Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Finding winning team compositions by jointly modeling player proficiency and team congruency in multiplayer online games

JOOYEON KIM

jooyeon.kim@unist.ac.kr

Ulsan National Institute of Science and Technology

Article

Keywords: Team composition, computational game analytics, longitudinal player modeling, multiplayer online battle arena, transformer, self-attention

Posted Date: September 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10625437/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Finding winning team compositions by jointly modeling player proficiency and team congruency in multiplayer online games

Jooyeon Kim[1*]

[1*]Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology (UNIST), South Korea.

Corresponding author(s). E-mail(s): jooyeon.kim@unist.ac.kr;

Abstract

Temporary teams often face a fundamental choice: should members take on roles at which they excel, or roles the team needs most? These aims are sometimes manifested as a dilemma; favoring one can compromise the other, making effective negotiation between individual proficiency and collective congruency crucial to team success. Here, we present a deep-learning framework that learns from large-scale records of team behavior to computationally identify winning compositions by accounting for both considerations. As a testbed, we compile data from approximately 26 million unique participants across 320 million five-on-five matches in a multiplayer online battle arena (MOBA) game named League of Legends, in which short-lived teams are continually assembled and evaluated through win-loss outcomes. We formulate a transformer architecture that jointly learns each participant's proficiency from their longitudinal choices and outcomes, while evaluating the current team composition through interactions among assigned roles within the team and across the opposing team. The proposed method substantially outperforms previously suggested approaches in predicting the winning composition before the actual play begins. Scaling analyses, ablation studies, and qualitative analyses further validate our construction. We release the dataset and code for replicability and future research.

Keywords: Team composition, computational game analytics, longitudinal player modeling, multiplayer online battle arena, transformer, self-attention

1 Introduction

Temporary teams—from weekend hackathons to pickup basketball teams—are increasingly asked to perform before their members synchronize and accumulate a shared past. In such teams, composition becomes an early condition of collective performance: the same people may possess different collective capabilities depending on how work is divided among them [1, 2]. This problem has grown more consequential as teamwork has become fluid, distributed, and short-lived [3, 4]. Differentiated roles can reduce ambiguity and provide a provisional basis for coordination among relative strangers [5, 6], but roles alone do not determine who should occupy them or whether the resulting arrangement will function as a coherent whole.

Team composition cannot therefore be characterized adequately by summing the abilities of individual members. A capable person may be poorly matched to a present responsibility, while a set of individually sensible assignments may collectively duplicate some capabilities and leave others uncovered. A hackathon team, for example, may have several strong coders but no one prepared to shape the presentation; a pickup basketball team may have several scorers but no one willing to defend or distribute the ball. A useful distinction is between *proficiency*—the fit between a member's prior experience and current assignment—and *congruency*—the complementarity of assignments across the team [7]. The two properties, however, need not move together. The importance of a member's capabilities can vary with the role they occupy [8], and the effects of prior collaborative experience can likewise differ across roles [9]. Deriving a winning composition is thus simultaneously a longitudinal problem, as relevant experience must be inferred from prior behavior. It is also a relational problem, because the value of any assignment depends partly on the assignments around it.

This conceptual tension also poses a challenge of measurement and representation. Organizational research has developed rich accounts of how member attributes combine at the team level [10], while computational studies have formalized team formation in terms of skill coverage and coordination costs [11]. Digital platforms now make transient team assembly observable at a scale rarely available in conventional organizational settings [12, 13]. These records have supported studies of self-assembly [14, 15], composition and accumulated experience in online games [16, 17], and learning-based representations of team membership and collaboration [18, 19]. Many commonly used summaries, however, compress the behavioral histories through which assignment-specific experience develops and prespecify which relations among members enter the analysis. The central methodological challenge is therefore to learn evidence relevant to member–assignment fit and collective configuration *jointly* from behavioral histories, before the newly formed team produces any behavior of its own.

We study the prospective assessment of temporary-team compositions: given records of how members previously acted under different assignments, what can be inferred at formation about a composition's prospects for collective performance? *League of Legends* provides a stringent empirical setting for this inquiry. Each match brings together two temporary teams of five participants, with every participant occupying one of five differentiated roles and selecting one of many playable characters. The eventual match outcome supplies a consistent observable indicator of relative team

performance, while recurring participants provide longitudinal evidence across characters, roles, teammates, opponents, and game versions. The game provider's internal matching process is designed to assemble contests among participants of broadly similar demonstrated performance, thereby limiting the broad ability differences that would otherwise make prospective assessment trivial.

Crucially, team compositions are evaluated only after roles and characters are assigned, but before the match actually begins. This is analogous to assessing two basketball lineups after position assignments are set but before tip-off. This timing shifts the focus entirely to whether players fit their roles and complement one another, effectively isolating composition quality from downstream gameplay factors. Nevertheless, such prospective assessments remain inherently challenging. Final outcomes are frequently shaped by subsequent events—such as injuries, momentary mistakes, or tactical shifts—that introduce in-game volatility, acting as noise that obscures the direct relationship between a team's initial makeup and its realized performance.

Our empirical record contains approximately 320 million five-on-five matches involving 26 million unique participants. This scale lets us investigate three related questions. First, can prior choices and outcomes reveal evidence relevant to how well a participant fits a present assignment? Second, how much information about collective performance resides in composition—the interplay among members—beyond what can be captured through mere aggregation of independently estimated individuals? Third, how do history depth and framework capacity affect the performance-relevant information recoverable at formation? All information is ordered chronologically, and only records preceding the focal match are admissible.

Methodologically, we develop a hierarchical framework for learned longitudinal and relational inference, implemented with transformers. For each participant, the framework interprets prior choices and outcomes in relation to the current character, role, and competitive context, allowing the relevance of a prior event to vary with the assignment being assessed. It then integrates the resulting participant representations across all ten positions to capture dependencies within each team and across the opposing teams. The proficiency–congruency distinction informs this structure: assignment-conditioned histories provide evidence relevant to individual fit, within-team interactions place that evidence in collective context, and across-team interactions incorporate opposition. Deep learning provides the capacity to learn these assignment-contingent and relational patterns from behavioral traces at scale.

Under a global chronological protocol, the learned representations distinguish eventual team performance more reliably than all evaluated skill-rating systems, manually constructed composition measures, recurrent sequence baselines, and previously proposed game-specific neural methods. Scaling experiments examine how this capability varies with history depth and model capacity, while ablation and composition-level analyses clarify the respective contributions of longitudinal evidence and joint relational modeling. Together, these results show that individual histories and joint composition contain information about eventual relative performance even within contests assembled on the basis of demonstrated performance.

Taken together, this study contributes a population-scale computational approach to the prospective assessment of temporary-team composition. It formulates the problem as joint longitudinal and relational learning and develops a scalable framework that interprets each member's history in light of the current assignment before modeling the formed compositions jointly. Applied to 320 million chronologically ordered matches involving 26 million participants, the approach characterizes how prospective assessment changes with data scale, historical context, and relational modeling under deliberately balanced team formation. More broadly, the study shows the ways in which large behavioral records and structured representation learning can be used to investigate how formation-time characteristics relate to subsequent collective performance. To support replicability and further research, we release the code and the processed, anonymized data.

2 Related work

2.1 Team assembly, composition, and analysis

Team-composition research treats collective performance as a property of both members and their configuration. The value of member attributes depends on how they are distributed and on the roles they occupy [2, 8, 10]. The same member may contribute differently across assignments, while individually sensible assignments may form an imbalanced whole. This issue is especially important in temporary teams, where roles scaffold coordination among strangers and specialization becomes more effective when paired with social and functional familiarity [5, 6, 9, 20]. Prior competition can also improve coordination when former opponents become teammates [21].

The proficiency–congruency distinction connects individual fit with collective configuration. Proficiency concerns whether a member's experience fits the present assignment, whereas congruency concerns whether the assignments complement one another across the team [7]. These concepts relate to person–job and person–group fit and the coordination of distributed expertise [22, 23]. Consistent with this account, role coordination is associated with performance and retention, and is facilitated by teammate familiarity [24]. Building on the distinction, we learn assignment-contingent evidence from longitudinal behavior and model the formed composition relationally.

Computational team-formation research generally selects members whose combined skills cover a task while limiting coordination costs [11]. Learning-based extensions recommend collaborators or represent teams through collaboration graphs [18, 19]. Our problem begins after membership and assignments have been fixed and instead assesses the resulting composition prospectively. Large-scale digital records make this possible by repeatedly observing the same individuals across assignments, teams, and outcomes [14, 15, 17]. Studies of ad hoc teams likewise show that individual and emergent team profiles can provide distinct information about subsequent behavior and performance [25].

2.2 Team composition and performance in multiplayer games

Skill-rating systems typically summarize each player's previous outcomes with a latent ability value [26]. Although useful for ranking and matchmaking, a single value compresses the contexts in which performance was observed and cannot explain why the same player may be more effective under one assignment than another. Multiplayer-game studies accordingly distinguish general ability from character strength and player–character-specific experience [27]. Behavioral profiling further identifies stable generalists, specialists, and mavericks whose performance and resilience differ as the game environment changes [28]. Context-aware representations and engineered measures such as character mastery likewise demonstrate that assignment-specific experience contains information beyond a global rating [29, 30]. These approaches, however, rely on factorized representations or pre-aggregated variables rather than learning directly from ordered behavioral histories.

Composition-level work considers how all participants' selections fit together. Large-scale studies associate team diversity and behavioral interdependence with effectiveness, while showing that these relationships vary across outcomes and levels of experience [16, 17]. Drafting methods model synergy among allies and opposition against the other team [31]. DraftRec is the closest prior neural formulation, encoding each player's recent history before modeling interactions among players and selected characters for personalized drafting [32]. Its history encoder produces a general player representation independently of the current assignment, whereas ours reinterprets prior behavior in relation to the present character–role assignment. For outcome prediction, DraftRec average-pools the relationally updated player representations within each team; ours instead uses a learned composition token that accumulates information through joint attention across all ten assignment-conditioned participants. The pair of fully specified compositions thereby becomes the direct object of inference.

Other esports methods focus on professional competition or use information generated during the focal match [33–35]. Such information is useful for skill estimation or live forecasting but is unavailable at team formation. Our protocol therefore derives every participant-level input from matches completed before the target match.

2.3 Longitudinal and relational representation learning

Target-conditioned attention allows the same behavioral history to be interpreted differently for different present targets [36, 37]. Rather than compressing past behavior into a fixed representation, it can emphasize the events most relevant to the assignment currently under consideration. Related models have extended attention-based learning to long, population-scale histories, but typically assess one individual trajectory at a time [38–40]. Our setting couples this longitudinal problem with a relational one: each participant's history must be interpreted for the current assignment, and the resulting representations must be evaluated jointly rather than aggregated independently. The framework therefore models dependencies within and across the two teams, integrating assignment-conditioned longitudinal inference with relational inference.

3 Methods

3.1 Problem formulation

3.1.1 Game background and team composition

League of Legends is a multiplayer online battle arena (MOBA) game in which two teams of five participants compete to destroy the opposing team's central base structure, called the Nexus [41]. The two sides are conventionally designated as the *blue team* and the *red team*; we retain these designations and define a blue-team victory as the positive outcome. Before play, the game provider's matchmaking process selects and divides ten participants so that their overall skill levels are broadly comparable.

Each team contains one participant in each of five roles—top, jungle, middle, attack-damage carry, and support—and we denote this role set by $\mathcal{K}$. The top and middle participants occupy the corresponding lanes, the jungle participant operates between lanes, and the attack-damage carry shares the bottom lane with the support. These differentiated responsibilities make a team more than an unordered collection of members and organize the two teams into corresponding role-labeled positions.

Within these roles, each participant selects one playable character, called a *champion*. Champions differ in their capacity to inflict or withstand damage, protect allies, restrict opponents, and move across the battlefield, giving them distinct strategic functions. The draft determines the ten selected champions and establishes the immediate strategic context of the match. This process yields two fully specified, role-labeled compositions before either team has produced any behavior in the actual gameplay.

Comparable overall skill, however, does not guarantee that assignment-specific proficiencies are well distributed within each team or that individually preferred champions form a coherent whole. Participants must therefore construct an effective composition themselves through their pregame champion choices within the assigned roles, thereby increasing their chance of winning before gameplay begins.

The prospective value of a composition consequently depends on both proficiency and congruency. A participant may be highly proficient with a particular champion or in a particular role but substantially less so with another; conversely, five individually suitable assignments need not form a coherent whole. Assessment must therefore consider how each participant's prior experience fits the present assignment, how assignments interact within each team, and how the two formed compositions match up against one another. The current patch further changes champion balance and strategic priorities, making the relevance of prior experience time-dependent.

3.1.2 Prospective assessment of formed compositions

Let m denote a target match and let $y_m \in \{0, 1\}$ denote its outcome, where $y_m = 1$ indicates a blue-team victory. We use $s \in \mathcal{S} = \{\mathrm{B}, \mathrm{R}\}$ to index the two sides and $k \in \mathcal{K}$ to index the five roles. The champion assigned to the participant in slot (s, k) is denoted by $c_{m,s,k}$, and the patch is denoted by p_m.

For each participant, we retrieve a chronologically ordered sequence of earlier match appearances. Let $\mathcal{H}_{m,s,k}$ denote the available history of the participant occupying slot (s, k) in match m. Each historical event records that participant's own champion, role,

side, outcome, and patch. The collection of histories for the assembled participants is $\mathcal{H}_m = \{\mathcal{H}_{m,s,k}\}_{s\in\mathcal{S},\,k\in\mathcal{K}}$. Every historical appearance satisfies $\mathrm{idx}(e) < \mathrm{idx}(m)$ for $e \in \mathcal{H}_{m,s,k}$, where $\mathrm{idx}(\cdot)$ returns the globally chronological match index.

We represent the two formed compositions as

$$\mathcal{C}_{m,s} = \{(k, c_{m,s,k}, \mathcal{H}_{m,s,k})\}_{k\in\mathcal{K}}\,, \qquad \mathcal{C}_m = (\mathcal{C}_{m,\mathrm{B}}, \mathcal{C}_{m,\mathrm{R}})\,. \tag{1}$$

The assessment task is then

$$\hat{y}_m = P_\theta\left(y_m = 1 \mid p_m, \mathcal{C}_m\right). \tag{2}$$

This formulation evaluates the resulting pair of compositions before gameplay begins, using the eventual outcome as an observable indicator of relative collective performance. Direct rank, matchmaking rating, champion-mastery scores, runes, summoner spells, and variables generated after gameplay begins are excluded. Participant identifiers are used only to retrieve preceding records and are neither supplied to the network nor represented by trainable identity embeddings. The framework must therefore infer assignment-relevant proficiency from observable histories and evaluate those representations jointly at the composition level.

3.2 Dataset and chronological protocol

3.2.1 Match corpus

We compile a corpus of 321,146,032 solo-queue matches from all supported regions through the game provider's API [42]. We arrange the matches into a single global chronology. Each record contains the game patch, the ten participating players, their selected champions and assigned roles, their team sides, and the match outcome. We map original player identifiers to internal integer indices.

The corpus comprises 26,093,819 player indices and 171 champions and spans Patch 10.25 through Patch 25.16. To represent the evolving game environment, we map patches to 98 categories. Versions through Patch 14.24 are represented individually. Later versions, which may appear under either the 15.xx or 25.xx naming convention, share a single terminal category. This grouping ensures that matches from later periods use a patch representation that can be estimated from the training data rather than a previously unseen category appearing only during validation.

Match timestamps establish the global ordering but do not enter the model. After sorting, the chronological match index serves as the temporal coordinate. Table 1 summarizes the corpus and its partitions.

3.2.2 Chronological protocol

After chronological sorting, the first 310 million matches form the training interval, and the remaining 11,146,032 matches form the validation interval. The split follows global match order without random mixing across periods, so every validation match occurs after the complete training interval. This protocol evaluates performance on

Table 1: Summary of the match corpus and chronological partitions

Property	**Value**
Corpus	
Total matches	321,146,032
Player indices	26,093,819
Champions	171
Patch categories	98
Patch coverage	10.25–25.16
Regions	All supported regions
Queue type	Solo queue
Chronological partitions	
Training matches	310,000,000
Validation matches	11,146,032

later matches amid temporal changes in player experience, population membership, game balance, and strategic practice.

The validation interval supports model selection and comparisons among methods. All methods use the same chronological partition.

For a target match with index m, each participant's history contains only appearances with indices smaller than m. During validation, histories advance with the global chronology: an earlier validation match may contribute to a later participant history, whereas the target match and all subsequent matches remain unavailable. Each newly observed match therefore extends the information available for prospective assessment of later compositions. All methods follow this chronological information boundary.

3.3 Joint longitudinal and relational modeling

3.3.1 Architecture overview

The framework assesses a formed pair of compositions through two coupled stages. The first constructs an assignment-conditioned representation of each participant from chronologically preceding records. The second models the ten two-of-five representations jointly, along with explicit relations between participants occupying corresponding roles on opposing teams. A learned composition token gathers information across this complete configuration and yields a single assessment of the compositions as a whole. The first stage operationalizes assignment-specific proficiency, while the second captures within-team congruency and between-team opposition.

Let d denote the latent representation width and L the maximum number of preceding appearances retained for each participant. All embeddings and tokens introduced below lie in $\mathbb{R}^d$. Their concrete values and the remaining implementation choices are reported in the experimental setup.

3.3.2 Assignment-conditioned participant representation

For the participant occupying side s and role k in target match m, we retain an ordered sequence of up to the L most recent preceding appearances from $\mathcal{H}_{m,s,k}$. Let $e_j \in \mathcal{H}_{m,s,k}$ denote the jth appearance in the retained history. It records the participant's selected champion c_j, game patch p_j, assigned role k_j, team side s_j, and match outcome o_j, expressed from that participant's perspective. We write $e_j = (c_j, p_j, k_j, s_j, o_j)$ and compute the corresponding latent representation as follows:

$$\mathbf{x}_j = \nu_{\text{evt}} \left(\mathbf{E}_c(c_j) + \mathbf{E}_p(p_j) + \mathbf{E}_k(k_j) + \mathbf{E}_s(s_j) + \mathbf{E}_o(o_j) \right) + \mathbf{P}(j), \tag{3}$$

where $\mathbf{E}_c$, $\mathbf{E}_p$, $\mathbf{E}_k$, $\mathbf{E}_s$, and $\mathbf{E}_o$ are learned embeddings of the corresponding attributes, ν_{evt} is a normalization operator. Note that the in-game role k_j refers to the role occupied in historical appearance e_j, not the participant's current role k in target match m. $\mathbf{P}(j)$ is a learned sequence-order embedding—commonly called a positional embedding in transformer terminology—that indicates where appearance e_j occurs within the retained chronological history. Note that this sequence "position" is not to be confused with the participant's in-game "position" or "role", represented by $\mathbf{E}_k(k_j)$. Histories shorter than L are padded and accompanied by a mask so that padded entries do not contribute to attention.

A context-independent summary would assign the same meaning to a history under every present assignment. We instead construct a query from a learned base token $\mathbf{q}_0$ and the current champion, role, and patch. Let $\mathcal{T}_{\text{hist}}$ denote the transformer encoder shared across all participants, and let $\mathbf{M}_{m,s,k}$ denote the length-$(L+1)$ padding mask whose query position remains unmasked. The assignment query, history summary, and final participant representation are

$$\begin{aligned}
\mathbf{q}_{m,s,k} &= \mathbf{q}_0 + \mathbf{E}_c(c_{m,s,k}) + \mathbf{E}_p(p_m) + \mathbf{E}_k(k), \\
\mathbf{h}_{m,s,k} &= \left[\mathcal{T}_{\text{hist}} \left([\mathbf{q}_{m,s,k}; \mathbf{x}_1; \ldots; \mathbf{x}_L], \mathbf{M}_{m,s,k} \right) \right]_0, \\
\mathbf{u}_{m,s,k} &= \nu_{\text{part}} \left(\mathbf{h}_{m,s,k} + \mathbf{E}_c(c_{m,s,k}) + \mathbf{E}_p(p_m) + \mathbf{E}_k(k) + \widetilde{\mathbf{E}}_s(s) \right).
\end{aligned} \tag{4}$$

Here, $[\cdot;\cdot]$ denotes sequence concatenation, $[\cdot]_0$ selects the output at the query position, $\widetilde{\mathbf{E}}_s$ is a learned embedding of the participant's current team side, and ν_{part} is a normalization operator. The historical side embedding $\mathbf{E}_s$ and current side embedding $\widetilde{\mathbf{E}}_s$ serve distinct purposes: the former describes an earlier appearance, whereas the latter locates the participant within the focal composition. Champion, patch, and role embeddings are shared between historical appearances and the present assignment.

Because the assignment query participates in self-attention with the historical events [43], $\mathbf{h}_{m,s,k}$ interprets the same preceding record differently as the current champion, role, or patch changes. The additional current-context terms keep the present assignment explicit in $\mathbf{u}_{m,s,k}$. The ten participant histories remain separate throughout this stage; interactions among participants enter only after these assignment-conditioned representations are formed.

3.3.3 Joint composition representation

The composition stage receives the ten participant representations while retaining their role and team identities. It then constructs a relational representation for each pair of participants occupying the same role on opposing teams. Let $\mathbf{b}_{m,k} = \mathbf{u}_{m,\mathrm{B},k}$ and $\mathbf{r}_{m,k} = \mathbf{u}_{m,\mathrm{R},k}$ denote the blue- and red-team representations in role k. We form

$$\begin{aligned} \mathbf{a}_{m,k} &= [\mathbf{b}_{m,k} \parallel \mathbf{r}_{m,k} \parallel (\mathbf{b}_{m,k} - \mathbf{r}_{m,k}) \parallel (\mathbf{b}_{m,k} \odot \mathbf{r}_{m,k})], \\ \mathbf{v}_{m,k} &= \nu_{\mathrm{pair}}\left(\mathbf{W}_2\sigma\left(\mathbf{W}_1\mathbf{a}_{m,k} + \mathbf{b}_1\right) + \mathbf{b}_2\right) + \mathbf{E}_k(k), \end{aligned} \tag{5}$$

where $\parallel$ denotes vector concatenation, $\odot$ denotes element-wise multiplication, σ is an element-wise nonlinear activation function, and ν_{pair} is a normalization operator. Here, $\mathbf{a}_{m,k} \in \mathbb{R}^{4d}$, $\mathbf{W}_1 \in \mathbb{R}^{d\times 4d}$, $\mathbf{W}_2 \in \mathbb{R}^{d\times d}$, and $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^d$ are learned parameters. The concrete choices of σ and the normalization operators are reported in the experimental setup. The ordered participant pair preserves team-specific information, the difference exposes relative features, the element-wise product expresses feature-level interactions, and $\mathbf{E}_k(k)$ identifies the role represented by $\mathbf{v}_{m,k}$.

Let $\mathbf{z}_{\mathrm{comp}} \in \mathbb{R}^d$ be a learned composition token. We arrange this token, the ten participant representations, and the five role-pair representations into

$$\begin{aligned} \mathbf{Z}_m^0 &= [\mathbf{z}_{\mathrm{comp}}; (\mathbf{u}_{m,\mathrm{B},k})_{k\in\mathcal{K}}; (\mathbf{u}_{m,\mathrm{R},k})_{k\in\mathcal{K}}; (\mathbf{v}_{m,k})_{k\in\mathcal{K}}], \\ \mathbf{Z}_m &= \mathcal{T}_{\mathrm{comp}}\left(\mathbf{Z}_m^0\right), \qquad \mathbf{z}_m = [\mathbf{Z}_m]_0, \\ f_\theta(m) &= \mathbf{w}_{\mathrm{out}}^\top \mathbf{z}_m + b_{\mathrm{out}}, \end{aligned} \tag{6}$$

where $\mathcal{T}_{\mathrm{comp}}$ is the composition-level transformer encoder, $\mathbf{z}_m$ is its output at the composition-token position, and $\mathbf{w}_{\mathrm{out}}$ and b_{out} are learned readout parameters. The symbol θ collects all learned parameters of the framework.

Self-attention in $\mathcal{T}_{\mathrm{comp}}$ allows every participant and role-pair representation to change in light of the complete pair of teams before the composition token is read out. The resulting score therefore concerns the paired compositions as a relational whole rather than an aggregation of independently assessed participants.

3.3.4 Learning objective and side symmetry

The eventual match outcome supplies supervision. During optimization, we randomly exchange the complete blue and red compositions and reverse the corresponding outcome label, while each participant's history remains attached to that participant. We learn θ by minimizing binary cross-entropy between y_m and $\mathrm{sigmoid}(f_\theta(m))$ over the resulting training examples, where sigmoid is the logistic function. At assessment, we enforce antisymmetry explicitly. Let $\mathcal{S}(m)$ denote match m with the blue and red compositions exchanged. From the unconstrained score $f_\theta(m)$, we compute

$$\widetilde{f}_\theta(m) = \frac{1}{2}\left[f_\theta(m) - f_\theta(\mathcal{S}(m))\right], \qquad \hat{y}_m = \mathrm{sigmoid}\Big(\widetilde{f}_\theta(m)\Big). \tag{7}$$

This construction guarantees $\widetilde{f}_\theta(\mathcal{S}(m)) = -\widetilde{f}_\theta(m)$ and therefore produces complementary probabilities when the two sides are exchanged. Exchanging the ordered pair of compositions consequently reverses the assessment.

3.4 Experimental setup

3.4.1 Comparison methods

We compare the proposed framework with methods that capture different aspects of the information available to assess the team compositions. All methods follow the chronological protocol in Section 3.2: participant histories and aggregate statistics for match m contain only appearances preceding m, and the same training and validation intervals apply throughout. Since all methods are evaluated on our newly proposed dataset, we do not use the results reported in the original papers. Instead, all results are obtained from implementations manually developed by the authors of this manuscript. We release the model and execution code for all comparison methods, enabling the reported results to be reproduced using the released dataset.

Historical counts and priors. The side prior uses the empirical blue-team win rate in the training interval. Participant activity compares the average amount of recorded experience across the two teams. Role and champion specializations, and their joint role–champion variant measure how frequently participants previously appear under their present assignments, either overall or conditional on the outcome. Each scalar measure is converted into a probability through a one-variable logistic model fitted on the training interval, allowing its direction and scale to be estimated than assumed.

General skill rating. TrueSkill maintains an online rating for each participant and updates it chronologically after every observed outcome [26]. The five ratings on each team are aggregated to obtain a blue-team victory probability. This comparison captures broad differences in participant ability without distinguishing proficiency across champions or roles.

Feature-based models. CatBoost receives the current champion selections and patch together with chronological summaries of participant activity, recent outcomes, role experience, and champion experience [44]. We also implement a player–champion network following Do *et al.* [30]. Because external champion-mastery scores are unavailable in our corpus, preceding player–champion appearance counts and win rates are supplied to a multilayer perceptron. These models learn nonlinear relationships among manually summarized features without representing the sequence of historical appearances directly.

DraftRec. We adapt DraftRec to the fully specified, role-labeled compositions available in our corpus [32]. Its history encoder forms a participant representation independently of the present assignment, after which interactions among participants and selected champions are modeled. The relationally updated participant representations are then average-pooled within each team for match-level assessment. Pick order, bans, and other inputs unavailable in the corpus are omitted.

Recurrent sequence model. The recurrent comparison uses a gated recurrent unit (GRU) to encode each participant's preceding appearances [45]. Its final recurrent state is combined with the current champion, role, side, and patch before match-level

Table 2: Implementation settings for the principal configuration.

Name	Notation	Value
Architecture		
Shared representation and operators		
Latent width	d	512
Attention heads	H	8
Feed-forward width	d_{ff}	$4d$
Activation	σ	GELU
Normalization	$\nu_{\text{evt}}, \nu_{\text{part}}, \nu_{\text{pair}}$	LayerNorm
Assignment-conditioned participant representation		
History input	L	32
Sequence-order encoding	$\mathbf{P}(j)$	Learned
History-encoder depth	N_{hist}	8
Joint composition representation		
Composition input	$\lvert\mathbf{Z}_m^0\rvert$	$1 + 10 + 5 = 16$
Composition-encoder depth	N_{comp}	8
Training and assessment		
Optimization		
Optimizer	—	AdamW
Optimizer coefficients	(β_1, β_2)	$(0.9, 0.95)$
Weight decay	λ_{wd}	0.03
Batch size	B	1024
Learning-rate schedule		
Learning-rate range	$(\eta_{\max}, \eta_{\min})$	$(10^{-4}, 5 \times 10^{-6})$
Iteration boundaries	$(T_{\text{warm}}, T_{\text{decay}}, T)$	$(5k,\ 800k,\ 900k)$
Symmetry and validation		
Training-side exchange	p_{swap}	0.5
Validation cadence	T_{eval}	2,000 iterations
Validation sample	N_{val}	$B \times 100$ matches

assessment. This model represents the ordering of prior behavior, but its historical summary is formed before the present assignment is introduced.

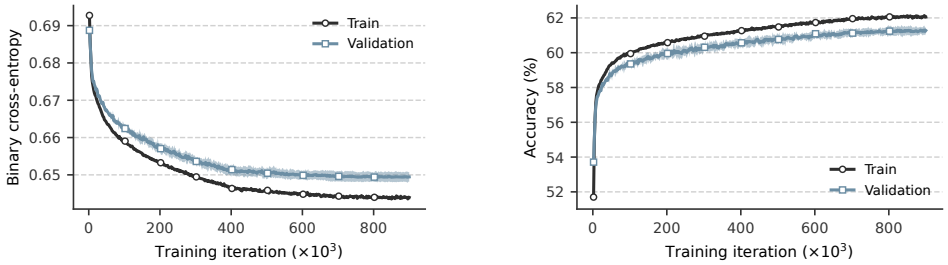
Proposed formulation. The complete framework interprets each participant's history in relation to the present assignment and subsequently models the ten participant representations and five role-aligned relations together. The paired compositions, rather than an aggregation of independently formed member-level assessments, are therefore represented before readout.

3.4.2 Implementation configuration

Table 2 maps the generic operators in Section 3.3 to their concrete implementation. For representation width d and encoder depths N_{hist} and N_{comp}, the parameter count is approximately

$$|\theta| \approx [12\,(N_{\text{hist}} + N_{\text{comp}}) + 5]\,d^2 + \mathcal{O}(Vd),$$

where each transformer layer contributes approximately $12d^2$ parameters, the role-pair projection contributes approximately $5d^2$, and V summarizes the embedding vocabularies. With $d = 512$ and $N_{\text{hist}} = N_{\text{comp}} = 8$, the complete configuration contains approximately 52 million trainable parameters.



(a) Binary cross-entropy.

(b) Prediction accuracy.

Fig. 1: Training dynamics of the proposed formulation. The left and right panels report binary cross-entropy and prediction accuracy, respectively. Shading denotes one standard deviation over five independent runs.

3.4.3 Training and reporting protocol

We optimize the proposed framework and neural comparison methods with AdamW and binary cross-entropy [46]. Training matches are sampled uniformly, with replacement, from the training interval, and their histories are constructed using only earlier match indices. For each training example, the complete blue and red compositions are exchanged with probability 0.5, with the outcome label reversed accordingly. Gradients are clipped to a maximum norm of 1.0.

The learning rate increases linearly during the first 5,000 iterations, follows cosine decay through iteration 800,000, and remains at 5.0×10^{-6} for the remaining iterations. The 900,000 iterations expose the optimizer to 921.6 million sampled examples, equivalent in number to approximately three traversals of the training interval, although sampling occurs with replacement.

Training uses distributed data parallelism across eight NVIDIA RTX PRO 6000 Blackwell GPUs with bfloat16 mixed-precision arithmetic. The implementation uses PyTorch 2.12.0 and CUDA 13.1.

We evaluate every 2,000 iterations on a fixed sample of 102,400 matches from the validation interval and retain the checkpoint with the highest validation accuracy. Accuracy is the primary measure, while binary cross-entropy assesses the quality of the estimated probabilities. Each stochastic method is evaluated over five independent runs, and deterministic methods are evaluated once. Comparisons, model-capacity analyses, history-length analyses, and ablation experiments follow the same chronological information boundary and reporting protocol.

4 Results

4.1 Predictive performance

Figure 1 displays the learning curves of our formulation. The left and the right panels show the loss, in binary cross entropy, and the prediction accuracy, respectively, on

Table 3: Pre-game prediction of winning team compositions using the validation dataset. Stochastic methods report one standard deviation over five runs. Boldface and underlining indicate the best and second-best means, respectively.

Method	Information represented	Accuracy (%)	BCE
Count-based measures			
Side prior	Training side-win rate	51.66	0.6929
Participant activity	# preceding appearances	51.49	0.6933
Role	Current-role frequency	52.70	0.6909
Champion	Champion frequency	53.09	0.6899
Role–champion	Role and champion frequencies	53.24	0.6895
Champion–outcome	Historical champion win rate	52.87	0.6912
Role–outcome	Historical role win rate	53.02	0.6907
Champion–role–outcome	Historical champion–role win rate	53.33	0.6891
General skill rating			
TrueSkill	Online participant ratings	53.88	0.7013
Feature-based models			
CatBoost	Aggregate chronological features	57.05 ± 0.07	0.6767 ± 0.0011
Player–champion MLP	Aggregate proficiency measures	56.29 ± 0.04	0.6833 ± 0.0005
History-based NN			
GRU	Histories and current assignment	57.01 ± 0.06	0.6757 ± 0.0010
DraftRec	Histories and current assignment	<u>58.31</u> ± 0.06	<u>0.6689</u> ± 0.0007
Proposed formulation		**61.33 ± 0.05**	**0.6490 ± 0.0008**

both train and validation sets. The results are averaged across five independent runs, represented alongside with one standard deviation.

Table 3 compares the proposed formulation with count-based measures, a general skill rating, feature-based models, and history-based neural models under the common chronological protocol. The nonlearned comparisons establish that simple historical signals are informative but limited. Measures tied to the current champion or role generally improve on side and activity priors, while TrueSkill provides the strongest accuracy among these comparisons but worse BCE than the simple priors. General participant ability therefore carries useful ranking information, yet a single context-independent rating does not provide a well-calibrated assessment of the present assignments and team compositions.

Learned representations produce a clear improvement over these manually summarized signals. CatBoost and the GRU perform similarly, indicating that replacing aggregate features with a recurrent history encoder is not sufficient by itself to produce a substantial gain. The player–champion MLP also improves on the nonlearned comparisons but remains behind the richer feature-based model. DraftRec is the strongest prior method on both measures, supporting the value of modeling interactions among participants and their selected champions rather than relying only on independent proficiency summaries.

The proposed formulation achieves the strongest accuracy and BCE. Because the strongest baseline already incorporates participant histories and relational updates, the remaining improvement is consistent with the two central design choices of our framework: interpreting each participant's preceding behavior in relation to the present assignment and representing the paired compositions jointly before readout. Overall,

Table 4: Scaling with model capacity and participant-history length. Layers list the history/composition encoder depths, and L is the number of preceding appearances retained per participant. Results are the mean ± one standard deviation over five runs. The principal model used throughout the paper as well as the other best performance values are highlighted in boldface.

Setting	Params.	Width	Layers	Context L	Acc. (%)	BCE
Model capacity						
Tiny	0.92M	128	2/2	32	58.64 ± 0.04	0.6705 ± 0.0006
Small	4.37M	256	2/3		59.11 ± 0.05	0.6662 ± 0.0005
Medium	13.3M	384	3/4		60.02 ± 0.07	0.6568 ± 0.0013
Principal	**51.9M**	**512**	**8/8**		**61.33** ± 0.05	**0.6490** ± 0.0008
Large	173M	768	12/12		**61.36** ± 0.06	**0.6502** ± 0.0011
History length						
Principal	51.9M	512	8/8	8	59.32 ± 0.06	0.6675 ± 0.0009
				16	60.14 ± 0.05	0.6626 ± 0.0010
				32	**61.33** ± 0.05	**0.6490** ± 0.0008
				64	**61.40** ± 0.07	**0.6478** ± 0.0010

the results favor composition-level inference over aggregating assignment-independent participant representations.

4.2 Scaling and historical context

We now investigate how model scaling and the length of historical context affect the model performance. Table 4 examines model capacity and the length of each participant's preceding history under the common chronological protocol. The capacity analysis holds the history length at $L = 32$, while the context analysis varies L with the remaining configuration fixed. All variants retain the same assignment-conditioned and joint composition construction.

Increasing capacity produces steady improvements through the principal configuration. Expanding the model further to 173 million parameters yields essentially unchanged accuracy and does not improve BCE, indicating that the benefit of additional width and depth has largely saturated. The principal configuration therefore provides the favorable operating point: it retains the predictive quality of the substantially larger model with less than one third of its parameters.

Longer participant histories likewise improve both measures. Extending the context from 8 to 32 preceding appearances produces substantial gains, whereas doubling it again to 64 yields marginal improvement. Thus, the relevant evidence is not confined to a few immediately preceding appearances, but the returns diminish once a moderately long history is available. We use $L = 32$ throughout the remaining experiments to balance predictive quality with the computation required to encode ten participant histories for every composition.

4.3 Ablation study

Table 5 isolates the longitudinal and relational components of the proposed formulation under the same chronological protocol. The complete formulation achieves the

Table 5: Ablation of the proposed formulation on the validation dataset. Each variant changes only the stated component while retaining the remaining configuration. Results report the mean ± one standard deviation over five runs.

Variant	Accuracy (%)	BCE
Full formulation	**61.33** ± 0.05	**0.6490** ± 0.0008
Assignment-conditioned participant representation		
Current composition only (no histories)	55.37 ± 0.03	0.6845 ± 0.0003
Assignment-independent history encoding	60.56 ± 0.08	0.6594 ± 0.0013
History without outcomes	60.85 ± 0.06	0.6588 ± 0.0008
Joint composition representation		
Without role-pair representations	59.62 ± 0.04	0.6610 ± 0.0004
Independent participant evaluation (mean aggregation)	58.19 ± 0.07	0.6713 ± 0.0006

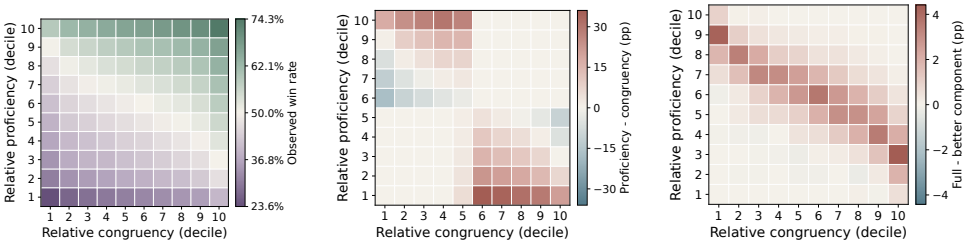
highest accuracy and lowest binary cross-entropy, with consistent differences averaged across five independent runs.

The two most restrictive variants separate the principal sources of evidence. The current-composition-only variant removes participant histories while retaining joint relational modeling, and consequently assesses the formed composition without participant-specific proficiency: *congruency*. Conversely, independent participant evaluation retains assignment-conditioned histories but replaces the joint composition representation with a shared proficiency head followed by mean aggregation over the ten output logits: *proficiency*. The respective gaps between these variants and the full formulation show that neither relational composition modeling without longitudinal proficiency nor aggregated proficiency without composition-level interactions is sufficient. The complete formulation benefits from modeling these two sources jointly.

The remaining variants clarify the construction of each stage. Encoding a participant's history independently of the present assignment weakens both metrics, supporting the use of the current champion, role, and patch to select relevant longitudinal evidence. Removing historical outcomes produces a further decline, indicating that prior choices alone do not fully characterize proficiency. At the composition stage, removing the role-pair representations also causes a substantial loss despite retaining joint attention among all ten participants. Direct, role-aligned comparisons therefore complement the more general within-team and cross-team interactions learned by the composition encoder.

4.4 Composition-level and qualitative analyses

Aggregate predictive performance does not reveal where proficiency and congruency provide useful evidence or where either component becomes insufficient. We therefore examine their relationship at different compositional levels. For each validation match, the outputs of the proficiency and congruency models provide two relative scores for the team composition. Specifically, we use *Current composition only (no histories)* model to measure the relative team congruency and *Independent participant evaluation (mean aggregation)* model to assess the relative team proficiency, both of which are discussed earlier in Table 5. Therefore, these are model-implied relative scores, rather than direct measurements. We then partition each score into marginal deciles and



(a) Observed win rate. (b) Component advantage. (c) Joint-model gain.

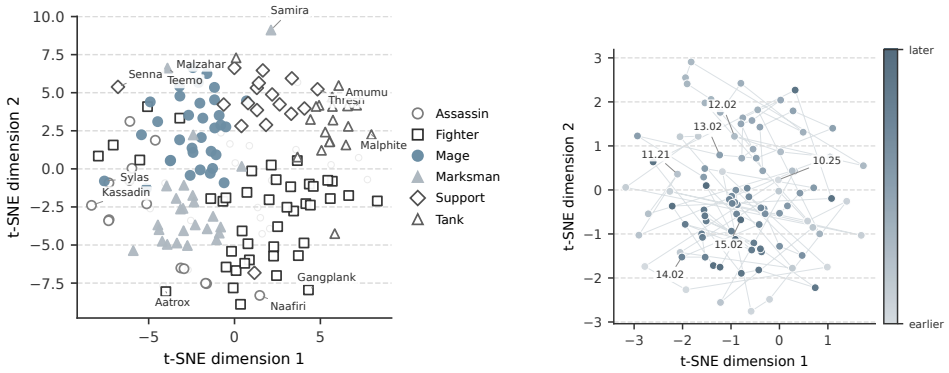
Fig. 2: Composition-level relationship among relative proficiency, relative congruency, and performance, with both component scores divided into deciles. (a) Observed team win rate within each cell. (b) Accuracy difference between the proficiency and congruency models, in percentage points; positive values favor proficiency and negative values favor congruency, exposing their complementary blind spots. (c) Accuracy difference between the full model and the better component model, highlighting gains from joint modeling where the component scores conflict.

cross the two partitions to obtain a common 10×10 grid. The axes thus represent relative proficiency and congruency with respect to the opposing composition.

The observed win-rate surface in Figure 2a relates the two learned quantities to the actual win-loss outcomes. Win rates rise most clearly when both proficiency and congruency are favorable; strength along only one axis does not produce the same advantage. This pattern is consistent with composition as a joint condition: individually suitable assignments are most valuable when they also form a coherent configuration, and a coherent configuration cannot fully compensate for weak assignment-specific proficiency.

Unlike the preceding panel, which reports observed win rates, Figure 2b compares the local prediction accuracies of the proficiency and congruency models. Where the two scores agree—the bottom-left region, in which both are low, and the top-right region, in which both are high—their predictive accuracy difference remains generally small. Clearer contrasts emerge where the scores disagree. Within the high-proficiency, low-congruency region, the congruency model is more accurate in the lower-left triangular portion, whereas the proficiency model becomes more accurate toward the upper-right. The pattern reverses in the low-proficiency, high-congruency region: proficiency is more informative in the lower-left portion, while congruency is more informative toward the upper-right. These shifting advantages expose complementary blind spots: neither component is uniformly sufficient, and the predictive value of each depends on the relative strength of both proficiency and congruency.

Figure 2c uses a demanding reference: within each cell, it compares predictive performance of the full model against whichever component model attains the higher accuracy. The largest gains appear along the anti-diagonal, where proficiency and congruency provide conflicting signals, and around the center, where neither component provides decisive evidence. These are precisely the regions in which the specialized models exhibit their complementary blind spots. The full model improves by jointly



(a) Champion embeddings.

(b) Patch embeddings.

Fig. 3: Qualitative structure of the learned representations. (a) t-SNE projection of champion embeddings, colored by functional tags published by the game provider and annotated with selected examples. The tags are used only for visualization and are not available to the model. (b) t-SNE projection of patch embeddings, colored by release order. The line connects consecutive patches, and selected patches are annotated.

interpreting the relative strength of both signals within the complete pair of compositions, rather than merely relying on the locally stronger component. Moreover, cells in which the full model performs comparatively worse are rare, indicating that joint modeling reduces these blind spots without introducing a corresponding weakness elsewhere in the grid.

We next examine the structure that emerges in two representations learned solely from the outcome-prediction objective: champion and patch embeddings. Figure 3 presents two-dimensional t-SNE projections of these embedding spaces [47]. Because t-SNE primarily preserves local neighborhoods, the projections serve as qualitative views of nearby relationships rather than metric-preserving maps of the complete spaces.

The champion space exhibits local organization by gameplay function (Figure 3a). Champions assigned similar functional tags frequently occupy neighboring regions even though these tags provide no training supervision. This organization arises from the contexts in which champions appear—participants' longitudinal choices and outcomes, their current roles, and the surrounding compositions. The embedding space thereby captures similarities that are relevant to both assignment-specific proficiency and relational congruency, rather than treating each champion as an unrelated identifier.

The patch space displays a complementary temporal structure (Figure 3b). Consecutive patches often remain close in the projected space, while the overall trajectory changes across the release sequence. Although patch identifiers enter the model categorically, their representations are learned from how the relevance of prior experience and the consequences of present assignments vary across game environments. The resulting organization is consistent with gradual evolution interrupted by larger shifts in champion balance and strategic conditions.

Together, the composition-level and representation analyses connect the full model's predictive advantage to both observable performance patterns and its internal organization. The grid analysis shows why proficiency and congruency must be considered jointly, while the embedding spaces show how binary outcome supervision organizes persistent champion function and changing game context. These results support interpreting prior behavior in relation to the present assignment before assessing the two compositions as a whole.

5 Discussion

This study examines how temporary teams transform individual experience into collective capability through the joint organization of proficiency and congruency. Using approximately 320 million five-on-five matches involving 26 million participants, we evaluate formed compositions before any behavior from the ensuing game becomes available. The proposed framework interprets each participant's longitudinal experience in relation to the present champion–role assignment and then represents the two opposing compositions jointly. It achieves the strongest predictive performance among the evaluated skill-rating systems, manually constructed composition measures, sequence models, and prior learning-based approaches. These results indicate that winning compositions cannot be characterized adequately by either accumulated individual experience or the current configuration alone.

The findings support two central aspects of the construction. First, proficiency is contextual rather than fixed. The same participant history carries different implications depending on the currently selected champion, assigned role, and prevailing patch. Making the history representation independent of the current assignment weakens prediction, as does removing historical outcomes. Second, composition requires relational modeling rather than independent evaluation followed by aggregation. Role-pair representations provide direct comparisons between corresponding assignments, while joint attention captures dependencies within and across the two teams. Removing these mechanisms or replacing the composition representation with independently estimated participant scores produces substantial losses. The distinction therefore extends the proficiency–congruency perspective [7] from manually specified summaries to representations learned directly from longitudinal and relational evidence.

The composition-level analyses further clarify why both components are necessary. When relative proficiency and congruency agree, the specialized models exhibit similar predictive performance. Their complementary blind spots emerge when the two signals conflict: proficiency alone can overlook weaknesses in the surrounding configuration, whereas congruency alone can overlook insufficient assignment-specific experience. The full model gains most clearly along this anti-diagonal and around the center of the grid, where neither component provides an unambiguous account of the composition. Importantly, these gains remain when the reference is the better component model within each cell. Joint modeling therefore contributes more than selecting whichever consideration appears locally stronger; it learns how the implications of one depend on the other within the complete pair of compositions.

This result has implications beyond multiplayer games. Temporary teams in settings such as hackathons, pickup sports, project work, and other short-lived collaborations often divide differentiated responsibilities before members develop a shared history. In such settings, prior success in a broadly similar task does not necessarily transfer to the responsibility currently assumed, and individually reasonable assignments do not necessarily form a coherent team. The present findings suggest that composition assessment should preserve both forms of structure: longitudinal evidence about who is prepared for each responsibility and relational evidence about how those responsibilities fit together. More generally, they caution against treating collective capability as the sum or average of independently measured individual attributes.

The pre-game setting makes this assessment particularly demanding. The game provider's matching process already limits broad differences in overall participant ability, leaving more subtle variation in assignment-specific proficiency and composition. At the same time, the framework observes no actions, events, or performance measurements from the game whose outcome it assesses. The ten participants subsequently compete for tens of minutes, during which execution, adaptation, and unobserved contingencies introduce substantial uncertainty. Restricting the inputs in this way is nevertheless central to the research question: it isolates the information contained in the composition itself rather than allowing evidence from unfolding play to dominate the assessment. The scaling results show that additional historical context and model capacity continue to improve this pre-game discrimination, while the learned champion and patch spaces display functional and temporal organization consistent with the model's intended construction.

Several boundaries qualify the interpretation. Proficiency and congruency are operationalized through specialized predictive models and are therefore model-implied, relative quantities rather than directly observed team attributes. Match outcomes also combine composition with execution and chance, so prediction errors do not necessarily indicate an incorrect assessment of the initial configuration. Moreover, the analysis is observational: the framework identifies associations that generalize chronologically but does not establish the causal effect of changing a participant's assignment. Finally, *League of Legends* provides unusually explicit roles, repeated participation, discrete outcomes, and extensive longitudinal records. These properties enable precise computational analysis but may not transfer directly to settings in which responsibilities remain fluid, performance is multidimensional, or prior behavior is sparsely recorded.

Future research can examine whether the same longitudinal–relational construction generalizes across games and non-game teams, accommodates variable team sizes and less formal role structures, or supports causal evaluation of alternative assignments. Another direction is to study how teams negotiate between proficiency and congruency during composition formation, rather than only assessing the configuration that results. Comparing pre-performance assessments with later interaction data may also separate the contribution of initial composition from the coordination processes that follow. We release the dataset and code to support replication, extension, and evaluation in related settings. Overall, the results establish team composition as a joint longitudinal and relational problem: effective teams depend not only on what their members can do, but on whether their present assignments form a coherent whole.

Data availability

The full dataset used to train and evaluate the proposed deep learning model and the comparison methods is publicly available in the Zenodo repository at https://doi.org/10.5281/zenodo.21873669. The repository also includes a README describing the dataset files and data structure.

Code availability

The custom code used to train and evaluate the models is available in the Supplementary Information.

References

[1] Kozlowski, S., Ilgen, D.: Enhancing the effectiveness of work groups and teams. Psychological Science in the Public Interest **7**(3), 77–124 (2006) https://doi.org/10.1111/j.1529-1006.2006.00030.x

[2] Stewart, G.: A meta-analytic review of relationships between team design features and team performance. Journal of Management **32**(1), 29–55 (2006) https://doi.org/10.1177/0149206305277792

[3] Wageman, R., Gardner, H., Mortensen, M.: The changing ecology of teams: New directions for teams research. Journal of Organizational Behavior **33**(3), 301–315 (2012) https://doi.org/10.1002/job.1775

[4] Mortensen, M., Haas, M.: Perspective—rethinking teams: From bounded membership to dynamic participation. Organization Science **29**(2), 341–355 (2018) https://doi.org/10.1287/orsc.2017.1198

[5] Bechky, B.: Gaffers, gofers, and grips: Role-based coordination in temporary organizations. Organization Science **17**(1), 3–21 (2006) https://doi.org/10.1287/orsc.1050.0149

[6] Valentine, M., Edmondson, A.: Team scaffolds: How mesolevel structures enable role-based coordination in temporary groups. Organization Science **26**(2), 405–422 (2015) https://doi.org/10.1287/orsc.2014.0947

[7] Kim, J., Keegan, B., Park, S., Oh, A.: The proficiency–congruency dilemma: Virtual team design and performance in multiplayer online games. In: CHI (2016). https://doi.org/10.1145/2858036.2858464

[8] Humphrey, S., Morgeson, F., Mannor, M.: Developing a theory of the strategic core of teams: A role composition model of team performance. Journal of Applied Psychology **94**(1), 48–61 (2009) https://doi.org/10.1037/a0012997

[9] Kim, S.-H., Song, H., Valentine, M.: Learning in temporary teams: The varying effects of partner exposure by team member role. Organization Science **34**(1), 433–455 (2023) https://doi.org/10.1287/orsc.2022.1585

[10] Mathieu, J., Tannenbaum, S., Donsbach, J., Alliger, G.: A review and integration of team composition models: Moving toward a dynamic and temporal framework. Journal of Management **40**(1), 130–160 (2014) https://doi.org/10.1177/0149206313503014

[11] Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: KDD (2009). https://doi.org/10.1145/1557019.1557074

[12] Contractor, N.: Some assembly required: Leveraging Web science to understand and enable team assembly. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **371**(1987), 20120385 (2013) https://doi.org/10.1098/rsta.2012.0385

[13] Carter, D., Asencio, R., Wax, A., DeChurch, L., Contractor, N.: Little teams, big data: Big data provides new opportunities for teams theory. Industrial and Organizational Psychology **8**(4), 550–555 (2015) https://doi.org/10.1017/iop.2015.82

[14] Guimerà, R., Uzzi, B., Spiro, J., Amaral, L.N.: Team assembly mechanisms determine collaboration network structure and team performance. Science **308**(5722), 697–702 (2005) https://doi.org/10.1126/science.1106340

[15] Wax, A., DeChurch, L., Contractor, N.: Self-organizing into winning teams: Understanding the mechanisms that drive successful collaborations. Small Group Research **48**(6), 665–718 (2017) https://doi.org/10.1177/1046496417724209

[16] Cheng, Z., Yang, Y., Tan, C., Cheng, D., Zhuang, Y., Cheng, A.: What makes a good team? a large-scale study on the effect of team composition in Honor of Kings. In: WWW (2019). https://doi.org/10.1145/3308558.3313530

[17] Carrasco-Farré, C., Hakobjanyan, N.: Experience shapes non-linearities between team behavioral interdependence, team collaboration, and performance in massively multiplayer online games. Scientific Reports **14**, 7850 (2024) https://doi.org/10.1038/s41598-024-57919-w

[18] Sapienza, A., Goyal, P., Ferrara, E.: Deep neural networks for optimal team composition. Frontiers in Big Data **2**, 14 (2019) https://doi.org/10.3389/fdata.2019.00014

[19] Carli, F., Foini, P., Gozzi, N., Perra, N., Schifanella, R.: Modeling teams performance using deep representational learning on graphs. EPJ Data Science **13**(1), 7 (2024) https://doi.org/10.1140/epjds/s13688-023-00442-1

[20] Ching, K., Forti, E., Rawley, E.: Extemporaneous coordination in specialist teams: The familiarity complementarity. Organization Science **32**(1), 1–17 (2021) https://doi.org/10.1287/orsc.2020.1376

[21] Ching, K., Forti, E., Rawley, E.: Competitive familiarity: Learning to coordinate by competing. Organization Science **35**(4), 1409–1426 (2024) https://doi.org/10.1287/orsc.2022.17068

[22] Kristof-Brown, A., Zimmerman, R., Johnson, E.: Consequences of individuals' fit at work: A meta-analysis of person–job, person–organization, person–group, and person–supervisor fit. Personnel Psychology **58**(2), 281–342 (2005) https://doi.org/10.1111/j.1744-6570.2005.00672.x

[23] Faraj, S., Sproull, L.: Coordinating expertise in software development teams. Management Science **46**(12), 1554–1568 (2000) https://doi.org/10.1287/mnsc.46.12.1554.12072

[24] Yang, A., Liu, D., Santhanam, R.: The impact of role coordination on virtual team performance and player retention in esports. In: Workshop on E-Business (2021). https://doi.org/10.1007/978-3-030-79454-5_11

[25] Bendell, R., Williams, J., Fiore, S., Jentsch, F.: Individual and team profiling to support theory of mind in artificial social intelligence. Scientific Reports **14**, 12635 (2024) https://doi.org/10.1038/s41598-024-63122-8

[26] Herbrich, R., Minka, T., Graepel, T.: TrueSkill: A bayesian skill rating system. In: NeurIPS (2006). https://dl.acm.org/doi/10.5555/2976456.2976528

[27] Chen, Z., Sun, Y., El-Nasr, M.S., Nguyen, T.-H.: Player skill decomposition in multiplayer online battle arenas. arXiv preprint arXiv:1702.06253 (2017)

[28] Jiang, J., Maldeniya, D., Lerman, K., Ferrara, E.: The wide, the deep, and the maverick. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW1), 1–26 (2021) https://doi.org/10.1145/3449290

[29] Jiang, J., Lerman, K., Ferrara, E.: Individualized context-aware tensor factorization for online games predictions. In: ICDMW (2020). https://doi.org/10.1109/ICDMW51313.2020.00048

[30] Do, T., Wang, S., Yu, D., McMillian, M., McMahan, R.: Using machine learning to predict game outcomes based on player–champion experience in League of Legends. In: International Conference on the Foundations of Digital Games (2021). https://doi.org/10.1145/3472538.3472579

[31] Chen, Z., Nguyen, T.-H., Xu, Y., Amato, C., Cooper, S., Sun, Y., El-Nasr, M.S.: The art of drafting: A team-oriented hero recommendation system for multiplayer online battle arena games. In: RecSys (2018).

https://dl.acm.org/doi/10.1145/3240323.3240345

[32] Lee, H., Hwang, D., Kim, H., Lee, B., Choo, J.: DraftRec: Personalized draft recommendation for winning in multi-player online battle arena games. In: WWW (2022). https://doi.org/10.1145/3485447.3512278

[33] Bisberg, A., Ferrara, E.: GCN-WP: Semi-supervised graph convolutional networks for win prediction in esports. In: IEEE Conference on Games (2022). https://ieeexplore.ieee.org/document/9893671

[34] De Bois, M., Parmentier, F., Puget, R., Tanti, M., Peltier, J.: PandaSkill: Player performance and skill rating in esports—application to League of Legends. arXiv preprint arXiv:2501.10049 (2025)

[35] Yang, Y., Qin, T., Lei, Y.-H.: Real-time esports match result prediction. arXiv preprint arXiv:1701.03162 (2016)

[36] Zhou, G., Song, C., Zhu, X., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H., Gai, K.: Deep interest network for click-through rate prediction. In: KDD (2018). https://arxiv.org/abs/1706.06978

[37] Zhou, G., Mou, N., Fan, Y., Pi, Q., Bian, W., Zhou, C., Zhu, X., Gai, K.: Deep interest evolution network for click-through rate prediction. In: AAAI (2019). https://arxiv.org/abs/1809.03672

[38] Li, Y., Rao, S., Solares, J.A., Hassaine, A., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: BEHRT: Transformer for electronic health records. Sci. Rep. **10**, 7155 (2020) https://doi.org/10.1038/s41598-020-62922-y

[39] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit. Med. **4**, 86 (2021) https://doi.org/10.1038/s41746-021-00455-y

[40] Shmatko, A., Jung, A., Gaurav, K., Brunak, S., Mortensen, L., Birney, E., Fitzgerald, T., Gerstung, M.: Learning the natural history of human disease with generative transformers. Nature **647**, 248–256 (2025) https://doi.org/10.1038/s41586-025-09529-3

[41] Riot Games: League of Legends: How to Play (2026). https://www.leagueoflegends.com/en-us/how-to-play/

[42] Riot Games: Riot Developer Portal: League of Legends APIs. Accessed: Jun. 25, 2026 (2026). https://developer.riotgames.com/apis

[43] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017).

https://arxiv.org/abs/1706.03762

[44] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A., Gulin, A.: CatBoost: Unbiased boosting with categorical features. In: NeurIPS (2018). https://arxiv.org/abs/1706.09516

[45] Cho, K., Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: EMNLP (2014). https://doi.org/10.3115/v1/D14-1179 . https://aclanthology.org/D14-1179/

[46] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019). https://arxiv.org/abs/1711.05101

[47] Maaten, L., Hinton, G.: Visualizing data using t-SNE. JMLR **9**, 2579–2605 (2008)

Author contributions

J.K. conceptualized the study, developed the methodology and software, collected and analyzed the data, interpreted the results, and wrote and revised the manuscript.

Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST)).

Competing Interests

The author declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- profcongSRsupp.zip