Journal of Software Engineering Research and Development, 2026, 14:23, doi: 10.5753/jserd.2026.7315

An Expert-Guided Method for Developer Compatibility Graphs and Bayesian Team-Fit Evaluation

Felipe Cunha [**Federal University of Campina Grande**| ***felipe.cunha@virtus.ufcg.edu.br***]
Mirko Perkusich [**Federal University of Campina Grande**| ***mirko@virtus.ufcg.edu.br***]
Danyllo Albuquerque [**Federal University of Campina Grande**| ***danyllo@copin.ufcg.edu.br***]
Emanuel Dantas Filho [**Federal Institute of Pernambuco**| ***emanuel.filho@jaboatao.ifpe.edu.br***]
Ademar Sousa Neto [**Federal University of Campina Grande**| ***ademar.sousa@virtus.ufcg.edu.br***]
Kyller Gorgônio [**Federal University of Campina Grande**| ***kyller@virtus.ufcg.edu.br***]
Angelo Perkusich [**Federal University of Campina Grande**| ***perkusic@virtus.ufcg.edu.br***]

Abstract

[Context]. Forming effective software teams remains challenging, particularly in large organizations managing multiple concurrent projects. Existing approaches to modeling developer compatibility often rely on subjective personality traits, purely structural proxies, or large volumes of historical interaction data, limiting both their interpretability and their alignment with how staffing decisions are made in practice. **[Objective]**. This paper proposes a lightweight, expert-guided method for modeling developer compatibility and operationalizing collaboration evidence for transparent and explainable team-fit evaluation. **[Method]**. We introduce a regression-calibrated Pair Compatibility (PC) model that combines two project-level indicators, Objective Success Factor (OSF) and Subjective Leadership Factor (SLF), with an expert-defined saturation effect for repeated collaborations. During multiple expert workshops, the expert validated the relevance and conceptual interpretation of OSF and SLF, their normalized scales, and their use in compatibility modeling. The model is calibrated using eight synthetic collaboration scenarios labeled by the expert and is then applied to an organizational analysis base parameterized from anonymized project-participation records and project-level OSF/SLF indicators, thereby producing a weighted developer compatibility graph. Pairwise compatibility evidence is then transformed into a size-agnostic collaboration profile and integrated into an expert-calibrated Bayesian Network (BN) that estimates overall Team Fit. **[Results]**. In an industrial organization with more than 300 professionals, the calibrated PC regression achieved a Mean Absolute Error (MAE) of 0.0812 and a Pearson correlation of 0.80 with expert expectations on the eight labeled scenarios. When applied to the organizational dataset of 192 developers, and after the $PC > 0.3$ filter, 773 valid developer pairs were retained; the resulting compatibility graph illustrates how the method produces cohesive collaboration clusters and identifies bridge developers to support managerial staffing decisions. The BN-based team evaluator produced Brier scores of 0.0788 and 0.0782 for Team Fit (AE) in scenario-based validation, reflecting the concentrated posterior distributions of the calibrated model ($\sigma = 0.05$) relative to the smoother expert-provided reference distributions. For Collaboration Fit (AC), the PC-tier distribution representation ranked collaboration quality in closer agreement with expert judgment than a mean-based encoding (Spearman $\rho = 0.882$ vs. $\rho = 0.812$). **[Conclusion]**. By combining expert-validated indicator definitions, organizational co-participation records, project-level OSF/SLF indicators, and a small set of synthetic expert-labeled scenarios, the proposed method yields an interpretable compatibility graph and a transparent probabilistic evaluator that bridges pairwise collaboration evidence and team-level decision-making. The results support the feasibility of expert-guided regression and Bayesian reasoning as a practical and explainable alternative to data-intensive black-box models for software team formation (STF), particularly in settings where interpretability and alignment with local staffing authority are central requirements.

Keywords: *Bayesian Networks, Software Team Formation, Software Engineering, Knowledge Engineering, Expert Systems.*

1 Introduction

Forming effective software development teams remains a persistent challenge for organizations managing multiple concurrent projects. As software work increasingly depends on collaboration across heterogeneous skills and roles, research has proposed a wide range of computational solutions to support team formation decisions (Cunha et al., 2025b; Costa et al., 2020; Hamidi Rad et al., 2023). A central difficulty, however, lies not only in searching for candidate teams but in defining how team quality should be assessed in a manner that is both valid and acceptable to human decision makers.

Graph-based models are among the most widely adopted representations in this context. In these solutions, developers are modeled as nodes and past interactions as weighted edges, derived from signals such as co-authorship (Lappas et al., 2009), co-participation in tasks (Hamidi Rad et al., 2021), or communication traces (Ruenin et al., 2024). While such representations are intuitive and reproducible, their edge semantics are typically inferred directly from available logs, assuming that historical interaction frequency or structural proximity is a sufficient proxy for collaboration quality. In practice, however, these signals are often sparse, noisy, or misaligned with the qualitative criteria that managers use to judge whether a team is likely to succeed.

This gap becomes particularly evident in formulations of the Software Team Formation Problem (STFP) that rely on search-based or heuristic optimization. Genetic algorithms and related techniques are frequently employed due to their ability to explore large combinatorial spaces efficiently (Costa et al., 2022; Cunha et al., 2022a; Harman et al., 2012). Yet, in most existing work, the objective (fitness) function is defined using fixed weights, structural proxies, or readily quantifiable attributes, with limited grounding in expert decision criteria. As a result, the fitness function effectively acts as the ruler of team quality: if this ruler does not faithfully represent how staffing authorities assess teams, optimization may converge toward solutions that are numerically optimal yet substantively misaligned with organizational judgment.

This limitation is not merely technical, but epistemic. Historical team allocations cannot be assumed to be optimal among all feasible alternatives, and software projects do not allow for realistic execution-level simulation to evaluate counterfactual team compositions. Consequently, the bottom-up construction of evaluation functions from available data alone provides a fragile foundation for both decision support and empirical validation. Industrial evidence suggests that such models are often distrusted unless their evaluation logic is transparent, auditable, and explicitly aligned with expert reasoning (Manzano et al., 2021; Rique et al., 2023).

To address this gap, we propose an expert-guided method for constructing and operationalizing developer compatibility as structured evidence for team-level evaluation. Rather than treating compatibility graphs as standalone artifacts or direct optimization targets, we calibrate the semantics of pairwise collaboration using a lightweight set of expert-labeled scenarios. This results in a weighted compatibility graph whose topology is derived from historical co-participation, but whose edge meanings are explicitly anchored in organizational judgment.

Further, Pair Compatibility (PC) estimates are systematically transformed into a size-agnostic collaboration profile that serves as evidence for an expert-informed Bayesian Network (BN). The BN encodes how decision makers reason about collaboration and technical fit under uncertainty, providing a transparent and auditable mechanism for team-project fit assessment. In this formulation, the BN functions as a principled evaluator of team quality—i.e., a validated ruler—capable of supporting downstream decision processes while preserving interpretability and traceability.

We evaluate our method in an industrial setting comprising more than 300 professionals. The regression-calibrated PC model is assessed against expert judgments elicited through synthetic collaboration scenarios and instantiated on an organizational dataset parameterized from anonymized developer–project participation records and project-level OSF/SLF indicators. The resulting compatibility graph illustrates cohesive clusters and bridge developers under the proposed modeling assumptions. We also demonstrate how PC-derived collaboration profiles can be systematically consumed by the BN evaluator, enabling a transparent rationale from organizational co-participation records and expert-validated project-level indicators to team-level fit assessment.

Our contributions are fourfold:

- We propose an expert-guided, regression-calibrated PC method that estimates the expected collaborative success of developer pairs using two project-level indicators, the OSF and the SLF, along with an expert-defined saturation effect. The semantics, normalization, and role of these indicators in compatibility modeling were reviewed and validated through multiple expert workshops.
- We construct an interpretable developer compatibility graph whose edge semantics are grounded in expert knowledge, enabling meaningful collaboration modeling in data-constrained industrial settings.
- We operationalize a systematic PC-to-BN bridge in which intra-team compatibility evidence is transformed into a size-agnostic collaboration profile that supports transparent team-project fit assessment and prepares the ground for interpretable fitness functions in optimization-based team formation.
- We demonstrate that expert-calibrated BNs can act as scenario-consistent and interpretable evaluators of team fit, achieving low probabilistic error against expert-provided reference distributions and preserving monotonic, semantically coherent behavior under unseen scenarios within the elicited domain.

This article extends our previous study on regression-calibrated developer compatibility graphs (Cunha et al., 2025a). While earlier work focused on modeling and validating pairwise collaboration via an expert-guided compatibility graph, the present study advances the proposed method by operationalizing evidence of collaboration at the team level through probabilistic reasoning. Specifically, we integrate the compatibility graph into a BN evaluator for explainable team-project fit assessment, introduce new research questions, and provide additional scenario-based validation using probabilistic consistency metrics.

The remainder of this paper is organized as follows: Section 2 discusses related work; Section 3 details the modeling and expert-guided regression process; Section 4 describes the study design and evaluation procedures; Section 5 presents and discusses the calibration results, the resulting compatibility graph, and the BN-based evaluation outcomes; Section 6 outlines implications for research and practice; Section 7 discusses threats to validity; and Section 8 concludes with final remarks and future directions.

2 Background and Related Work

This section positions the proposed method within the broader literature on Software Team Formation (STF) and developer compatibility modeling. Section 2.1 reviews the main families of existing approaches, including trait- and capability-based models, data-driven and graph-based models, search-based and heuristic methods, and neural or learning-based models. Section 2.2 synthesizes these approaches through a structured conceptual comparison, highlighting their trade-offs in terms of interpretability,

data requirements, expert involvement, team-quality criteria, and suitability for data-constrained industrial settings. Section 2.3 then discusses the technical foundations that support our method, namely compatibility graphs and BNs as expert-informed evaluators. Section 2.4 motivates our treatment of *Collaboration Fit* (AC) as a relational construct derived from pairwise collaboration evidence. Finally, Section 2.5 summarizes the research gap and clarifies how the proposed method contributes to interpretable, organization-specific team-fit evaluation.

2.1 Families of Team Formation Approaches

Early approaches modeled compatibility using personality and role-based assessments, such as psychometric inventories and preference surveys (Vishnubhotla et al., 2020; Strnad and Guid, 2010; Tuarob et al., 2021). While useful in certain contexts, these methods rely heavily on subjective and self-reported data, exhibit limited empirical linkage to observable project outcomes, and pose significant challenges to scalability and maintainability in large, dynamic organizational settings (Costa et al., 2020).

More recent approaches infer developer compatibility from historical collaboration data, typically representing developers as nodes and prior interactions as weighted edges (Cunha et al., 2025b; Kamel et al., 2011; Ruenin et al., 2024). While such models enhance transparency and reproducibility, they remain highly dependent on the availability, completeness, and expressiveness of historical interaction logs. In practice, these signals are often sparse, noisy, or misaligned with the qualitative criteria that staffing authorities use to evaluate collaboration quality. Our approach addresses these limitations by calibrating edge semantics through expert-labeled scenarios, enabling knowledge-informed graph construction that remains applicable in data-scarce and organizationally grounded settings.

Search-based and heuristic methods, including genetic algorithms, have been widely applied to optimize team formation objectives such as skill coverage, workload balance, and role allocation (Baghel and Bhavani, 2018; Cunha et al., 2022b; Costa et al., 2022; Cunha et al., 2025b). Despite their effectiveness in exploring large solution spaces, these approaches typically depend on handcrafted objective functions and fixed weighting schemes. Such formulations limit interpretability, hinder transferability across organizational contexts, and often obscure the decision rationale that practitioners require to trust and adopt the resulting recommendations (Strnad and Guid, 2010; Alves et al., 2021). More fundamentally, the objective function acts as the ruler of "team quality": if it is not a faithful proxy for the decision maker's criteria, search can optimize toward solutions that are numerically "best" under an invalid ruler, undermining both the validity of the recommendations and their evaluation. This issue is exacerbated in team formation, where historical data rarely captures all factors considered by decision makers, and past allocations cannot be assumed optimal among all feasible alternatives. Since software projects do not admit realistic execution-level simulation to assess counterfactual team assignments, purely bottom-up, data-driven fitness construction provides a weak foundation for optimization.

Finally, neural and learning-based models have been shown to capture latent patterns in skills, collaboration, and team dynamics (Hamidi Rad et al., 2023; Rad et al., 2022; Hamidi Rad et al., 2020). However, these models typically require large volumes of historical interaction data and operate as opaque predictors, offering limited interpretability. As a result, their applicability in managerial decision-making contexts is constrained, particularly in settings where transparency, traceability, and the ability to incorporate contextual or policy-driven adjustments are essential (Rique et al., 2023).

2.2 Conceptual Comparison and Positioning

The approaches discussed above occupy distinct positions in the design space of team formation methods. Personality-based models (Vishnubhotla et al., 2020; Strnad and Guid, 2010; Tuarob et al., 2021) offer explicit criteria but rely on self-reported data with limited empirical linkage to project outcomes and high maintenance cost due to temporal instability of personality constructs. Search-based methods (Costa et al., 2022; Lappas et al., 2009; Saeedi et al., 2025a) produce readable objective functions but typically adopt hand-crafted weights without structured grounding in managerial reasoning, optimizing for skill coverage rather than perceived collaboration quality. Learning-based models (Hamidi Rad et al., 2023; Ruenin et al., 2024; Rad et al., 2022) capture latent patterns at scale but operate as black-box predictors that require large volumes of interaction logs and offer limited transparency for managerial decision-making (Rique et al., 2023). Table 1 summarizes these trade-offs across five dimensions relevant to organizationally grounded, interpretable team formation.

Overall, Table 1 shows that the proposed method should not be interpreted as a replacement for all STF techniques, but as a complementary decision-support approach for settings in which interpretability, limited data availability, and alignment with local staffing authority are central requirements.

2.3 Compatibility Graphs and Probabilistic Team Evaluation

Compatibility graphs provide a natural representation for modeling collaboration in STF. Formally, a developer compatibility graph can be defined as $G = (V, E, w)$, where V is the set of vertices, each vertex $v \in V$ represents a developer, and each weighted edge $(u, v) \in E$ encodes an estimated relationship between two developers (Lappas et al., 2009; Saeedi et al., 2025b). In graph-based team-formation research, such edges commonly represent prior collaboration, communication cost, co-participation, social proximity, or interaction strength, while candidate teams are modeled as subgraphs satisfying project or skill constraints (Saeedi et al., 2025b). Adjacent work has also used compatibility-aware interaction graphs, where edge weights encode coordination cost, inverse compatibility, or historical synergy among candidate team members (Li et al., 2026).

These representations expose relational structure: strongly connected subgroups, isolated candidates, bridging mem-

Table 1. Conceptual comparison of team formation approaches across five dimensions. Taxonomy grounded in Saeedi et al. (2025a) and Costa et al. (2020).

Approach	Interpretability	Data requirements	Expert involvement	Team quality criterion	Industrial fit (scarce data)
Personality-based models (Vishnubhotla et al., 2020; Strnad and Guid, 2010; Tuarob et al., 2021)	Criteria are explicit but rely on self-reported, subjective input with limited empirical linkage to project outcomes	Surveys are cheap to run once but traits and perceptions change over time, requiring periodic reassessment to remain valid	A specialist may know 10–20 developers well; assessing hundreds requires multiple evaluators with inconsistent criteria, or falls back to self-report	Profile compatibility; does not directly model whether the team will deliver a specific project	Self-report bias, temporal instability, and scaling cost may limit its practicality for ongoing staffing decisions
Search-based (graph + heuristic) (Costa et al., 2022; Lappas et al., 2009; Saeedi et al., 2025a)	Objective function is readable, but weights are hand-crafted without explicit grounding in managerial reasoning	Requires co-participation logs or skill profiles; more stable than surveys but must cover enough interactions to build a meaningful graph	Weights in the objective function are typically set ad hoc; expert judgment does not enter the model in a structured or auditable way	Skill coverage; if the objective function is misaligned, optimization produces technically valid but practically poor teams	Feasible when skill data is available, but practical value depends entirely on whether the objective function reflects real organizational priorities
Machine-learning and neural models (Hamidi Rad et al., 2023; Ruenin et al., 2024; Rad et al., 2022)	Black-box predictors; embeddings cannot be inspected or explained to a manager making a staffing decision	Requires large-scale interaction logs; most mid-sized organizations do not have data at the required density or granularity	Mostly data-driven; the model learns patterns present in the data, which may not fully reflect explicit organizational values	Latent data patterns; cannot explain what it is optimizing for in terms a manager can evaluate or override	Requires data volume and granularity that most industrial organizations do not have; opacity further limits managerial trust
Proposed method (PC + graph + BN) [this work]	Regression coefficients are directly interpretable; BN inference path from pairwise evidence to team fit is fully auditable and traceable	Multiple expert workshops, eight synthetic expert-labeled scenarios, anonymized developer–project participation records, and project-level OSF/SLF indicators used to parameterize the organizational analysis base	CTO as modeling anchor with real decision authority and 25+ years of experience; expert knowledge enters in a structured, documented, and recalibratable form	Team-project fit; explicitly models how the decision maker reasons about team suitability, with an auditable and adjustable criterion aligned with actual staffing authority	Designed for data-constrained settings; experimental instantiation combines expert-labeled scenarios with organizational participation records and project-level indicators, whereas deployment in a new context requires collecting or deriving comparable OSF/SLF-like indicators and recalibrating the model locally

bers, and potential cohesion patterns can be inspected directly or used by downstream algorithms. However, the practical value of a compatibility graph depends on the semantics of its edge weights. In many existing approaches, edge weights are determined by the traces available in the data, such as co-authorship, co-participation frequency, communication frequency, or graph distance. Although these signals are reproducible, they do not necessarily capture how staffing authorities assess collaboration quality in practice. A higher edge weight may indicate more frequent interaction, but not necessarily better collaboration or higher suitability for a future project. This distinction is central to our work: we treat the compatibility graph not merely as a structural artifact, but as an expert-calibrated evidence layer in which each edge represents an estimated PC value grounded in organizational judgment.

Compatibility graphs alone, however, do not fully solve the team-evaluation problem. When used as standalone analytical artifacts or direct optimization targets, they tend to emphasize pairwise or structural properties without explicitly representing uncertainty or explaining how collaboration evidence combines with other decision factors. For staffing decisions, the relevant question is not only whether developers have strong pairwise ties, but whether a candidate team is suitable for a specific project under both technical and collaborative criteria. This requires a team-level evaluator capable of integrating heterogeneous evidence in a transparent and auditable way.

BNs provide such a mechanism since they are well-suited for modeling decision processes under uncertainty and for encoding expert knowledge in domains where data are incomplete or interpretability is essential (Mendes, 2014; Fenton et al., 2007). In software engineering, prior work has used BNs to model and reason about project, process, and teamwork-related constructs under uncertainty (Freire et al., 2018, 2022). In our method, the BN plays the role of an expert-informed evaluator: it receives technical evidence and collaboration evidence, propagates uncertainty through an explicit dependency structure, and produces a probabilistic assessment of Team Fit.

This design also connects to a broader methodological pattern beyond software engineering and team formation. In knowledge-driven optimization and decision support, expert or designer knowledge can be compiled into explicit evaluators that guide search, ranking, or recommendation when direct ground-truth quality labels are unavailable, incomplete, or insufficient. For example, Setayandeh and Babaei (Setayandeh and Babaei, 2020) use fuzzy preference functions derived from designer experience to transform a constrained multiobjective aircraft-design problem into a single-objective function optimized by a genetic algorithm. Although the domain and modeling formalism differ from ours, the abstraction is relevant: expert knowledge is translated into an auditable mathematical evaluator used by an optimization pipeline. Our work follows this knowledge-driven rationale in the STF domain, but instantiates the evaluator as an expert-elicited BN rather than a fuzzy preference system, and uses compatibility-graph evidence as input to probabilistic team-fit inference.

This offline, expert-compiled evaluator design differs from several related model-based optimization paradigms. In estimation-of-distribution algorithms and probabilistic model-building genetic algorithms, probabilistic models are typically learned from promising candidate solutions and then used to generate new candidates (Pelikan et al., 2002). In surrogate-assisted evolutionary optimization, models are learned from previously evaluated solutions to approximate an expensive objective function (Brownlee et al., 2013). In interactive and preference-based Search-Based Software Engineering, decision-maker preferences are often incorporated through repeated feedback or scoring during the search process (Ramírez et al., 2019; Ferreira et al., 2017). In contrast, our evaluator is elicited once, offline, from the staffing authority's decision reasoning and remains fixed during team evaluation. Our prior work instantiates this pattern in a search-based setting by using an expert-elicited BN approval model as the fitness function for genetic algorithm-based STF (Cunha et al., 2026); the present paper focuses on the construction and validation of the underlying compatibility-evidence and probabilistic evaluation pipeline.

A key requirement for integrating compatibility graphs with BN-based evaluation is transforming pairwise collaboration evidence into team-level evidence without losing the relational structure of the team. Rather than collapsing collaboration into a single mean score, our method maps PC values into an ordinal collaboration profile that captures the distribution of compatibility tiers within a candidate team. This profile is then transformed into AC evidence and injected into the BN together with technical evidence, enabling transparent inference of overall Team Fit.

Because the resulting variables are ordinal, the Ranked Nodes Method (RNM) provides a suitable mechanism for constructing Conditional Probability Tables (CPTs) from expert knowledge while preserving monotonic behavior (Fenton et al., 2007). This makes BNs a natural complement to compatibility graphs: the graph captures relational collaboration evidence, while the BN provides an auditable probabilistic evaluator that combines collaboration quality with technical coverage to support team-project fit assessment.

2.4 Collaboration Fit as a Relational Construct

AC captures dimensions of teamwork quality (Hoegl and Gemuenden, 2001), emphasizing behavioral and interpersonal dynamics that emerge from interactions among team members. Instead of modeling this construct at the individual level through soft skills or personality traits, we adopt a relational strategy grounded in historical team performance. Empirical evidence has questioned the predictive validity of personality- and soft-skill-based models in explaining team outcomes (Vishnubhotla et al., 2020; Vishnubhotla and Mendes, 2024), reinforcing the need for relational, performance-derived proxies.

Performance records from past projects are routinely collected by mature organizations, providing a pragmatic and readily available data source. Leveraging this availability, we construct a pairwise compatibility graph in which nodes represent individuals and edge weights encode historical collaboration outcomes, such as project productivity and indicators of teamwork quality (Hoegl and Gemuenden, 2001; Marsicano et al., 2020). These pairwise signals are subsequently aggregated into an ordinal collaboration profile used to derive AC evidence, which is then injected into the BN for Team Fit inference.

2.5 Research Gap and Contribution

Despite substantial advances, existing solutions exhibit complementary but unresolved limitations. Trait-based and learning-based models struggle with subjectivity, data requirements, and interpretability; graph-based and data-driven models depend heavily on the availability and semantic adequacy of historical logs; and search-based methods rely on opaque objective functions that rarely reflect organizational decision policies. Moreover, prior work largely treats compatibility graphs as static artifacts or direct optimization targets, leaving unaddressed the question of how pairwise collaboration evidence should be systematically transformed into explainable, team-level decision support under uncertainty.

To address this gap, we propose an expert-guided compatibility modeling method that (i) grounds pairwise collaboration semantics in expert judgment, (ii) preserves the relational structure of collaboration through an ordinal, size-agnostic profile, and (iii) integrates this evidence into a BN evaluator for transparent and auditable Team Fit inference. This hybrid method unifies empirical signals, expert reasoning, and probabilistic decision modeling, advancing the state of the art in interpretable and organizationally aligned STF.

3 Proposed Method

This section introduces an expert-guided, regression-calibrated method for Bayesian team-fit evaluation, organized as a six-stage workflow (S1–S6). The method is designed to support staffing and STF decisions in settings where collaboration logs are sparse or heterogeneous, and where decision makers require transparent, auditable

reasoning rather than black-box predictions.

The workflow starts by defining the scope of the decision problem and the variables of interest through expert-guided modeling (S1), followed by the construction and discretization of labeled collaboration scenarios that anchor the semantics of compatibility (S2). These scenarios are then used to calibrate a pairwise collaboration estimator via regression (S3), producing an interpretable model of expected collaboration quality between pairs of developers.

In the next stages, the calibrated pairwise model is applied to historical collaboration records to estimate expected collaboration potential for all relevant developer pairs (S4). These estimates are used to construct a weighted compatibility graph (S5), in which nodes represent developers and edge weights encode expected collaboration quality under realistic organizational conditions.

Finally, graph-derived pairwise evidence is transformed into a size-agnostic collaboration profile and provided as structured input to a BN (S6). The BN performs team-level probabilistic inference, integrating collaborative evidence with other decision factors to support transparent and auditable team-fit evaluation. Importantly, the method does not aim to predict short-term project outcomes, but to operationalize expert reasoning about collaboration and team suitability in a probabilistic and reusable form.

Figure 1 summarizes this six-stage process. Each stage is detailed in the following subsections, which describe the modeling assumptions, data transformations, and validation procedures underpinning the proposed method.

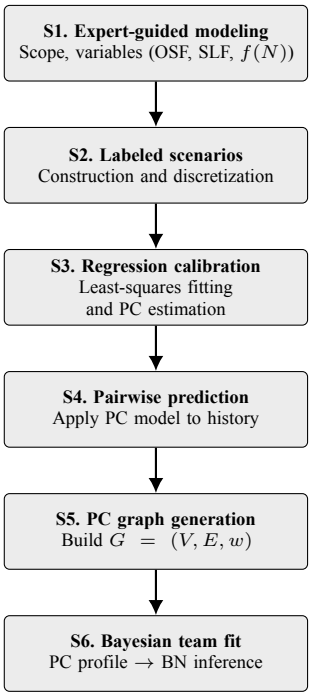

Figure 1. Method workflow (S1-S6) for regression-calibrated PC modeling and BN-based Team Fit evaluation.

S1: Expert-Guided Modeling and Scope

Stage S1 defines the modeling scope and the variables used to estimate the expected pairwise compatibility between two developers d_i and d_j, denoted $PC(d_i, d_j)$. Based on expert elicitation, PC is assumed to depend on three components derived from past joint projects:

- **OSF**: an outcome-oriented indicator reflecting external project success;
- **SLF**: an internal indicator capturing perceived leadership, communication, and coordination quality;
- **Collaboration-history saturation**: a function of the number of projects previously shared by the pair.

The selection and interpretation of OSF and SLF were examined during multiple workshops with the elicitation expert. During these workshops, the expert assessed the organizational relevance of both indicators, validated their intended interpretations, and confirmed their normalization to the $[0, 1]$ interval for compatibility modeling. OSF was defined as an outcome-oriented indicator of external project success, whereas SLF was defined as an internal indicator of collaborative effectiveness, particularly regarding leadership, communication, and coordination. While OSF reflects delivery success from an external perspective, SLF captures internal perceptions of collaborative effectiveness. The expert emphasized that both dimensions should contribute comparably to compatibility assessment. The concrete organizational operationalization of OSF and SLF is described in Section 4.2.

Saturation Assumption. A key modeling assumption is that repeated collaboration exhibits diminishing returns. Up to a certain point, additional shared projects increase familiarity and efficiency; beyond that point, excessive repetition may reduce learning, diversity, or adaptability. This effect is operationalized through a saturation function $f(N)$, where N denotes the number of projects previously shared by the pair:

$$f(N) = \begin{cases} \frac{N}{4}, & \text{if } N \leq 4, \\ \max(0.0, 1 - 0.1 \cdot (N - 4)), & \text{if } N > 4. \end{cases} \quad (1)$$

Figure 2 illustrates the behavior of the saturation function $f(N)$ across different values of N. In the growing phase ($N \leq 4$), each additional shared project increases the collaboration benefit linearly, reaching its peak at $f(4) = 1.0$. For instance, a pair that has collaborated in two projects yields $f(2) = 0.50$, while a pair at the saturation point yields $f(4) = 1.00$. Beyond this threshold, the function captures diminishing returns: a pair with five shared projects yields $f(5) = 0.90$, one with six yields $f(6) = 0.80$, and a pair with fourteen shared projects yields $f(14) = 0.00$. This design reflects the expert's observation that, in the studied organization, collaboration benefits peak around four shared projects and then gradually decline due to reduced learning, lower diversity exposure, and increasing collaboration fatigue. The inflection point $N = 4$ is a context-dependent hyperparameter that should be re-elicited when applying the method to organizations with different project durations, team rotation policies, or staffing cultures.

It is important to note that $f(N) = 1.0$ represents the saturation ceiling of the collaboration-history component, not a

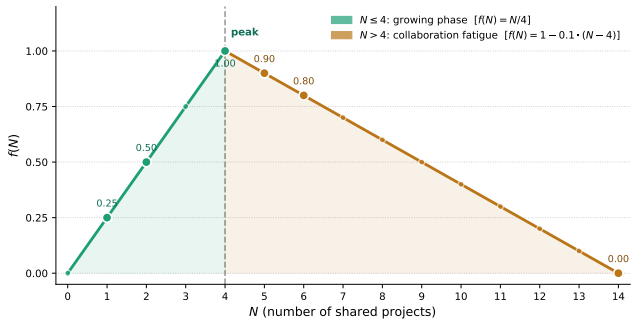

Figure 2. Saturation function $f(N)$: growing phase ($N \leq 4$) and collaboration fatigue phase ($N > 4$). Key values are annotated directly on the curve.

ceiling on PC itself. Since $f(N)$ enters the model as a multiplicative factor over the OSF/SLF-based score (Eq. (3)), two pairs at the saturation point ($N = 4$, $f(N) = 1.0$) can still yield different PC values depending on their OSF and SLF values. For instance, in the organizational analysis base used in this study, the pair with the highest computed compatibility ($PC = 0.581$, $OSF = 0.775$, $SLF = 0.662$) and another pair with lower computed compatibility ($PC = 0.550$, $OSF = 0.645$, $SLF = 0.696$) both reached the saturation point, yet their PC values differ because their OSF/SLF profiles differ. Thus, $f(N) = 1.0$ removes the saturation penalty but does not by itself guarantee a high PC score: compatibility remains dependent on the OSF and SLF values associated with the projects shared by the pair.

S2: Construction and Discretization of Scenarios

Following the expert validation of the indicators and their normalized scales in S1, Stage S2 constructs a compact calibration set of eight synthetic but realistic scenarios spanning different combinations of OSF and SLF. Both indicators are discretized into five ordinal levels: Very Low (VL), Low (L), Medium (M), High (H), and Very High (VH).

For each scenario (i.e., each unique combination of OSF and SLF), the expert was asked to estimate the likelihood of each of the five possible levels of PC. Responses were provided using a Likert-type scale with both textual and numeric mappings, using the verbal scale provided by Renooij and Witteman (1999) (e.g., "Probable" → 85%).

These responses were transformed into probability distributions over the five PC states. Each row in the resulting matrix was normalized to ensure that the probabilities summed to 1. For each scenario, we computed:

- **Mode**: showing the central tendency of the resulting probability distribution;
- **Heatmaps**: showing distribution across states;
- **Mode extraction**: highlighting the most likely outcome.

This information helped the expert visualize and validate the elicited distributions (e.g., via their modal tier). Later, we used the expert-elicited expected value as the reference label for the regression analysis. To this end, each scenario row was transformed into a tuple (OSF, SLF, PC), where PC denotes the expected value computed from the expert's tier distribution. We mapped the ordinal values for OSF and SLF into numerical values following the rules in Table 2. The collected data and the analysis procedures are provided in the Supplementary Material.

Table 2. Ordinal scale to numerical scale for OSF and SLF.

Verbal Expression	Numerical Equivalent
Very Low	0.1
Low	0.3
Medium	0.5
High	0.7
Very High	0.9

Having defined labeled calibration points in S2, Stage S3 fits an interpretable regression model that operationalizes the expert's judgments as a reusable compatibility function.

S3: Regression Calibration and PC Estimation

Stage S3 calibrates a regression model that estimates expected pairwise compatibility from OSF, SLF, and collaboration-history saturation. The compatibility between developers d_i and d_j is modeled as:

$$PC_{\text{base}}(d_i, d_j) = \beta_0 + \alpha \cdot SLF_{ij} + \beta \cdot OSF_{ij} + \gamma \cdot (SLF_{ij} \cdot OSF_{ij}), \quad (2)$$

$$PC(d_i, d_j) = PC_{\text{base}}(d_i, d_j) \cdot f(N_{ij}), \quad (3)$$

where β_0 denotes the intercept (the baseline compatibility when $SLF = OSF = 0$, before applying the saturation adjustment), α and β weight the leadership-based (SLF) and outcome-based (OSF) signals, respectively, and γ captures a synergistic interaction between the two when both are simultaneously high. PC_{base} is estimated via Ordinary Least Squares (OLS) on the eight expert-labeled calibration scenarios (S2), which by construction have no collaboration-history dimension; the saturation function $f(N_{ij})$ (Eq. (1)) is therefore applied afterward, as a multiplicative adjustment over PC_{base}, rather than jointly estimated as a regression term. This formulation encodes the expert's assumption that compatibility increases with successful outcomes, effective leadership, and their synergy, while insufficient or excessive repeated collaboration should modulate the resulting compatibility estimate.

The multiplicative form was adopted because it matches the intended semantics of the saturation function. When $f(N_{ij}) = 1.0$, the OSF/SLF-based compatibility estimate remains unchanged; when $f(N_{ij}) < 1.0$, compatibility is proportionally discounted. An additive formulation would not preserve this interpretation, because the collaboration-history term could increase PC even when the OSF/SLF profile is weak. Thus, $f(N_{ij})$ acts as a saturation adjustment over the calibrated compatibility estimate, rather than as an independent source of compatibility.

The linear specification for PC_{base} was intentionally adopted to preserve interpretability. Each coefficient corresponds directly to a modeling assumption that can be inspected, discussed, and revised as organizational priorities evolve. Rather than replacing expert judgment, the regression operationalizes it into a parametric form that enables consistent and scalable application.

Model calibration was performed using OLS, implemented with `statsmodels` in Python, with the eight expert-labeled scenarios serving as the training dataset. The resulting coefficients define the calibrated PC model, which is subsequently applied to developer pairs for which OSF, SLF, and N can be instantiated. Calibration fidelity is assessed by comparing expert-expected compatibility values ($PC_{expected}$) with model-predicted values ($PC_{predicted}$) using both the Mean Absolute Error (MAE) and the Pearson correlation coefficient (r). While MAE captures the average magnitude of deviation, the correlation coefficient quantifies the directional and linear agreement between expert judgment and model output, providing a complementary view of calibration quality.

Once calibrated, the model is applied at scale in Stage S4 to compute $\widehat{PC}_{ij}$ for valid developer pairs and to support the construction of the compatibility graph.

S4: PC Estimation on the Organizational Analysis Base

Stage S4 applies the calibrated compatibility model to developer pairs derived from the organizational project-participation structure. For each pair (d_i, d_j), we:

1. Identify the projects shared by the pair from the anonymized organizational records;
2. Compute the average project-level OSF and SLF values associated with those shared projects;
3. Count the number of shared projects N_{ij};
4. Compute $f(N_{ij})$ using Eq. (1);
5. Apply Eq. (3) to estimate $PC(d_i, d_j)$.

This process produced computed $PC(d_i, d_j)$ values for all valid developer pairs represented in the organizational analysis base, forming the basis for the compatibility graph analyzed in Stage S5.

Team Success Factor. Given a team T, we compute the Team Success Factor (SF) by aggregating all pair compatibilities within the team:

$$SF_{\text{team}}(T) = \sum_{i<j,\, d_i, d_j \in T} PC(d_i, d_j) \tag{4}$$

Although Eq. 4 provides a compact aggregate view of intra-team compatibility, it does not preserve how compatibility is distributed across pairs. Two teams may present similar aggregate scores while exhibiting very different internal cohesion patterns. For this reason, in S6, we operationalize a size-agnostic collaboration profile by computing the proportion of pairs in each PC tier (VL–VH) and transforming this profile into an AC evidence value through an expert-calibrated aggregation rule. Before that, S5 turns all predicted compatibilities into a graph artifact that supports interpretability and downstream analysis.

S5: Generation of a PC Graph

Stage S5 transforms the predicted pair compatibilities $\widehat{PC}_{ij}$ into a weighted undirected graph $G = (V, E, w)$:

- **Nodes (V):** developers, annotated with identifiers and attributes;
- **Edges (E):** valid pairs (i, j) retained when $PC > 0.3$, each storing:
  - the numeric compatibility weight $\widehat{PC}_{ij}$;
  - the discretized PC tier (VL–VH);
  - the historical collaboration count N_{ij}.

The graph structure was serialized in JavaScript Object Notation format and visualized using a force-directed layout. The resulting PC graph provides an interpretable structural view of the organization's collaboration potential. It enables identification of strongly connected pairs, cohesive subgroups, and bridging developers, and it serves as an intermediate artifact between pairwise estimation and team-level evaluation. In S6, the same graph also serves as the computational substrate for extracting team-level collaboration evidence in a size-agnostic manner.

S6: Bayesian Modeling of Team Fit Using Compatibility-Graph Inputs

Stage S6 integrates graph-derived compatibility evidence into an expert-guided BN to evaluate team–project fit. The BN summarizes team evaluation through three ordinal constructs on a five-level scale (VL, L, M, H, VH): Technical Fit (AT), AC, and Team Fit (AE). The BN structure is shown in Figure 3. It encodes how technical evidence is aggregated into AT and how AT and AC jointly determine AE, enabling an auditable inference path from PC-derived collaboration evidence and technical coverage to the likelihood of final approval.

Instead of directly eliciting numerical probabilities, we used a verbal-numerical scale to reduce cognitive burden and bias (Renooij and Witteman, 1999). For the ranked target nodes AT and AE, the expert evaluated representative parent configurations and assigned likelihoods to ordinal outcomes. These elicited values were normalized to form valid probability distributions, and the corresponding CPTs were calibrated using the RNM, enabling coherent monotonic behavior and supporting alternative aggregation functions: weighted mean (WMEAN), weighted minimum (WMIN), weighted maximum (WMAX), and mixed minimum–maximum (MIXMINMAX) (Fenton et al., 2007). In the BN structure, AC is computed externally from the intra-team PC-tier profile and injected into the BN as collaboration evidence.

PC-to-AC Evidence Construction. AC depends on how well all members of a team get along with each other, not just one pair. It cannot, therefore, be represented by a single PC value. We instead transform the full set of pairwise PC values within a candidate team into a single, size-agnostic AC evidence value for the BN, following the formulation used in the BN–Genetic Algorithm (GA) pipeline. This is done in four steps.

First, for a team with n developers, all $\binom{n}{2}$ pairwise PC scores are classified into five tiers: VL $[0, 0.2)$, L $[0.2, 0.4)$, M $[0.4, 0.6)$, H $[0.6, 0.8)$, and VH $[0.8, 1.0]$.

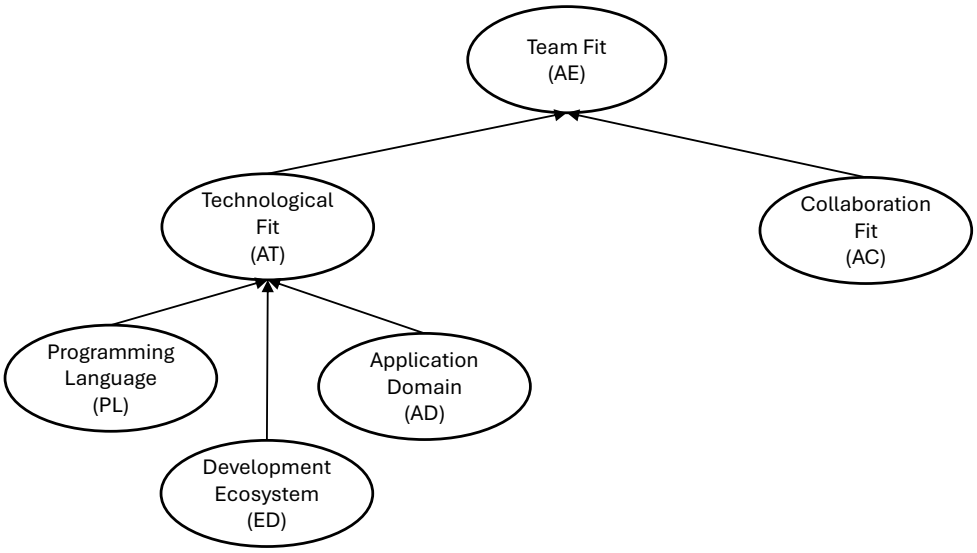

Figure 3. BN structure for estimating team-project fit.

Second, we compute the proportion of pairs in each tier $(p_{VL}, p_L, p_M, p_H, p_{VH})$, giving a profile of the team's collaboration quality instead of a single averaged number.

Third, this profile is combined into one collaboration index m using expert-calibrated tier weights $(1, 1, 4, 4, 5)$:

$$m = \sum_{i \in \{VL,L,M,H,VH\}} w_i \cdot p_i.$$

These weights come from 30 independent GA-based RNM calibration runs on expert-labeled scenarios (originally used to parameterize the AC component of the full BN) and are reused here as a compact, external PC-to-AC transformer, preserving the same calibrated semantics without the extra BN nodes.

Fourth, m is converted into the AC evidence value. Since the weights are applied to a distribution (the p_i sum to 1), m always falls between the smallest weight (1) and the largest (5), i.e., $m \in [1, 5]$. We rescale this to the $[0.1, 0.9]$ scale already used for the BN states (Table 2):

$$pc\_score = 0.1 + 0.2 \cdot (m - 1),$$

where $(m - 1)$ is how far m has moved from its minimum (1), and 0.2 converts each unit of m into the matching step on the $[0.1, 0.9]$ scale (0.8 range over 4 units of m). The nearest centroid among $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ then gives the final AC label (VL, L, M, H, or VH).

The resulting AC value is injected into the BN as collaboration evidence. Technical evidence is injected through the domain, ecosystem, and language nodes to infer AT; AT and AC are then combined to infer AE.

Illustrative Example. Consider a team composed of three developers, A, B, and C, with pairwise compatibility values $PC(A, B) = 0.7$, $PC(A, C) = 0.6$, and $PC(B, C) = 0.5$. The first two values fall into the H tier, while the last one falls into the M tier. The resulting profile is therefore $p_H = 2/3$, $p_M = 1/3$, and $p_{VL} = p_L = p_{VH} = 0$.

Using the calibrated weights $(1, 1, 4, 4, 5)$, the collaboration index is:

$$m = 4 \cdot \frac{2}{3} + 4 \cdot \frac{1}{3} = 4.0.$$

The mapped score is:

$$pc\_score = 0.1 + 0.2 \cdot (4.0 - 1) = 0.7.$$

Thus, the team receives AC = H, which is then used as evidence in the BN. This example shows how the method preserves the internal distribution of pairwise compatibility while producing a compact and interpretable collaboration signal for team-level probabilistic inference.

Team-Level Interpretation. Table 3 compares two hypothetical candidate teams whose three developer pairs have different individual PC values but the same arithmetic mean (0.50). Team 1 has a homogeneous profile: all three pairs fall in the M tier, giving the tier profile $(p_{VL}, p_L, p_M, p_H, p_{VH}) = (0, 0, 1, 0, 0)$. Team 2 has a heterogeneous profile: one pair falls in L, one in M, and one in VH, giving $(0, \frac{1}{3}, \frac{1}{3}, 0, \frac{1}{3})$. The mean-based encoding (*C2-Mean* in Section 4.3) maps the arithmetic mean of the pairwise PC values, not m, to the nearest centroid, and would therefore collapse both teams to the same AC evidence, since both means are identical. The calibrated tier-weighted transformer (*C1-Weighted*) instead applies the calibrated weights $(1, 1, 4, 4, 5)$ directly to each team's tier profile, not to the raw PC values, producing different weighted sums m (4.00 for Team 1 versus 3.33 for Team 2) and, consequently, different AC evidence: H for Team 1 versus M for Team 2.

This limitation of mean-based encoding is not incidental but structural: the arithmetic mean discards information by construction, so any two teams whose pairwise PC values share the same mean are indistinguishable to a method that consumes only that mean, regardless of how different their underlying collaboration profiles are. Team 1 and Team 2 in Table 3 are constructed to share an identical mean (0.50)

Table 3. Two candidate teams with identical arithmetic mean PC (0.500) but different tier compositions, and the resulting AC evidence under the calibrated weighted-tier transformer. Tier weights $(w_{VL}, w_L, w_M, w_H, w_{VH}) = (1, 1, 4, 4, 5)$ apply to both teams.

	Team 1	**Team 2**
Pairwise PC values	0.45, 0.50, 0.55	0.30, 0.40, 0.80
Tier profile $(p_{VL}, p_L, p_M, p_H, p_{VH})$	$(0, 0, 1, 0, 0)$	$(0, \frac{1}{3}, \frac{1}{3}, 0, \frac{1}{3})$
Weighted sum $m = \sum_i w_i p_i$	$4 \cdot 1 = 4.00$	$1 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 5 \cdot \frac{1}{3} \approx 3.33$
Mean PC (arithmetic, for comparison)	0.500	0.500
$pc\_score = 0.1 + 0.2(m - 1)$	0.700	0.567
AC evidence	**H**	**M**

while differing in tier composition, guaranteeing that *C2-Mean* assigns them the same AC evidence. Team 2 contains a poorly compatible pair (L tier) that this mean-based score fully masks, since it is offset by the strong pair (VH tier) in the same average. Because the tier-weighted transformer operates on the tier *proportions* rather than on the raw PC values, and the lowest tiers (VL, L) receive weight 1 while M, H, and VH receive weights 4, 4, and 5 respectively, a team with an unevenly distributed profile is penalized relative to a team whose pairs are uniformly moderate, even when both teams share the same mean. In practice, a manager reviewing Team 2 would see the lower AC = M evidence as a signal to inspect the compatibility graph and identify which specific pair is weighing the team down, before finalizing its composition.

Current BN Configuration. The compact BN contains three input-level technical dimensions (domain, ecosystem, and language), the collaboration evidence AC, and two ranked target nodes: AT and AE. The current calibrated configuration is summarized in Table 4. AT is computed from domain, ecosystem, and language using WMEAN with weights $(3, 1, 5)$, reflecting the higher importance assigned to programming language. AE is computed from AT and AC using MIXMINMAX, with weight 5 assigned to the minimum component and weight 1 assigned to the maximum component, reflecting a conservative approval policy in which weak evidence should strongly constrain AE while still allowing strong evidence to contribute.

Table 4. Current PC-to-AC transformer and BN configuration.

Component	**Function**	**Parameters / Use**
PC-to-AC transformer	Weighted tier aggregation	PC_VL=1, PC_L=1, PC_M=4, PC_H=4, PC_VH=5; selected from 30 independent GA-based RNM calibration runs originally used to parameterize the AC component and now reused to map intra-team PC-tier distributions into AC evidence
AT	WMEAN	Dom=3, Eco=1, Ling=5; $\sigma = 0.05$
AE	MIXMINMAX	minimum component=5, maximum component=1; $\sigma = 0.05$

Stage S6: PC-to-BN Integration Pipeline. The final component of Stage S6 operationalizes how the compatibility graph supports team-level evaluation. The integration pipeline comprises four steps:

1. **Extraction of intra-team PC values**. For each candidate team T, all internal developer pairs (d_i, d_j) are enumerated and their continuous PC values are recovered from the compatibility graph.
2. **Construction of the PC-tier profile**. Each pairwise PC value is mapped to an ordinal tier (VL, L, M, H, VH), and the proportion of pairs in each tier is computed.
3. **PC-to-AC transformation**. The PC-tier profile is converted into a single AC evidence value using the calibrated weights $(1, 1, 4, 4, 5)$ and mapped to the nearest BN state centroid.
4. **BN inference of Team Fit**. The AC evidence and the technical evidence are injected into the compact BN. Probabilistic inference is then performed to estimate the posterior distribution of AE.

Overall, the proposed method provides a coherent pipeline from expert-grounded pairwise compatibility estimation to probabilistic team-fit inference, producing both interpretable intermediate artifacts (a calibrated PC function, a compatibility graph, and a PC-to-AC profile) and actionable team-level outputs (BN posterior distributions over AE).

4 Study Design

This section details the empirical study design used to evaluate the proposed compatibility modeling and BN-based team-evaluation pipeline. The focus is not on reiterating how the model is constructed, but on making explicit how its outputs are assessed under realistic decision-support requirements: (i) agreement with expert judgment at the pairwise level, (ii) adequacy of the chosen representation of intra-team collaboration for BN inference, and (iii) generalization to unseen but plausible collaboration and team-fit situations. Accordingly, we specify the evaluation goals and research questions (RQs), characterize the industrial context and empirical basis, and describe the scenario-based validation strategy and metrics used throughout the study.

4.1 Evaluation Goals

The primary goal of this evaluation is to assess whether the outputs produced by the compatibility graph can be operationalized as reliable, interpretable, and useful evidence of

collaboration within a BN for the team evaluation. Specifically, we investigate whether the proposed pipeline (i) faithfully reproduces expert judgments at the pairwise level, (ii) captures intra-team collaboration more effectively when compatibility is modeled as a distribution rather than as a single aggregate value, and (iii) generalizes to previously unseen but plausible collaboration scenarios.

These goals are investigated through the following RQs:

- **RQ1 (PC modeling).** To what extent does the expert-calibrated regression model reproduce the expected PC judgments under OSF, SLF, and collaboration-history saturation?
- **RQ2 (Intra-team collaboration representation).** Does the calibrated PC-tier weighted aggregation transformer rank intra-team collaboration quality in closer agreement with expert-elicited AC judgments than a mean-PC baseline, as measured by Spearman rank correlation with the expert's expected AC value?
- **RQ3 (BN predictive accuracy).** To what extent does the compact BN, with expert-calibrated CPTs, generalize to previously unseen scenarios by accurately predicting AE from AT and AC evidence, as quantified by the Brier score?

Together, these RQs assess complementary aspects of the proposed method. RQ1 focuses on the fidelity of the regression-calibrated PC model at the pairwise level, evaluating whether expert expectations regarding collaboration can be reliably approximated from OSF, SLF, and collaboration-history saturation. RQ2 examines the adequacy of different representations of intra-team collaboration, testing whether preserving the internal distribution of pairwise compatibilities provides a more faithful and informative signal for team-level evaluation than a single aggregate summary. Finally, RQ3 evaluates the generalization capability of the expert-informed BN, assessing whether the calibrated probabilistic model can produce accurate and stable predictions of AE from AT and AC evidence for previously unseen but plausible scenarios. Collectively, these questions establish whether the proposed pipeline is not only accurate at the pair level, but also interpretable and operational at the team level. Because all three RQs are evaluated against expert-provided reference scenarios rather than independent, real-world outcome data, they should be read as assessing decision support. Section 4.3 discusses this scope and the associated circularity risk in detail.

4.2 Industrial Context and Empirical Basis

The study was conducted at a Brazilian university-affiliated research and innovation center in information and communication technology, employing over 300 professionals and delivering more than 30 projects per year across domains such as embedded systems, artificial intelligence, web/mobile development, hardware, and cybersecurity. Projects typically range from 3 months to more than 3 years in duration (average of 22 months) and involve teams of 5 to 17 members (average of 9.2). Within this environment, 45% of developers participate in a single project, while 27% appear in four or more projects, indicating a recurring core of professionals with repeated collaboration patterns. The center serves clients ranging from startups to large enterprises, is funded through a combination of public grants and industry contracts, and adopts mature agile practices (Scrum/Kanban) supported by internally developed project management tools, despite not holding formal process certifications such as Capability Maturity Model Integration (CMMI) or ISO 9001.

Historical project data were retrieved from the organization's internal management system and constitute the empirical basis of this study. Two collaboration-related indicators were extracted and normalized to the $[0, 1]$ range: (i) OSF, derived from project-level NPS evaluations collected from the client representative at project closure (Reichheld, 2003); and (ii) SLF, derived from internal post-project evaluations completed by the project manager responsible for the project. The SLF evaluation captures leadership, communication, and coordination effectiveness. These dimensions are aligned with teamwork constructs discussed by Marsicano et al. (2020), although SLF corresponds to the organization's internal evaluation process rather than to a direct administration of the full TPA questionnaire. Operationally, the project manager assigns an SLF value in the $[0, 1]$ range by considering these teamwork dimensions. The resulting dataset comprises records from 47 completed projects conducted between 2020 and 2024, totaling over 300 developer–developer collaboration records. Prior to analysis, all data were anonymized to protect individual identities and to ensure that the evaluation focuses on the behavior of the proposed method rather than on specific developers.

For Stages S4 and S5 (Section 3), the calibrated PC model was applied to an organizational analysis base of 192 developers constructed to reflect the project-participation structure, OSF/SLF ranges, and collaboration patterns observed in the organization, rather than to the organization's full production dataset. Accordingly, the resulting compatibility graph should be interpreted as an organizationally grounded demonstration of the method's behavior, not as a direct empirical reconstruction of the organization's complete collaboration network. This design choice is discussed further in Section 7.

Expert knowledge played a central role throughout the study. The elicitation expert was the center's Chief Technology Officer (CTO), who is responsible for staffing and team allocation decisions across all projects. The expert has more than 25 years of experience in corporate and research-oriented software development, with extensive expertise in project delivery and personnel management. His contributions encompassed problem formulation, variable selection, definition of modeling assumptions (including the collaboration saturation effect), and validation activities. This expert-in-the-loop involvement is consistent with Design Science Research principles, ensuring that the proposed method is grounded in practical decision-making and aligned with organizational realities (Wieringa, 2014).

The single-expert design was therefore a deliberate organization-specific modeling choice. The goal was not to infer a consensus view of team formation across multiple stakeholders, but to formalize the decision policy of the person who actually holds staffing authority in the studied or-

ganization. This design improves contextual validity for the target setting, but it also means that the calibrated PC model, PC-to-AC transformer, and BN evaluator may encode the expert's own assumptions and biases. We therefore do not treat the resulting model as a universal model of STF. Instead, we view it as an auditable representation of local decision logic: by externalizing tacit judgment into explicit variables, functions, and probabilistic dependencies, the method enables the organization to inspect, challenge, and recalibrate its own decision process. This view is aligned with prior work showing that BN-based modeling of expert knowledge can support reflection on measurement interpretation (Saraiva et al., 2020), and with context-driven software engineering research, which emphasizes industrially grounded problems and concrete organizational needs over context-independent generalization (Briand et al., 2017).

4.3 Evaluation Procedures and Metrics

Given the absence of ground-truth measurements of future collaboration quality, the evaluation adopts a scenario-based strategy grounded in expert judgment. This approach is applied consistently across all three RQs, using expert-defined reference scenarios as proxies for expected system behavior. Rather than treating evaluation as a purely statistical exercise, this strategy prioritizes interpretability and alignment with managerial reasoning, ensuring that performance metrics reflect decision-support quality under realistic organizational assumptions.

This strategy carries an inherent circularity risk: because the same expert defined the modeling assumptions and supplied the reference labels, agreement between model output and expert expectation reflects internal consistency with that expert's own reasoning, not an independent measure of predictive accuracy against real project outcomes. We therefore frame the goal of this evaluation as assessing *decision support*, whether the model faithfully operationalizes the expert's stated reasoning in a way that can be applied consistently across many candidate teams, rather than *real-world outcome prediction*, for which independent ground-truth data collected after actual STF decisions would be required.

Three design choices reduce, without eliminating, this risk. *First*, calibration and validation were elicited in separate sessions: the expert labeled the eight training scenarios of S2 without access to the model's later predictions, and subsequently assessed the 15 RQ2 collaboration scenarios and the two RQ3 Team Fit configurations without access to the model's predictions; the model was not revised after the expert's validation responses were collected. This reduces the risk that the researchers or the expert could tune the model or the reference labels toward agreement. *Second*, the semantic-consistency checks reported in Section 5 (extreme VL-VL, VL-VH, VH-VL, VH-VH configurations) evaluate whether the calibrated function preserves the intended ordinal semantics beyond the specific calibration labels. *Third*, the RQ2 comparison between the distribution-based and mean-based AC encodings is a *relative*, not absolute, evaluation: both encodings are scored against the same expert-provided distribution, which reduces the influence of global rating tendencies in the expert's judgments and focuses the analysis on representational differences between the two encodings. These safeguards do not remove the circularity threat, but they clarify the scope of the evidence provided by the scenario-based evaluation.

RQ1 Procedure. To address RQ1, we calibrate the regression model defined in Eq. (3) using the expert-labeled scenarios described in S2 (Section 3). For each scenario, we compute the predicted PC value and compare it against the expert-expected PC target. Predictive accuracy is assessed using the MAE, defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|PC^{(i)}_{expected} - PC^{(i)}_{predicted}\right|, \qquad (5)$$

and the Pearson correlation coefficient, capturing both absolute deviation and directional agreement between expert judgments and model predictions. In addition, we analyze the calibrated regression coefficients and the separate saturation adjustment $f(N)$ to verify whether the resulting PC function preserves the expert's intended monotonic and directional behavior.

Additionally, to estimate the generalization capability of the regression model calibrated with only eight scenarios, we applied leave-one-out cross-validation (LOOCV). We adopted LOOCV due to the small calibration set (n=8), following prior guidance that recommends LOOCV for smaller datasets (Kocaguneli and Menzies, 2013; Afzal et al., 2012). In this approach, the model is recalibrated eight times, each time excluding one of the scenarios, and the MAE is calculated on the excluded point. This technique provides a more conservative estimate of the expected error when the model is applied to scenarios not seen during calibration.

RQ2 Procedure. To answer RQ2, we compare two alternative representations of intra-team collaboration against the expert-provided AC judgments, using 15 expert-labeled collaboration scenarios. Each scenario specifies a distribution of pairwise compatibility tiers (the proportion of developer pairs falling into each tier, PC_VL–PC_VH) together with the expert's elicited AC distribution over the five ordinal states.

In the first condition (*C1-Weighted*), the PC-tier profile is aggregated into a single collaboration score using the calibrated tier weights $(1, 1, 4, 4, 5)$, as described in Stage S6. In the second condition (*C2-Mean*), collaboration is summarized by the arithmetic mean of the pairwise compatibility values, mapped to the same $[0.1, 0.9]$ scale. For each scenario, we derive the expert's expected AC value as the centroid-weighted mean of the elicited AC distribution, providing a scalar reference on the same scale.

We assess whether each representation *orders* the scenarios in agreement with the expert. We therefore compute the Spearman rank correlation (ρ) between each representation's score and the expert's expected AC value across the 15 scenarios. A higher ρ under *C1-Weighted* would indicate that preserving the tier distribution yields a collaboration signal that tracks the expert's ordinal judgment of team collabora-

tion quality more faithfully than collapsing it into a single mean value.

RQ3 Procedure. To assess BN generalization beyond the elicitation set, we perform scenario-based validation of the AE node with previously unseen cases defined by the expert. Validation scenarios combine different ordinal configurations of AT and AC as evidence; for each scenario, we compare the BN-predicted AE distribution against the expert-provided target distribution and compute the Brier score. Consistently low scores provide evidence that the expert-calibrated CPTs behave coherently on unseen but plausible AT/AC configurations within the elicited reasoning domain.

5 Results and Discussion

This section reports and discusses the empirical results obtained with the proposed method, organized according to the three RQs defined in Section 4. We first validate the regression-based PC model calibrated from expert knowledge (RQ1). We then analyze how pairwise PC estimates behave on realistic data and how they yield an interpretable compatibility graph. Finally, we evaluate the integration of PC-derived evidence into the BN, compare it with baseline representations of collaboration (RQ2), and assess BN predictive accuracy within the elicited reasoning domain (RQ3).

5.1 PC Modeling (RQ1)

This subsection addresses **RQ1**, which investigates whether a regression model calibrated with expert knowledge can accurately reproduce expected pairwise collaboration judgments under OSF, SLF, and collaboration-history saturation. We report the calibration results, assess quantitative fit against expert expectations, analyze the behavior of the model on empirical samples, and examine its semantic consistency and structural implications for the compatibility graph construction.

Regression calibration and coefficient analysis. The regression model was calibrated using the expert-labeled scenarios described in Stage S2, with the objective of estimating the parameters of the PC formulation. Table 5 reports the resulting coefficients together with the model's prediction error.

Table 5. Calibrated regression coefficients

Parameter	Value
Intercept	0.230232
α (SLF)	0.179206
β (OSF)	0.230728
γ (SLF · OSF)	0.104040
MAE	0.0812

The estimated parameters indicate that all predictors contribute substantively to the PC score. Both SLF and OSF exhibit positive coefficients ($\alpha = 0.179$ and $\beta = 0.230$, respectively), supporting the expert's assessment that objective project outcomes and leadership perceptions jointly shape collaborative effectiveness. This balance suggests that the model does not rely on a single signal, but reflects a combined view of collaboration quality.

The interaction term is also positive ($\gamma = 0.104$), indicating a mild synergy when OSF and SLF are simultaneously high. The saturation effect identified during expert elicitation is operationalized separately by the collaboration-history component (e.g., the $f(N)$ factor and associated rules), preventing repeated pairings from being overestimated even when historical indicators remain favorable.

From a fit standpoint (on the elicited scenarios), the The model achieves an MAE of 0.0812 on the normalized [0,1] scale, corresponding to 8.12% of the full scale range. Given the small number of elicited scenarios and the ordinal nature of the underlying judgments, this error level indicates strong alignment between model outputs and expert reasoning.

These results suggest that the regression-calibrated PC formulation provides an interpretable, scenario-consistent approximation of expert judgment within the elicited calibration domain. The learned coefficients preserve the intended semantic relationships among predictors while embedding safeguards against overestimation. This combination makes the PC metric suitable for large-scale automated exploration of team configurations—such as downstream optimization-based team formation—without sacrificing transparency or conceptual fidelity.

Expert vs Predicted Pair Compatibility. Figure 4 compares the expert-expected pairwise compatibility ($PC_{expected}$), derived from expert probability distributions, with the corresponding regression-based predictions ($PC_{predicted}$).

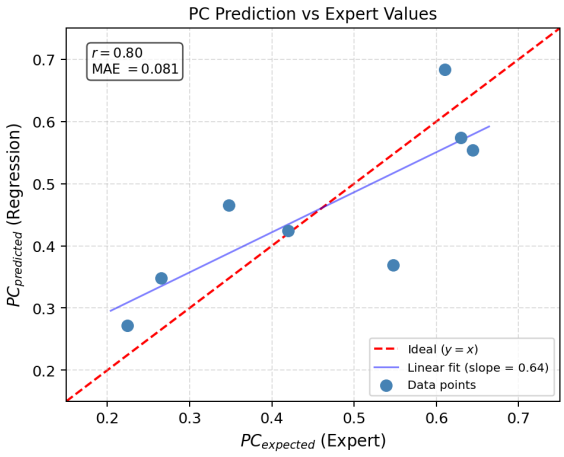

Figure 4. Comparison between $PC_{expected}$ and $PC_{predicted}$.

The observations follow a clear positive trend relative to the identity line ($y = x$), yielding a strong Pearson correlation of $r = 0.80$ and an in-sample MAE of 0.081. A linear fit between predicted and expected values results in a slope of 0.64 (intercept 0.16), indicating a compressed response relative to the ideal case. Overall, these results support that the calibrated model captures key aspects of the expert's judgment structure on the elicited scenarios.

Two scenarios exhibit deviations greater than 0.10: one

overprediction (OSF high, SLF low) and one underprediction (OSF moderate, SLF low). Both occur under asymmetric OSF/SLF combinations, where one indicator is substantially lower than the other. Rather than indicating a systematic bias, this behavior is consistent with the compressed mapping observed in Figure 4 (slope $= 0.64 < 1$), where the model tends to slightly overpredict lower expert values and underpredict higher ones.

LOOCV revealed an MAE of 0.176 (±0.114) and a Pearson correlation of 0.312 (p=0.453). The difference between the fitting metrics and the LOOCV metrics was expected given the small number of calibration scenarios (n=8), and reflects the sensitivity of the regression coefficients to the exclusion of individual points. This result quantifies the uncertainty inherent in parameterizing the expert's knowledge with such a reduced set of examples. The in-sample correlation (r=0.80) indicates that the model reproduces the ordinal structure of the expert's judgments within the elicited scenarios, while the LOOCV results (r=0.312, p=0.453) confirm that this adjustment should not be interpreted as evidence of broad generalization: with n=8 calibration scenarios, the model is best understood as a faithful operationalization of expert reasoning within the elicited domain, rather than a general predictive model. Overall, the alignment observed in Figure 4 supports that the regression-calibrated PC formulation captures the qualitative structure of expert expectations on the scenarios for which it was designed.

PC Scores in Sample Data. To illustrate the operational behavior of the calibrated model, PC scores were computed for a representative sample of developer pairs from the organizational analysis base. Table 6 reports, for each pair, the number of shared projects (N), the corresponding saturation factor $f(N)$, the normalized OSF and SLF values, and the resulting PC estimate.

Table 6. PC computed for developer pairs (top pairs by weight)

Pair	N	$f(N)$	OSF	SLF	PC
Dev3 - Dev327	4	1.000	0.775	0.662	0.581
Dev3 - Dev198	4	1.000	0.721	0.712	0.577
Dev168 - Dev175	4	1.000	0.801	0.605	0.574
Dev282 - Dev467	4	1.000	0.687	0.721	0.570
Dev53 - Dev300	4	1.000	0.660	0.723	0.562
Dev53 - Dev137	4	1.000	0.707	0.660	0.560
Dev369 - Dev363	4	1.000	0.712	0.643	0.557

The sample highlights the model's ability to discriminate collaboration quality in a realistic setting. Among the top pairs shown in Table 6, PC values span a range of approximately 0.557 to 0.581; across the full compatibility graph (773 valid pairs), PC ranges from 0.309 to 0.581 (mean = 0.424, as reported in Section 5). This distribution indicates that the metric avoids the common ceiling effect observed in simpler heuristics, where most collaborations tend to be rated uniformly high near the upper bound of the scale. This spread supports the use of PC as a ranking signal rather than as a binary or threshold-based indicator.

The effect of collaboration-history saturation is observable in the expected direction. Developer pairs with more than four shared projects consistently receive lower PC scores than comparable pairs at or below the inflection point, with an average reduction of roughly one percentage point. This behavior demonstrates that the saturation function operationalizes the expert's expectation of diminishing returns from repeated collaborations.

The calibrated regression coefficients further indicate that both inputs contribute meaningfully to PC: the fitted weights of OSF ($\beta = 0.2307$) and SLF ($\alpha = 0.1792$) are of the same order of magnitude, showing that objective project outcomes and leadership evaluations both exert substantial influence on the final score, with a moderate additional contribution from their interaction ($\gamma = 0.1040$). Importantly, the resulting ranking offers actionable managerial insight: pairs with similar OSF and SLF values can be ranked differently depending on whether they have exceeded the saturation threshold, providing a transparent, data-driven rationale for deciding when to preserve successful pairings and when to deliberately separate them when forming new teams.

Semantic consistency under extreme scenarios. To assess whether the calibrated model preserves the expert's intended ordinal semantics, we evaluated its behavior under four extreme input configurations: VL-VL, VL-VH, VH-VL, and VH-VH, where "VL" and "VH" correspond to the lowest and highest discretized levels of OSF and SLF, respectively. The resulting PC estimates exhibit strictly monotonic behavior. The VH-VH configuration yields a high compatibility score (PC = 0.93), whereas the VL-VL configuration results in a markedly low value (PC = 0.12). The two asymmetric cases (VL-VH and VH-VL) produce intermediate scores (0.46 and 0.49), reflecting balanced but imperfect collaboration potential.

These outcomes are fully aligned with the qualitative rankings articulated by the domain expert during elicitation, indicating that the regression coefficients encode numeric accuracy and the intended conceptual ordering of collaboration quality (i.e., *excellent*, *poor*, and *mixed* pairings). Importantly, this analysis indicates that the model behaves coherently under extreme OSF/SLF configurations beyond the specific calibration scenarios, avoiding counterintuitive behavior in extreme regions of the input space. Such semantic robustness is critical for downstream optimization and decision-support processes, as it ensures that automated team formation algorithms can rely on PC scores without introducing artifacts that contradict practitioner expectations.

Compatibility graph construction. Finally, the calibrated PC scores were used to construct a weighted developer compatibility graph. Figure 5 illustrates the top 20 highest-weight edges of the resulting compatibility graph (192 developers, 773 valid pairs), annotated with the estimated pairwise compatibility w and the number of shared projects N. Across the full graph, edge weights range from 0.309 to 0.581 (mean = 0.424), indicating moderate levels of expected collaborative performance.

The graph structure reveals meaningful collaboration patterns. Developers with multiple high-weight connections emerge as structurally central nodes, acting as bridges be-

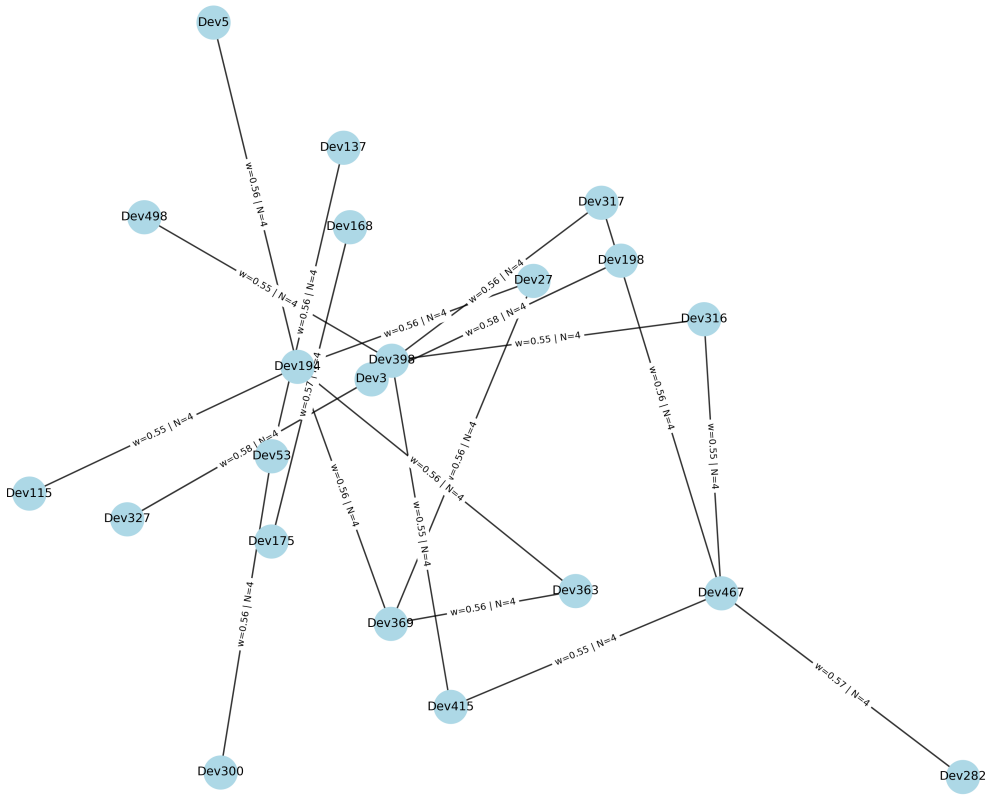

Figure 5. Developer pair-compatibility graph.

tween otherwise weakly connected subgroups. The top-weighted subgraphs also reveal cohesive collaboration units, characterized by dense local connectivity and relatively homogeneous compatibility scores. The strongest observed tie in the network is the pair `Dev3-Dev327` ($w = 0.581$, $N = 4$), which is consistent with the reported edge-weight range of 0.309 to 0.581.

These structural properties illustrate how the compatibility graph provides an interpretable, relational view of collaboration potential that goes beyond isolated pairwise scores. By exposing cohesive clusters and bridge developers, the graph serves as an effective intermediate representation for downstream team formation and analysis, supporting systematic reasoning about team composition and redeployment strategies grounded in historical collaboration evidence.

RQ1: PC Modeling

RQ1 examined whether a regression model calibrated with expert knowledge can reproduce expected pairwise collaboration judgments under OSF, SLF, and collaboration-history saturation. The results support calibration fidelity within the elicited scenario set. The calibrated model assigns meaningful importance to both objective project outcomes (OSF) and internal collaboration evaluations (SLF), while the positive interaction term captures a mild synergy when both signals are high. The expert-identified saturation effect of repeated collaborations is operationalized separately through the history-based factor $f(N)$. Quantitatively, the model achieves a low in-sample prediction error (MAE = 0.0812) and a strong positive correlation with expert expectations ($r = 0.80$). However, the LOOCV results indicate that broader statistical generalization should be interpreted cautiously.

5.2 Intra-team collaboration representation (RQ2)

This subsection addresses RQ2 by testing whether representing intra-team collaboration as a distribution of pairwise compatibility tiers ranks collaboration quality in closer agreement with expert judgment than collapsing it into a single mean. Across the 15 expert-labeled scenarios, the distribution-based transformer (*C1-Weighted*) achieved a Spearman rank correlation of $\rho = 0.882$ ($p < 0.001$) with the expert's expected AC value, compared to $\rho = 0.812$ ($p < 0.001$) for the mean-PC baseline (*C2-Mean*) and $\rho = 0.814$ ($p < 0.001$) for a majority-tier baseline. The distribution-based representation therefore tracks the expert's ordinal assessment of collaboration quality more closely than either baseline, as summarized in Table 7.

We note that this advantage is specific to ordinal agreement. In terms of absolute error against the expert's expected AC value, the mean-PC baseline attains a slightly lower deviation, reflecting a mild systematic optimism in the tier-weighted aggregation. However, since the operational role of the collaboration signal is to *rank* candidate teams by collaboration quality rather than to reproduce an exact scalar value, ordinal agreement is the more relevant criterion, and it is under this criterion that the distribution-based representation shows a consistent advantage.

A key reason for this improvement is that scenarios with similar mean PC values may encode markedly different internal compatibility compositions. The tier distribution retains this structure (e.g., concentration on VH versus mixtures across tiers), enabling the transformer to map distinct collaboration profiles to different AC labels, which scalar

Table 7. RQ2: Spearman rank correlation between each intra-team collaboration representation and the expert's expected AC value, across 15 expert-labeled scenarios (higher is better).

Representation	Spearman ρ	p-value
PC tier distribution (C1-Weighted)	0.882	< 0.001
Mean-PC (C2-Mean)	0.812	< 0.001
Majority-tier baseline	0.814	< 0.001

summaries cannot capture.

RQ2: Intra-team collaboration representation

RQ2 examined whether encoding intra-team collaboration as a distribution of pairwise compatibility tiers (PC_VL–PC_VH) ranks collaboration quality in closer agreement with expert judgment than collapsing it into a single mean tier. Across 15 expert-labeled scenarios, the distribution-based transformer achieved a higher Spearman rank correlation with the expert's expected AC value ($\rho = 0.882$) than both the mean-PC baseline ($\rho = 0.812$) and a majority-tier baseline ($\rho = 0.814$). This indicates that preserving the heterogeneity of pairwise relationships yields a collaboration signal that tracks the expert's ordinal assessment of team collaboration more faithfully than scalar aggregation.

5.3 BN Scenario-Based Validation (RQ3)

This subsection addresses RQ3 by examining whether the expert-calibrated BN generalizes beyond the elicitation set and accurately predicts AE from AT and AC evidence under previously unseen but plausible configurations. Validation is conducted through scenario-based assessments defined by the expert. The use of held-out scenarios allows us to evaluate the goodness of fit and the robustness of the calibrated CPTs when applied to novel combinations of inputs.

Team Fit validation. For the *Team Fit* node, we evaluated two representative validation scenarios that combine different ordinal configurations of AT and AC:

- **Scenario 1**: AT = VH, AC = M, yielding a Brier score of 0.0788;
- **Scenario 2**: AT = M, AC = VH, yielding a Brier score of 0.0782.

In both cases, the BN assigns the highest probability mass to the central states (M and H), consistent with the conservative MIXMINMAX(5,1) aggregation policy, which weights the weaker of the two inputs more heavily. The Brier scores reflect a structural characteristic of the calibrated model: with $\sigma = 0.05$, the Ranked Node formulation produces concentrated posterior distributions, assigning most probability mass to one or two adjacent states. The expert-provided reference distributions, by contrast, are smoother, spreading probability across all five states. This mismatch in distributional sharpness contributes to the observed Brier values independently of whether the BN's modal prediction is correct.

The relative ordering of the two scenarios is consistent with expert reasoning: both configurations produce similar Brier scores, reflecting the symmetric trade-off encoded by MIXMINMAX between high AT with moderate AC (Scenario 1) and moderate AT with high AC (Scenario 2). These results provide evidence that the expert-calibrated CPTs preserve monotonic and semantically coherent behavior when applied to unseen AT/AC configurations within the elicited domain. As such, the BN serves not only as a descriptive model of expert reasoning, but also as an evaluative mechanism suitable for supporting decision-making in dynamic organizational settings, complementing the transformer-level validation reported under RQ2.

To illustrate the practical use of these outputs, consider a staffing scenario in which a manager must compare two candidate teams for a project requiring both technological coverage and sustained collaboration. Suppose Team X is represented by the evidence configuration AT = VH and AC = M, while Team Y is represented by AT = M and AC = VH. The validation results in Scenarios 1 and 2 show that the BN captures both configurations coherently with respect to the expert-expected AE distributions, assigning the highest probability mass to the states consistent with the MIXMINMAX aggregation policy. In practice, the manager would not use the Brier score to choose between teams; the Brier score is a validation metric. The operational output is the AE posterior. Team X would be interpreted as a technically strong option with moderate collaboration evidence, whereas Team Y would be interpreted as a collaboration-strong option with moderate technical evidence. By making these trade-offs explicit in the AE posterior, the BN supports consistent and auditable comparison of candidate teams under the same expert-calibrated evaluation policy.

RQ3: BN predictive accuracy

RQ3 examined whether the expert-calibrated BN produces expert-consistent AE estimates for previously unseen scenarios when predicting AE from AT and AC evidence.
The results provide evidence of semantic consistency on the held-out scenarios within the elicited domain. Across the two held-out validation scenarios, the BN produced Brier scores of 0.0788 and 0.0782. These values reflect the structural mismatch between the concentrated posterior distributions produced by the calibrated model ($\sigma = 0.05$) and the smoother expert-provided reference distributions, rather than a failure of ordinal alignment. The BN assigns the highest probability mass to the states consistent with the MIXMINMAX(5,1) aggregation policy in both scenarios, preserving the intended monotonic trade-off between technical and collaborative evidence. Together with the transformer-level validation reported under RQ2, these findings indicate that the Ranked Node calibration yields a semantically consistent evaluator for AE within the elicited domain.

6 Implications

The empirical results obtained for RQ1-RQ3 lead to concrete implications for both industrial practice and future research on developer compatibility modeling and STF decision support. Importantly, these implications are grounded in qualitative alignment with expert reasoning as well as in quantitative performance indicators observed throughout the study.

Implications for Practice. The results of this study offer actionable guidance for organizations seeking to systematically incorporate evidence on collaboration into STF decisions, while maintaining transparency, scalability, and alignment with managerial judgment.

- **Lightweight expert-calibrated models can approximate managerial judgment with high fidelity (RQ1)**. The regression-calibrated PC model achieved a low prediction error (MAE = 0.0812 on a normalized [0,1] scale) and a strong Pearson correlation with expert expectations ($r = 0.80$). These results indicate that organi-

zations can operationalize collaboration assessment using a small set of expert-labeled scenarios, without the need for data-hungry or opaque learning models.

- **Balanced consideration of outcome-based and leadership-based signals is supported by the expert-calibrated model (RQ1).**
  The calibrated coefficients show comparable weights for OSF ($\beta = 0.2307$) and SLF ($\alpha = 0.179$), quantitatively confirming that collaboration quality depends on both objective delivery outcomes and leadership or coordination perceptions. This supports staffing practices that explicitly integrate both dimensions instead of privileging purely technical or productivity-based metrics.
- **Explicit history-based saturation supports sustainable team rotation policies (RQ1)**. The saturation effect identified during expert elicitation is operationalized through the collaboration-history component (e.g., $f(N)$), which penalizes pairs with excessive repeated collaborations (e.g., more than four shared projects). This provides managers with a quantitative mechanism to discourage excessive long-term pairings while still recognizing accumulated collaboration experience.
- **Preserving the distribution of pairwise compatibilities improves agreement with expert-ranked collaboration quality (RQ2).** For 15 expert-labeled scenarios, the distribution-based transformer achieved a higher Spearman rank correlation with the expert's expected AC value ($\rho = 0.882$) than the mean-PC baseline ($\rho = 0.812$), indicating that scalar summaries obscure the cohesion structure that the expert uses to judge collaboration quality.
- **Probabilistic evaluators provide stable and explainable team approval signals (RQ3)**. Across the two held-out AE validation scenarios, the BN produced Brier scores of 0.0788 and 0.0782, consistent with the concentrated posterior distributions generated by the calibrated model ($\sigma = 0.05$). The BN preserved the intended monotonic trade-off between technical and collaborative evidence in both scenarios, supporting its use as a transparent and semantically consistent decision-support tool in team formation processes.
- **Adoption costs are front-loaded in calibration, not in ongoing use**. Instantiating the method in a new organizational context requires: (i) a structured elicitation session with a decision-making expert to define OSF/SLF semantics and re-elicit context-dependent assumptions such as the saturation inflection point (Section 3); (ii) access to OSF/SLF-like signals that are already collected or can be derived from existing project management practices, avoiding new data-collection infrastructure when comparable indicators are available; and (iii) a practitioner familiar with a BN tool or probabilistic inference library to instantiate the compact BN configuration (Table 4). Once calibrated, applying the model to new candidate teams requires only computing the intra-team PC profile, transforming it into AC evidence, injecting the technical and collaboration evidence into the BN, and running standard probabilistic inference. Thus, routine use does not require repeated expert elicitation or manual CPT construction for each staffing decision. The learning curve is reduced by three design choices: the verbal-numerical elicitation scale reduces the cognitive burden on the expert; the RNM reduces the number of CPT parameters that must be elicited directly; and the compact BN configuration limits the number of modeled constructs. In addition, the RNM/GA calibration and validation procedures are encapsulated in the scripts and reference configurations provided as supplementary material, so practitioners do not need to reimplement these components from scratch. Organizations without readily available OSF/SLF-like indicators, without access to a decision-making expert, or without technical familiarity with BN tools would face a higher initial adoption cost.

Overall, these implications indicate that organizations can move beyond ad hoc or purely technical STF heuristics and adopt evidence-based, interpretable evaluation pipelines that reflect how staffing decisions are actually made in practice.

Implications for Research. From a research perspective, the findings open several directions for advancing the state of the art in developer compatibility modeling, explainable team analytics, and knowledge-guided decision-support systems in software engineering.

- **Expert-guided calibration can provide useful fidelity while preserving interpretability (RQ1)**. Achieving MAE = 0.0812 and $r = 0.80$ with only eight expert-labeled scenarios suggests that expert-informed regression can approximate expert judgment with useful calibration fidelity, without requiring the data volumes typical of learning-based models. This encourages further research on hybrid, knowledge-guided modeling solutions in data-scarce software engineering contexts, though direct empirical comparison against learning-based baselines remains an open direction (Section 7).
- **Saturation effects should be explicitly modeled and empirically evaluated (RQ1)**. The observed quantitative impact of the saturation term on PC scores demonstrates how the expert-elicited non-linear saturation assumption affects PC scores. Future studies should explore alternative saturation functions and validate their influence across different organizational settings.
- **Collaboration should be studied as a distributional, not scalar, construct (RQ2)**. The higher Spearman rank correlation achieved by the distribution-based transformer ($\rho = 0.882$ vs $\rho = 0.812$ for the mean-PC baseline) provides empirical evidence that collaboration quality emerges from the internal composition of pairwise relationships, and that this structure is better captured by preserving tier heterogeneity than by collapsing it into a scalar mean. This motivates further investigation of distributional and relational representations in team analytics.
- **BNs are a promising interpretable fitness function candidate (RQ3)**. The scenario-based validation of the AE node (Brier scores of 0.0788 and 0.0782) provides evidence of semantic consistency within the elicited domain: the BN preserves the monotonic trade-off be-

tween AT and AC encoded by the MIXMINMAX aggregation policy under previously unseen configurations. The observed Brier values reflect the distributional sharpness of the calibrated model ($\sigma = 0.05$) relative to the smoother expert-provided reference distributions; future work should investigate evaluation criteria, such as ordinal rank alignment, that are better suited to concentrated posterior distributions and to the intended use of the BN as a relative ranking tool across candidate teams.

- **Scenario-based probabilistic validation is a viable strategy when ground truth is unavailable (RQ3)**. The use of expert-defined validation scenarios combined with probabilistic accuracy metrics (Brier score) provides evidence of internal consistency and generalization within the elicited reasoning domain, though not of predictive accuracy against real-world outcomes (Section 7). This validation strategy can be adopted by future studies where real project outcome data is limited or delayed.

These implications suggest a research agenda that prioritizes interpretability, expert alignment, and probabilistic validation, positioning knowledge-guided and hybrid approaches as a viable foundation for next-generation team formation methods in software engineering.

7 Threats to Validity

This section discusses potential limitations that may impact the robustness, interpretation, and generalizability of our findings, organized according to the classification proposed Wohlin et al. (2012).

Construct Validity. A primary construct-validity threat concerns the reliance on a limited set of expert-labeled scenarios (eight in total) to calibrate the regression model. Although these scenarios were deliberately designed to span the OSF–SLF space and were reviewed iteratively with the expert, they may not fully capture the diversity and nuance of real-world collaboration patterns. Additional scenarios could reveal non-linearities, asymmetries, or interaction patterns not represented in the current elicitation design.

A second threat concerns the operationalization of OSF and SLF as proxies for prior collaboration quality. These constructs are intended to capture complementary aspects of previous team performance, but any concrete instantiation of them may omit relevant factors that influence collaboration, such as informal communication, interpersonal conflict, workload pressure, organizational politics, or project-specific constraints. This limitation is particularly relevant for OSF when instantiated through NPS-based feedback: NPS reflects client-perceived satisfaction with project delivery, but it does not necessarily capture technical quality, maintainability, internal process efficiency, or longer-term project outcomes. Therefore, OSF and SLF should be interpreted as pragmatic proxies for compatibility modeling, not as complete measures of collaboration quality.

A third threat concerns the use of a single expert. The elicitation expert is the senior authority responsible for staffing decisions in the studied organization; therefore, the model intentionally captures this authority's decision policy rather than an organization-wide consensus. This improves contextual validity for the target setting, but it also means that the calibrated PC model, PC-to-AC transformer, and BN evaluator may encode the expert's own assumptions, preferences, and biases. We therefore interpret the resulting model as an auditable representation of local decision logic, not as a universal model of software team formation.

We mitigated these threats by constructing scenarios that cover extreme, intermediate, and asymmetric OSF–SLF configurations, validating the ordinal mappings directly with the expert, and modeling collaboration relationally at the pair level rather than through individual traits. Nevertheless, future work should expand the elicitation set, incorporate additional experts, and evaluate alternative or complementary indicators for operationalizing OSF and SLF in different organizational contexts.

RQ2 introduces an additional construct-validity consideration because it evaluates alternative encodings of collaboration evidence rather than alternative team-formation methods. RQ2 is designed as a controlled ablation on evidence encoding: the BN parameterization is kept fixed and only the collaboration representation is varied (tier distribution vs. mean-tier one-hot). This isolates representational loss due to aggregation, rather than conflating it with alternative elicitation or CPT reparameterization. Therefore, RQ2 supports claims about representational adequacy under a fixed, expert-calibrated BN, not about the best possible BN calibration for each encoding.

Internal Validity. The regression model assumes a linear structure for PC_{base}, combining OSF, SLF, and their interaction, followed by a multiplicative saturation adjustment through $f(N)$. While the calibration results show strong alignment with expert expectations on the elicited scenarios (MAE = 0.0812, $r = 0.80$), non-linear effects or interaction patterns beyond those explicitly modeled may exist in practice. Furthermore, the calibration scenarios are synthetic by construction, which introduces potential bias if the elicited configurations do not adequately reflect the diversity of real collaboration situations.

Stages S4–S5 were instantiated on an organizational analysis base derived from anonymized project-participation records and project-level OSF/SLF indicators. Although this base reflects actual organizational co-participation structures and project-level evaluation signals, the resulting compatibility graph should not be interpreted as a complete empirical reconstruction of the organization's collaboration quality. The graph represents compatibility as defined by the proposed model and depends on the available records, the operationalization of OSF and SLF, and the expert-elicited saturation function. It may therefore omit relevant collaboration factors not captured in the organizational data, such as informal communication patterns, interpersonal conflict, workload pressure, or project-specific constraints.

We mitigated these threats by incorporating the SLF$\times$OSF interaction term, by applying the expert-elicited saturation function $f(N)$ as a separate multiplicative adjustment, by analyzing monotonicity under extreme configurations, and by complementing MAE with correlation analysis to cap-

ture directional agreement. However, the saturation function was not cross-validated against longitudinal outcome data, and the compatibility graph cannot substitute for validation against future project outcomes. Future studies should explore alternative functional forms, validate saturation effects using longer-term empirical collaboration trajectories, and instantiate the graph construction process across additional organizations and datasets.

External Validity. The study was conducted within a single organization with a specific management culture, evaluation practices, project portfolio, and staffing process. As a result, the calibrated model should not be interpreted as directly transferable to organizations with different norms, team sizes, project types, incentive structures, or decision-making practices. The elicitation expert also holds a senior technical leadership role, and judgment criteria may differ across managerial levels or organizational contexts.

To mitigate this threat, the proposed method emphasizes recalibration rather than reuse of fixed parameters. In particular, the operationalization of OSF and SLF, the saturation threshold in $f(N)$, the coefficients of PC_{base}, the PC-to-AC tier weights, and the BN structure and CPT parameters may all depend on the target organization. These elements reflect local assumptions about what counts as successful collaboration, how repeated collaboration should affect compatibility, and how technical and collaborative evidence should be balanced in team approval decisions. Therefore, applying the method in another organization would require re-eliciting these assumptions with the relevant local decision maker.

A further external-validity limitation concerns practical adoption across organizations. The method assumes that the organization can instantiate OSF/SLF-like indicators, has access to a local decision-making expert, and has at least minimal technical familiarity with BN tools or probabilistic inference libraries. Organizations whose staffing decisions are distributed across multiple stakeholders, or whose project records do not support comparable compatibility indicators, may face higher adoption costs. Similarly, organizations without prior experience in BN-based modeling may require additional technical support to instantiate and validate the compact BN configuration. These limitations do not affect the internal behavior of the method in the studied setting, but they constrain how easily the procedure can be transferred to other organizational contexts.

A related limitation concerns the absence of an empirical comparison with alternative STF methods. In this revision, we address this issue through the structured conceptual comparison presented in Section 2.2, rather than through a new experimental baseline. This choice reflects the scope of the study: the proposed method evaluates expert-guided compatibility modeling and team-fit assessment, whereas existing approaches often optimize different constructs, such as personality compatibility, skill coverage, graph connectivity, communication cost, or latent collaboration patterns. A fair empirical comparison would require adapting competing methods to the same organization-specific team-project fit criterion and evaluating them under comparable data and decision assumptions. We therefore treat direct empirical comparison with alternative STF methods as an important direction for future work.

Conclusion Validity. Predictive accuracy was primarily assessed using MAE, Pearson correlation, and Brier scores computed against expert-labeled expectations rather than independent ground-truth project outcomes. Consequently, the results demonstrate alignment with expert reasoning rather than direct causal impact on real project success. This introduces a *circularity* risk, since the same expert who defined the modeling assumptions also supplied the reference labels used for validation (Section 4.3); scenario-based validation can therefore only assess internal consistency within the elicited reasoning domain, not correlation with actual project outcomes, which would require longitudinal data collected after real STF decisions. Statistical inference is also limited by the small size of the calibration dataset.

The LOOCV analysis showed high variability in the prediction error (LOOCV MAE = 0.176) compared to the fitting error (MAE = 0.081), which highlights the inherent limitation of calibrating a regression model with only eight points. Although the main goal is the parameterization of expert knowledge and not traditional statistical inference, this instability in the coefficients is a limitation of the study. We mitigated this issue by not relying exclusively on the regression for team evaluation, but rather by integrating its output into a compatibility graph and a BN, whose behavior was further evaluated using held-out expert-defined scenarios.

We partially mitigated conclusion-validity threats by employing complementary evaluation metrics (MAE, Pearson correlation, and Brier score), by analyzing semantic consistency under extreme configurations, by comparing the PC-tier distribution representation against a mean-PC baseline in RQ2, and by validating AE inference on held-out expert-defined scenarios. Nonetheless, these evaluations do not establish superiority over alternative STF methods. Future longitudinal studies should compare BN-predicted Team Fit with actual project performance indicators and evaluate the proposed method against alternative approaches under a common organizational decision criterion.

8 Final Remarks

This paper proposes a lightweight, expert-guided method for estimating developer compatibility and supporting explainable STF decisions. The method combines two project-level indicators, OSF and SLF, with an expert-defined collaboration-history saturation function and expert judgments on eight synthetic scenarios. These labels calibrate a linear regression model that accounts for diminishing returns in repeated collaborations. The resulting PC scores form a weighted graph, which is embedded in a BN to evaluate team fit probabilistically.

The findings address three core research questions. For RQ1, the regression model closely reproduces expert expectations, achieving an MAE of 0.0812 and a Pearson correlation of 0.80, while preserving monotonicity and modeling the saturation effect beyond four projects. Regarding intra-team collaboration modeling, the results of RQ2 provide evidence that representing collaboration as a distribution of pairwise compatibility tiers leads to a more faithful PC-to-AC transformation than collapsing collaboration into a single mean

value. By preserving heterogeneity among developer interactions, the distribution-based representation ranked collaboration quality in closer agreement with expert judgment (Spearman $\rho = 0.882$ versus 0.812 for the mean-PC baseline), reinforcing the importance of distributional collaboration modeling in team formation systems. RQ3 provides evidence of semantic consistency for AE within the elicited domain: across the two held-out AT/AC configurations tested, the BN produced Brier scores of 0.0788 and 0.0782, preserving the monotonic trade-off between technical and collaborative evidence encoded by the MIXMINMAX(5,1) aggregation policy. The Brier values reflect the concentrated posterior distributions of the calibrated model ($\sigma = 0.05$) relative to the smoother expert-provided reference distributions, a structural characteristic discussed further in Section 7.

These results offer both practical and research-oriented implications. From a managerial standpoint, the proposed method enables interpretable, auditable, and scalable team recommendations using limited historical data. The saturation model supports reasoning about team rotation, and the compatibility distribution allows transparent inspection of collaboration structure. From a research perspective, the work suggests that expert-calibrated regression combined with Bayesian inference offers a viable complementary direction to data-intensive black-box models in settings where outcome data is limited and interpretability is required.

Several opportunities for future work emerge from this study. First, the calibration process can be strengthened by expanding the set of elicited scenarios and incorporating multiple experts to assess inter-expert variability and improve robustness. Second, alternative saturation functions and nonlinear regression formulations could be explored and empirically compared across different organizational contexts. Third, future work should develop a tool-supported adoption framework to reduce the practical burden of applying the method, including guided interfaces for expert elicitation, structured forms for collecting judgments used in CPT calibration, assisted BN configuration, automated consistency and validation checks, and encapsulated RNM/GA calibration procedures. Fourth, the compatibility graph and BN evaluator can be embedded more tightly into search-based optimizers, such as genetic algorithms, to enable fully automated yet interpretable team assembly. Fifth, future studies should compare the proposed evaluator with alternative STF approaches under a common decision criterion, including graph-based, heuristic, and learning-based methods adapted to the same organizational setting. Finally, longitudinal and multi-organizational studies are needed to empirically assess the relationship between BN-predicted Team Fit and actual project outcomes, thereby strengthening conclusions and external validity.

Data Availability

All data and scripts used in this study are publicly available in a GitHub repository[1]. The repository contains the expert calibration spreadsheet, Python scripts for regression fitting, pairwise compatibility computation, and compatibility graph generation, the BN engine and calibrated configuration together with the validation scenarios and scripts used in RQ1, RQ2 and RQ3, as well as the datasets and artifacts required to reproduce the reported results. No sensitive or personally identifiable information is included.

[1] `https://github.com/isevirtus/graph_validation`

Acknowledgements

This work has been partially funded by the project “iSOP Base: Investigação e desenvolvimento de base arquitetural e tecnológica da Intelligent Sensing Operating Platform (iSOP)” supported by CENTRO DE COMPETÊNCIA EMBRAPII VIRTUS EM HARDWARE INTELIGENTE PARA INDÚSTRIA - VIRTUS-CC, with financial resources from the PPI HardwareBR of the MCTI grant number 055/2023, signed with EMBRAPII.

References

Afzal, W., Torkar, R., and Feldt, R. (2012). Resampling methods in software quality classification. *International Journal of Software Engineering and Knowledge Engineering*, 22(02):203–223.

Alves, L., Ricardo, V., and Xavier, L. (2021). Beyond Subjectiviness: Assessing Abilities and Preferences to Create Software Development Teams. In *Anais do I Workshop Brasileiro de Engenharia de Software Inteligente (ISE 2021)*, pages 7–12. Sociedade Brasileira de Computação.

Baghel, V. S. and Bhavani, S. D. (2018). Multiple Team Formation Using an Evolutionary Approach. In *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–6. IEEE.

Briand, L., Bianculli, D., Nejati, S., Pastore, F., and Sabetzadeh, M. (2017). The case for context-driven software engineering research: Generalizability is overrated. *IEEE Software*, 34(5):72–75.

Brownlee, A. E. I., McCall, J. A. W., and Zhang, Q. (2013). Fitness modeling with markov networks. *IEEE Transactions on Evolutionary Computation*, 17(6):862–879.

Costa, A., Ramos, F., Perkusich, M., Dantas, E., Dilorenzo, E., Chagas, F., Meireles, A., Albuquerque, D., Silva, L., Almeida, H., and Perkusich, A. (2020). Team formation in software engineering: A systematic mapping study. *IEEE Access*, 8:145687–145712.

Costa, A., Ramos, F., Perkusich, M., Neto, A. D. S., Silva, L., Cunha, F., Rique, T., Almeida, H., and Perkusich, A. (2022). A genetic algorithm-based approach to support forming multiple scrum project teams. *IEEE Access*, 10:68981–68994.

Cunha, F., Perkusich, M., Albuquerque, D., Filho, E. D., Gorgônio, K., and Perkusich, A. (2025a). Constructing developer compatibility graphs for team formation: An industrial case study. In *Anais do IV Workshop Brasileiro de Engenharia de Software Inteligente*, pages 31–36, Porto Alegre, RS, Brasil. SBC.

Cunha, F., Perkusich, M., Albuquerque, D., Gorgônio, K., Almeida, H., and Perkusich, A. (2025b). A genetic algorithm with convex combination crossover for software team formation: Integrating technical and collaboration

skills. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 146–153.

Cunha, F., Perkusich, M., Albuquerque, D., Santos, R. N. d., Gorgônio, K. C., and Perkusich, A. (2026). Integrating bayesian reasoning and evolutionary search for knowledge-driven team formation. In *Proceedings of the 2026 IEEE/ACM 2nd International Workshop on Neuro-Symbolic Software Engineering*, NSE '26, page 9–16, New York, NY, USA. Association for Computing Machinery.

Cunha, F., Perkusich, M., Almeida, H., Gorgônio, K., and Perkusich, A. (2022a). Teamplus: A decision support system for software team formation. In *Proceedings of the XXXVI Brazilian Symposium on Software Engineering*, pages 118–123.

Cunha, F., Rique, T., Perkusich, M., Gorgônio, K., Almeida, H., and Perkusich, A. (2022b). A data-driven framework to support team formation in software projects. In *Workshop Brasileiro de Engenharia de Software Inteligente (ISE)*, pages 7–12. SBC.

Fenton, N. E., Neil, M., and Caballero, J. G. (2007). Using ranked nodes to model qualitative judgments in bayesian networks. *IEEE transactions on knowledge and data engineering*, 19(10):1420–1432.

Ferreira, T. N., Vergilio, S. R., and de Souza, J. T. (2017). Incorporating user preferences in search-based software engineering: A systematic mapping study. *Information and Software Technology*, 90:55–69.

Freire, A., Neto, M., Perkusich, M., Costa, A., Gorgónio, K., Almeida, H., and Perkusich, A. (2022). A literature-based thematic network to provide a comprehensive understanding of agile teamwork (106). *International Journal of Software Engineering and Knowledge Engineering*, 32(05):645–659.

Freire, A., Perkusich, M., Saraiva, R., Almeida, H., and Perkusich, A. (2018). A bayesian networks-based approach to assess and improve the teamwork quality of agile teams. *Information and Software Technology*, 100:119–132.

Hamidi Rad, R., Bagheri, E., Kargar, M., Srivastava, D., and Szlichta, J. (2021). Retrieving Skill-Based Teams from Collaboration Networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2015–2019. ACM.

Hamidi Rad, R., Fani, H., Bagheri, E., Kargar, M., Srivastava, D., and Szlichta, J. (2023). A variational neural architecture for skill-based team formation. *ACM Transactions on Information Systems*, 42(1):1–28.

Hamidi Rad, R., Fani, H., Kargar, M., Szlichta, J., and Bagheri, E. (2020). Learning to Form Skill-based Teams of Experts. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2049–2052. ACM.

Harman, M., Mansouri, S. A., and Zhang, Y. (2012). Search-based software engineering: Trends, techniques and applications. *ACM Computing Surveys (CSUR)*, 45(1):1–61.

Hoegl, M. and Gemuenden, H. G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization science*, 12(4):435–449.

Kamel, K., Tubaiz, N., AlKoky, O., and AlAghbari, Z. (2011). Toward forming an effective team using social network. In *2011 International Conference on Innovations in Information Technology*, pages 308–312. IEEE.

Kocaguneli, E. and Menzies, T. (2013). Software effort models should be assessed via leave-one-out validation. *Journal of Systems and Software*, 86(7):1879–1890.

Lappas, T., Liu, K., and Terzi, E. (2009). Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 467–476. ACM.

Li, Y., Yin, X., Luo, K., Gai, Y., and Yang, X. (2026). Optimizing competitive team synergy through compatibility-aware graph search in distributed platforms. *Journal of Cloud Computing*.

Manzano, M., Ayala, C., Gómez, C., Abherve, A., Franch, X., and Mendes, E. (2021). A method to estimate software strategic indicators in software development: an industrial application. *Information and Software Technology*, 129:106433.

Marsicano, G., da Silva, F. Q., Seaman, C. B., and Adaid-Castro, B. G. (2020). The teamwork process antecedents (tpa) questionnaire: developing and validating a comprehensive measure for assessing antecedents of teamwork process quality. *Empirical Software Engineering*, 25(5):3928–3976.

Mendes, E. (2014). Expert-based knowledge engineering of bayesian networks. In *Practitioner's Knowledge Representation: A Pathway to Improve Software Effort Estimation*, pages 73–105. Springer.

Pelikan, M., Goldberg, D. E., and Lobo, F. G. (2002). A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, 21(1):5–20.

Rad, R. H., Seyedsalehi, S., Kargar, M., Zihayat, M., and Bagheri, E. (2022). A neural approach to forming coherent teams in collaboration networks. In *Proceedings of the 25th International Conference on Extending Database Technology (EDBT)*, pages 440–444. OpenProceedings.org.

Ramírez, A., Romero, J. R., and Simons, C. L. (2019). A systematic review of interaction in search-based software engineering. *IEEE Transactions on Software Engineering*, 45(8):760–781.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard business review*, 81(12):46–55.

Renooij, S. and Witteman, C. (1999). Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, 22(3):169–194.

Rique, T., Perkusich, M., Dantas, E., Albuquerque, D., Gorgônio, K., Almeida, H., and Perkusich, A. (2023). On adopting software analytics for managerial decision-making: A practitioner's perspective. *IEEE Access*, 11:73145–73163.

Ruenin, P., Choetkiertikul, M., Supratak, A., and Tuarob, S. (2024). Terekg: A temporal collaborative knowledge graph framework for software team recommendation.

Knowledge-Based Systems, 289:111492.

Saeedi, M., Hosseini, H., Wong, C., and Fani, H. (2025a). A survey of subgraph optimization for expert team formation. *ACM Comput. Surv.*, 57(12).

Saeedi, M., Hosseini, H., Wong, C., and Fani, H. (2025b). A Survey of Subgraph Optimization for Expert Team Formation. *ACM Computing Surveys*, 57(12):1–40.

Saraiva, R., Medeiros, A., Perkusich, M., Valadares, D., Gorgônio, K. C., Perkusich, A., and Almeida, H. (2020). A bayesian networks-based method to analyze the validity of the data of software measurement programs. *IEEE Access*, 8:198801–198821.

Setayandeh, M. R. and Babaei, A.-R. (2020). Multidisciplinary design optimization of an aircraft by using knowledge-based systems. *Soft Computing*, 24(16):12429–12448.

Strnad, D. and Guid, N. (2010). A fuzzy-genetic decision support system for project team formation. *Applied soft computing*, 10(4):1178–1187.

Tuarob, S., Assavakamhaenghan, N., Tanaphantaruk, W., Suwanworaboon, P., Hassan, S.-U., and Choetkiertikul, M. (2021). Automatic team recommendation for collaborative software development. *Empirical Software Engineering*, 26(4):64.

Vishnubhotla, S. D. and Mendes, E. (2024). Examining the effect of software professionals' personality & additional capabilities on agile teams' climate. *Journal of Systems and Software*, 214:112054.

Vishnubhotla, S. D., Mendes, E., and Lundberg, L. (2020). Investigating the relationship between personalities and agile team climate of software professionals in a telecom company. *Information and Software Technology*, 126:106335.

Wieringa, R. (2014). *Design science methodology for information systems and software engineering*. Springer.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., Wesslén, A., et al. (2012). *Experimentation in software engineering*, volume 236. Springer.