BEYOND INDEPENDENCE: MODELING CORRELATED ADVERSARIAL TRANSFER IN MODEL ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Ensembles are widely believed to resist black-box transfer attacks: if adversarial examples transfer imperfectly, compromising several members at once should be rare. This argument silently assumes that transfer events across members are *independent*. We test that assumption directly, treating transfer against an ensemble as a vector of correlated binary events and measuring per-model transfer rates, pairwise correlations, and the overdispersion of the fooled count for six-model ensembles on CIFAR-10, GTSRB and MNIST under ℓ_∞ PGD transfer from a ResNet surrogate. In the regime where the ensemble argument is actually invoked, at small budgets where per-model transfer is a few percent and majority compromise is meant to be rare, independence is rejected wherever the data are well powered, including by an exact permutation test ($p < 5 \times 10^{-5}$): the observed rate of compromising a majority exceeds the independence prediction by more than an order of magnitude ($28\times$ on CIFAR-10, $615\times$ on GTSRB, $1842\times$ on MNIST at the smallest budget with events). As the budget rises and transfer saturates the gap closes and eventually reverses, exactly as it must; the danger zone for a defender is the low-budget corner, and that is where independence fails hardest. The failure is correctable with a single parameter: an exchangeable beta-binomial calibrated on the measured overdispersion matches the observed risk in every well-powered configuration, where independence fails in all of them, and, fit at one ensemble size, predicts majority-vote risk at other sizes to within a percentage point. The dependence is not an artifact of our training pipeline: it survives replacing distilled students with independently trained ones and the teacher surrogate with an unrelated model, and it persists under three transfer-optimized attacks. It also survives the standard fix: adversarial training lowers how often members are fooled but not how correlated their failures are, so the residual risk must still be read through the dependence structure. Ensemble robustness is governed by the dependence structure of failures, not by per-model transfer rates alone.

1 INTRODUCTION

Deep networks are vulnerable to adversarial examples, imperceptible perturbations that flip a prediction (Szegedy et al., 2013; Goodfellow et al., 2014), and these perturbations often *transfer*: an example crafted on one model fools another with different architecture and parameters (Papernot et al., 2016a; Liu et al., 2016). Transfer is what makes black-box attacks practical, since an adversary attacks a surrogate and deploys the result against a target it never sees (Papernot et al., 2017).

The standard defense against this is an *ensemble*, and its appeal rests on a single probabilistic intuition. If one model is fooled by a transferred example with probability p, then fooling a majority of m models requires many unlikely events to coincide, so majority compromise should be exponentially rarer than p and should fall off sharply as m grows. This reasoning underlies ensemble adversarial defenses (Pang et al., 2019; Kariyappa & Qureshi, 2019; Yang et al., 2020; Tramèr et al., 2017a), which build member diversity precisely so that no single perturbation defeats the whole ensemble. **The intuition is not a theorem but a calculation, and the calculation is valid only if the members fail independently.** That premise is never stated, never measured, and, we argue, false.

It should be suspected on its face. Every leading account of *why* adversarial examples transfer says that models trained on the same data fail on the *same inputs*: they share adversarial subspaces

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Tramèr et al., 2017b), aligned gradients (Demontis et al., 2019), and non-robust features (Ilyas et al., 2019). Independence is exactly what these mechanisms deny. Yet the two literatures that could have caught this both look elsewhere. The transferability literature reports surrogate-to-target success rates and works to make transfer stronger (Dong et al., 2018; Xie et al., 2019), while the ensemble-defense literature reports aggregate accuracy. Neither measures the one quantity that decides whether the ensemble argument holds: how often different members fail on the same input.

This paper measures that quantity, in the regime where the ensemble defense is used. We treat transfer against an m-model ensemble as a vector of correlated binary events and measure the per-model transfer rate $\bar{p}$, the mean pairwise failure correlation $\bar{\rho}$, and the overdispersion D of the fooled count relative to independence, on CIFAR-10, GTSRB and MNIST with six independently trained students each, sweeping ϵ from $1/255$ to $16/255$. The defense is only invoked where per-model transfer is small and majority compromise is *supposed* to be negligible, so that low-budget regime, not the saturated one where every member is already fooled, is where the assumption must be tested. Our findings are fourfold. *(i)* Independence fails worst exactly where it is trusted: at the smallest budget with majority-level failures it underestimates the true rate by $28\times$ (CIFAR-10), $615\times$ (GTSRB) and $1842\times$ (MNIST), and an assumption-free permutation test rejects it at $p < 5 \times 10^{-5}$; as transfer saturates the gap closes and reverses, confining the error to the rare-event corner. *(ii)* A one-parameter exchangeable beta-binomial (Williams, 1975; Crowder, 1978), calibrated on D and never refitted, matches the observed rate in every well-powered configuration and predicts majority-vote risk at other ensemble sizes to within a percentage point. *(iii)* The dependence is a property of the problem, not our pipeline: it survives replacing distilled students with from-scratch ones, replacing the teacher surrogate with an unrelated model, and switching to three transfer-optimized attacks (MI-FGSM, DI^2-FGSM, TI-FGSM). *(iv)* It survives the standard fix: adversarial training lowers how often members are fooled but not how correlated their failures are, so the residual risk must still be read through D. Ensemble robustness is governed by the dependence structure of failures, not by per-model transfer rates alone.

2 Related Work

Transferability. Adversarial transferability is a central property enabling black-box attacks: perturbations generated on a surrogate model can remain effective against unseen target models (Papernot et al., 2016a; 2017). Early work established transfer across architectures and training procedures (Liu et al., 2016), while subsequent methods substantially increased transfer success through momentum-based optimization (Dong et al., 2018), input diversity (Xie et al., 2019), and feature-level objectives (Huang et al., 2019). More recent work has continued to improve transfer by exploiting model structure, intermediate representations, and input or model diversity. For example, transferable attacks have been specifically adapted to Vision Transformers (Naseer et al., 2021; Wei et al., 2022), while later methods exploit path augmentation (Zhang et al., 2023), model alignment (Ma et al., 2024), and diverse surrogate ensembles (Tang et al., 2024; Fang et al., 2024). Recent work has also demonstrated that diversity across checkpoints of a single training trajectory can provide a strong source of transferability without requiring independently trained surrogate architectures (Li et al., 2025). A recent survey provides a comprehensive synthesis of this rapidly expanding transferability literature (Gu et al., 2023).

The focus of this literature is primarily the *marginal probability* that a perturbation transfers from a source to an individual target model, or on mechanisms for increasing that probability. Our question is different: given one transferred perturbation and an ensemble of targets, *how statistically dependent are the target failures?* Pairwise source-to-target transfer rates do not determine the probability that several targets fail simultaneously. Our analysis therefore treats transfer against an ensemble as a joint distribution of binary failure events rather than as a collection of independent source-to-target success rates.

Why transfer happens. Several lines of work provide mechanisms that naturally imply dependence among transfer events. Tramèr et al. (2017b) showed that models can share substantial adversarial subspaces, while Demontis et al. (2019) connected transferability to gradient alignment and model complexity. The shared-feature hypothesis of Ilyas et al. (2019) further suggests that models trained on the same data distribution can rely on common non-robust features, creating vulnerabilities that are shared across models.

More recent studies have provided additional evidence that transferability is associated with similarity in the representations and local geometry learned by different models. Wiedeman & Wang (2022) found that models trained on the same task can extract highly correlated features even when their parameters differ substantially, and showed that feature decorrelation can reduce adversarial transferability. Yang et al. (2021) theoretically related adversarial transferability to gradient diversity and model smoothness, showing that gradient orthogonality alone is insufficient to guarantee low transferability. Qin et al. (2023) subsequently proposed a transferability metric based on the singular structure of model Jacobians and studied its relationship to ensemble robustness; like the other works in this line, however, it characterizes a property of the *models* (their Jacobians) rather than the joint distribution of their failures. More recently, Ma et al. (2024) showed that deliberately aligning independently trained models can increase the transferability of adversarial perturbations, providing further evidence that the geometric relationship between models affects transfer.

These results motivate the dependence hypothesis tested in this work. However, existing analyses generally measure or regularize *model similarity*, for example gradients, features, Jacobians, or decision boundaries, as a proxy for transferability. We instead measure the quantity that directly enters ensemble risk: the empirical dependence between the *failure events themselves*. This distinction is important because model-level similarity neither necessary nor sufficient to specify the joint probability of simultaneous failures: two models with nearly aligned gradients may still fail on disjoint inputs, and two with near-orthogonal gradients may fail together, so a scalar similarity between models does not pin down the joint law of their failures that governs ensemble risk.

Ensembles and diversity. Ensembles have long been used to improve predictive accuracy and uncertainty estimation (Dietterich, 2000; Lakshminarayanan et al., 2017). In adversarial robustness, a substantial body of work constructs ensembles whose members are intentionally diversified so that a transferred perturbation is less likely to fool all members simultaneously. DVERGE (Yang et al., 2020) diversifies non-robust vulnerabilities across ensemble members, while diversified simultaneous training (Huang et al., 2021) encourages disagreement in high-level representations and gradients. Bui et al. (2021) explicitly regularize transferability between members through secure and insecure sets, and Shukla et al. (2025) introduce a pairwise adversarially robust loss that directly penalizes shared vulnerability between every pair of members.

The relationship between diversity and transferability has subsequently received more formal treatment. TRS (Yang et al., 2021) derives conditions relating gradient diversity and model smoothness to transferability and trains ensembles to reduce transfer between members. Zhang et al. (2022) formulate robust ensemble construction through margin boosting, while Qin et al. (2023) propose a Jacobian-based transferability measure for explicitly isolating transfer between ensemble submodels. Recent work has continued to exploit diversity as a defense mechanism, including methods based on differentiated ensemble predictions (Zhang et al., 2023) and negative correlation among ensemble members (Zhao et al., 2025). Conversely, Tang et al. (2024) demonstrate that ensemble diversity can be exploited by the *attacker*: diverse surrogate models can improve black-box transferability when their perturbations are appropriately combined.

Most closely related to our perspective, Yao et al. (2024) decompose transferable-attack error into a vulnerability term and an ensemble-diversity term across 54 models. Their decomposition, however, serves the *attacker*: it explains and improves how a diverse surrogate ensemble transfers onto a target. We ask the *defender's* question on the other side of the same phenomenon: given a fixed target ensemble and attack, what is the joint distribution of member failures, and by how much does it depart from the independence model used to argue that majority compromise is rare? This growing literature establishes that the dependence and diversity of ensemble members matter, but it generally uses diversity or transferability as a *training objective or proxy*. Our contribution is therefore complementary to diversity-based ensemble training: rather than proposing another mechanism to reduce correlation, we quantify the correlation that remains and determine its consequences for ensemble risk.

Adversarial training and distillation. Adversarial training remains a central approach for improving robustness to both white-box and transferred attacks (Madry et al., 2017; Tramèr et al., 2017a). Subsequent work has studied how adversarial training changes representation geometry and transferability. For example, Terzi et al. (2021) showed that adversarial training changes the information content of learned representations and can affect transferability, while Wiedeman & Wang

(2022) demonstrated that decorrelating representations can reduce transfer between independently trained models. In the ensemble setting, adversarial training has also been combined with explicit mechanisms for promoting diversity or suppressing transferability (Bui et al., 2021; Yang et al., 2021; Zhang et al., 2022). Consistent with this line, we find (Section 5.7) that adversarial training sharply lowers the marginal transfer rate but does not, on its own, remove the failure correlation wherever that correlation can still be measured.

Distillation has likewise been studied as a mechanism for transferring information between models, although its use as an adversarial defense has important limitations (Hinton et al., 2015; Papernot et al., 2016b; Carlini & Wagner, 2017). In our experiments, distillation is not proposed as a defense. Instead, it provides a controlled procedure for constructing multiple independently trained student models sharing a common task and teacher, allowing us to study whether correlated transfer failures arise from the training pipeline itself. Crucially, our experiments replace distilled students with independently trained models and replace the teacher surrogate with an unrelated model, while observing the same qualitative dependence structure. Thus, the phenomenon we study is not tied to a particular distillation procedure.

Our work differs from these defenses in its target of analysis. Existing methods generally seek to *reduce* transferability, increase member diversity, or improve aggregate robust accuracy. We instead quantify the statistical structure of transfer failures across ensemble members. In particular, we compare the observed majority-compromise probability against the independence prediction, measure pairwise failure correlations and count overdispersion, and use an exchangeable beta-binomial model to translate the measured dependence into calibrated majority-vote risk. This makes explicit the probabilistic assumption underlying ensemble robustness arguments and tests whether that assumption is empirically justified.

3 Transfer as a Vector of Correlated Events

The ensemble argument is a claim about a count: how many of the m members one transferred perturbation fools at once, and whether a majority is rare. That conclusion holds only if the count behaves like a sum of independent coin flips. We therefore treat it as a sum of m possibly-correlated binary failure events and measure their dependence directly. Because every statistic below is a function of the per-example failure vector, we retain that vector rather than only its summary rates.

Setup. Fix an ensemble of m target models and a surrogate. For an input x correctly classified by all targets, let x' be the adversarial example crafted on the surrogate, and define

$$X_i = \mathbf{1}\{\text{model } i \text{ misclassifies } x'\}, \qquad i = 1, \dots, m, \qquad S = \sum_{i=1}^{m} X_i.$$

Write $p_i = \Pr(X_i = 1)$ and $\bar{p} = \frac{1}{m}\sum_i p_i$. An ensemble requiring k agreeing members is compromised when $S \geq k$; unless stated otherwise we use the majority threshold $k = \lceil m/2 \rceil$, so $k = 3$ for $m = 6$. We also report *hard-vote* success, where the plurality label must differ from the ground truth, and *logit-averaging* success, where member logits are averaged before the argmax.

The independence baseline. If the X_i were independent, S would follow a Poisson–binomial law and

$$P_{\text{ind}}(S \geq k) = \sum_{s \geq k} \sum_{\substack{A \subseteq [m] \\ |A| = s}} \prod_{i \in A} p_i \prod_{i \notin A} (1 - p_i). \tag{1}$$

This is the quantity implicitly invoked whenever an ensemble is argued to be robust because its members are individually hard to fool. It depends only on the marginals, which is exactly why it is easy to compute and easy to get wrong.

Measuring dependence. We use two complementary statistics. The first is the mean pairwise Pearson correlation of transfer events, $\bar{\rho} = \binom{m}{2}^{-1} \sum_{i<j} \mathrm{Corr}(X_i, X_j)$, an average over the 15 member pairs of a six-model ensemble. The second is threshold-free and is the one our model

consumes: the *overdispersion ratio*

$$D = \frac{\mathrm{Var}(S)}{\sum_{i=1}^{m} p_i(1-p_i)}, \tag{2}$$

the observed variance of the fooled count relative to its variance under independence. $D = 1$ means independent failures; $D > 1$ means members fail together more than chance allows. For exchangeable members the two agree, $D \approx 1 + (m-1)\bar{\rho}$; empirically, on CIFAR-10 at $\epsilon = 1/255$, $\bar{\rho} = 0.258$ gives $1 + 5\bar{\rho} = 2.29$ against a measured $D = 2.29$. Note that D is size-dependent by construction while $\bar{\rho}$ is not; Section 5.5 exploits this.

A one-parameter correction. Independence is a point on a line, not the only tractable model. Let $\rho = (D-1)/(m-1)$ be the implied intraclass correlation and consider the exchangeable beta-binomial (Williams, 1975; Crowder, 1978), in which a per-example latent susceptibility $\pi \sim \mathrm{Beta}(\alpha, \beta)$ is drawn with $\alpha = \bar{p}(1-\rho)/\rho$ and $\beta = (1-\bar{p})(1-\rho)/\rho$, and $S \mid \pi \sim \mathrm{Binomial}(m, \pi)$. Then

$$P_{\mathrm{BB}}(S \geq k) = \sum_{s \geq k} \binom{m}{s} \frac{B(s+\alpha,\, m-s+\beta)}{B(\alpha, \beta)}, \tag{3}$$

which recovers Eq. 1 as $\rho \to 0$. The interpretation is deliberately literal: some inputs are intrinsically more vulnerable across the whole ensemble than others, and ρ measures how much. This adds exactly one parameter over independence, and that parameter is already available from any evaluation that records per-example outcomes.

Statistical protocol. Ensemble-level failures are rare events, and rare-event estimates are easy to over-read. We report 95% Wilson intervals (Wilson, 1927) on observed rates; test independence both parametrically (exact one-sided binomial against $N \cdot P_{\mathrm{ind}}$) and non-parametrically (permutation, Section 5.6); bootstrap $\bar{\rho}$ and D over examples; and compare standard against robust ensembles with Fisher's exact test. We flag every configuration with fewer than 10 ensemble-level events as underpowered and draw no conclusions from it.

4 Experimental Setup

Benchmarks. We evaluate three tasks that span the label space and the visual-confusability axis. **CIFAR-10** (Krizhevsky et al., 2009), ten classes at 32×32, teacher ResNet-50 (He et al., 2016). **GTSRB** (Stallkamp et al., 2012), 43 traffic-sign classes at 32×32, teacher ResNet-50; students are trained as full 43-class classifiers. **MNIST** (LeCun et al., 1998), ten digit classes, teacher ResNet-50. Every evaluation set passes the four data-quality checks of Appendix D (full label coverage, no degenerate members, an unbiased shuffled single-pass scan, and no repeated images); the audit is not decorative—an earlier iteration of this study was invalidated by a label-ordered loader that wrapped past the test set and by a collapsed distillation run that produced a constant-classifier member, and Appendix D documents both failure modes and the checks that now catch them. Inputs are in $[0, 1]$ with each model's training-time preprocessing.

Models. For each teacher we train $m = 6$ compact CNN students by knowledge distillation (Hinton et al., 2015), independently initialized and trained. Robust students are produced by PGD adversarial training at $\epsilon = 8/255$ (Madry et al., 2017). To rule out distillation as the source of the measured dependence, we additionally train, for each benchmark, six students *from scratch* on labels (disjoint seeds) and an independent third-model surrogate of a different architecture (Section 5.3).

Evaluation set. Adversarial examples are generated on, and evaluated over, images correctly classified by all six target students, drawn by a shuffled single-pass scan and capped at $2{,}000$ eligible images; the cap binds well before the scan exhausts the test set on every benchmark ($\texttt{total_seen} \leq 2304 \ll$ the test-set size), so the retained set is a sample and not a truncation.

Threat model and attack. Black-box transfer: adversarial examples are crafted on the surrogate and evaluated on the targets without access to their gradients. Attacks are ℓ_∞ PGD with 20 steps and step size $2/255$, at $\epsilon \in \{1, 2, 4, 8, 16\}/255$. Section 5.3 adds MI-FGSM, DI^2-FGSM and TI-FGSM. Robust ensembles are attacked from the same standard surrogate, so standard and robust settings differ only in the targets.

Benchmark	ϵ	$\bar{p}$ (%)	$\bar{\rho}$	D	Obs. $P(S{\geq}3)$ (%)	Indep. (%)	Beta-bin. (%)	indep. test
CIFAR-10 $N{=}2000$	1	3.533	+0.258	2.286	**2.300** [1.73, 3.05]	0.0807	2.394	$p{=}8{\times}10^{-50}$
	2	13.250	+0.448	3.237	13.200 [11.79, 14.75]	3.4001	13.162	$p{=}5{\times}10^{-77}$
	4	37.625	+0.557	3.785	41.150 [39.01, 43.32]	40.6499	40.592	$p{=}0.33$
	8	76.242	+0.531	3.658	81.400 [79.64, 83.04]	96.8695	81.193	$p{=}1.00$
	16	94.617	+0.380	2.897	97.350 [96.55, 97.97]	99.9886	97.319	$p{=}1.00$
GTSRB $N{=}2000$	1	0.883	+0.516	3.567	0.800 [0.49, 1.30]	0.0013	0.852	$p{=}2{\times}10^{-39}$
	2	3.492	+0.644	4.203	3.650 [2.91, 4.56]	0.0777	3.596	$p{=}1{\times}10^{-93}$
	4	13.067	+0.771	4.852	**13.500** [12.07, 15.07]	3.2713	13.658	$p{=}2{\times}10^{-84}$
	8	29.708	+0.783	4.916	31.200 [29.21, 33.26]	25.0161	31.011	$p{=}3{\times}10^{-10}$
	16	50.383	+0.777	4.887	52.750 [50.56, 54.93]	66.3493	52.352	$p{=}1.00$
MNIST $N{=}2000$	1	0.017	+1.000	2.001	0.000† [0.00, 0.19]	0.0000	0.008	—
	2	0.017	+1.000	2.001	0.000† [0.00, 0.19]	0.0000	0.008	—
	4	0.100	+0.537	3.666	0.150† [0.05, 0.44]	0.0000	0.097	$p{=}9{\times}10^{-15}$
	8	0.525	+0.666	4.233	**0.500** [0.27, 0.92]	0.0003	0.538	$p{=}6{\times}10^{-30}$
	16	18.892	+0.703	4.500	20.400 [18.69, 22.22]	8.4051	19.876	$p{=}3{\times}10^{-62}$

Table 1: Teacher-surrogate transfer against $m{=}6$ standard-trained student ensembles, $k{=}3$. Observed rates carry 95% Wilson intervals; † marks configurations with fewer than 10 ensemble-level failures, which are underpowered. Bold marks the smallest well-powered budget on each benchmark, the rare-event operating point where the ensemble argument is invoked. At the smallest budget that produces any majority events, independence underestimates the true rate by $28\times$ on CIFAR-10 ($\epsilon{=}1$), $615\times$ on GTSRB ($\epsilon{=}1$), and $1842\times$ on MNIST ($\epsilon{=}8$). "Indep." is the exact Poisson–binomial prediction (Eq. 1); "Beta-bin." is the one-parameter correction (Eq. 3). The last column gives an exact one-sided binomial test of the independence null, omitted where no events occur. Adversarially trained counterparts appear in Table 3.

5 Results

5.1 Independence underestimates ensemble risk in the rare-event regime

Table 1 contrasts the observed $P(S \geq 3)$ with the independence prediction of Eq. 1. At the low budgets where the ensemble argument is invoked, the observed rate exceeds it by more than an order of magnitude on every benchmark: $28\times$ on CIFAR-10 ($\epsilon = 1$), $615\times$ on GTSRB ($\epsilon = 1$), and $1842\times$ on MNIST at its first well-powered budget ($\epsilon = 8$). Exact binomial tests reject independence at $p < 10^{-29}$ there, and the permutation test of Section 5.6 at the floor $p < 5 \times 10^{-5}$.

The gap is largest exactly where a defender cares, and closes only when the defense is moot. As ϵ rises the per-model rate climbs toward one and majority compromise stops being rare: on CIFAR-10 the ratio falls from $28\times$ ($\epsilon = 1$) to unity ($\epsilon = 4$), and by $\epsilon = 8$ independence *over*-predicts (96.9% against an observed 81.4%), because at $\bar{p} = 0.76$ the Poisson–binomial already puts almost all its mass above threshold while correlation holds some inputs back. This crossover is the signature of the phenomenon, not a flaw in the measurement: the independence calculation is an argument about *rare* conjunctions, and its error is an order of magnitude precisely in the rare-event corner the argument reasons about. At $\epsilon \leq 2/255$ on MNIST fewer than three inputs transfer to any member; those rows are underpowered and we draw no conclusion from them.

5.2 A single dispersion parameter restores calibration

The independence gap is a missing parameter, not noise. Substituting the measured D into Eq. 3 yields predictions that track the observed rate at every budget on every benchmark (Figure 1): across all well-powered configurations the beta-binomial falls inside the 95% Wilson interval and independence falls inside none. The correction is not tuned to the threshold; it is fixed by the variance of the fooled count and then evaluated at $k = 3$.

A skeptic will note that D and $P(S \geq 3)$ are both functionals of the same joint distribution, so agreement at $m = 6$ is partly built in. The decisive test is extrapolation. Fixing $\bar{p}$ and D from the

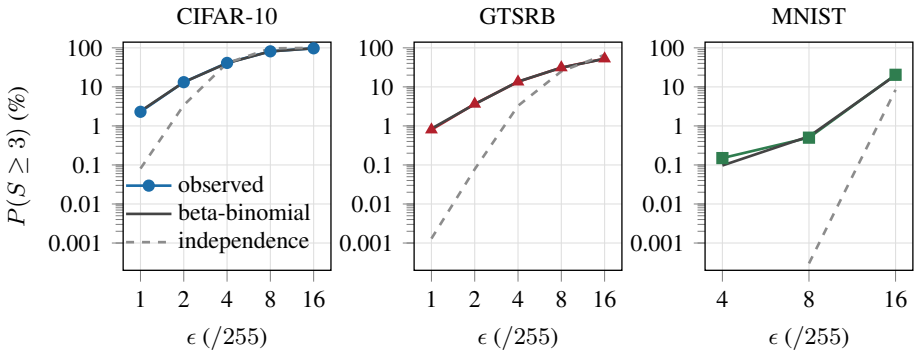


Figure 1: Observed $P(S \geq 3)$, the one-parameter beta-binomial prediction, and the Poisson–binomial independence prediction against the ℓ_∞ budget (log scale, standard training, $m = 6$). At low budgets independence sits one to three orders of magnitude below the observed rate while the beta-binomial tracks it; where transfer saturates ($\epsilon \geq 8$ on CIFAR-10) independence crosses above the observed rate, mispredicting in the other direction, and the beta-binomial continues to track. The per-model transfer rate $\bar{p}$ and mean pairwise correlation $\bar{\rho}$ that determine these curves are given per configuration in Table 1; underpowered points ($\epsilon \leq 2$ on MNIST) are omitted from the observed series.

Benchmark	configuration	$\bar{p}$ (%)	$\bar{\rho}$	D	Obs. (%)	Indep. (%)
CIFAR-10 ϵ=1/255	distilled, teacher surrogate	3.533	0.258	2.29	2.30	0.081
	from scratch, teacher surr.	3.310	0.183	1.92	1.35	0.067
	distilled, *indep. surrogate*	3.800	0.244	2.23	2.70	0.099
GTSRB ϵ=4/255	distilled, teacher surrogate	13.067	0.771	4.85	13.50	3.271
	from scratch, teacher surr.	10.000	0.719	4.58	10.10	1.560
	distilled, *indep. surrogate*	17.520	0.811	5.05	18.35	7.077

Table 2: The dependence survives removing distillation and removing the shared surrogate, shown at each benchmark's rare-event budget. From-scratch students (six independent label-trained models, no teacher) and an independent third-model surrogate both retain a large $\bar{\rho}$ and the independence underestimate.

six-model ensemble, we predict attack success for every smaller ensemble under its own majority rule ($m = 2, 3, 4, 5$) and compare against empirical rates over all size-m subsets (Appendix B). At the rare-event budgets the beta-binomial tracks the held-out rates while independence understates them by up to an order of magnitude. A defender can therefore measure dispersion once, on a small ensemble, and obtain calibrated risk for the size they deploy.

5.3 The dependence is not an artifact of distillation or the shared surrogate

A natural worry is that our students are correlated because they are distilled from one teacher, or attacked from that same teacher. Both are ruled out. Replacing the six distilled students with six trained *from scratch* on labels, with disjoint seeds and no teacher, leaves the correlation essentially unchanged ($\bar{\rho} = 0.183$ vs. 0.258 on CIFAR-10 at $\epsilon = 1/255$; 0.719 vs. 0.771 on GTSRB at $\epsilon = 4/255$), both far from the independence value of zero, and independence still underestimates majority compromise by more than an order of magnitude on GTSRB (10.10% observed vs. 1.56%). Replacing the teacher surrogate with an independently trained model of a different architecture, holding the distilled targets fixed, if anything *raises* the dependence ($\bar{\rho} = 0.244$ on CIFAR-10, 0.811 on GTSRB). The co-failure is a property of the data and the attack geometry, not of our pipeline: models that share a data distribution share the directions along which they are fragile, whether or not they share a teacher.

5.4 Dependence persists under transfer-optimized attacks

Our main attack is plain PGD, but the phenomenon is not specific to it. Under momentum (MI-FGSM), input-diversity (DI^2-FGSM) and translation-invariant (TI-FGSM) attacks crafted on the same surrogate and evaluated on the same targets, the dependence is of the same order and independence underestimates by the same margin. On CIFAR-10 at $\epsilon = 1/255$, $\bar{\rho}$ is 0.261 (MI), 0.309 (DI) and 0.171 (TI) against 0.258 for PGD; on GTSRB at $\epsilon = 4/255$ it is 0.789, 0.822 and 0.812 against 0.771. The transfer-optimized attacks raise the marginal rate somewhat, as intended, but they do not decorrelate the failures; the beta-binomial calibrated on each attack's own dispersion continues to match its observed majority rate.

5.5 How many members do we need?

$\bar{\rho}$ is an average over member pairs. Averaged over all size-m subsets of a fixed pool it is *identically* the pool's own $\bar{\rho}$ for every $m \geq 2$—a combinatorial identity, not an empirical finding. What is empirical is how the estimate behaves: the subset-to-subset standard deviation of $\bar{\rho}$ shrinks roughly as $1/m$ (Appendix B), so enlarging the ensemble buys precision on the dependence parameter and nothing else. The overdispersion D climbs with m only because $D = 1 + (m-1)\bar{\rho}$ by construction; the implied $\rho = (D-1)/(m-1)$ recovers $\bar{\rho}$ at every size. Six members are enough to estimate the parameter and small enough that majority-level events remain frequent enough to test; beyond six, $P(S \geq k)$ becomes unobservable at these sample sizes while the estimate barely tightens.

5.6 Distribution-free validation

Because we retain the per-example indicator matrix $Z \in \{0,1\}^{N \times 6}$, the central claim can be checked without any parametric assumption. Shuffling each column of Z independently (20,000 permutations) preserves every marginal p_i exactly while destroying all dependence, giving an exact finite-sample null for $P(S \geq 3)$. The test rejects independence at the resolution floor $p < 5 \times 10^{-5}$ at every well-powered rare-event budget—$\epsilon = 1/255$ on CIFAR-10, ϵ through $8/255$ on GTSRB, and $\epsilon = 8/255$ on MNIST—and, correctly, fails to reject once transfer saturates and the observed rate falls below independence ($p = 1.0$ at $\epsilon \geq 8/255$ on CIFAR-10 and $\epsilon = 16/255$ on GTSRB) and on the underpowered low-budget MNIST rows. Appendix C gives the full table with bootstrap intervals.

5.7 Adversarial training lowers the rate, not the dependence

Adversarial training is the strongest single-model defense, and the natural question is whether it also fixes the ensemble argument by making failures independent. It does not; it lowers the transfer rate so far that the question rarely arises (Table 3). PGD adversarial training collapses per-model transfer to near zero on all three benchmarks (e.g. $3.5\% \to 0.04\%$ on CIFAR-10 at $\epsilon = 1$), and majority compromise with it, Fisher-rejected at every budget. But where the robust ensemble still transfers enough to measure, the failures remain about as correlated as before: on GTSRB the robust $\bar{\rho}$ is 0.48, 0.64, 0.72 at $\epsilon = 4, 8, 16$ against the standard 0.77–0.78, and on MNIST at $\epsilon = 16$ it is unchanged; D stays well above 1 throughout, so the beta-binomial is still needed and independence still underestimates (GTSRB $\epsilon = 16$: 7.9% observed against 0.75%). Only CIFAR-10 appears decorrelated, and there the robust transfer rate is too low to produce ten majority events, so we read no dependence conclusion from it. Hardening and decorrelation are orthogonal levers: adversarial training shrinks how often members are fooled, but the inputs that still transfer are shared across them, so the residual risk must be read through D, not the independence formula.

6 Limitations

Rare-event power and modeling assumptions. The effect is a rare-event effect by construction, so majority-level events are scarce at the low budgets where independence fails hardest; we report 95% Wilson intervals, flag every configuration with fewer than ten events, and draw no conclusion from flagged rows. The dependence estimates $\bar{\rho}$ and D use all N examples and are unaffected. The beta-binomial assumes exchangeable members, which the diagnostics of Appendix support (no flagged outlier member; $1 + (m-1)\bar{\rho}$ reproduces D to within a few percent); a block-structured

Benchmark	ϵ	$\bar{p}$ std→adv (%)	$\bar{\rho}$ std→adv	D std→adv	$P(S{\geq}3)$ (%)		Fisher	verdict
					std	adv		
CIFAR-10	1	$3.53 \to 0.000$	$0.26 \to$ n/a	$2.29 \to 1.00$	2.30	$0.000^{\dagger}$	$2{\times}10^{-14}$	—
CIFAR-10	2	$13.25 \to 0.001$	$0.45 \to$ n/a	$3.24 \to 1.00$	13.20	$0.000^{\dagger}$	$6{\times}10^{-84}$	—
CIFAR-10	4	$37.62 \to 0.002$	$0.56 \to 0.03$	$3.79 \to 1.19$	41.15	$0.000^{\dagger}$	≈ 0	—
CIFAR-10	8	$76.24 \to 0.004$	$0.53 \to 0.11$	$3.66 \to 1.55$	81.40	$0.150^{\dagger}$	≈ 0	—
CIFAR-10	16	$94.62 \to 0.009$	$0.38 \to 0.27$	$2.90 \to 2.35$	97.35	0.700	≈ 0	adv. better
GTSRB	1	$0.88 \to 0.001$	$0.52 \to 0.10$	$3.57 \to 1.25$	0.80	$0.000^{\dagger}$	$3{\times}10^{-5}$	—
GTSRB	2	$3.49 \to 0.002$	$0.64 \to$ n/a	$4.20 \to 1.07$	3.65	$0.000^{\dagger}$	$1{\times}10^{-22}$	—
GTSRB	4	$13.07 \to 0.012$	$0.77 \to 0.48$	$4.85 \to 3.41$	13.50	1.200	$2{\times}10^{-57}$	adv. better
GTSRB	8	$29.71 \to 0.036$	$0.78 \to 0.64$	$4.92 \to 4.18$	31.20	3.800	$9{\times}10^{-128}$	adv. better
GTSRB	16	$50.38 \to 0.076$	$0.78 \to 0.72$	$4.89 \to 4.58$	52.75	7.900	$6{\times}10^{-227}$	adv. better
MNIST	16	$18.89 \to 0.002$	$0.70 \to 0.76$	$4.50 \to 4.67$	20.40	$0.200^{\dagger}$	$3{\times}10^{-125}$	—

Table 3: Effect of PGD adversarial training on the student ensemble, standard (std) versus robust (adv), same surrogate and evaluation protocol, $m{=}6$, $k{=}3$. Adversarial training collapses the per-model transfer rate $\bar{p}$ to near zero on all three benchmarks, and with it the majority-compromise rate. But where the robust ensemble still transfers enough to measure ($\bar{\rho}$, D columns), the failure correlation largely persists: on GTSRB at $\epsilon = 16/255$ the robust ensemble still has $\bar{\rho} = 0.72$ against the standard 0.78, and on MNIST $\bar{\rho}$ is unchanged. † marks robust-arm configurations with fewer than ten majority events, where the compromise rate is bounded by its Wilson interval but the dependence estimate rests on few events; the "verdict" column reports Fisher's exact test only where both arms are well powered. Robust rows where $\bar{\rho}$ is "n/a" have too few transferred inputs for a defined pairwise correlation.

model would suit a heterogeneous ensemble better, and the model targets threshold rules rather than logit averaging.

Attacks and architectures. The headline results use ℓ_∞ PGD; Section 5.4 adds three transfer-optimized attacks and finds the same dependence, but stronger or adaptive attacks (Croce & Hein, 2020; Tramer et al., 2020) may differ, and best practice argues for a broader suite (Carlini et al., 2019). Our six students per benchmark share an architecture family, which isolates dependence arising from the data and attack rather than from architectural identity but is also a best case for the ensemble argument; a heterogeneous-architecture study is the natural next step.

No defense baseline. We evaluate the measurement framework, not a new defense, and do not measure D for existing diversity-promoting defenses (Pang et al., 2019; Kariyappa & Qureshi, 2019; Yang et al., 2020). Doing so is the most direct application of this work and would test whether those methods move the quantity they are designed to move.

7 Conclusion

The ensemble defense against transfer attacks rests on an unstated premise, that member failures are independent, and that premise is false where the defense is used. At the low budgets where majority compromise is supposed to be negligible, it exceeds the independence prediction by $28\times$ on CIFAR-10, $615\times$ on GTSRB and $1842\times$ on MNIST, and a permutation test rejects independence at $p < 5 \times 10^{-5}$; the error closes only once transfer saturates and the ensemble is compromised anyway. One dispersion parameter, free from any evaluation that records per-example outcomes, repairs the estimate, and it is a property of the data and attack rather than the pipeline or ensemble size, so it transfers across sizes. We recommend reporting transfer correlation and overdispersion alongside attack success, and scoring diversity-promoting defenses on D directly: higher aggregate accuracy is consistent with both decorrelation and plain member hardening, and only the former delivers the multiplicative benefit ensembles are chosen for.

ETHICS STATEMENT

This work studies the failure modes of a defensive technique. Its offensive content is limited to standard, published attacks (PGD and three transfer-optimized variants) applied to public academic benchmarks; we introduce no new attack and provide no capability not already available. The practical effect of our findings is defensive: practitioners who estimate ensemble risk under an independence assumption are underestimating it in the regime that matters, and we give them a cheap correction. No dataset used contains personal or sensitive data. We report negative and boundary results that is, budgets at which independence cannot be rejected, budgets at which it reverses, and the data-quality failures our own audit caught rather than selecting favorable settings.

REPRODUCIBILITY STATEMENT

Section 4 specifies benchmarks, architectures, distillation and adversarial-training procedures, the shuffled single-pass evaluation filter, surrogates, and the full attack configuration. Section 3 gives closed-form definitions of every reported statistic (Eqs. 1–3) and states the statistical protocol, including exact tests and the 10-event power threshold. Table 1 reports per-configuration N, from which every rate converts back to an integer event count. Appendix C details the permutation and bootstrap procedures including resample counts and seed; Appendix D states the four data-quality criteria applied to every run and documents the two failure modes an earlier iteration exhibited. Code, per-example outcome matrices, and evaluation logs will be released.

AI USE STATEMENT

We used a large language model to edit prose. All experiments, training, and attacks were performed by the authors, who reviewed all AI-assisted output and take full responsibility for this work.

REFERENCES

Anh Tuan Bui, Trung Le, He Zhao, Paul Montague, Olivier deVel, Tamas Abraham, and Dinh Phung. Improving ensemble robustness by collaboratively promoting and demoting adversarial robustness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6831–6839, 2021.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Martin J Crowder. Beta-binomial anova for proportions. *Applied statistics*, pp. 34–37, 1978.

Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security 19)*, pp. 321–338, 2019.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193. IEEE, 2018.

Zhengwei Fang, Rui Wang, Tao Huang, and Liping Jing. Strong transferable adversarial attacks via ensembled asymptotically normal distribution learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24841–24850. IEEE, 2024.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Bo Huang, Zhiwei Ke, Yi Wang, Wei Wang, Linlin Shen, and Feng Liu. Adversarial defence by diversified simultaneous training of deep ensembles. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7823–7831, 2021.

Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4733–4742, 2019.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Sanjay Kariyappa and Moinuddin K Qureshi. Improving adversarial robustness of ensembles with diversity training. *arXiv preprint arXiv:1901.09981*, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Shixin Li, Chaoxiang He, Xiaojing Ma, Bin Benjamin Zhu, Shuo Wang, Hongsheng Hu, Dongmei Zhang, and Linchen Yu. Enhancing adversarial transferability with checkpoints of a single model's training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20685–20694. IEEE, 2025.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Avery Ma, Amir-massoud Farahmand, Yangchen Pan, Philip Torr, and Jindong Gu. Improving adversarial transferability via model alignment. In *European conference on computer vision*, pp. 74–92. Springer, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*, 2021.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pp. 4970–4979. PMLR, 2019.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016a.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pp. 582–597. IEEE, 2016b.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Ruoxi Qin, Linyuan Wang, Xuehui Du, Shuxiao Ma, Xingyuan Chen, and Bin Yan. An adversarial transferability metric based on svd of jacobians to disentangle the correlation with robustness. *Applied Intelligence*, 53(10):11636–11653, 2023.

Shubhi Shukla, Subhadeep Dalui, Manaar Alam, Shubhajit Datta, Arijit Mondal, Debdeep Mukhopadhyay, and Partha Pratim Chakrabarti. Guardian of the ensembles: Introducing pairwise adversarially robust loss for resisting adversarial attacks in dnn ensembles. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7205–7214. IEEE, 2025.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Bowen Tang, Zheng Wang, Yi Bin, Qi Dou, Yang Yang, and Heng Tao Shen. Ensemble diversity facilitates adversarial transferability. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24377–24386. IEEE, 2024.

Matteo Terzi, Alessandro Achille, Marco Maggipinto, and Gian Antonio Susto. Adversarial training reduces information and improves transferability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2674–2682, 2021.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017a.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017b.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2668–2676, 2022.

Christopher Wiedeman and Ge Wang. Disrupting adversarial transferability in deep neural networks. *Patterns*, 3(5), 2022.

DA Williams. 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, pp. 949–952, 1975.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2725–2734. IEEE, 2019.

Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawhich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, and Hai Li. Dverge: diversifying vulnerabilities for enhanced robust generation of ensembles. *Advances in Neural Information Processing Systems*, 33: 5505–5515, 2020.

Zhuolin Yang, Linyi Li, Xiaojun Xu, Shiliang Zuo, Qian Chen, Pan Zhou, Benjamin Rubinstein, Ce Zhang, and Bo Li. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. *Advances in Neural Information Processing Systems*, 34:17642–17655, 2021.

Wei Yao, Zeliang Zhang, Huayi Tang, and Yong Liu. Understanding model ensemble in transferable adversarial attack. *arXiv preprint arXiv:2410.06851*, 2024.

Dinghuai Zhang, Hongyang Zhang, Aaron Courville, Yoshua Bengio, Pradeep Ravikumar, and Arun Sai Suggala. Building robust ensembles via margin boosting. In *International Conference on Machine Learning*, pp. 26669–26692. PMLR, 2022.

Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang, Yuxin Su, and Michael R Lyu. Improving the transferability of adversarial samples by path-augmented method. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8173–8182. IEEE, 2023.

Yunce Zhao, Wei Huang, Wei Liu, and Xin Yao. Negatively correlated ensemble against transfer adversarial attacks. *Pattern Recognition*, 161:111155, 2025.

A Aggregation Rule

The $S \geq k$ abstraction is agnostic to how member outputs are combined, and that matters, in a direction that depends on the label space. At the rare-event budget under standard training the three rules give:

	3-of-6 threshold	hard voting	logit averaging
CIFAR-10 (ϵ=1)	2.300%	1.600%	1.200%
GTSRB (ϵ=4)	13.500%	10.500%	10.900%
MNIST (ϵ=8)	0.500%	0.150%	0.050%

Three members being individually fooled is necessary but not sufficient for the deployed system to fail, and the slack depends on the task. On MNIST logit averaging is an order of magnitude safer than the threshold rule, because ten well-separated classes make coincidental agreement on a single wrong label unlikely; on GTSRB the margin nearly disappears, because confusable signs make members agree on *which* wrong class to prefer. $P(S \geq k)$ should therefore be read as an upper bound on soft-voting risk, and a defender should measure dispersion under the rule they deploy.

B Ensemble-Size Scaling

For each m we enumerate all $\binom{6}{m}$ subsets of the six students, evaluate each on the same N examples under its own majority threshold $k = \lceil m/2 \rceil$, and report the mean and standard deviation across subsets. The beta-binomial column is a single prediction computed once from the $m = 6$ values of $\bar{p}$ and D and never refitted. Three features stand out. First, $\bar{\rho}$ is *identical* across m (a combinatorial identity, since it is a mean over member pairs), while its subset-to-subset spread shrinks as roughly $1/\sqrt{m}$: enlarging the ensemble buys precision on the dependence parameter and nothing else. Second, D climbs with m exactly as $1 + (m-1)\bar{\rho}$ predicts. Third, the beta-binomial, calibrated only at $m = 6$, tracks the observed majority rate at every held-out size to within a percentage point, whereas independence departs in either direction as the threshold $k = \lceil m/2 \rceil$ steps. Values are at $\epsilon = 16/255$; the $m = 6$ rows match Table 1.

Benchmark	Train	ϵ	$\bar{p}$ (%)	$\bar{\rho}$	D	Obs. (%)	Beta-bin. (%)
CIFAR-10	Std	1	3.533	+0.258	2.286	2.300	2.394
		2	13.250	+0.448	3.237	13.200	13.162
		4	37.625	+0.557	3.785	41.150	40.592
		8	76.242	+0.531	3.658	81.400	81.193
		16	94.617	+0.380	2.897	97.350	97.319
GTSRB	Std	1	0.883	+0.516	3.567	0.800	0.852
		2	3.492	+0.644	4.203	3.650	3.596
		4	13.067	+0.771	4.852	13.500	13.658
		8	29.708	+0.783	4.916	31.200	31.011
		16	50.383	+0.777	4.887	52.750	52.352
MNIST	Std	1	0.017	+1.000	2.001	0.000	0.008
		2	0.017	+1.000	2.001	0.000	0.008
		4	0.100	+0.537	3.666	0.150	0.097
		8	0.525	+0.666	4.233	0.500	0.538
		16	18.892	+0.703	4.500	20.400	19.876

Table 4: Complete per-configuration results, m=6, k=3, standard training, teacher surrogate (N=2000 throughout). The independence prediction and the exact per-row test appear in Table 1; the confound and transfer-optimized-attack runs of Sections 5.3–5.4 are in the released logs.

Benchmark	m (k)	Obs. (%)	Indep. (%)	Beta-bin. (%)	$\bar{\rho}$	D
CIFAR-10	1 (1)	94.62 ± 0.47	94.62	94.62	—	—
	2 (1)	97.78 ± 0.25	99.71	97.78	0.380 ± 0.040	1.38 ± 0.04
	3 (2)	95.95 ± 0.29	99.16	95.89	0.380 ± 0.026	1.76 ± 0.05
	4 (2)	97.47 ± 0.17	99.94	97.44	0.380 ± 0.018	2.14 ± 0.05
	5 (3)	96.45 ± 0.13	99.86	96.29	0.380 ± 0.011	2.52 ± 0.04
	6 (3)	97.35	99.99	97.32	0.380	2.90
GTSRB	1 (1)	50.38 ± 1.47	50.38	50.38	—	—
	2 (1)	55.98 ± 1.06	75.39	55.95	0.777 ± 0.015	1.78 ± 0.01
	3 (2)	50.61 ± 0.65	50.58	50.39	0.777 ± 0.008	2.56 ± 0.02
	4 (2)	53.59 ± 0.64	69.33	53.29	0.777 ± 0.005	3.33 ± 0.02
	5 (3)	50.70 ± 0.36	50.72	50.39	0.777 ± 0.003	4.11 ± 0.01
	6 (3)	52.75	66.35	52.35	0.777	4.89
MNIST	1 (1)	18.89 ± 4.61	18.89	18.89	—	—
	2 (1)	23.69 ± 3.25	34.26	23.49	0.703 ± 0.057	1.70 ± 0.06
	3 (2)	18.56 ± 2.30	9.28	18.39	0.703 ± 0.036	2.40 ± 0.08
	4 (2)	21.25 ± 1.84	16.31	20.80	0.703 ± 0.025	3.10 ± 0.08
	5 (3)	18.33 ± 1.24	4.85	18.24	0.703 ± 0.015	3.80 ± 0.06
	6 (3)	20.40	8.41	19.88	0.703	4.50

Table 5: Ensemble-size scaling at $\epsilon = 16/255$, standard training. Means and standard deviations are over all $\binom{6}{m}$ subsets of the six students, evaluated on identical examples; $\bar{\rho}$ and D are the mean and subset spread of the dependence statistics. The beta-binomial is fit once at $m = 6$ and applied to every smaller size without refitting. $\bar{\rho}$ is constant across m by construction and its spread shrinks as $\sim 1/\sqrt{m}$; the beta-binomial tracks the observed rate at every size while independence departs in both directions as k steps (e.g. MNIST $m = 5$: observed 18.3%, beta-binomial 18.2%, independence 4.9%).

C Distribution-Free Procedures

Because we retain the per-example indicator matrix $Z \in \{0,1\}^{N\times 6}$, the central claim can be checked without any parametric assumption. **Permutation test.** Permuting each column of Z independently ($B = 20{,}000$ replicates, seed 0) preserves every marginal p_i exactly while destroying all dependence, giving an exact finite-sample null for $P(S \geq 3)$; the reported p-value is $(1 + \#\{b : \hat{P}_b \geq \hat{P}_{\text{obs}}\})/(B+1)$, with floor 5×10^{-5}. **Bootstrap.** We resample examples (rows of Z) with replacement (2,000 resamples) and recompute $\bar{\rho}$ and D, reporting percentile intervals.

Table 6 confirms the parametric results without their assumptions. The permutation test rejects independence at the floor at every well-powered rare-event budget ($\epsilon = 1/255$ on CIFAR-10 and GTSRB, $\epsilon \geq 4/255$ on GTSRB, $\epsilon = 8/255$ on MNIST) and, correctly, does *not* reject once transfer saturates and the observed rate falls below independence ($p = 1.0$ at $\epsilon \geq 8/255$ on CIFAR-10 and $\epsilon = 16/255$ on GTSRB). The bootstrap interval for $\bar{\rho}$ excludes zero at every budget with events, on every benchmark.

Benchmark	ϵ	events	$\bar{\rho}$ [95% boot.]	D [95% boot.]	perm. p
CIFAR-10	1	46	0.258 [0.209, 0.309]	2.29 [2.04, 2.53]	5×10^{-5}
	2	264	0.448 [0.414, 0.480]	3.24 [3.07, 3.40]	5×10^{-5}
	4	823	0.557 [0.535, 0.578]	3.78 [3.67, 3.89]	0.22
	8	1628	0.531 [0.506, 0.556]	3.66 [3.53, 3.78]	1.00
	16	1947	0.380 [0.332, 0.429]	2.90 [2.66, 3.14]	1.00
GTSRB	1	16	0.516 [0.371, 0.631]	3.57 [2.84, 4.12]	5×10^{-5}
	2	73	0.644 [0.577, 0.702]	4.20 [3.86, 4.49]	5×10^{-5}
	4	270	0.771 [0.743, 0.797]	4.85 [4.71, 4.98]	5×10^{-5}
	8	624	0.783 [0.763, 0.801]	4.91 [4.81, 5.00]	5×10^{-5}
	16	1055	0.777 [0.759, 0.794]	4.88 [4.80, 4.97]	1.00
MNIST	1	0	—	—	—
	2	0	—	—	—
	4	3†	0.537 [0.378, 1.00]	3.66 [1.00, 5.00]	5×10^{-5}
	8	10	0.666 [0.500, 0.779]	4.23 [3.26, 4.85]	5×10^{-5}
	16	408	0.703 [0.678, 0.726]	4.50 [4.37, 4.61]	5×10^{-5}

Table 6: Distribution-free analysis on the standard ensemble, m=6, k=3. The permutation test rejects independence at the floor (5×10^{-5}) at every well-powered rare-event budget and does not reject once transfer saturates and the observed rate drops below independence ($p = 1.0$), matching the parametric verdicts of Table 1 without their assumptions. The bootstrap $\bar{\rho}$ interval excludes zero wherever events occur. Rows with no events ($\epsilon \leq 2/255$ on MNIST) have no defined majority rate to test. † marks the MNIST $\epsilon = 4/255$ row, whose three events make its bootstrap interval very wide; it is shown for completeness and no conclusion rests on it.

D Data-Quality Audit

Rare-event ensemble statistics are sensitive to defects in the evaluation set that aggregate accuracy numbers hide. We apply four checks to every run before reporting it, and every benchmark in this paper passes all four; we describe them together with two failure modes an earlier iteration of this study exhibited, because they are the reason the checks exist.

1. **Label coverage.** The retained evaluation set must span a representative fraction of the label space. Concentration on a few mutually confusable classes inflates both transfer and correlation for reasons unrelated to ensemble structure. All three benchmarks pass: CIFAR-10 and MNIST retain all ten classes and GTSRB all 43, with the three most frequent classes accounting for at most 32% of retained images.
2. **Non-degenerate targets.** Each target must be a non-trivial classifier. A model that collapses to a near-constant prediction passes a joint clean-correctness filter on exactly the examples matching its constant output, while carrying no information, and registers a transfer rate of zero that reflects degeneracy, not robustness. Every student here emits the full label set under attack (ten labels on CIFAR-10 and MNIST, 41–43 on GTSRB); no member is degenerate.
3. **Unbiased filtering.** If the eligible-example cap binds before the scan completes and the data are traversed in label order, the retained set is a truncation of the label space rather than a sample. Our loader shuffles once with a fixed seed and never wraps: the cap of 2,000 binds after a `total_seen` of at most 2304, far below every test-set size, so the retained set is a random sample.

4. **Distinct examples.** The retained set must contain no repeated images. A loader that wraps around the test set to satisfy the cap inflates N without adding information, narrowing confidence intervals that should be wider. All retained sets contain 2000 unique images (0% repeats).

Two failure modes this audit caught. An earlier iteration of the GTSRB evaluation used a label-ordered loader that, to satisfy the cap, wrapped past the $12{,}630$-image test set: the retained set contained more than a quarter repeated images and collapsed to two speed-limit classes, which passes no eye test but inflates N and every dependence statistic. Regenerating under the shuffled single-pass loader above restores full label coverage and removes the repeats, and is the configuration reported throughout this paper. Separately, an earlier MNIST run contained a distilled student that had collapsed during training to a constant classifier emitting a single label; the joint-correctness filter then retained only images of that one class, and the summary statistics ($\bar{p}$, $\bar{\rho}$, D) looked entirely normal while being computed on a one-class evaluation set. Check 2 catches this—the constant member emits a single distinct label—but no summary statistic does, which is the point: a defender computing $\bar{p}$, $\bar{\rho}$ and D from an outcome matrix cannot distinguish a healthy ensemble from a degenerate one, because joint-correctness filtering is defined by the very models under evaluation and a degenerate member silently rewrites the evaluation set in its own image. Retraining the collapsed student and re-checking prediction diversity before the filter is applied removes the pathology; the MNIST run reported here passes all four checks.