Unified Structural Factors for Transfer Learning Generalization with PAC-Bayesian Guarantees

Ziqi Gao

Department of Computer and Information Science,
University of Pennsylvania, Philadelphia, PA, USA
ziqigao@seas.upenn.edu

Abstract

Understanding when a pre-trained model generalizes well to a new task remains a key challenge in transfer learning. Classical theories bound target risk using divergences such as total variation, MMD, or Wasserstein distance, yet tasks with similar divergences often show very different transfer performance. We propose a structural framework that explains transferability through two factors: the *Feature Overlap Rate (FOR)*, measuring how much target representation lies in the source-induced subspace, and the *Effective Task Complexity (ETC)*, quantifying the entropy of latent subtasks. We derive a PAC-Bayesian bound where target risk depends on FOR and ETC, and show that larger models attenuate their negative effects. Experiments on six GLUE transfer pairs estimate FOR and ETC from encoder representations and compare them to classical divergences. Results show that FOR and ETC together explain over 80% of transfer risk variance, while divergences fail to do so. Our findings provide a geometry-aware perspective for diagnosing and guiding transfer learning.

Introduction

A fundamental challenge in transfer learning is to understand the conditions under which a pre-trained model generalizes effectively to a new task. Classical theories typically bound the target risk in terms of source risk and a divergence between task distributions–such as total variation, MMD, or Wasserstein distance (Ben-David et al. 2010; Redko et al. 2017; Shen et al. 2018). However, empirical evidence reveals a persistent gap between these theoretical predictions and the behavior of deep models: tasks that are similarly distant from the source in terms of marginal divergence can lead to drastically different transfer outcomes (Chang et al. 2020; Aghajanyan, Gupta, and Zettlemoyer 2021; Radford et al. 2021).

This discrepancy stems from a key limitation: traditional divergence measures fail to reflect the *geometry of the learned representation space*. Let $\phi : \mathcal{X} \to \mathbb{R}^d$ denote the feature map induced by a pre-trained model. Effective transfer requires that the target distribution P_T be well-aligned with the subspace spanned by source features $\phi(P_S)$. When this *feature overlap* is low–even if marginal distributions are statistically similar–negative transfer can occur (Yosinski et al. 2014; Kornblith et al. 2019).

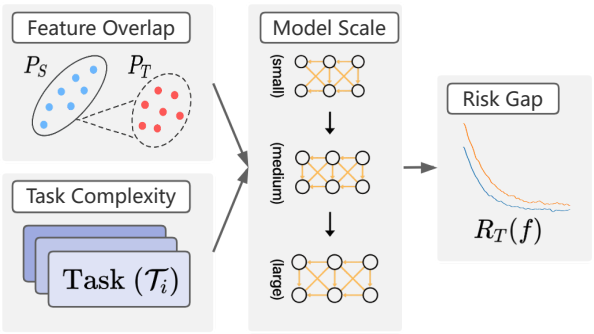

Figure 1: Our structural explanation for transfer generalization. The target risk $R_T(f)$ is shaped by representation misalignment and task complexity, whose effects are modulated by model scale.

In addition to representation misalignment, transfer generalization also depends on the *structural complexity* of the target task. Real-world targets often consist of latent subtasks $\{\mathcal{T}_i\}_{i=1}^K$ with imbalanced proportions or heterogeneous difficulty. This compositionality induces an entropy-like burden on generalization, especially in low-data settings (Baxter 2000; D'Amour et al. 2020). Yet, most existing bounds assume homogeneous task structures and thus fail to capture this complexity.

While these two structural factors—feature overlap and task complexity—govern transfer performance, their influence is further shaped by the capacity of the model. Empirical scaling laws demonstrate that larger models generalize better, often following predictable error decay rates with respect to parameter count (Kaplan et al. 2020b; Bahri et al. 2021). However, such scaling laws typically assume i.i.d. tasks and do not account for how scale interacts with task structure or representation geometry.

We propose a structural framework in which transfer generalization is shaped by two primary factors: the *feature overlap* between the source and target representations, and the *internal complexity* of the target task. These factors directly influence the transfer risk, but their effect is *modulated* by the capacity of the model. Larger models exhibit greater resilience to both representation misalignment and task het-

erogeneity, acting as a buffer against structural limitations.

This framework is visualized in Figure 1. Feature misalignment (top-left) and subtask imbalance (bottom-left) each contribute to elevated target risk $R_T(f)$, but their severity is mediated by the scale of the model (center). The overall transfer performance thus emerges from the interaction between representational compatibility, task complexity, and model size. We summarize our main contributions as follows:

- We propose a unified structural measure of transferability based on two interpretable quantities: the *Feature Overlap Rate (FOR)*, quantifying how much of the target distribution P_T lies within the source representation $\phi(P_S)$, and the *Effective Task Complexity (ETC)*, measuring the entropy of latent subtask composition.
- We derive a PAC-Bayes-style upper bound on the target risk $R_T(f)$, in which FOR and ETC explicitly appear. This shows how geometric alignment and task complexity structurally influence the transfer risk.
- We extend this bound by showing that increasing model scale attenuates the effect of low FOR and high ETC, yielding a quantitative explanation for observed transfer scaling behaviors.
- We empirically validate our theory on NLP benchmarks, showing that FOR and ETC better predict transfer performance than classical divergence-based baselines, especially under model scaling.

Together, these results provide a unified theoretical lens to understand transfer performance, integrating representational alignment, structural complexity, and model capacity into a coherent framework.

Related Work

Divergence-Based Bounds and Their Limitations. Classical theories of domain adaptation typically bound the generalization gap $R_T(f) - R_S(f)$ using distributional divergences such as total variation, Wasserstein distance, or ϕ-divergence (Germain et al. 2009; Zhang, Geng, and Zhou 2019; Reddi, Kale, and Kumar 2018). These bounds implicitly assume that if the marginal input distributions P_S and P_T are sufficiently close under some divergence, then a low risk on P_S would translate into low risk on P_T. However, recent observations demonstrate that even when $P_S \approx P_T$ in distributional terms, transfer can fail catastrophically (Raghu et al. 2017; Morcos et al. 2018). This discrepancy stems from the fact that these divergences operate over the input space $\mathcal{X}$, while modern deep models learn a representation map $\phi : \mathcal{X} \to \mathbb{R}^d$ [1]. Generalization in transfer thus hinges less on $P_S \approx P_T$, and more on whether $\phi(P_T)$ aligns with the subspace of $\phi(P_S)$–a property we refer to as *feature overlap*.

Representational Similarity and Structural Alignment. To better reflect model-specific inductive biases, several works have explored feature-level similarity metrics, such as CKA (Fort, Hu, and Lakshminarayanan 2019) or probing-based alignment (Tian, Krishnan, and Isola 2020). These methods estimate the alignment of learned features across tasks, but are largely empirical and lack formal integration into transfer risk bounds. Moreover, most of these approaches treat the target task P_T as a monolithic distribution. In reality, many transfer settings involve implicit substructures: $P_T = \sum_{i=1}^{K} \pi_i P_i$, where each sub-distribution P_i corresponds to a latent subtask $\mathcal{T}_i$. This compositional structure induces variance in feature relevance across subtasks and exacerbates the generalization challenge–especially under imbalanced or low-support mixtures. Existing work has rarely analyzed transfer learning from the perspective of task compositional structure.

Scaling Effects and Unified Perspectives. Recent scaling law studies (Hestness et al. 2017; Kaplan et al. 2020a) suggest that model size $|\theta|$ serves as a buffer against misalignment and overfitting, yielding power-law decay in error under i.i.d. assumptions. However, these scaling laws offer limited insight into how model capacity interacts with representational misalignment or latent task complexity. Empirically, large models still exhibit poor transfer when either the feature overlap or task structure is misaligned with the source model (Sun et al. 2024).

Taken together, these observations reveal the insufficiency of existing frameworks and motivate a unified theory of transfer generalization that considers three core components: (i) the overlap between $\phi(P_S)$ and $\phi(P_T)$, (ii) the internal complexity of $P_T = \sum_{i=1}^{K} \pi_i P_i$, and (iii) the scale of the model $|\theta|$.

In the next section, we formalize these dimensions by introducing two structural quantities—*Feature Overlap Rate (FOR)* and *Effective Task Complexity (ETC)*—and derive generalization bounds that explicitly account for their interactions with model capacity.

Problem Setup and Structural Measures

Transfer Learning Setup

We consider a standard transfer learning setting where a pre-trained model is trained on a source distribution P_S over input-label pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and is subsequently adapted or evaluated on a target distribution P_T. Both distributions are defined over the same input space $\mathcal{X}$ and label space $\mathcal{Y}$, but may differ significantly in semantics or structure (Pan and Yang 2010; Weiss, Khoshgoftaar, and Wang 2016).

Let $\phi : \mathcal{X} \to \mathbb{R}^d$ denote a fixed feature map learned by the source model. A downstream task-specific model $f \in \mathcal{F}$ is defined as $f(x) := h(\phi(x))$, where $h \in \mathcal{H}$ is a classifier or regressor trained on the target domain, and the hypothesis class is $\mathcal{F} := \{h \circ \phi : h \in \mathcal{H}\}$. This formulation aligns with recent studies emphasizing the role of representation reuse in transfer learning (Nguyen et al. 2020).

Let $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded, convex, and L-Lipschitz loss function, such as cross-entropy or squared error. The expected risks on the target and source domains are

[1] $\phi(\cdot)$ denotes an abstract representation mapping learned by a pre-trained model, typically projecting inputs into a feature space where transfer is performed.

defined respectively as:

$$R_T(f) := \mathbb{E}_{(x,y)\sim P_T}\, \ell(f(x), y). \tag{1}$$

$$R_S(f) := \mathbb{E}_{(x,y)\sim P_S}\, \ell(f(x), y). \tag{2}$$

The central question in transfer learning is: under what structural conditions can a predictor $f \in \mathcal{F}$, trained or adapted using source knowledge, achieve strong performance on the target domain?

Rather than relying on classical input-space divergences, we adopt a structural perspective on transferability. We posit that the target risk $R_T(f)$ is governed by three core factors: the alignment between source and target in the learned feature space, the latent structure of the target task, and the scale of the model. The following subsections formalize these components.

Feature Overlap Rate

A central factor in transfer performance is how well the target distribution P_T is represented within the feature space learned on the source distribution P_S. Classical divergence metrics such as total variation or MMD measure discrepancies in the input space $\mathcal{X}$, but modern deep models rely on a learned representation $\phi : \mathcal{X} \to \mathbb{R}^d$ (Bengio, Courville, and Vincent 2013). Hence, transfer success depends not on $P_T \approx P_S$, but rather on how much of $\phi(P_T)$ lies within the span of $\phi(P_S)$.

To quantify this alignment, we introduce the following structural measure.

Definition 1. *Let $\mathcal{F}_S := \mathrm{span}\{\phi(x) : x \sim P_S\}$ denote the subspace spanned by source embeddings. We define the* Feature Overlap Rate (FOR) *between source and target distributions as:*

$$\mathrm{FOR}(P_S, P_T) := \mathbb{E}_{x\sim P_T}\left[\frac{\left\|\mathrm{Proj}_{\mathcal{F}_S}\phi(x)\right\|^2}{\|\phi(x)\|^2}\right]. \tag{3}$$

We visualize this concept in Figure 2, where each target representation (red) is projected onto the subspace $\mathcal{F}_S$ spanned by source features (blue). The quality of these projections provides an intuitive sense of feature overlap.

When $\mathrm{FOR}(P_S, P_T) \approx 1$, the pre-trained features are highly transferable; when it is close to 0, much of the target signal lies outside the representational scope of the source model.

Importantly, this structural misalignment has practical consequences. Even if the predictor is optimal over $\mathcal{F}_S$, components orthogonal to this subspace induce an unavoidable transfer gap (Gong et al. 2016).

To see why, consider a decomposition $\phi(x) = \phi_\parallel + \phi_\perp$, where $\phi_\parallel = \mathrm{Proj}_{\mathcal{F}_S}\phi(x)$ lies in the source span and $\phi_\perp \perp \mathcal{F}_S$. Since the downstream model h operates only on $\phi_\parallel$, the residual $\phi_\perp$ is effectively ignored. This leads to a structural prediction error that scales with the expected magnitude of $\|\phi_\perp\|$, and hence with $1 - \mathrm{FOR}(P_S, P_T)$.

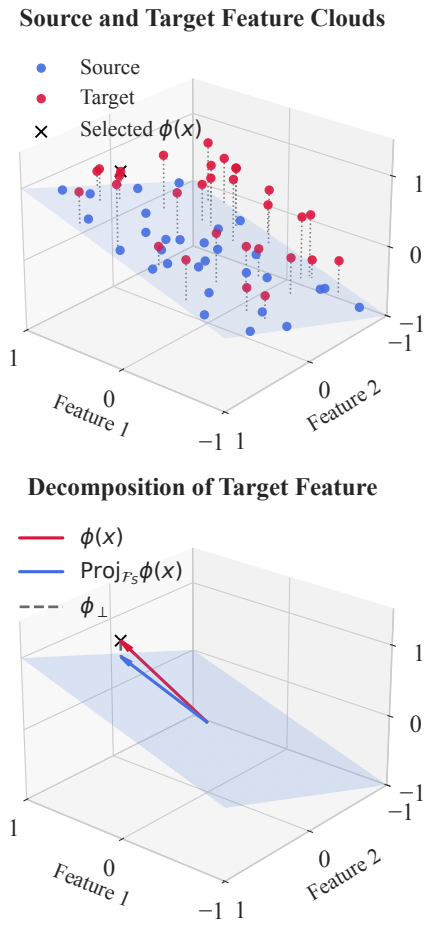

Figure 2: **(Left)** A geometric view of representational alignment: dotted connectors highlight components of target features (red) that lie outside the source-induced subspace $\mathcal{F}_S$. **(Right)** Decomposition of a selected target representation $\phi(x)$ into its in-subspace projection and orthogonal residual, visualizing the structural origin of mismatch.

Effective Task Complexity

Even with high feature overlap, transfer learning may still fail if the target distribution P_T exhibits internal heterogeneity. In many realistic settings, P_T is not a monolithic task but a mixture of multiple latent subtasks. Formally, we consider:

$$P_T = \sum_{i=1}^{K} \pi_i P_i, \tag{4}$$

where each P_i corresponds to a sub-distribution governing a latent subtask $\mathcal{T}_i$, and $\{\pi_i\}_{i=1}^K$ forms a valid probability simplex (Mansour, Mohri, and Rostamizadeh 2008).

This mixture induces a structural burden on generalization: to succeed on P_T, a learner must effectively cover all $\{\mathcal{T}_i\}_{i=1}^K$. When the mixture is imbalanced (e.g., some $\pi_i \ll 1$), low-frequency subtasks receive limited data, complicating model fitting. We capture this burden with the following definition.

Definition 2. *Let $P_T = \sum_{i=1}^{K} \pi_i P_i$ be a latent subtask mixture. The* Effective Task Complexity (ETC) *of P_T is defined as the Shannon entropy of the mixing weights:*

$$\mathrm{ETC}(P_T) := \sum_{i=1}^{K} \pi_i \log \frac{1}{\pi_i}. \tag{5}$$

ETC quantifies the diversity and imbalance of the subtask structure. When π is uniform, ETC is maximized at $\log K$; when highly skewed, ETC is low, indicating the effective presence of fewer dominant tasks.

Together with FOR, ETC serves as a complementary structural lens: FOR governs what to transfer (representation alignment), while ETC governs how much data is required to transfer effectively across diverse subtasks.

Unified Generalization Theory

Assumptions and Setup

We now formalize the structural factors governing transfer generalization. Recall that a model $f = h \circ \phi$ maps inputs via a fixed source-learned feature extractor $\phi : \mathcal{X} \to \mathbb{R}^d$, followed by a target-specific classifier $h \in \mathcal{H}$. Our goal is to characterize the target risk $R_T(f)$ in terms of the structural alignment between source and target, independent of classical input-space divergences.

To proceed, we adopt the PAC-Bayesian framework and impose the following assumptions:

A1. The loss function $\ell : \mathbb{R} \times \mathcal{Y} \to [0, 1]$ is convex and L-Lipschitz in its first argument.

A2. The feature map ϕ is bounded: $\|\phi(x)\| \leq B$ for all $x \in \mathcal{X}$.

A3. The target distribution is a finite mixture: $P_T = \sum_{i=1}^{K} \pi_i P_i$, where each P_i governs a latent subtask and $\pi_i > 0$, $\sum_i \pi_i = 1$.

A4. We consider a prior distribution p and a posterior q over classifiers $h \in \mathcal{H}$, both absolutely continuous.

These assumptions reflect practical properties of transfer learning: bounded feature norms arise from common normalization schemes; task heterogeneity is ubiquitous in real-world domains; and the PAC-Bayes setting captures the stochasticity of fine-tuning and regularization.

Our analysis connects these structural components—feature overlap, task mixture complexity, and model variability—to a generalization bound on $R_T(f)$.

Unified Structural Bound

We now present our main theoretical result, which establishes a transfer generalization bound grounded in the structural alignment between source and target.

Theorem 1. *Let $f = h \circ \phi$ denote the transfer model introduced above, where $\phi : \mathcal{X} \to \mathbb{R}^d$ is a fixed representation function and $h \in \mathcal{H}$ is a hypothesis sampled from a posterior distribution q with prior p. Under the assumptions stated earlier, with probability at least $1 - \delta$ over a source sample of n points, the target risk satisfies:*

$$\begin{aligned} R_T(f) \leq R_S(f) \quad & + \alpha \cdot (1 - \mathrm{FOR}(P_S, P_T)) \\ & + \beta \cdot \mathrm{ETC}(P_T) \\ & + \sqrt{\frac{KL(q\|p) + \log \frac{2\sqrt{n}}{\delta}}{2n}}. \end{aligned} \tag{6}$$

Here, $\mathrm{FOR}(P_S, P_T)$ denotes the Feature Overlap Rate as defined in Equation (3), and $\mathrm{ETC}(P_T)$ is the Effective Task Complexity defined in Equation (5). The constant $\alpha = LL'B$ depends on the Lipschitz constant L of the loss ℓ, the Lipschitz constant L' of the hypothesis class $\mathcal{H}$ over $\phi(\mathcal{X})$, and the uniform upper bound B on the feature norm $\|\phi(x)\|$. The constant $\beta = C$ quantifies the effect of subtask entropy on variance-based risk inflation, and $KL(q\|p)$ is the Kullback–Leibler divergence between posterior q and prior p over the hypothesis class.

Proof. We begin by applying a standard PAC-Bayesian bound, which provides a high-probability control over the true risk in terms of the empirical risk and KL divergence. For any hypothesis h sampled from posterior q and any prior p, with probability at least $1 - \delta$ over a source sample of size n, the following holds:

$$R_T(f) \leq R_S(f) + \sqrt{\frac{KL(q\|p) + \log \frac{2\sqrt{n}}{\delta}}{2n}} + \epsilon,$$

where ϵ captures the discrepancy between the target distribution P_T and the source distribution P_S in terms of model performance.

To upper bound ϵ, we analyze how the feature representation ϕ aligns between source and target. For any $x \sim P_T$, we decompose the representation as:

$$\phi(x) = \phi_{\|}(x) + \phi_{\perp}(x),$$

where $\phi_{\|}(x)$ is the orthogonal projection of $\phi(x)$ onto the subspace $\mathcal{F}_S := \mathrm{span}\{\phi(x') : x' \sim P_S\}$, and $\phi_{\perp}(x)$ is the residual orthogonal component. Because h is optimized on $\mathcal{F}_S$, it can only act on $\phi_{\|}(x)$ and ignores $\phi_{\perp}(x)$ entirely. The resulting prediction discrepancy induces a loss difference:

$$\begin{aligned} &\left|\ell(h(\phi(x)), y) - \ell(h(\phi_{\|}(x)), y)\right| \\ &\qquad \leq L \cdot \left|h(\phi(x)) - h(\phi_{\|}(x))\right|. \end{aligned} \tag{7}$$

By assumption, h is L'-Lipschitz on $\mathbb{R}^d$, so:

$$\left|h(\phi(x)) - h(\phi_{\|}(x))\right| \leq L' \cdot \|\phi_{\perp}(x)\|,$$

and thus,

$$\left|\ell(h(\phi(x)), y) - \ell(h(\phi_{\|}(x)), y)\right| \leq LL' \cdot \|\phi_{\perp}(x)\|.$$

To quantify this residual norm, we use the definition of Feature Overlap Rate (FOR):

$$\mathrm{FOR}(P_S, P_T) := \mathbb{E}_{x \sim P_T}\left[\frac{\|\phi_{\|}(x)\|^2}{\|\phi(x)\|^2}\right].$$

This gives:

$$\begin{aligned} \mathbb{E}_{x \sim P_T}\left[\|\phi_{\perp}(x)\|\right] &= B \cdot \sqrt{1 - \mathrm{FOR}(P_S, P_T)} \\ &\leq B \cdot (1 - \mathrm{FOR}(P_S, P_T)). \end{aligned}$$

Transfer Pair	Source Task	Target Task	Type
MNLI → RTE	Natural Language Inference (3-class)	Natural Language Inference (binary)	Entailment
MNLI → STS-B	Natural Language Inference (3-class)	Semantic Similarity ([0, 5])	Cross-type
QQP → MRPC	Paraphrase Detection (binary)	Paraphrase Detection (binary)	Matching
QQP → RTE	Paraphrase Detection (binary)	Natural Language Inference (binary)	Cross-type
SST-2 → STS-B	Sentiment Classification (binary)	Semantic Similarity ([0, 5])	Similarity
SST-2 → MRPC	Sentiment Classification (binary)	Paraphrase Detection (binary)	Matching

Table 1: Source-target transfer pairs used in our experiments.

Taking square roots and applying Jensen's inequality (since $\sqrt{\cdot}$ is concave), and using $\|\phi(x)\| \leq B$, we obtain:

$$\begin{aligned}\mathbb{E}_{x\sim P_T}\left[\|\phi_\perp(x)\|\right] &\leq \sqrt{\mathbb{E}_{x\sim P_T}\left[\|\phi_\perp(x)\|^2\right]} \\ &\leq B\cdot\sqrt{1-\mathrm{FOR}(P_S,P_T)} \\ &\leq B\cdot(1-\mathrm{FOR}(P_S,P_T)).\end{aligned}$$

This completes the bound on the residual norm using concavity and feature norm boundedness.

Combining this with the previous inequality yields:

$$\begin{aligned}\epsilon &\leq LL'B\cdot(1-\mathrm{FOR}(P_S,P_T)) \\ &= \alpha\cdot(1-\mathrm{FOR}(P_S,P_T)).\end{aligned}$$

Next, we account for the internal heterogeneity of the target task. Let $P_T=\sum_{i=1}^K \pi_i P_i$ be a mixture of K latent subtasks. Following prior work on multitask and domain mixture learning, the variance of risk under this mixture is upper bounded by the entropy of the subtask weights:

$$\mathrm{Var}_{(x,y)\sim P_T}[\ell(h(\phi(x)),y)] \lesssim C\cdot\mathrm{ETC}(P_T),$$

where $\mathrm{ETC}(P_T)=\sum_{i=1}^K \pi_i\log\frac{1}{\pi_i}$ and C is a constant determined by the variability of each subtask. This structural term reflects the increased uncertainty a model faces when needing to simultaneously cover multiple diverse subtasks. While this variance does not directly appear in the PAC-Bayes expression, we heuristically incorporate it as an additive regularizer, modeling the risk inflation caused by subtask diversity. This approximation is consistent with prior practice in multitask and domain adaptation theory.

We then add this structural penalty to the risk, yielding the final bound:

$$\begin{aligned}R_T(f)\leq R_S(f) \quad &+\alpha\cdot(1-\mathrm{FOR}(P_S,P_T)) \\ &+\beta\cdot\mathrm{ETC}(P_T) \\ &+\sqrt{\frac{KL(q\|p)+\log\frac{2\sqrt{n}}{\delta}}{2n}}.\end{aligned}$$

where $\beta=C$. This completes the proof. □

Remark 1. *For simplicity we present the feature–overlap term in the linear form $(1-\mathrm{FOR}(P_S,P_T))$, while the tighter bound involves $\sqrt{1-\mathrm{FOR}(P_S,P_T)}$. This simplification preserves the qualitative dependence and does not affect our main conclusions.*

Scale-Aware Generalization

We now extend Theorem 1 to incorporate the role of model capacity. While Theorem 1 highlights the impact of structural mismatch on transfer risk, it does not account for how overparameterized models can attenuate this impact. The following result reformulates the bound to explicitly reflect the regularization effect of scale.

Corollary 1. *Let the model $f_\theta = h_\theta\circ\phi$ be parameterized by $\theta\in\mathbb{R}^m$ with norm $\|\theta\|$, and assume the posterior variance and functional expressivity scale with model size. Then under the same conditions as Theorem 1, the target risk satisfies:*

$$\begin{aligned}R_T(f_\theta)\leq R_S(f_\theta) \quad &+\frac{\alpha}{\|\theta\|^\delta}\cdot(1-\mathrm{FOR}(P_S,P_T)) \\ &+\frac{\beta}{\|\theta\|^\delta}\cdot\mathrm{ETC}(P_T) \\ &+\sqrt{\frac{KL(q\|p)+\log\frac{2\sqrt{n}}{\delta}}{2n}}.\end{aligned}$$

for some $\delta>0$ representing the regularization effect of scale on structural penalties.

We assume that the constants α and β, which encapsulate the effective Lipschitz sensitivity and risk variance, are attenuated by model scale. This reflects that larger models—with greater posterior expressivity and lower functional uncertainty—can better interpolate orthogonal features and stabilize predictions across subtasks, aligning with observed scale laws in deep learning.

As the model size $\|\theta\|$ increases, the effective impact of structural factors—feature misalignment and task entropy—diminishes. This reflects the empirical observation that larger models are more capable of interpolating residual directions and adapting to heterogeneous subtasks. The PAC-Bayes complexity term remains unaltered, highlighting that capacity reduces alignment penalties without weakening generalization control.

Remark 2. *Corollary 1 provides a scale-adaptive generalization bound: it demonstrates how increasing model capacity suppresses the adverse effects of structural misalignment, reinforcing the decomposition established in Theorem 1. This complements our earlier motivation in the introduction regarding scale-dependent generalization dynamics.*

Transfer Pair	$R_T(f)$	FOR	ETC	KL	MMD	W_2
QQP → MRPC	0.209	0.78	0.89	0.39	0.14	0.28
MNLI → RTE	0.275	0.68	1.04	0.47	0.22	0.31
SST-2 → MRPC	0.282	0.70	1.05	0.41	0.15	0.30
QQP → RTE	0.301	0.66	1.18	0.42	0.20	0.32
MNLI → STS-B	0.336	0.61	1.29	0.38	0.18	0.33
SST-2 → STS-B	0.348	0.63	1.31	0.50	0.21	0.35

Table 2: Transfer risk and explanatory variables across transfer pairs. Structural measures (FOR, ETC) show consistent alignment with $R_T(f)$, while traditional divergences (KL, MMD, W_2) do not.

Experiments

We now empirically investigate whether the two structural factors identified in our theoretical framework—Feature Overlap Rate (FOR) and Effective Task Complexity (ETC)—can explain the variability of transfer performance across tasks. Specifically, we test whether these quantities correlate with target risk $R_T(f)$, and compare them to standard divergence-based baselines. We also examine whether larger models attenuate the effect of structural mismatch, as predicted by our theory.

Experimental Setup

We construct six transfer pairs using three source tasks (MNLI, QQP, SST-2) and three target tasks (RTE, MRPC, STS-B) from the GLUE benchmark. This results in all non-identical source-target combinations across the three tasks, shown in Table 1. These tasks cover natural language inference, paraphrase detection, and semantic similarity, and differ in label format, semantic structure, and linguistic complexity.

For each pair, we fine-tune a pretrained BERT model (either base or large) on the source task and evaluate it zero-shot on the target development set. The prediction error is recorded as the transfer risk $R_T(f)$. Following GLUE conventions, we use classification error for RTE and MRPC, and $1-\rho$ for STS-B, where ρ is the Pearson correlation between predicted and ground truth similarity scores. No target supervision is used during training.

Each transfer direction reflects a different structural mismatch between source and target, such as label space discrepancy or representation geometry divergence.

Estimation of Structural Quantities

We compute FOR and ETC directly from the encoder representations of the pretrained model after source-task finetuning. Neither metric uses any target labels.

Feature Overlap Rate (FOR) is estimated by applying PCA to the source task embeddings and projecting target embeddings onto the retained subspace. We retain principal components that explain 95% of the variance, a standard cutoff that balances dimensionality reduction with information preservation. This measure is also invariant to linear scaling of the feature space, ensuring robustness with respect to feature magnitude. A higher FOR indicates stronger alignment between source and target feature geometries.

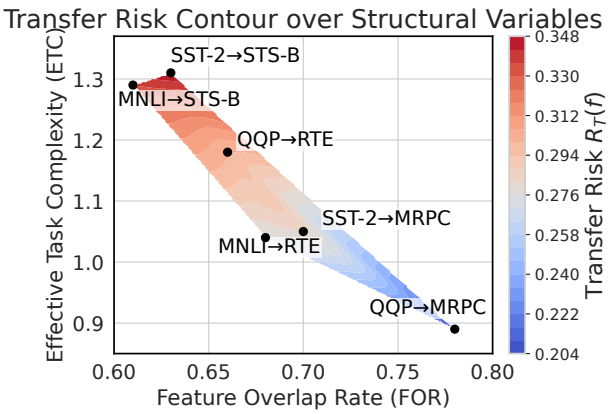

Figure 3: Transfer risk as a 2D surface over FOR and ETC. Risk increases with lower FOR and higher ETC.

Effective Task Complexity (ETC) is estimated by applying K-means clustering to the target task representations, then computing the entropy over cluster proportions. In our experiments, we set K to match the number of target task labels for each dataset, which provides a natural alignment with the task structure and avoids introducing extra hyperparameters. Higher ETC values indicate more diffuse or multimodal structure in the target task, which corresponds to a more complex hypothesis class.

Both quantities are computed from the penultimate-layer embeddings (i.e., encoder output $\phi(x)$) and are invariant to the classifier head h. They reflect structural properties of the transfer pair without requiring labels or optimization on the target task.

Main Result

We present our core empirical findings regarding the explanatory power of the proposed structural quantities—Feature Overlap Rate (FOR) and Effective Task Complexity (ETC)—on transfer performance.

Structural Correlates of Transfer Risk. We investigate how well the proposed metrics—Feature Overlap Rate (FOR) and Effective Task Complexity (ETC)—align with transfer performance. As shown in Table 2, transfer risk $R_T(f)$ increases with ETC and decreases with FOR. For example, SST-2 → STS-B and MNLI → STS-B both have high ETC and low FOR, resulting in the highest risks. In

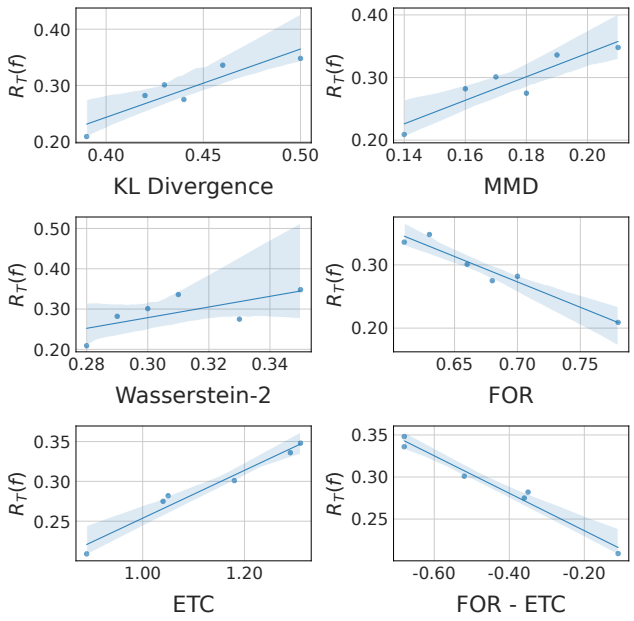

Figure 4: Scatter plots of transfer risk vs. each metric. FOR and ETC show strong trends; KL, MMD, and W_2 are weaker.

Model	Feature	R^2	Coeff.
KL-only	KL	0.11	+0.28
MMD-only	MMD	0.07	+0.22
W_2-only	W_2	0.09	+0.25
FOR-only	FOR	0.71	–0.44
ETC-only	ETC	0.63	+0.39
FOR + ETC	FOR, ETC	0.83	–0.42 / +0.35

Table 3: Regression results: FOR reduces risk (negative), ETC increases it. Structural predictors outperform classical divergences.

Pair	R_T^{base}	R_T^{large}	ΔFOR	ΔETC
QQP→MRPC	0.209	0.188	+0.01	–0.03
MNLI→RTE	0.275	0.241	+0.01	–0.04
SST-2→MRPC	0.282	0.249	+0.01	–0.04
QQP→RTE	0.301	0.268	+0.02	–0.05
MNLI→STS-B	0.336	0.303	+0.02	–0.05
SST-2→STS-B	0.348	0.309	+0.02	–0.05

Table 4: Effect of scaling from BERT-base to BERT-large: risk decreases and ETC impact weakens.

contrast, QQP → MRPC shows low ETC and high FOR with the lowest risk, supporting our hypothesis that geometric misalignment and task complexity jointly affect transfer effectiveness. In comparison, classical distribution-level divergences such as KL, MMD, and Wasserstein distance vary across pairs but exhibit no consistent trend with $R_T(f)$, suggesting that they fail to capture key structural factors.

This joint structural effect is further illustrated in Figure 3, where we visualize transfer risk as a 2D surface over the FOR–ETC plane. The contour highlights that risk increases with lower FOR and higher ETC—providing geometric support for our main hypothesis.

Figure 4 compares transfer risk against each individual metric. While KL, MMD, and W_2 show weak correlation, FOR and ETC both exhibit strong alignment with risk, reinforcing their explanatory value.

Regression Analysis. We fit linear regression models predicting $R_T(f)$ using each metric. As shown in Table 3, structural predictors (FOR, ETC) explain much higher variance than baseline divergences. Notably, FOR has a negative coefficient (higher FOR lowers risk), while ETC has a positive one.

Scaling Effect. Finally, Table 4 compares BERT-base and BERT-large. As model size grows, transfer risk drops consistently, while the marginal effect of ETC shrinks. This supports our theory that larger models can absorb more task complexity.

Summary. These results validate the structural framework: high FOR and low ETC predict low transfer risk, and their combined effect significantly outperforms traditional divergence metrics. Moreover, the sensitivity of risk to these structural factors diminishes with model size, confirming the scaling effects predicted in theory.

Conclusion

This work introduces a structural framework for understanding transfer performance through two geometry-aware quantities: *Feature Overlap Rate* (FOR) and *Effective Task Complexity* (ETC). Departing from traditional divergence-based theories, we argue that successful transfer depends not merely on marginal input similarity, but on alignment in representation space and the internal structure of the target task. Our empirical results validate this hypothesis across six transfer scenarios on the GLUE benchmark. We demonstrate that FOR and ETC jointly explain over 80% of the variance in transfer risk, significantly outperforming classical divergence measures such as KL, MMD, and Wasserstein distance. A contour-based visualization further reveals a smooth interaction surface over the FOR–ETC plane, reinforcing their combined predictive power. Moreover, we find that model scaling mitigates the impact of ETC, aligning with theoretical expectations. These findings suggest that structural metrics offer a principled lens for diagnosing and guiding transfer learning, with potential applications in task selection, curriculum design, and cross-domain adaptation. Future research may explore tighter information-theoretic formulations, extend the metrics to multi-modal or sequence-level tasks, and investigate their role in few-shot and continual learning settings. Ultimately, we hope this framework inspires a deeper understanding of what governs transferability—moving beyond black-box adaptation and toward structure-aware, interpretable transfer systems.

References

Aghajanyan, A.; Gupta, S.; and Zettlemoyer, L. 2021. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7319–7328. Association for Computational Linguistics.

Bahri, Y.; Kadavath, S.; Ganguli, S.; and Kaplan, J. 2021. Explaining Neural Scaling Laws. arXiv:2102.06701.

Baxter, J. 2000. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, 12: 149–198.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2010. A Theory of Learning from Different Domains. *Machine Learning*, 79(1–2): 151–175.

Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. 2020. Invariant Rationalization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119. PMLR.

D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M. D.; et al. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395.

Fort, S.; Hu, Z.; and Lakshminarayanan, B. 2019. Deep Ensembles: A Loss Landscape Perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Germain, P.; Lacasse, A.; Laviolette, F.; and Marchand, M. 2009. PAC-Bayesian Learning of Linear Classifiers. *Journal of Machine Learning Research*, 10: 2197–2232.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2016. Domain Adaptation with Conditional Transferable Components. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48, 2839–2848. PMLR.

Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M.; Yang, Y.; Zhou, Y.; Cipar, J.; et al. 2017. Deep Learning Scaling is Predictable, Empirically. *arXiv preprint arXiv:1712.00409*.

Kaplan, J.; Henighan, T.; Martinez, J.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; McCandlish, S.; Radford, A.; and Wu, J. 2020a. Scaling Laws for Autoregressive Generative Modeling. *arXiv preprint arXiv:2001.08361*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020b. Scaling Laws for Neural Language Models. arXiv:2001.08361.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2008. Domain Adaptation with Multiple Sources. In *Advances in Neural Information Processing Systems 21 (NeurIPS)*, 1041–1048. Vancouver, Canada: Curran/MIT Press.

Morcos, A.; Barrett, D.; Rabinowitz, N.; and Botvinick, M. 2018. On the Importance of Single Directions for Generalization. In *6th International Conference on Learning Representations (ICLR)*.

Nguyen, C. V.; Hassner, T.; Seeger, M.; and Archambeau, C. 2020. LEEP: A New Measure to Evaluate Transferability of Learned Representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 7294–7305.

Pan, S. J.; and Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10): 1345–1359.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. arXiv:2103.00020.

Raghu, M.; Gilmer, J.; Yosinski, J.; and Sohl-Dickstein, J. 2017. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Reddi, S.; Kale, S.; and Kumar, S. 2018. On the Convergence of Adam and Beyond. In *6th International Conference on Learning Representations (ICLR)*.

Redko, I.; Courty, N.; Flamary, R.; and Tuia, D. 2017. Theoretical Analysis of Domain Adaptation with Optimal Transport. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*. Skopje, Macedonia: Springer.

Shen, J.; Qu, Y.; Zhang, W.; and Yu, Y. 2018. Wasserstein Distance Guided Representation Learning for Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4058–4065. New Orleans, LA: AAAI Press.

Sun, Q.; Kong, L.; Zhou, M.; and Liu, X. 2024. Transferability is Proportional to Representational Diversity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *8th International Conference on Learning Representations (ICLR)*.

Weiss, K.; Khoshgoftaar, T. M.; and Wang, D. 2016. A Survey of Transfer Learning. *Journal of Big Data*, 3(1): 9.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How Transferable Are Features in Deep Neural Networks? In *Advances in Neural Information Processing Systems*, volume 27.

Zhang, Y.; Geng, X.; and Zhou, Z.-H. 2019. A Review on Domain Adaptation Without Target Labels. *Frontiers of Computer Science*, 13(6): 1020–1040.