Disentangling Model Multiplicity in Deep Learning

Ari Heljakka [1] **Martin Trapp** [2] **Juho Kannala** [2] **Arno Solin** [2]

Abstract

Model multiplicity is a well-known but poorly understood phenomenon that undermines the generalisation guarantees of machine learning models. It appears when two models with similar training-time performance differ in their predictions and real-world performance characteristics. This observed 'predictive' multiplicity (PM) also implies elusive differences in the internals of the models, their 'representational' multiplicity (RM). We introduce a conceptual and experimental setup for analysing RM by measuring activation similarity via singular vector canonical correlation analysis (SVCCA). We show that certain differences in training methods systematically result in larger RM than others and evaluate RM and PM over a finite sample as predictors for generalizability. We further correlate RM with PM measured by the variance in i.i.d. and out-of-distribution test predictions in four standard image data sets. Finally, instead of attempting to eliminate RM, we call for its systematic measurement and maximal exposure.

1. Introduction

Machine learning (ML) models are typically underdetermined by data. This fact is often poorly understood, imprecisely conceptualized, and only superficially measured. Consequently, one often mistakes the observed success of a model for proof that it has somehow 'captured' the relevant features of its target, and that one has converged upon the single 'right' model. Due to this oversight, surprising practical problems with real-world model generalization may appear. As ML models are increasingly deployed into real-world environments, it has become common to find that a model that worked well in the tests faces various failures in real-world scenarios. The primary textbook explanation for

this disconnect is that the training data of the model and the deployment-time data were not generated from the same distribution ('non-i.i.d.'), rendering the predictions unreliable.

Our focus is on the more elusive yet prevalent case in which model variants with similar training-time performance behave very *differently* from each other on individual held-out test data samples. Here, the data available at training time is insufficient to justify sound model selection among the model variants. This phenomenon has been approached with concepts such as non-identifiability, underspecification (D'Amour et al., 2020), Rashomon set (Breiman, 2001; Fisher et al., 2019), arbitrariness-1 (Heljakka, 2014), and predictive multiplicity (Marx et al., 2020). Even though the *model selection* problem appears well-grounded in data analysis (Ding et al., 2018), the established techniques assume a well-defined likelihood measure, which in the case of neural networks often does not exist. In short, the problem occurs when risk-equivalent model variants have different internal representations of the same data. We call these differences *representational multiplicity* (RM) and propose to measure RM in terms of correlations of activations across model variants. These differences, in turn, may bring about the observed spread of predictions across the variants of the same model, the definition of predictive multiplicity (PM, Marx et al., 2020). The multiplicity may result from seemingly random factors, which limits the reliability of inferences for explanatory purposes and undermines the very notion of a 'best model for the data'.

For deterministic pathways within a network, variations in predictive outputs imply variance in the intermediate representations. Hence, the presence of PM implies the presence of RM. The opposite is not the case since, for a given sample, two different representations can still lead to the same prediction. PM only measures the observable predictive differences *as a function of the test samples at hand*, constrained by RM. Even in the absence of observed PM, any presence of RM implies that there exist other potential inputs that would also yield observable PM. This pivotal distinction fails to align with traditional notions of uncertainty defined in terms of observables (see Fig. 1a).

Many prior works approach model multiplicity as a nuisance, calling for its *elimination* or *reduction to uncertainty*. In contrast, we approach each internal representation as an

[1]GenMind Ltd, Helsinki, Finland (Contributed all theory. Work done while in Aalto University.) [2]Department of Computer Science, Aalto University, Espoo, Finland. Correspondence to: Ari Heljakka <heljakka@iki.fi>.
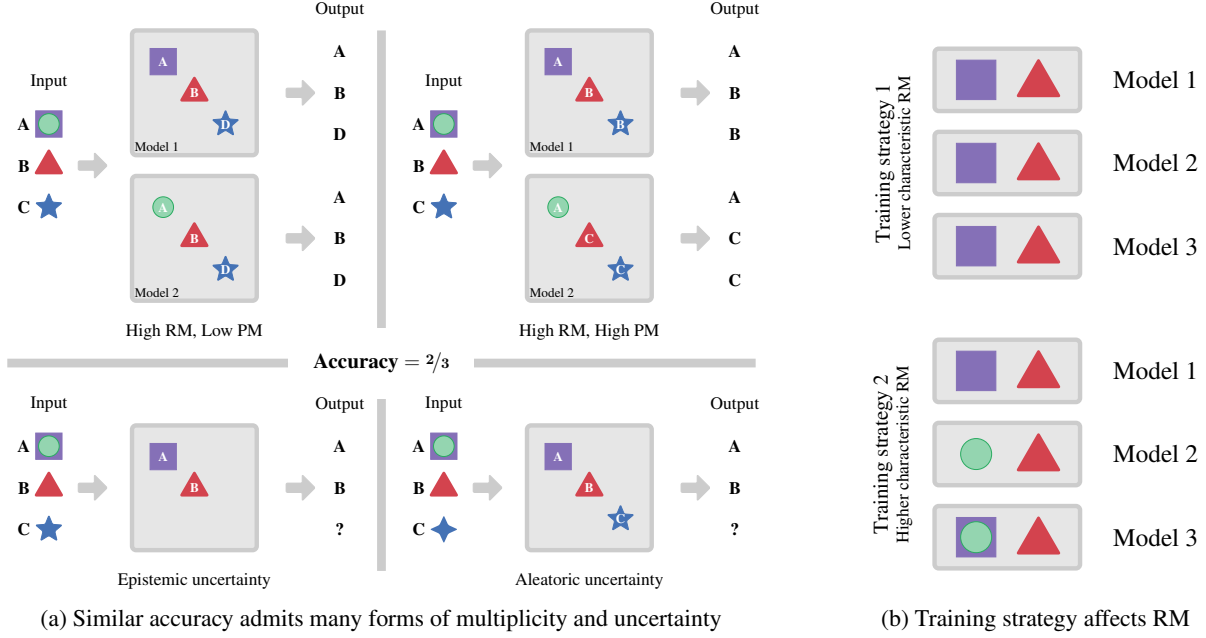
Figure 1. (a) Even under *identical training* setup, and controlled for the same empirical risk, model variants may represent the input data differently (high RM) due to different dependencies on learned features (⭐, 🟪, 🟢, 🔺) when classifying the three input samples with true class labels A, B, C and the true features indicated under 'Input' column. The 3rd input has true feature ⭐ with true class C *etc.* These differences may or may not be observable at the outputs (low or high PM, respectively). (b) Different training strategies may systematically yield different degrees of RM. Higher RM indicates correspondingly higher *hidden risk* for generalization performance.

alternative 'compression' of the data, encapsulating hitherto unleveraged additional information. RM can be considered a *hidden risk* for future generalization performance. As such, for model variants with equivalent *observed* empirical risk, RM should be maximally *exposed* rather than minimized.

We look for possible systematic relationships between certain training strategies and the magnitude of RM (suggested also in Raghu et al., 2017; Morcos et al., 2018). Our vision data set results confirm a strong correlation between RM and learning rate / batch size. Furthermore, irrespective of the well-known association between larger batch sizes and worse generalization via sharper minima (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Elad Hoffer, 2017), larger batch sizes appear strongly correlated with lower PM.

Perhaps due to the combination of extra computational costs and somewhat muddled prior conceptual treatment, these empirically observable differences are often ignored (see Fig. 1b). Our emphasis on the irreducibility of RM aims to challenge the discourse based on a one-dimensional notion of 'the best' model for the data and equivalent-sounding notions like 'the predictor that encodes the *right* structure' (D'Amour et al., 2020), '[models that contain] the *true* data generating process' (Fisher et al., 2019) or even 'true data generating model' (Ding et al., 2018). This terminology falsely suggests one can cross and dissolve the categorical gap between the model and the modelled, leading to a single

'correct' representation. However, in general, representational multiplicity can only be hidden, not eliminated.

We summarize the contributions of this paper as follows. *(i)* We present a well-defined conceptual and experimental setup for analyzing the critical phenomenon of disentangled representational and predictive multiplicity (RM - PM). *(ii)* We show the significance of observing both RM and PM through experiments and find various regularities between hyper-parameter values and RM. We design an experiment to empirically show that RM cannot be reduced to PM. *(iii)* We introduce the *confabulation matrix* as a straightforward tool for visualizing multiplicity.

2. Related Work

The phenomenon of representative multiplicity in models has been studied as, for example, the underdetermination of theories in the philosophy of science and epistemology (*e.g.*, Duhem, 1954; Newton-Smith, 1980; Laudan & Leplin, 1991), and system non-identifiability in statistics and economics (*e.g.*, Dawid, 1979; Rothenberg, 1971; Gelfand & Sahu, 1999). The implications for modern (very large) computational models are less obvious. From a modern standpoint, the seminal paper by Breiman (2001) contrasts two conceptions of models. First, the models that are structurally designed and selected for explanatory purposes, and second, the models optimized for prediction. In the latter case, the

data admit a set of equally valid model parametrizations—the *Rashomon set*— and alternative competing explanations.

In computer science, the problem has been directly approached in the context of loss landscape analysis (*e.g.*, Garipov et al., 2018; Izmailov et al., 2018; Chaudhari et al., 2019; Nakkiran et al., 2021), overparametrization (*e.g.*, Belkin et al., 2019), model averaging (*e.g.*, Izmailov et al., 2018; Wilson & Izmailov, 2020), domain adaption and interpretability, and ensembling (Lakshminarayanan et al., 2017; Fort et al., 2019). Risk-equivalent modes have also been examined from the point of view of accuracy-diversity trade-offs (Fort et al., 2019) as well as accuracy-reproducibility trade-offs (Shamir et al., 2020). During recent years, empirical consequences of multiplicity have been examined in pathological cases such as spurious correlations and shortcut learning (Schölkopf, 2022; Arjovsky et al., 2019). D'Amour et al. (2020) examine the prevalence and consequences of multiplicity in practical ML pipelines that admit 'underspecification'. Underspecification and Rashomon sets have directly been leveraged in Semenova et al. (2019) with the goal of selecting models within the set based on criteria such as sparsity or monotonicity along an important set of features. In a similar vein, multiplicity can also be leveraged to estimate variable importance (Fisher et al., 2019).

The alternative notions of 'underspecification', 'identifiability', and Rashomon sets are closely related to the same scenario. To capture the internal and external aspects of the problem, the notion of representational multiplicity (RM) was introduced in this work to supplement *predictive* multiplicity (PM) introduced in Marx et al. (2020), a distinction often obscured in prior works. Further, our PM definition accounts for the actual variance of predictions, whereas, *e.g.*, D'Amour et al. (2020) only measure multiplicity as the *variation in total accuracy*, ignoring the distribution in the *kinds* of errors the model variants make.

For large models, 'underspecification' as a concept problematically seems to imply that a 'full' specification is possible. In D'Amour et al. (2020), 'credible intuitive bias' is effectively defined as a way to eliminate the underspecification. This concept may be valuable in terms of segmenting the ML workflow, but it appears circular: we 'solve' underspecification – originally defined in terms of unobserved downstream environment – by improving our pipeline so as to deal with the *observed* problems. Shifting previously unobserved issues into the realm of observed issues surely fails to exhaust the realm of remaining unobserved issues.

Prior works often intrinsically posit the existence of a unique correct solution model, rendering any observed multiplicity into nuisance or noise to be minimized. This position can be defended if one can plausibly unearth the underlying causal properties via designing a set of interventions to that end. See D'Amour et al. (2020) for a large-scale empirical

evaluation and Romeijn & Williamson (2018) for theoretical analysis. Yet, this rarely happens except in the very theory of causal inference itself (see, *e.g.*, Schölkopf, 2022). In Bayesian and ensemble analyses (*e.g.*, Barber & Bishop, 1998; Fort et al., 2019; Izmailov et al., 2018), the variance is characteristically treated in aggregate, usually in parameter space. Finally, as a matter of standard ML practice, the variance is reduced to a scalar for generalization error, even in the in-depth analysis of the topic in Raghu et al. (2017).

Principled approaches to model selection exist in Bayesian statistics, with methods such as BIC and AIC routinely used to compare models. However, in practice, they are not applicable to deep neural networks (Ding et al., 2018) that perform equally well on the test data set. The same holds for techniques such as majority voting or randomization procedures (Germain et al., 2016). Prior works that focus on *mitigating* the multiplicity in practice (D'Amour et al., 2020; Arjovsky et al., 2019; Schölkopf, 2022) are valuable as long as we can assume that the information is available to resolve the multiplicity. Our work considers the general scenario in which we have no way of knowing whether this is the case nor of distinguishing one model from another, which renders all RM variations informative. Following the same line of reasoning, while it may be useful to marginalize over all model variants in parameter space (Wilson & Izmailov, 2020), that may risk throwing away relevant information.

To avoid this, the quantification of RM in the network activation space is critical. We chose SVCCA (Raghu et al., 2017) to efficiently compare the similarity of distributed representations across networks. SVCCA was shown to be capable of comparing different architectures and surfacing detailed learning dynamics, such as the order in which the layers 'solidify' during learning. In comparison to regular canonical correlation analysis (CCA), SVCCA provides the importance of each direction in the original activation space. Kornblith et al. (2019) improve the reliability of the method when the number of evaluation data points is small.

The treatment of predictive multiplicity (PM) in Marx et al. (2020) is closest in spirit to our work (but also see the 'prediction variation' metric in Chen et al. (2020) and related Dusenberry et al. (2020)). They call for reporting PM in quantitative measurements in the same way as the canonical test error. However, while they build upon 'the Rashomon effect' (effectively equivalent to RM), they focus on the practical consequences for predictions and do not quantitatively address the PM-RM distinction. Our work provides a more complete yet concise picture that draws from the methodology of D'Amour et al. (2020), the explicit metrics for RM from Raghu et al. (2017) and the conceptual import for PM from Marx et al. (2020). Most importantly, we hope to motivate incorporating these conceptual distinctions in regular ML practises.
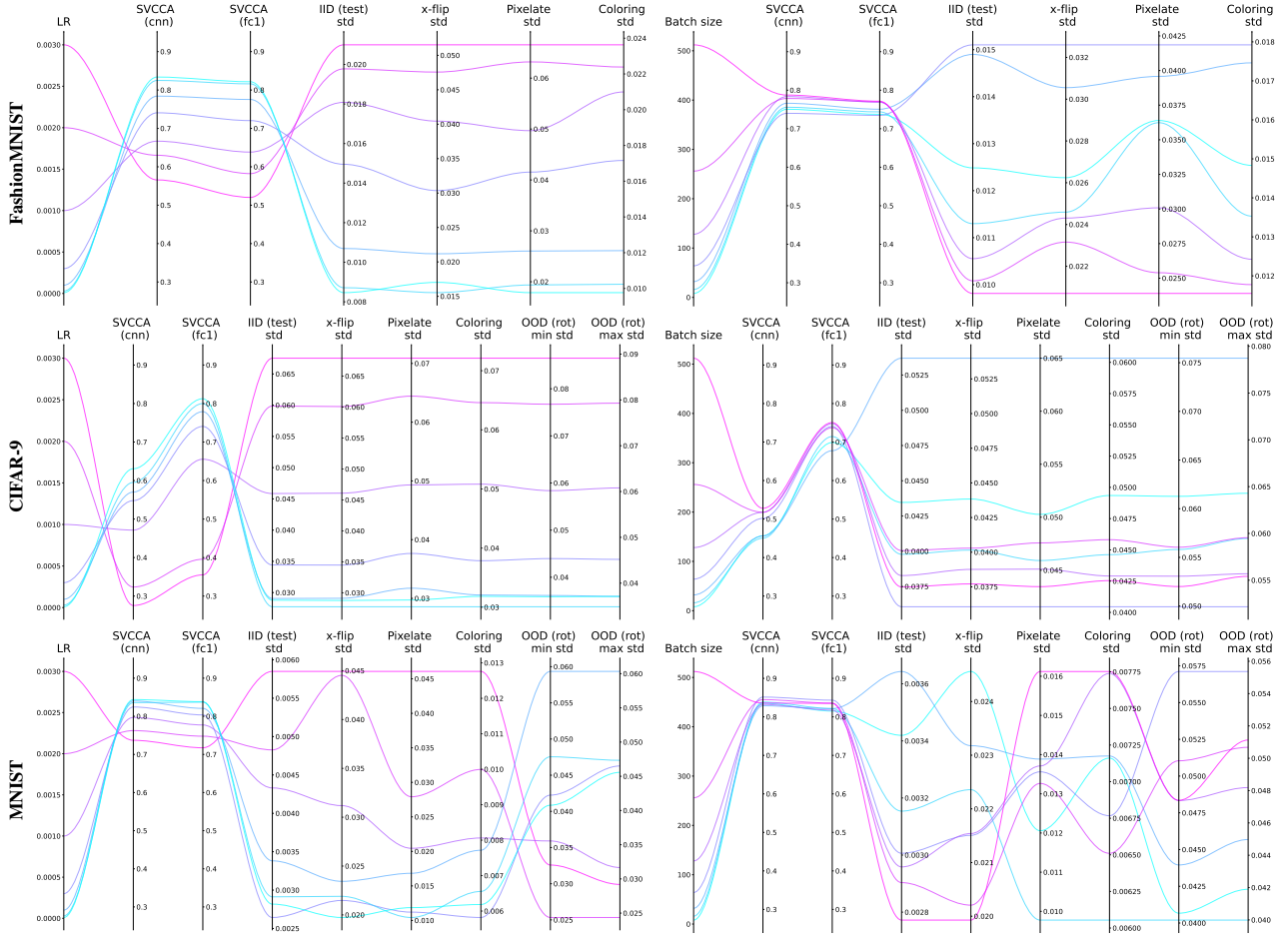
Figure 2. Associating representative (RM) and predictive (PM) multiplicity in two hyper-parameter regimes (**Left:** learning rate, **Right:** batch size) for three data sets, with identical model architectures. SVCCA is the measure of inverted RM at two feature layers, while the prediction 'std' columns measure PM for the i.i.d. and various OOD distributions. data sets: FashionMNIST (top), CIFAR-9 (middle), and MNIST (bottom). Low learning rate and high batch size correlate with higher SVCCA (lower RM) and smaller variance (lower PM). Thus, across the datasets, RM has a peculiar yet predictable dependence on LR and batch size.

3. Methods

In order to illustrate representational multiplicity from various perspectives, we now proceed to find a sufficiently precise definition for it and make it measurable. We then evaluate its sensitivity to seemingly irrelevant hyper-parameters of the training strategy under a fixed model. Such hyper-parameters could involve training batch size, learning rate, initialization, *etc.*. Finally, we show how to disentangle the predictive power of RM and PM.

Problem Setup We start with the classical setup for a supervised prediction problem, in which one aims to learn a function h that maps inputs $\boldsymbol{x} \in \mathcal{X}$ to targets $y \in \mathcal{Y}$, *i.e.*, $h\colon \mathcal{X} \to \mathcal{Y}$. To learn h, one minimizes the empirical risk $E(\ell(\boldsymbol{x}_i, y_i))$ based on a training set $\boldsymbol{X}_{\text{train}} = \{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{y}_{\text{train}} = \{y_i\}_{i=1}^N$. To leverage h as a predictor on a final deployment target system, one usually assumes the test data

$\boldsymbol{X}_{\text{test}}$ to be drawn from the same distribution and satisfy the common i.i.d. assumption.

Predictive Multiplicity Let $H = \{h_k\}_{k=1}^K$ be a set of classifiers such that the difference in empirical risk $|E(h_k) - E(h_j)| < \epsilon$ for all $h_k, h_j \in H$ with $k \neq j$ and a given error tolerance $\epsilon > 0$. We define Predictive Multiplicity (PM) over a data set S and H as:

$$\text{PM}(S, H) = \mathbb{E}_{\boldsymbol{x} \in S}\left[\sqrt{\mathbb{V}\text{ar}_{h \in H}\{h(\boldsymbol{x})\}}\right]. \quad (1)$$

In words, we measure the extent to which the risk-equivalent model variants assign conflicting predictions over the data set. Note that while we follow the concept definition of Marx et al. (2020), we diverge from their mathematical definition.

Representational Multiplicity Representational multiplicity can be defined in terms of three constraints. **(1)** It

measures the variation across the internal representations of risk-equivalent models. This is in contrast to the variation in *predictions*. The representations are expressed by activations in the function space. Unlike evaluation in the weight space, the (function) activation space yields a similarity estimate informed by the available data. Hence, we focus on dissimilarities across the learned functions rather than across their parametrizations. The latter can be diverse even if the underlying functions are not (Garipov et al., 2018). **(2)** The multiplicity and risk must be defined in terms of specific data, typically the training or testing data already available in the context. **(3)** The relevant variation across the representations must be asymmetric in terms of trivial transformations such as affine transformations. This prevents us from, *e.g.*, treating the different order of the same weights in latent space as different neural network representations. Importantly, this excludes the use of straightforward measures such as standard deviations in the activation space. Still, the notion of what constitutes a 'trivial' transformation leaves room for interpretation. Here, we only consider the invariance of the output activations to linear transformations. Thus construed, without exhausting every possible trivial transformation, *one can only ever arrive at a lower bound for similarity and upper bound for RM* (also see Ainsworth et al. (2022) for possible deeper symmetries).

A suitable approach for this purpose is the singular vector canonical correlation analysis (SVCCA) metric (Raghu et al., 2017), a method for assessing activation similarity via combining Singular Value Decomposition with Canonical Correlation Analysis (CCA, Hardoon et al., 2004).

With SVCCA, each neuron $j = 1, 2, \ldots, M$ is represented as a vector of its *output activations* $z_j(\boldsymbol{x}) = [z_j(\boldsymbol{x}_1), \ldots, z_j(\boldsymbol{x}_N)]^\top$ across all the data points. We then compare a single layer of M neurons between networks 1 and 2 as matrices $\boldsymbol{Z}_1 \in \mathbb{R}^{M \times N}$ and $\boldsymbol{Z}_2 \in \mathbb{R}^{M \times N}$ composed of the output activation vectors. This setup satisfies the criteria **(1)** and **(2)** mentioned above. To run the comparison, SVCCA first performs SVD on $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ and picks the subspaces $\boldsymbol{Z}'_1 \subset \boldsymbol{Z}_1$ and $\boldsymbol{Z}'_2 \subset \boldsymbol{Z}_2$ to explain the desired degree (*e.g.*, 99%) of the variance in each subspace. Finally, one runs the CCA to find a linear transformation of each subspace so as to maximize the correlations ρ_i between them. The average of the (maximized) correlations is, then, the SVCCA similarity.

For a single metric for representational multiplicity, we average the largest correlations. It was observed in Raghu et al. (2017) that taking the top 25 vectors (we found 20 sufficient) accounted for almost all the variance across image data sets, allowing us to discard the rest. Finally, defining multiplicity as the negated expectation over correlations, we arrive at

representational multiplicity given as

$$\text{RM}(S, f_1, f_2) = -\frac{1}{T} \sum_{i=1}^{T} \rho_i(f_1, f_2), \qquad (2)$$

s.t. $\rho_1(f_1, f_2) \geq \rho_2(f_1, f_2) \geq \ldots$ for a single similarity measure of data set S and (sub-)networks f_1 and f_2 with SVCCA correlations $\rho(f_1, f_2)$. Since CCA is invariant to affine transformations, it satisfies the criteria **(3)** above, making the method well-suited for RM estimation. For SVCCA among the comparison set of $K > 2$ network variants, we paired every variant u with every other and took the mean of the triangular matrix produced by all pairings:

$$\text{RM}(S, \mathbf{f}) = \binom{K}{2} \sum_{u=1}^{K} \sum_{v=1}^{u} \text{RM}(S, f_u, f_v). \qquad (3)$$

We considered SVCCA separately for the last CNN layer ('cnn') and the first fully connected one ('fc1').

Assessing Patterns in RM and PM One can now assess the dependency of RM and PM on hyper-parameters as follows. First, for a given choice of hyper-parameter ('training strategy'), we need to train K variants of our target models, changing nothing but the random seed, yielding variants $h_k \in H, k = 1, 2, \ldots, K$. We stop the training in each case once we reach the predictive accuracy that we presume can be achieved by all the variants and all the hyper-parameter values. This typically requires extra training runs. Second, for each variant h_k, we need to look at the *individual* predictions for each $\boldsymbol{x}_i$ in the data set. This yields the variance of predictions across the variants h_k of H and hence PM (Eq. (1)). Third, we do the same for the activations and apply SVCCA to get the RM (Eq. (3)). Finally, we select the hyper-parameter we wish to evaluate ('regime'), and repeat the steps above for M different values of it ('training strategies'). Within each regime, we track whether one strategy yields higher PM or RM than another (Fig. 2).

Irreducibility of RM Computing RM is relatively expensive, while PM can be computed directly from inferred outputs. This raises the question of whether we could just use PM everywhere. But as we have seen earlier, for a finite sample, we should expect that some samples will exhibit RM without changing model outputs and, therefore, without causing any observable PM. In other words, if the subset is not large enough, the PM of the subset does not predict the PM of the full data set. However, since RM is not a function of PM, and RM is expected to be predictive of PM, RM over the subset should predict the PM over the full data set. Essentially, we want to express that for some subsets of data, the RM over the subset is a better predictor of the 'real' PM (measured overall data) than PM over the subset. This can be formalized as the following hypothesis.

Hypothesis 1. Given data set S of size N, consider a subset $s_i \subset S$ with size $N' \ll N$, and K model variants h_{ij} with $j = 1, \ldots, K$ for each training strategy H_i. Define the estimation errors $\mathcal{E}_{RM}(s_i, S, H_i) = [c \cdot RM(s_i, H_i) - c' \cdot PM(S, H_i)]$ and $\mathcal{E}_{PM}(s_i, S, H_i) = [c'' \cdot PM(s_i, H_i) - c' \cdot PM(S, H_i)]$, with constants c, c', and c'' scaling each type of a measure to 1 ($c = 1/\max_i(RM(s_i, H_i))$, *etc.*). Now, for each i, we can pick s_i in a way that is independent of RM measures, such that the expectation over L training strategies is $\mathbb{E}_{i=1,\ldots,L} \, \mathcal{E}_{RM}(s_i, S, H_i) < \mathbb{E}_{i=1,\ldots,L} \, \mathcal{E}_{PM}(s_i, S, H_i)$.

If the hypothesis holds, then for RM and PM over a finite sample size, it is possible for the RM to be a better predictor of the full set PM. Hence, RM is not reducible to PM. In order to examine the hypothesis, we can pick s_i in the most obvious way, by taking the samples that minimize PM over H_i. As this sample-picking method is uninformed by RM, it serves as proof by example that a finite sample *can* allow such situations to occur. We show that in all data sets, the hypothesis clearly holds (Fig. 3).

Arbitrary Predictions In the traditional best-model based analysis of classification models, it is common to review the samples most prone to classification errors *for that model*, often presented as a confusion matrix. Now, admitting the possibility of RM, one can similarly ask which samples are most prone to being *differently classified across various equally accurate variants of the model*. This yields a sort-of 'second-order' confusion analysis that we call *confabulation analysis*, visualized in terms of the inputs that yield the highest multiplicity values (Fig. 4). The terminology is motivated by the fact that the choices the model makes for those samples appear arbitrary, or *confabulated*, irrespective of how confident a specific model instance is about them.

There is a major conceptual difference between confusion and confabulation. Confusion informs us which kind of samples our model is *currently wrong about*, suggesting we should do something about it. Confabulation informs us which samples our models might be *fundamentally unable to classify correctly*, so ambivalent or devoid of relevant information that they should perhaps be discarded all together. In a disturbing contrast, the common approach to force one's model to classify even the worst of these samples 'correctly' may be akin to learning to classify noise.

4. Experiments

We conduct three types of experiments to support our case. First, we relate RM, PM, and OOD generalization in various data sets, under various training scenarios. Second, we show hypothesis 1 to hold on all the examined data sets, showing that one cannot simply reduce RM to PM. Third, we visualize the concrete manifestations of high RM.

Data Sets and Architectures We examined RM on classifiers trained on four typical small vision data sets: MNIST, FashionMNIST (Xiao et al., 2017), CIFAR-9 (CIFAR-10 (Krizhevsky, 2009) with one dropped class), and SVHN (Netzer et al., 2011). We used SVHN as an OOD data set for MNIST and STL9 (STL10 (Coates et al., 2011) with one dropped class to match CIFAR-9) as an OOD data set for CIFAR-9. Our focus was on typical scenarios, therefore, we picked a simple and identical CNN architecture for each data set and controlled for the same acceptable, but not state-of-the-art, accuracy within the data set (see App. B for training and architecture details).

Peculiar Regularities in Training Strategies We focused on relevant measurable differences in predictions and representations of risk-equivalent (*i.e.*, equally accurate) models trained with the same data, model structure, and hyper-parameters. The baseline assumption was that such models should behave equivalently, despite hyper-parameter differences during training. To measure the empirical divergence from this assumption, for each strategy, we first fix all hyper-parameters and train 10 variants while only changing the random seed. We measure the SVCCA similarities and output variance, separately for manual augmentations that transform the data set to OOD and for the matching separate OOD data sets. In Fig. 2, a single continuous line represents a single strategy. We then repeat the measurements for 7 values of each hyper-parameter, separately for batch size and learning rate. This yields $2{\times}7{\times}10$ training runs per data set (the 7 continuous lines).

To ensure that the results within each data set are risk-equivalent, we first had to determine the maximum (test set) accuracy that can be reached using any of the hyper-parameter values in the tested range. For the actual training sessions, we then stopped the training in each case at this value. Following the training to the highest achievable accuracy would not have made a difference to the conclusions (see App. C).

For each line, one should look for any suspiciously small SVCCA (high RM) and large i.i.d. test set prediction variance (high PM) values, whether the two are correlated, and how they relate to the OOD prediction variances. Across the lines, the first interesting question is how widely dispersed they are. Any changes in their SVCCA values reflect the effect of the hyper-parameter on the multiplicity of potential solutions. Finally, the correlation between the varied hyper-parameter and the magnitude of the ensuing PM or RM is examined. Across the data sets, the altered hyper-parameters correlate with RM. The larger batch size and the smaller learning rate correlate strongly with higher SVCCA in both examined layers (Fig. 2) and with smaller PM both in the i.i.d. and OOD test sets. (See also Fig. 5 in App.)

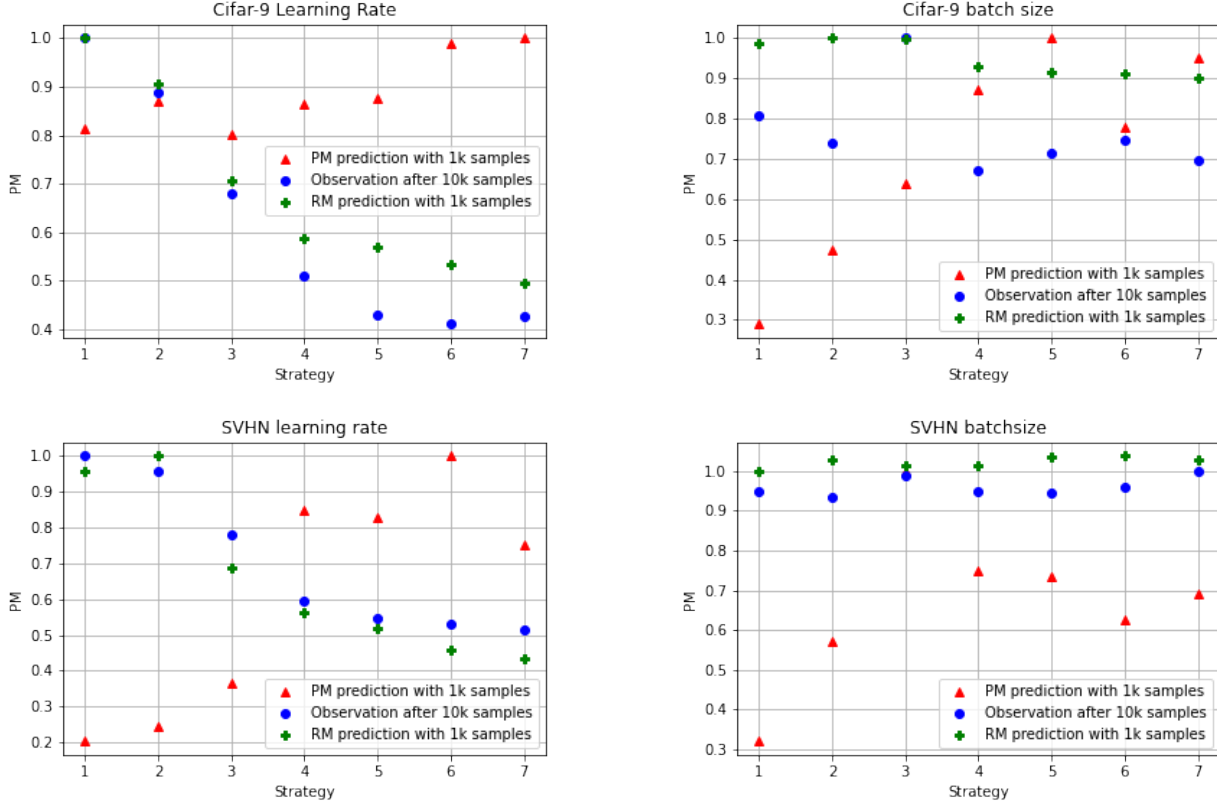Of the four data sets included, the only distinctly different

Figure 3. Relevance of RM. We compare the predictive power of RM vs PM in two data sets under two hyper-parameter regimes, each. In order to show that there *exist* cases where RM cannot be reduced to PM, we first select a subset of 1000 samples for each data set such that those samples have the lowest overall PM. For each strategy (descending learning rates: $0.003, 0.002, 0.001, 0.0003, 0.0001, 0.00003, 0.00001$ or increasing batch sizes: $8, 16, 32, 64, 128, 256, 512$), we then plot the PM of those 1000 samples for each strategy (red, PM-1k), the actual final PM after observing all 10k or 26k samples (blue), and the RM computed from the same subset of 1000 samples (RM-1k). This comparison surfaces the strong correlation between RM-1k and the PM of the full data set, whereas the PM-1k and the PM of the full data set are uncorrelated. This confirms the intuition that, over a small sample with little observed PM, RM can be superior to PM in predicting the final PM that will become observable with a larger sample.

behaviour was observed for MNIST under the *batch size strategy*. This is likely due to the small range in SVCCA values, especially since MNIST under the *LR strategy*, with a wide SVCCA range of $[0.737, 0.849]$, shows the same regularities as observed with other data sets (Fig. 2). As an alternative to SVCCA, we also attempted the CKA metric (Kornblith et al., 2019) for the same purpose, but failed to observe any correlation between CKA and PM values, hence we chose not to use it for the analysis. Measurements of Pearson correlation coefficient (Table 1) across data sets support the same conclusions.

Relevance of RM as Predictor Given the correlations between RM and PM, one might well ask, do we really need RM in the first place, when PM often appears to be more easily accessible? Conceptually, RM must necessarily precede PM, but at the limit of infinite test data, all the interesting variation in RM must necessarily be perceived as PM. On the other hand, under limited data, some of the

internal variation (RM) will still be collapsed at the final layers, at least under operations such as softmax activation. Hence, following Hypothesis 1 in Sec. 3, we demonstrate the difference empirically, by constructing the *existence* proof for a case where the two are not equivalent.

To this end, we first measured PM for the complete training data sets of Cifar-9 (10,000 samples) and SVHN (26,000 samples). Then, we picked a subset of 1000 samples with the lowest PM, as these items are likely to yield the least amount of information about RM. Using the same samples, we measured RM. Finally, we asked, is the PM or RM in the 1k sample predictive of the PM for the *whole* data set, when measured for different training strategies? As seen in Fig. 3, after scaling all values to $[0, 1]$ range, the RM values for the 1k subset are strongly correlated with the PM values of the whole data set, whereas the PM of the 1k test set is uncorrelated. Note that the 1k sample was in no way selected to favour the utility of RM values as such. Hence,

Figure 4. Top-16 confabulation inputs for CIFAR-9, SVHN, FashionMNIST, and MNIST. Confabulation measures the diversity of predictions assigned to each sample, across different training runs of the same model. In other words, the highly confabulated items are those that will most likely be classified differently under different runs. This measure is more informative than the aggregate classification error. Considering some of these results, it seems that forcing the model to learn their 'correct' class amounts to learning noise.

Table 1. The Pearson correlation coefficient (PCC) between negative representational multiplicity (RM) measured using SVCCA metric, and predictive multiplicity (PM) measured using standard deviation between OOD predictions across the data sets (LR = learning rate). The columns 'x-flip' (horizontal flip), 'Pixelate' and 'Coloring' represent the OOD predictions, followed by an exposition of the range of RM values for the data set and the strategy (Δ SVCCA). Strong negative correlation values indicate that PM and RM align.

Data set	x-flip		Pixelate		Coloring		Δ SVCCA (fc1)	
	batch size	LR	b. size	LR	b. size	LR	b. size	LR
F-MNIST	$-.750$	$-.991$	$-.872$	$-.985$	$-.767$	$-.978$	.037	.301
CIFAR-9	$-.623$	$-.990$	$-.571$	$-.980$	$-.609$	$-.985$	.073	.445
SVHN	$-.422$	$-.986$	.191	$-.985$	.089	$-.985$	.033	.234
MNIST	$-.482$	$-.977$	.221	$-.933$	.063	$-.932$	.041	.167

in each case, we have shown that RM captures information that is not reducible to the information captured by PM.

Confabulator Inputs We visualize the confabulation matrices in Fig. 4 for all data sets. The confabulators appear similar to what one might expect in confusion matrices, but here, they reflect not just a single training run, but a range of independent sessions. (Note that the confabulations are computed across all training strategies, and could further be examined for each recipe separately.)

5. Conclusion and Limitations

We propose that practitioners consider explicitly addressing RM-PM distinction as a routine aspect of any ML problem setup. As a preliminary example, considerable representational and predictive multiplicity appeared in all four data sets studied. Within the network architectures considered, certain choices of basic training hyper-parameters appear to correlate with higher occurrence of multiplicity. Given the apparent ubiquity of these phenomena, we suggest that, in general, model evaluation should either include such measures or explicitly address their absence, so as to not overstate the power of the resulting models.

The prevalence of RM raises questions about the ML practises and discourse based on the one-dimensional notion of 'the best' model for the data, and its equally imaginary counterpart 'unbiased model'. Thus construed, 'the best' appears as if transcending the model–data gap altogether and, in some unconditional sense, becoming one with the objective patterns in the data, even in cases of large models and data sets. The inevitable failure to really close the gap then leaves room for *ethical* questions in the form of *biases* that cannot be resolved, as if RM had resulted from procedural malpractice.

The experiments rely on the correlations across the top 20 SVCCA vectors as a proxy for RM. While the observed regularities lend credence to this decision, one could search for better proxies for RM. CCA-based methods are limited by their reliance on linear transformations, and do not support comparing different topologies or initialization conditions. For the latter, methods such as centred kernel alignment (CKA) might be a better fit (Kornblith et al., 2019).

In terms of computational demands, the training time of models dominates. SVD computation in SVCCA, for an $m \times n$ matrix, does have runtime complexity of $\mathcal{O}(m\,n\,\min(m, n))$, but remains feasible as the analysis is done on a per-layer basis. To observe trends across several data sets and still ensure reproducibility with the over 500 training runs, we chose to use only low-resolution image data sets. The experiments in classification models could also be repeated for, *e.g.*, generative models. While large image data sets and architectures were not examined, the same principles apply to any parameterized model. Only two training strategies were compared, and we did not study the effects of normalization or different optimizers (ADAM by Kingma & Ba, 2015, was used). While these limitations indicate that a larger research effort is called for, our key results are sufficiently consistent across four separate data sets so as to serve as a starting point for follow-up studies.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Barber, D. and Bishop, C. M. Ensemble learning in Bayesian neural networks. *Nato ASI Series F Computer and Systems Sciences*, 168:215–238, 1998.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Breiman, L. Statistical modeling: the two cultures. *Statistical Science*, 16(3):199–231, 2001.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Chen, Z., Wang, Y., Lin, D., Cheng, D., Hong, L., Chi, E., and Cui, C. Beyond point estimate: Inferring ensemble prediction variation from neuron activation strength in recommender systems. *arXiv preprint arXiv:2008.07032*, 2020.

Coates, A., Lee, H., and Ng, A. Y. An analysis of single layer networks in unsupervised feature learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 215–223, 2011.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., and Sculley, D. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Dawid, A. P. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

Ding, J., Tarokh, V., and Yang, Y. Model selection techniques: An overview. *IEEE Signal Processing Magazine, 35 (6)*, pp. 16–34, 2018.

Duhem, P. M. M. *The aim and structure of physical theory*. Princeton University Press, 1954.

Dusenberry, M. W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G., Heller, K., and Dai, A. M. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 204–213, 2020.

Elad Hoffer, Itay Hubara, D. S. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017.

Fisher, A., Rudin, C., and Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31. Curran Associates, Inc., 2018.

Gelfand, A. E. and Sahu, S. K. Identifiability, improper priors, and gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445): 247–253, 1999.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. PAC-Bayesian theory meets Bayesian inference. *Advances in Neural Information Processing Systems (NIPS)*, 29, 2016.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Heljakka, A. *Model Zero: Why You Are Obsolete at Almost Every Level and Live Largely in Fiction*. CreateSpace Independent Publishing Platform; 2nd edition, 2014.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *Proceedings of the 34th International Conference on Machine Learning (ICLR)*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Machine Learning (ICLR)*, 2015.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30. Curran Associates, Inc., 2017.

Laudan, L. and Leplin, J. Empirical equivalence and underdetermination. *The journal of philosophy*, 88(9):449–472, 1991.

Marx, C. T., du Pin Calmon, F., and Ustun, B. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Morcos, A. S., Raghu, M., and Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31. Curran Associates, Inc., 2018.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Newton-Smith, W. The underdetermination of theory by data. *Rationality in Science: Studies in the Foundations of Science and Ethics*, pp. 91–110, 1980.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30. Curran Associates, Inc., 2017.

Romeijn, J.-W. and Williamson, J. Intervention and identifiability in latent variable modelling. In *Minds & Machines 28*, pp. 243–264, 2018.

Rothenberg, T. J. Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pp. 577–591, 1971.

Schölkopf, B. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 765–804. 2022.

Semenova, L., Rudin, C., and Parr, R. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*, 2019.

Shamir, G. I., Lin, D., and Coviello, L. Smooth activations and reproducibility in deep networks. *arXiv preprint arXiv:2010.09931*, 2020.

Wilson, A. G. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4697–4708. Curran Associates, Inc., 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
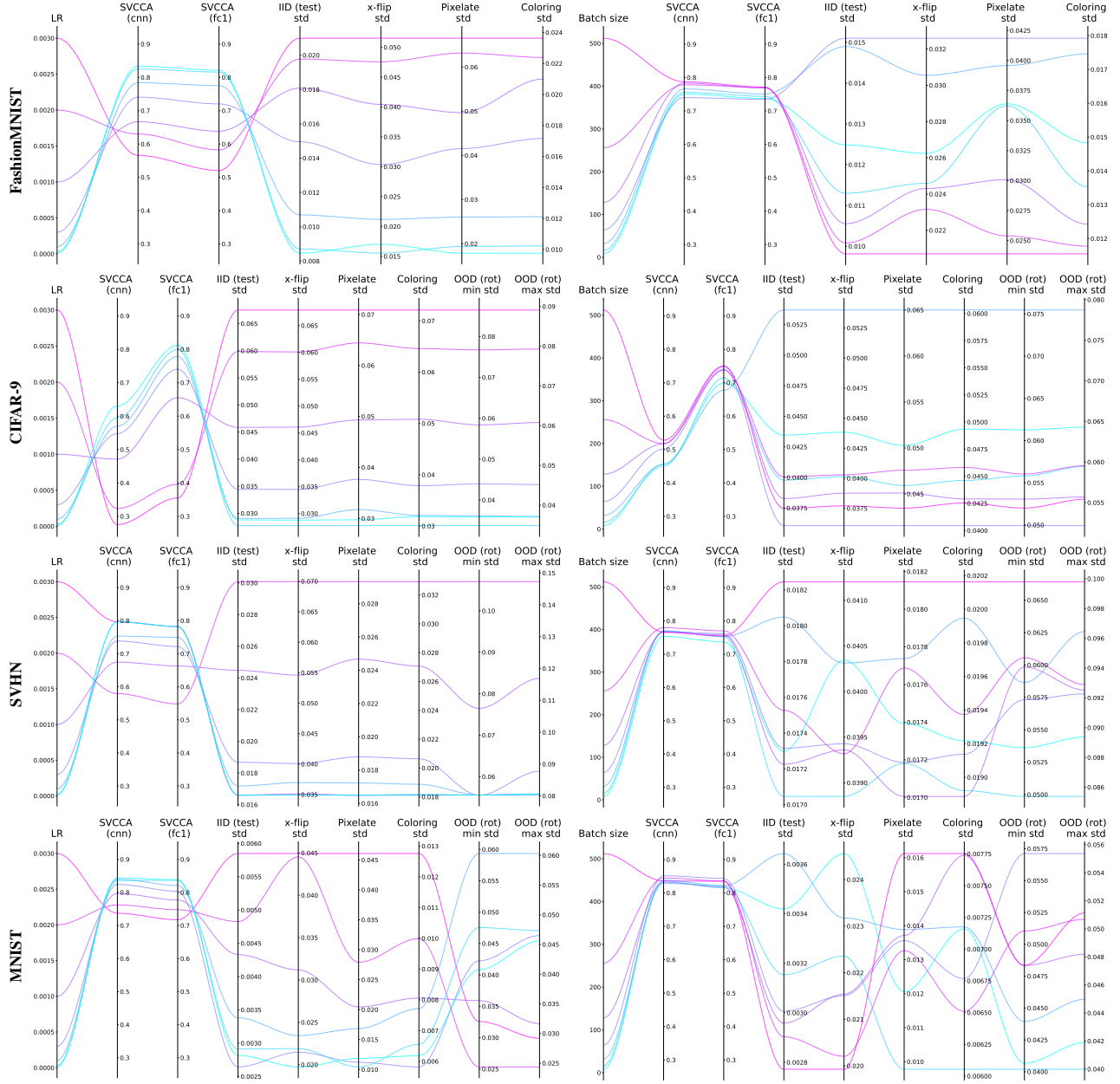
A. Extended Results

Figure 5. More comparisons of representative (RM) and predictive (PM) multiplicity, with SVHN data set included to showcase that SVHN in the learning rate regime breaks some of the trends observed with other data sets. Two hyper-parameter regimes are shown (**Left:** learning rate, **Right:** batch size) for four data sets. Identical model architectures were used for every hyper-parameter of a data set. SVCCA is the measure of inverted RM at two feature layers, while the prediction 'std' columns measure PM for the i.i.d. and various OOD distributions. data sets: FashionMNIST (top), CIFAR-9 (middle), SVHN, and MNIST (bottom). Low learning rate and high batch size strongly correlate with higher SVCCA (lower RM) and smaller variance (lower PM).

B. Training Details

Architectures   We trained MNIST, SVHN and FashionMNIST with the typical convolutional architecture $conv((1, 48, 3) - ReLU - MaxPool(2, 2) - conv(48, 96, 3) - ReLU - MaxPool(2, 2) - conv(96, 80, 3) - ReLU - conv(80, 96, 3) - ReLU - FC(96, 512) - FC(512, 10)$ and Cifar9 with $conv((3, 48, 3) - ReLU - MaxPool(2, 2) - conv(48, 96, 3) -$

Table 2. Test set accuracy statistics for the data sets used in this work.

Data set (HP regime)	Mean $\pm$ Std
CIFAR-9 (batch size)	0.709 ± 0.001
CIFAR-9 (learning rate)	0.643 ± 0.005
FashionMNIST (batch size)	0.940 ± 0.007
FashionMIST (learning rate)	0.930 ± 0.015
MNIST (batch size)	0.979 ± 0.001
MNIST (learning rate)	0.981 ± 0.002
SVHN (batch size)	0.905 ± 0.001
SVHN (learning rate)	0.894 ± 0.002

$ReLU - MaxPool(2,2) - conv(96,80,3) - ReLU - conv(80,96,3) - ReLU - FC(384,512) - FC(512,9).$

Training   Apart from the architecture, the same training methods and hyper-parameters were used for each data set and run, with the exception of the batch size and learning rate. In learning rate regime, we used batch size 64 while using the learning rates $[0.003, 0.002, 0.001, 0.0003, 0.0001, 0.00003, 0.00001]$. In the batch size regime, we used learning rate 0.0001 and batch sizes $[8, 16, 32, 64, 128, 256, 512]$. Cross entropy loss was used, with ADAM optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$).

For each data set, We first found the maximum test set accuracy comfortably achieved by every strategy for each hyper-parameter regime of that data set. We then trained 10 variants for each of them, varying only the random seed between the runs. We then used the checkpoints at the target levels for our evaluations. The accuracies were reached as in Table 2. Note that it was not our goal to achieve state-of-the-art accuracy for the experiments of this paper.

data sets   MNIST, FashionMNIST, SVHN, and CIFAR-9 were used for training. STL-9 was additionally used as an extra OOD test data set. CIFAR-9 was made by dropping the class 'frog' and STL-9 by dropping the class 'monkey', after which the class order of the two sets was made to match so that STL-9 served as the OOD data set for CIFAR-9 trained models. The standard training/validation split was used for each data set. Note that since we explicitly focused on our specific hyper-parameter strategies, there was no need for a separate validation and test data sets, hence, for each data set, the held-out data was treated as the i.i.d. 'test' data.

For out of distribution (OOD) data sets, we cross-matched data sets as SVHN against MNIST, and CIFAR-9 against STL-9. In addition, other OOD data sets (see Fig. 2) were created from the original data sets in the following manner. The 'x-flip' was created by random horizontal flip with .9 probability, 'color jitter' by randomly changing brightness uniformly with a factor 0.0 to 0.3 and hue by a factor -0.1 to $+0.1$. Pixelation was done by downscaling and upsampling by a factor of 2. 'OOD rot' refers to rotations done to the original data set by uniformly random increments of $0 \ldots 20$ degrees, $20 \ldots 30$ degrees, *etc.*, up to $90 \ldots 110$ degrees, to yield 10 OOD test cases. The 'min std' and 'max std' in the figure refer to the smallest and the largest rotation range, respectively.

The data sets contained no person identifiable data. All the data sets allow non-commercial usage for research purposes.

C. Training to Maximum Accuracy

We take as a given that there are less model variants that achieve the top performance than there are variants that achieve sub-par performance (for empirical comparisons, see Fort 2020). In other words, the increase in validation accuracy should correlate with decreasing diversity of possible solutions.

One should then ask, first, is it possible that the decreased diversity of solutions reached by some hyper-parameter strategies can be explained by the greater achieved accuracy? By controlling for the test accuracy across compared variants (Fig. 2), we preclude this explanation.

Second, one can further ask, whether the differences could still be explained by the greater *achievable* accuracy, that is, perhaps there is less diversity with the solutions that are on the right path to achieve maximum accuracy? Though the hypothesis appears contrived, we addressed it by checkpointing *but not stopping* the main experiments at the equivalent-risk level. Instead, we also followed each experiment until convergence, and identified the pseudo-maximum accuracy for each

training strategy, so that the maximum accuracy was achieved in at least 5 of the runs. (The absolute maximum accuracy would of course, strictly speaking, be only achieved by a single variant, preventing any dispersion analysis.)

In Fig. 5, we show the same results as in Fig. 4, but measured for the variants that achieve the pseudo-maximum accuracy for that strategy. Although one can see some repeating patterns, the decreased diversity of solutions certainly does not correlate with greater achieved accuracy. Hence, it is clear that the variation across the experiments is not explained by differences in either the achieved or potentially achievable accuracy.
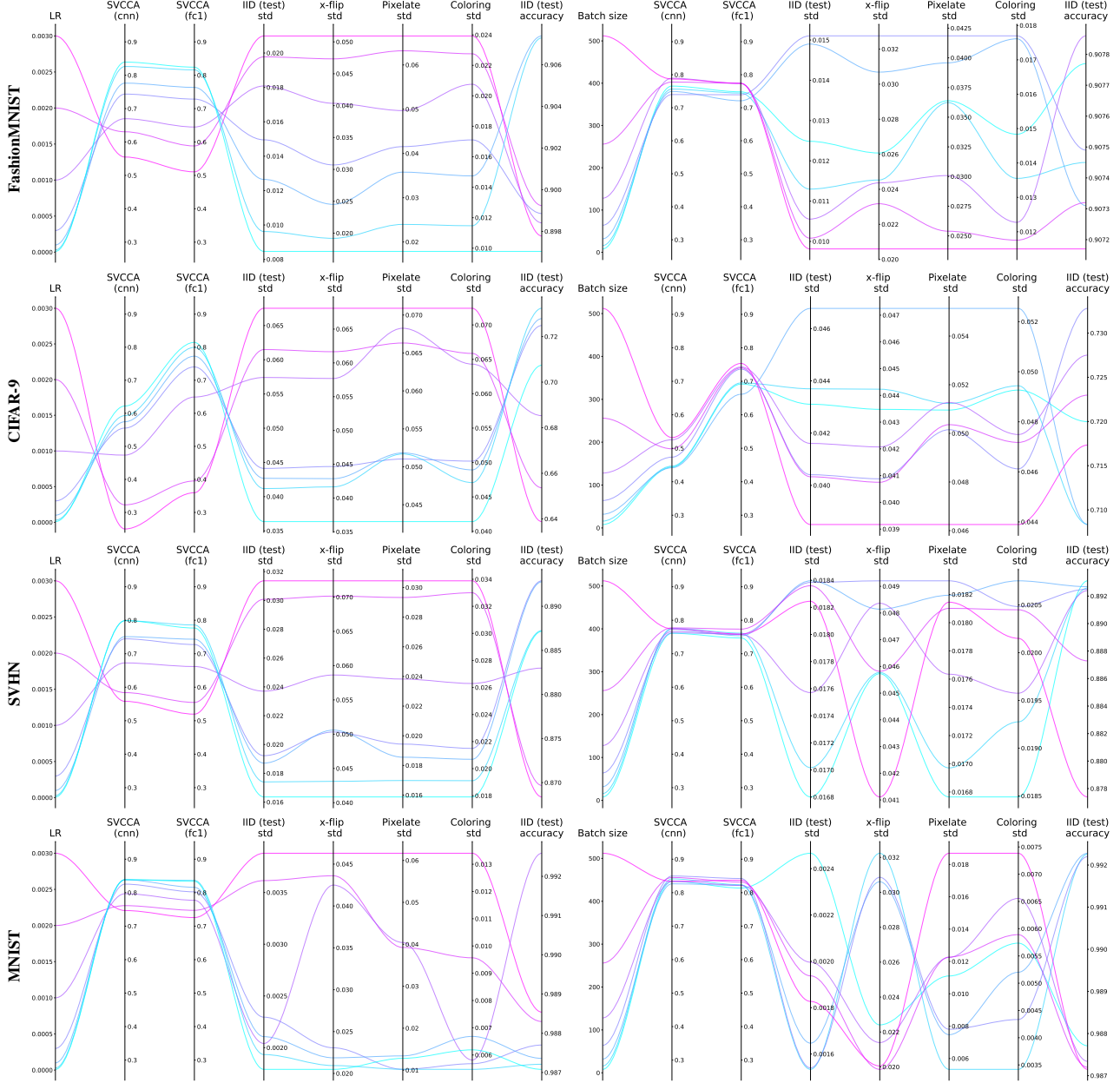

Figure 6. Results using the the pseudo-maximum test set accuracy.