data of this form, and individuals may act with an expectation that they will remain anonymous, without knowing how revealing their data is. Moreover, there is a quantitative dimension to this—it is not simply all-or-nothing, but a question of how accurately these determinations can be made as a function of the data available. Only through a careful analysis of stylometry can we determine the level of risk in any given domain, and as such it is an important part of making people aware of privacy risks. Our work is following this principle in illuminating the power of stylometry from behavioral traces. By deliberately investigating and quantifying these risks in the domain of chess, we can begin to develop this understanding before something similar is attempted in higher-stakes domains.

The above is particularly important since our methodology can be applied to other potentially more concerning domains. The framework of mapping sequences of decisions into a vector space, and then using an appropriate stylometry loss function, is not unique to chess. We hope that the research community continues to develop both the methodological and ethical frameworks necessary to ensure behavioral stylometry techniques enhance human-AI compatibility and collaboration.

### 6 Conclusion

Motivated by the goal of algorithmically characterizing human behavior, we propose a method for performing behavioral stylometry—identifying people down to the level of specific individuals from their patterns of decisions—and demonstrate the effectiveness of these methods in the domain of chess. We adapt techniques developed for speaker verification tasks to this decision-making setting, and construct a neural network architecture that can reliably identify individual-level decision-making style. Our architecture uses residual blocks to extract position- and move-level features, then combines these with a position encoding and inputs them to a modified Vision Transformer that learns how to accurately aggregate this information into a game-level representation. We use a GE2E loss framework to maximize both intra-player compactness and inter-player discrepancy in the resulting embedding space. The final model is then capable of performing highly accurate few-shot classification of unseen players, out-performing both the previous state of the art and a competitive opening baseline. The resulting embedding space generalizes in several ways—it can identify high-ranked players even though it was only trained on amateurs, it identifies unseen players equally well as players it was trained on, and maintains its strong performance when expanded to tens of thousands of players.

Our methodology has several limitations that raise natural questions for future work. First, while our methodology far out-performs the baseline on top players, the lower performance level suggests there is room to improve in capturing top-class, complex styles. Second, our methodology is geared towards identifying individuals from their behavior alone, which could lead to privacy concerns. While this critique pertains to any stylometry task, including hand-writing recognition and speaker verification, our work raises similar questions in behavioral domains. Future work could investigate whether the benefits of methodologies capable of stylometry could be achieved without the privacy risks of uniquely identifying people. Finally, our methodology could be extended to other behavioral domains where fruitful human-AI collaboration is on the horizon.

**Acknowledgments.** Supported in part by Microsoft, NSERC, CFI, a Simons Investigator Award, a Vannevar Bush Faculty Fellowship, MURI grant W911NF-19-0217, AFOSR grant FA9550-19-1-0183, ARO grant W911NF19-1-0057, a Simons Collaboration grant, and a grant from the MacArthur Foundation.
Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default [TODO] to [Yes], [No], or [N/A]. You are strongly encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss them in the Conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discuss the privacy risks in the Conclusion.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Our supplement contains the code, the model files are too large for the supplement so will be released when the paper goes public
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We have data and methods sections discussing this along with more information in the supplement
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] We ran over thousands of players to get error estimates, but did not train or test multiple times
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Our methods and supplement discuss this

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Our code release links to the code we build on
   (b) Did you mention the license of the assets? [Yes] The code we are using is GPL or MIT licensed and is mentioned in the code release
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We plan to release the code and models after publication
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes] We discuss this in the Data section, all data used is public and not sensitive
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We discuss this in the Data section, there is no PII in our data
5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
References


Figure 4: Validation loss when using a 10-move window versus a 32-move window during model training.

7 Supplement

This supplement includes additional details about the models discussed in the main text and their stylometry performance, comparisons against other models we trained (e.g., on smaller subsets of the data), and a summary of the artifacts we are releasing.

7.1 Training Data Parameters

Move window. As mentioned in our methodology (Section 3), we randomly sample a 32-move sequence from each game to speed up the training process. In a given training iteration, this sequence is comprised of moves from one game of one player, but different parts of the game may be sampled during different training iterations. In our earlier experiments, we used a 10-move window since this required no zero-padding to be added to the games (all games have at least 10 moves). However, we later observed a benefit from using longer sequences of the input, even though some games have fewer than 32 moves and require zero-padding. This comparison is shown in Figure 4.

Starting move. The models discussed in the main text are trained on any 32-move window from each game. During inference time, we distinguished between the easier task of using the opening moves to perform stylometry (starting at move \( k = 0 \)) versus the harder task of using moves later on during the games (e.g., starting at move \( k = 15 \)), when players are making decisions in instances they (and our models) have not seen before. However, this leaves the question of how opening moves impact training unanswered. To investigate this, we additionally train a model using 32-move windows that only start at move 15 or later. The stylometry accuracy of this model is 0.390 on a candidate pool \( |C| = 232 \) and evaluation pool \( |E| = 232 \). When evaluated on the same exact task, our main model (which does include the opening moves in its training data) achieves a significantly higher accuracy of 0.9927. This reveals that information about the opening moves is quite important during model training: our main model likely draws on this information when evaluated on later moves, whereas the model trained only on moves 15 or later does not have access to such information directly.

7.2 Model Architecture Variants

LSTM vs Transformer. In our initial experiments, we used the same LSTM network that was used in the original GE2E work [49]. We subsequently observed a small performance gain when using a Bi-directional LSTM. This is likely due to the transformer being better able to focus on the early game as show in figure 3b. Figure 5 shows that a Transformer model consistently outperforms a
Bidirectional LSTM throughout the training process. Table 6 also shows the LSTM’s accuracy when fully trained.

**ArcFace vs GE2E.** As we discussed in related work (Section 2), there are two families of loss functions that are most relevant to our behavioral stylometry task: ArcFace [14] and GE2E [49]. During our investigation, we experimented with one loss function from each of these families. However, holding everything else constant, we observed that ArcFace consistently failed to converge in the settings we evaluated.

**Embedding size.** The embedding size is also an important hyperparameter that we explored. We experimented with both a 128 and 512-dimensional embedding, as shown in Figure 6. Even though the two models perform similarly during the first 10K training steps, the validation loss for the 512-dimensional model continues to drop beyond that point, while the 128-dimensional model roughly stays the same (Figure 6). All of our models thus use an embedding size of 512.

### 7.3 Number of Games Played

We can partition the players by the number of games they have played to match the partitioning used by [31]. Table 4 shows the accuracy of our main model as we vary the evaluation pool of the number

---

**Figure 5:** Validation loss of a Bi-directional LSTM compared to the Transformer model.

**Figure 6:** Validation loss when using a 512 vs. 128-dimension embedding.
Table 4: Model accuracy on different sets of Lichess players based on number of games played. Each model is given games starting from the 15th move to create game vectors. The candidate pool is the entire set of Lichess players with over 10K games in our dataset; 100 games from each player’s test set is used for both the reference and query sets of each player.

<table>
<thead>
<tr>
<th>Evaluation Pool (E)</th>
<th>E</th>
<th>C</th>
<th>Baseline</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>10k – 40k+</td>
<td>578</td>
<td>2844</td>
<td>0.0004</td>
<td>0.244</td>
</tr>
<tr>
<td>20k – 40k+</td>
<td>112</td>
<td>2844</td>
<td>0.0004</td>
<td>0.286</td>
</tr>
<tr>
<td>30k – 40k+</td>
<td>40</td>
<td>2844</td>
<td>0.0004</td>
<td>0.325</td>
</tr>
<tr>
<td>40k+</td>
<td>20</td>
<td>2844</td>
<td>0.0004</td>
<td>0.300</td>
</tr>
</tbody>
</table>

Table 5: Comparison of baseline and our model on a large dataset of players. Candidate and evaluation pool are the same set and include players from the seen and unseen sets, all games are from the test set. 100 games were used for query and reference sets.

<table>
<thead>
<tr>
<th>Evaluation Pool (E)</th>
<th>k</th>
<th>E</th>
<th>C</th>
<th>Baseline</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k – 40k+</td>
<td>15</td>
<td>9465</td>
<td>9465</td>
<td>0.171</td>
<td>0.737</td>
</tr>
<tr>
<td>1k – 40k+</td>
<td>15</td>
<td>41184</td>
<td>41184</td>
<td>0.0849</td>
<td>0.540</td>
</tr>
</tbody>
</table>

7.4 Expanded Dataset Results

The most important results for the large data set are covered in Section 4.3, but we also ran other tests that are summarized in Table 5. The results for the baseline and our models on the task of identifying players from a large set can be broken down by the player’s category. This task is constructed the same way as in the main paper with \( k = 15 \). Looking at players with 5K or greater games, the baseline achieves 0.171 accuracy while our model achieves 0.737 accuracy. The evaluation and candidate pools are both the same (of size 9465) since we use all players. When we look at the largest possible set of Lichess players—those with at least 1K games—we see 0.0849 accuracy for the baseline and 0.540 for our model.

7.5 Alternative Models Comparison

Table 6 shows the accuracy of an alternative baseline model which uses all positions instead of a 5 move window and an LSTM model. The LSTM model uses an LSTM instead of a transformer for the variable input length, and was trained on the same data as the main model. The all moves model while slightly better has higher overhead so for the 41k players test (section 4.3) was infeasible.

7.6 Accuracy on Rating Subsets

Table 7 shows the accuracy on sets of players with different ratings. Note that the rating is the minimum and the buckets are of width 100, e.g., 1100 refers to players rated [1100, 1200).
<table>
<thead>
<tr>
<th>Evaluation Pool (E)</th>
<th></th>
<th></th>
<th></th>
<th>Random</th>
<th>All Moves</th>
<th>LSTM</th>
<th>Transformer (Final Model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10k – 40k+</td>
<td>578</td>
<td>2844</td>
<td>15</td>
<td>0.0004</td>
<td>0.2768</td>
<td>0.7803</td>
<td>0.8599</td>
</tr>
<tr>
<td>20k – 40k+</td>
<td>112</td>
<td>2844</td>
<td>15</td>
<td>0.0004</td>
<td>0.3750</td>
<td>0.8036</td>
<td>0.8482</td>
</tr>
<tr>
<td>30k – 40k+</td>
<td>40</td>
<td>2844</td>
<td>15</td>
<td>0.0004</td>
<td>0.4500</td>
<td>0.8750</td>
<td>0.8750</td>
</tr>
<tr>
<td>40k+</td>
<td>20</td>
<td>2844</td>
<td>15</td>
<td>0.0004</td>
<td>0.4500</td>
<td>0.9000</td>
<td>0.9000</td>
</tr>
</tbody>
</table>

Table 6: Comparison of alternative models not discussed in the main text

<table>
<thead>
<tr>
<th>k</th>
<th></th>
<th></th>
<th>Rating</th>
<th>Model Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>19</td>
<td>19</td>
<td>1100</td>
<td>1.000</td>
</tr>
<tr>
<td>15</td>
<td>36</td>
<td>36</td>
<td>1200</td>
<td>0.972</td>
</tr>
<tr>
<td>15</td>
<td>54</td>
<td>54</td>
<td>1300</td>
<td>0.926</td>
</tr>
<tr>
<td>15</td>
<td>106</td>
<td>106</td>
<td>1400</td>
<td>0.981</td>
</tr>
<tr>
<td>15</td>
<td>162</td>
<td>162</td>
<td>1500</td>
<td>0.988</td>
</tr>
<tr>
<td>15</td>
<td>297</td>
<td>297</td>
<td>1600</td>
<td>0.976</td>
</tr>
<tr>
<td>15</td>
<td>340</td>
<td>340</td>
<td>1700</td>
<td>0.968</td>
</tr>
<tr>
<td>15</td>
<td>487</td>
<td>487</td>
<td>1800</td>
<td>0.926</td>
</tr>
<tr>
<td>15</td>
<td>522</td>
<td>522</td>
<td>1900</td>
<td>0.925</td>
</tr>
</tbody>
</table>

Table 7: Accuracy on sets of players with different ratings. 1100 means players in [1100, 1200), 1200 players in [1200, 1300), ....

7.7 Training on Other Players

In the main text, we train our model on a large set of amateurs on Lichess (Section 4.2). Here, we report on two additional experiments in which we train on a separate set of interesting players. In the first, we train our model only on the 400 players used in [31]. This is an experiment with training on orders of magnitude fewer players, and also facilitates a direct comparison with previous work (and our main model's performance on the 400 players from previous work). In the second, we train on strong, high-ranking players from both Lichess and Chess.com, the two largest online chess platforms. This is also a smaller training set, and helps us understand whether our training methodology generalizes to high-level players. Both models performed worse on their respective tasks than our main model due to the small training set size, and are included here for comparison and completeness.
| Task | Evaluation Pool | $|E|$ | $|C|$ | Baseline Model | Our Model | High Model | 400 Model | Model |
|------|-----------------|------|------|----------------|------------|------------|------------|--------|
| 10k – 40k+ | all | 15 | 2844 | 2844 | 0.268 | 0.854 | 0.202 | 0.515 |
| 10k – 40k+ | seen | 15 | 2266 | 2844 | 0.274 | 0.853 | 0.204 | 0.520 |
| 10k – 40k+ | unseen | 15 | 578 | 2844 | 0.244 | 0.860 | 0.195 | 0.497 |
| 20k – 40k+ | seen | 15 | 400 | 2844 | 0.355 | 0.870 | 0.216 | 0.573 |
| 20k – 40k+ | unseen | 15 | 112 | 2844 | 0.286 | 0.848 | 0.273 | 0.607 |
| 30k – 40k+ | seen | 15 | 105 | 2844 | 0.457 | 0.886 | 0.258 | 0.590 |
| 30k – 40k+ | unseen | 15 | 40 | 2844 | 0.325 | 0.875 | 0.282 | 0.725 |
| 40k+ | seen | 15 | 19 | 2844 | 0.368 | 0.895 | 0.294 | 0.737 |
| 40k+ | unseen | 15 | 20 | 2844 | 0.300 | 0.900 | 0.263 | 0.700 |
| High-ranked $\cup$ 10K–40K+ | all | 15 | 4911 | 4911 | 0.158 | 0.595 | 0.136 | 0.325 |
| High-ranked $\cup$ 10K–40K+ | seen | 15 | 3925 | 4911 | 0.160 | 0.590 | 0.137 | 0.327 |
| High-ranked $\cup$ 10K–40K+ | unseen | 15 | 986 | 4911 | 0.151 | 0.614 | 0.130 | 0.318 |
| High-ranked, high only $C$ | seen | 15 | 1659 | 2157 | 0.026 | 0.273 | 0.107 | 0.108 |
| High-ranked, high only $C$ | unseen | 15 | 408 | 2157 | 0.039 | 0.308 | 0.106 | 0.118 |
| High-ranked | seen | 15 | 1659 | 4911 | 0.019 | 0.265 | 0.080 | 0.097 |
| High-ranked | unseen | 15 | 408 | 4911 | 0.027 | 0.301 | 0.081 | 0.108 |
| McIlroy-Young et al. | seen | 15 | 400 | 400 | 0.477 | 0.953 | 0.480 | 0.825 |

Table 8: Behavioral stylometry performance on Lichess and high-ranked players with all games starting at move $k = 15$. The candidate pool is the entire set of Lichess players with over 10K games on the top, and various other combinations as noted on the bottom. The evaluation pool is as described with players in either seen set, unseen set, or both (“all”). 100 games were used for both reference and query sets of each player.

| Task | Evaluation Pool | $|E|$ | $|C|$ | Baseline Model | Our Model | High Model | 400 Model | Model |
|------|-----------------|------|------|----------------|------------|------------|------------|--------|
| 10k – 40k+ | all | 0 | 2844 | 2844 | 0.930 | 0.982 | 0.455 | 0.810 |
| 10k – 40k+ | seen | 0 | 2266 | 2844 | 0.930 | 0.982 | 0.464 | 0.815 |
| 10k – 40k+ | unseen | 0 | 578 | 2844 | 0.929 | 0.979 | 0.419 | 0.791 |
| 20k – 40k+ | seen | 0 | 400 | 2844 | 0.948 | 0.988 | 0.497 | 0.890 |
| 20k – 40k+ | unseen | 0 | 112 | 2844 | 0.893 | 1.000 | 0.446 | 0.812 |
| 30k – 40k+ | seen | 0 | 105 | 2844 | 0.933 | 1.000 | 0.533 | 0.924 |
| 30k – 40k+ | unseen | 0 | 40 | 2844 | 0.850 | 1.000 | 0.475 | 0.850 |
| 40k+ | seen | 0 | 19 | 2844 | 0.947 | 1.000 | 0.684 | 1.000 |
| 40k+ | unseen | 0 | 20 | 2844 | 0.900 | 1.000 | 0.500 | 0.900 |
| High-ranked $\cup$ 10K–40K+ | all | 0 | 4911 | 4911 | 0.789 | 0.799 | 0.345 | 0.557 |
| High-ranked $\cup$ 10K–40K+ | seen | 0 | 3925 | 4911 | 0.785 | 0.800 | 0.347 | 0.558 |
| High-ranked $\cup$ 10K–40K+ | unseen | 0 | 986 | 4911 | 0.806 | 0.795 | 0.338 | 0.552 |
| High-ranked, high only $C$ | seen | 0 | 1659 | 2157 | 0.614 | 0.569 | 0.283 | 0.259 |
| High-ranked, high only $C$ | unseen | 0 | 408 | 2157 | 0.651 | 0.554 | 0.295 | 0.278 |
| High-ranked | seen | 0 | 1659 | 4911 | 0.605 | 0.565 | 0.243 | 0.244 |
| High-ranked | unseen | 0 | 408 | 4911 | 0.649 | 0.549 | 0.263 | 0.266 |
| McIlroy-Young et al. | seen | 0 | 400 | 400 | 0.990 | 0.995 | 0.740 | 0.973 |

Table 9: Behavioral stylometry performance on Lichess and high ranked players with all games starting at move $k = 0$. The candidate pool is the entire set of Lichess players with over 10K games on the top, and various other combinations as noted on the bottom. The evaluation pool is as described with players in either unseen set, seen set, or both (“all”). 100 games were used for both reference and query sets of each player.
Figure 7: Two-dimension t-SNE projection of high-ranked player vectors generated from games starting at move $k = 15$.

7.8 Additional Embedding visualizations

Figure 7 shows the same embedding layout as figure 3c but colored by the moves made in response to white’s first move.

7.8.1 400 Player Model

We train a 400-player model on the same 400 players that [31] uses for its stylometry analysis. In contrast with the main model, which uses 16,000 players for training, this one is drastically reduced to 400 players.

The results for this model are shown in the “400” column of Tables 8 and 9. The main comparison task for this model is the McIlroy-Young et al. dataset which is the same set of players (but not games) it was trained on. It outperforms the baseline on the $k = 15$ tasks (Table 8), but not on the $k = 0$ tasks (Table 9). Additionally, it does not generalize well and performs badly on the high-ranked players task. These results suggest that more than 400 players are needed to learn the space of chess-playing style.

7.8.2 High-Ranking Player Model

We also trained a high-ranking player model with the 1,813 high-ranked training players. Having a model that performs well on high-skill players is of interest to the chess community. While our final model presented in the main text performs well above the baseline, we were interested in evaluating whether directly training on high-skill players would perform even better.

The results are shown in the “High” column of Tables 8 and 9. Similarly to the 400 model, it under-performs our final model by a large margin. Notably, the model explicitly trained on high-ranked players performs worse than our main model on the high-ranked players task, suggesting that high-ranked players are a harder set to learn from. This is consistent with our observation that they are harder to distinguish.
7.9 Artifacts for Release

7.9.1 Code

Our public code release can be found at [github.com/CSSLab/behavioral-stylometry](https://github.com/CSSLab/behavioral-stylometry). As there are ethical concerns regarding releasing the model and full training code it may not be publicly accessible. You can contact our corresponding author at: reidmcy@cs.toronto.edu to request access.

We incorporate three pieces of open source code in ways that are not trivially separated from our own code. One is GPL licensed and handles the board representations, the second and third are MIT licensed. One provides help with the transformer implementation and the other is for the training pipeline. All are noted in this code release.

7.9.2 Data

The data used for training are all from the Lichess database which is free and released under a Creative Commons CC0 license. The code for processing and analysis is included with the code release.

7.9.3 Interactive embedding viewer

We used Embedding Projector to view the embedding vectors. Instructions on how to view it can be found with our code release at [github.com/CSSLab/behavioral-stylometry](https://github.com/CSSLab/behavioral-stylometry)