
ID and OOD Performance Are Sometimes Inversely Correlated on Real-world Datasets

Damien Teney^{1,3} Seong Joon Oh² Ehsan Abbasnejad³

¹Idiap Research Institute ²University of Tübingen ³Australian Inst. for Machine Learning

Abstract

Context. Several studies have empirically compared in-distribution (ID) and out-of-distribution (OOD) performance of various models. They report frequent positive correlations on benchmarks in computer vision and NLP. Surprisingly, they never observe *inverse* correlations suggesting necessary trade-offs. This matters to determine whether ID performance can serve as a proxy for OOD generalization.

Findings. This short paper shows that inverse correlations between ID and OOD performance do happen in real-world benchmarks. They may have been missed in past studies because of a biased selection of models. We show an example of the pattern on the WILDS-Camelyon17 dataset, using models from multiple training epochs and random seeds. Our observations are particularly striking on models trained with a regularizer that diversifies the solutions to the ERM objective [19].

Implications. We nuance recommendations and conclusions made in past studies.

- High OOD performance may sometimes require trading off ID performance.
- Focusing on ID performance alone may not lead to optimal OOD performance: it can lead to diminishing and eventually negative returns in OOD performance.
- Our example reminds that empirical studies only chart regimes achievable with existing methods: care is warranted in deriving prescriptive recommendations.

1 Introduction

Past observations. This paper complements existing studies that empirically compare in-distribution (ID) and out-of-distribution¹ (OOD) performance of deep learning models [1, 3, 8–10, 18, 22]. It has long been known that models applied to OOD data suffer a drop in performance, e.g. in classification accuracy. The above studies show that, despite this gap, **ID and OOD performance are often positively correlated**² across models on benchmarks in computer vision [10] and NLP [9].

Past explanations. Frequent positive correlations are surprising because nothing forbids opposite, inverse ones. Indeed, ID and OOD data contain different associations between labels and features. One could imagine e.g. that an image background is associated with class C_1 ID and class C_2 OOD. The more a model relies on the presence of this background, the better its ID performance but the worse its OOD performance, resulting in an inverse correlation. Never observing inverse correlations has been explained with the possibility that **real-world benchmarks might contain only mild distribution shifts** [8]. We will show that such observations can also be an artefact of study design.

Correspondence to: damien.teney@idiap.ch.

¹We use “OOD” to generally refer to test data conforming to a covariate shift [17] w.r.t. the training data.

²We use “correlation” to refer both to linear and non-linear relationships.

| | Positive transfer | Underspecification | Misspecification | No transfer |
|--|--|---|--------------------------|--|
| Distribution shift | Mild | → (Too?) Severe | | |
| Typical pattern (toy representation) | Positive correlation | Vertical line/no clear trend | Negative correlation | Low horizontal line |
| Valid approaches for improving OOD performance | Simply focus on improving ID performance | Task-relevant inductive biases e.g. arch., regularizers. Data augmentation with task-relevant transformations. Non-i.i.d. training data e.g. multiple training domains. | | ??? (Open question) |
| Example benchmark | ImageNet → ImageNet v2 [13] | PACS sketch → photograph [7] | WILDS- Camelyon17 [5] | DomainNet infograph → quickdraw [5] |

Figure 1: Various patterns of ID vs OOD performance occur at different levels of distribution shift. This paper reminds that negative correlations are possible, though absent from past studies. They stem from misspecification, i.e. a conflict between the ERM objective (driving ID performance) and the goal of higher OOD performance.

A recent large-scale study. Wenzel et al. [22] show that not all datasets display a clear positive correlation. They also observe patterns that reveal underspecification [2, 6, 21], and severe shifts that prevent any training / test transfer. Surprisingly, they never observe inverse correlations:

“We did not observe any trade-off between accuracy and robustness, where more accurate models would overfit to spurious features that do not generalize.” [22]

On the contrary, we do observe such cases, which we showcase on a dataset also used in [22].

Explaining inverse correlations. We term the underlying cause a **misspecification**, by extension of an underspecification [2, 6, 21] previously used to explain why models with similar ID performance can vary in OOD performance. In cases of misspecification, the standard ERM objective (empirical risk minimization) is not aligned with the goal of OOD performance, such that ID and OOD metrics can vary independently and inversely to one another. This leads to an ordering of possible ID / OOD patterns according to the severity of the underlying distribution shift (see Figure 2).

The contributions of this paper are three-fold.

- An example of inverse correlation between ID / OOD performance on Camelyon17 [5] (Section 3).
- Explanations why past studies could miss this because of a biased sampling of models (Section 4).
- A revision of conclusions and recommendations made in past studies (Section 5).

2 Previously-observed patterns of ID vs OOD performance

Past studies conclude that ID and OOD performance tend to vary jointly across models on many real-world datasets [3, 10, 18]. Miller et al. [10] report an almost-systematic linear correlation³ between probit-scaled ID and OOD accuracies. Mania et al. [8] explain this trend with the fact that real-world benchmarks contain only mild distribution shifts.⁴ Andreassen et al. [1] find that pre-trained models perform “above the linear trend” in the early stages of fine-tuning. Their OOD accuracy rises more quickly than their ID accuracy early on, even though the final accuracies agree with a linear trend.

Most recently, the large-scale study of Wenzel et al. [22] is more nuanced: they observe a linear trend only on some datasets. Their setup consists in fine-tuning an ImageNet-pretrained model on a chosen dataset and evaluating it on matching ID and OOD test sets. They repeat the procedure with a variety of datasets, architectures, and implementation options such as data augmentations.

The scatter plots of ID vs OOD accuracy in [22] show four typical patterns (Figure 2).

³The linear correlation in [1, 10] is not really linear: it applies to probit-scaled accuracies (a non-linear transform).

⁴Mania et al. [8] explain the linear trend with (i) certain data points having similar probabilities of occurring in ID and OOD data, and (ii) the probability being low that a model correctly classifies some points that a higher-accuracy model classifies incorrectly.

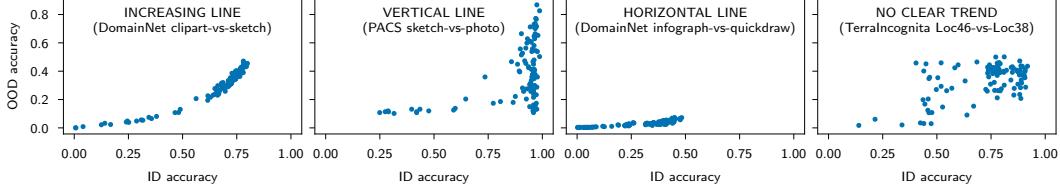


Figure 2: Typical patterns of ID vs OOD accuracy observed in [22] (reproduced with permission).

- Increasing line.** ID and OOD accuracies are positively correlated, indicating a **mild distribution shift**. Focusing on classical (ID) generalization brings concurrent OOD improvements.
- Vertical line.** Different models obtain a similar high ID performance but different OOD performance. This indicates **underspecification** [2, 6, 21]: the objective of high ID performance does not sufficiently constrain the learning. Typically, multiple features in the data (a.k.a. biased or spurious features) can be used to obtain high ID performance, but not all of them are equally reliable on OOD data. To improve OOD performance, additional task-specific information is necessary, e.g. additional supervision or inductive biases (custom architectures, regularizers, etc.).
- Horizontal line of low OOD accuracy.** No model performs well OOD. This indicates a **severe distribution shift** that prevents any transfer between training and OOD test data. The task needs to be significantly more constrained e.g. with task-specific inductive biases.
- No clear trend.** Models show a variety of ID and OOD accuracies, typically because of underspecification. The difference with (2) is the wider variety along the ID axis, e.g. because a difficult learning task yields solutions of lower ID accuracy from local minima of the ERM objective.

The study [22] notes the absence of decreasing patterns, even though they are possible in theory.

- Decreasing line.** The highest accuracy ID and OOD are achieved by different models. This indicates a **misspecification**: optima of the ERM objective – expected to be optima in ID performance – do not correspond to optima in OOD performance. This implies a necessary trade-off: higher OOD performance is possible at the cost of lower ID performance.

What causes inverse correlations between ID and OOD performance?

The cause is typically a pattern in the data that is predictive in one distribution and misleading in the other. For example, object classes C_1 and C_2 are respectively associated with image backgrounds B_1 and B_2 in ID data, and respectively B_2 in B_1 (**swapped**) in OOD data. Relying on the background can improve performance on either distribution but not both simultaneously. While such severe shifts might be rare, the next section presents an actual example thereof.

3 New observations: inversely correlated ID and OOD performance

We use the WILDS-Camelyon17 dataset [5] similarly to Wenzel et al. [22]. They compare models of different architectures, assuming that their different inductive biases will cover a range of ID/OOD accuracies. For simplicity, we rely instead on different random seeds since D’Amour et al. [2] showed this to be sufficient to cover a variety of ID/OOD accuracies on this dataset. To increase this variety even further without manually picking alternative architectures, we also train models with the diversity regularizer of Teney et al. [19](details below).

Experimental details. We use DenseNet-121 models trained by the authors of the dataset with 10 different seeds. For each of these 10 models, we retrain the last linear layer from a random initialization while keeping the rest of the model frozen. We record the ID and OOD accuracy after each training epoch from 1 to 10. These are referred to below as ERM-trained models. In addition, we repeat this training with the “diversity regularizer” from [19]. This method trains multiple copies of the same model (the last linear layer) in parallel with an additional objective that encourages each copy to rely on different features, by minimizing the mutual alignment of input gradients.⁵ We use 24

⁵The diversity regularizer was extended in [21] but we use the basic version from [19] for simplicity.

copies and a regularizer weight of 10, which were selected for giving a wide range of ID accuracies across the copies in each run. All training runs are repeated with 10 different seeds.

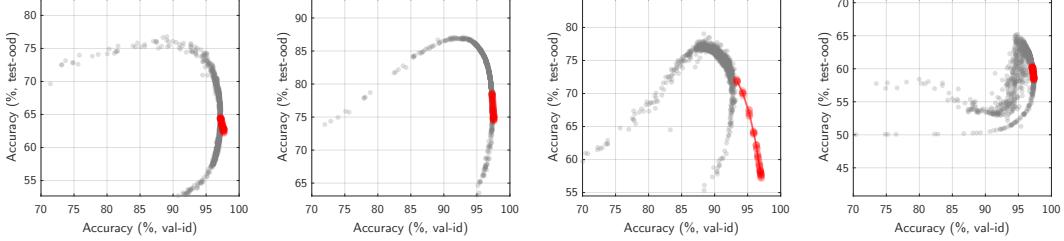


Figure 3: Our new observations show that higher OOD accuracy can be sometimes be traded off for lower ID accuracy. Each panel corresponds to a different pretraining seed. Each dot represents a linear classifier on frozen features, retrained either with standard ERM (red dots ●) or with an additional diversity regularizer [19] (gray dots ○). The latter includes models with clearly higher OOD accuracy but lower ID accuracy.

Results with ERM-trained models. In Figure 3 we plot the ID vs OOD accuracy of ERM-trained models as red dots (●). Each panel corresponds to a different pretraining seed. As noted in [5], OOD performance varies across seeds (note the different Y-axis limits) though the ID accuracy is similar. New observations are visible *within* each panel. Each dot corresponds a linear classifier on frozen features from a different seed or number of epochs. The seed makes little difference but the number of epochs creates patterns of decreasing trend i.e. negative correlation between ID and OOD performance. Despite the narrow ID variation (X axis), careful inspection confirms that the pattern appears in nearly all cases (see Appendix A for zoomed-in plots).

Results with diversity regularizer. We plot models trained with the diversity regularizer [19] as gray dots (○). They cover a wider range of accuracies and form patterns that extend those of ERM-trained models. The decreasing trend is now obvious. It is also clearly jointed to a rising trend where ID / OOD performance are positively correlated. This suggests a point of highest OOD performance beyond which the model overfits to ID data. Appendix A shows similar results on other pretrained models. The patterns are not always discernible however because large regions of the performance landscape aren't observed, even with the diversity regularizer. We further discuss this issue next.

4 Why past studies missed negative correlations: a biased sampling of models

We identified factors that can explain the discrepancy between our observations and past studies.

- ERM-trained models alone may not form clear patterns (red dots ● in Figure 3). In our case, **including models trained with a diversity regularizer** (gray dots ○) was key in making the suspected patterns more obvious, by covering a wider range of accuracies.
- The ID / OOD trade-off varies during training [1] and models of different architectures or seeds are not necessarily comparable, so **each** model should be observed at **different training stages**.
- Decreasing patterns are not equally apparent across **different pretraining seeds**. They sometimes require careful examination of zoomed-in plots (Appendix A). This reminds that empirical studies must randomize stochastic factors as much as possible, given their large effect on model behaviour.

To demonstrate these points, we plot our data again in Figure 4, keeping only ERM-trained models at 10 epochs, with all seeds on the same panel. These small changes reproduce the vertical line observed by Wenzel et al. [22], which completely misses the inverse correlations patterns from

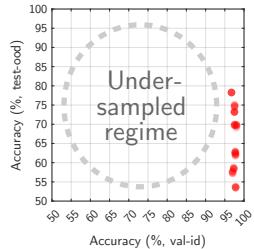


Figure 4: We plot again the ERM-trained models of Figure 3 (red dots ●) but **only include models trained for a fixed number of epochs** and combine pretraining seeds in one plot. This reproduces the vertical line from [22], which completely misses the patterns of inverse correlation.

A sufficient overall explanation is that past studies **undersample regions of the ID/OOD performance space**. They usually consider a variety of architectures in an attempt to sample this space. However, different architectures do not necessarily behave very differently from one another (see box below). We lack methods to reliably identify models of high OOD performance, but the diversity regularizer that we use yields models spanning a wide range of the performance spectrum.

Why isn't it sufficient to compare a variety of architectures?

Different architectures do not necessarily induce radically different behaviour. Even CNNs and vision transformers show similar failure modes [12]. Distinct architectures can induce similar inductive biases e.g. related to SGD, such as the simplicity bias [15, 16] or neural anisotropies [11].

Scimeca et al. [15] showed e.g. that architectures do not affect preferences for simple predictive features. Independently trained models as in Wenzel et al. [22] are not necessarily diverse despite a variety of architectures. Hence ID and OOD performance may only vary along similar directions.

5 Revisiting recommendations from past studies

We have established that observations in past studies were incomplete. We now bring nuance to some recommendations and conclusions made in these studies.

- **Focusing on a single metric.**

“We see the following potential prescriptive outcomes (...) correlation between OOD and ID performance can simplify model development since we can focus on a single metric.” [10]

We demonstrated that inverse correlations are a possibility, hence there exist scenarios where an ID metric would be misleading. In general, relying on a single metric during model development is ill-advised [20]. Even more so here since it cannot capture trade-offs along multiple axes. A model with a suboptimal ID performance may have learned features that enable better OOD generalization. Our recommendation is to track multiple metrics e.g. performance on multiple distributions or qualitative interpretable predictions on representative test points.

- **Improving ID performance for OOD robustness.**

“If practitioners want to make the model more robust on OOD data, the main focus should be to improve the ID classification error. (...) We speculate that the risk of overfitting large pre-trained models to the downstream test set is minimal, and it seems to be not a good strategy to, e.g., reduce the capacity of the model in the hope of better OOD generalization.” [22]

This recommendation assumes the persistence of a positive correlation. On the opposite, we saw that a positive correlation can precede a regime of inverse correlation (Figure 3, left panels). If the goal is to improve OOD, focusing on ID performance is a blind alley since this goal requires to increase ID performance at times, and reduce it at other times.

- **Future achievable OOD performance.**

As obvious as it is, it feels necessary to point out that empirical studies only chart regimes achievable with existing methods. Observations have limited predictive power, hence more care seems warranted when deriving prescriptive recommendations from empirical evidence.

For example, our own observations of an ID/OOD trade-off will not necessarily hold beyond the current Pareto front. Future models could achieve our best ID and OOD accuracies *simultaneously*.

6 Discussion

This paper showed that inverse correlations between ID/OOD performance are not only theoretically possible, but actually happen in real-world data. **We do not know how frequent this situation is.** We present a single counterexample to past claims that *positive* correlations are ubiquitous. This suffices to show that one cannot know a priori where a given task falls on the spectrum of Figure 1. It is thus **ill-advised to blindly make the assumption of a positive correlation** proposed in the past.

Can we avoid inverse correlations with a larger training set? Scaling alone without data curation [4] seems unlikely to prevent inverse correlations. They stem from biases in the training (ID) distribution (e.g. a class \mathcal{C}_1 appearing more frequently with image background \mathcal{B}_1 than any other). Biases in a distribution do not vanish with more samples. More data will cover more of the support of the distribution in the input space, but this coverage will remain uneven, i.e. biased. We then face a *subpopulation shift* [14] rather than a usual distribution shift, but this remains similarly challenging.

Training full networks with the diversity regularizer. We showed that inverse correlations occur with standard ERM-trained models, and also showed more striking observations with linear classifiers trained with a diversity regularizer [19]. It would be interesting to confirm these observations with entire networks trained with this regularizer. This was not done because of the computational expense. We do not foresee any reason why the observations would not hold.

Qualitative differences along the Pareto frontier. We focused on quantitative measures of performance. It would be interesting to examine models with various ID/OOD performance trade-offs and observe differences in the features they rely on e.g. with interpretability methods.

References

- [1] Andreassen, A., Bahri, Y., Neyshabur, B., Roelofs, R.: The evolution of out-of-distribution robustness throughout fine-tuning. arXiv preprint arXiv:2106.15831 (2021)
- [2] D’Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D., et al.: Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395 (2020)
- [3] Djolonga, J., Yung, J., Tschanne, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D’Amour, A., Moldovan, D., et al.: On robustness and transferability of convolutional neural networks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2021)
- [4] Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., Schmidt, L.: Data determines distributional robustness in contrastive language image pre-training (CLIP). arXiv preprint arXiv:2205.01397 (2022)
- [5] Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Beery, S., et al.: Wilds: A benchmark of in-the-wild distribution shifts. arXiv preprint arXiv:2012.07421 (2020)
- [6] Lee, Y., Yao, H., Finn, C.: Diversify and disambiguate: Learning from underspecified data. arXiv preprint arXiv:2202.03418 (2022)
- [7] Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 5542–5550 (2017)
- [8] Mania, H., Sra, S.: Why do classifier accuracies show linear trends under distribution shift? arXiv preprint arXiv:2012.15483 (2020)
- [9] Miller, J., Krauth, K., Recht, B., Schmidt, L.: The effect of natural distribution shift on question answering models. In: Proc. Int. Conf. Mach. Learn. (2020)
- [10] Miller, J.P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P.W., Shankar, V., Liang, P., Carmon, Y., Schmidt, L.: Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In: Proc. Int. Conf. Mach. Learn. (2021)
- [11] Ortiz-Jimenez, G., Salazar-Reque, I.F., Modas, A., Moosavi-Dezfooli, S.M., Frossard, P.: A neural anisotropic view of underspecification in deep learning. In: Proc. Int. Conf. Learn. Representations (2021)
- [12] Pinto, F., Torr, P.H., Dokania, P.K.: An impartial take to the cnn vs transformer robustness contest. arXiv preprint arXiv:2207.11347 (2022)
- [13] Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: Proc. Int. Conf. Mach. Learn. (2019)

- [14] Santurkar, S., Tsipras, D., Madry, A.: BREEDS: Benchmarks for subpopulation shift. arXiv preprint arXiv:2008.04859 (2020)
- [15] Scimeca, L., Oh, S.J., Chun, S., Poli, M., Yun, S.: Which shortcut cues will dnns choose? a study from the parameter-space perspective. In: Proc. Int. Conf. Learn. Representations (2022)
- [16] Shah, H., Tamuly, K., Raghunathan, A., Jain, P., Netrapalli, P.: The pitfalls of simplicity bias in neural networks. arXiv preprint arXiv:2006.07710 (2020)
- [17] Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference (2000)
- [18] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. Proc. Advances in Neural Inf. Process. Syst. (2020)
- [19] Teney, D., Abbasnejad, E., Lucey, S., van den Hengel, A.: Evading the simplicity bias: Training a diverse set of models discovers solutions with superior OOD generalization. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2022)
- [20] Teney, D., Kafle, K., Shrestha, R., Abbasnejad, E., Kanan, C., van den Hengel, A.: On the value of out-of-distribution testing: An example of Goodhart's law. In: Proc. Advances in Neural Inf. Process. Syst. (2020)
- [21] Teney, D., Peyrard, M., Abbasnejad, E.: Predicting is not understanding: Recognizing and addressing underspecification in machine learning. arXiv preprint arXiv:2207.02598 (2022)
- [22] Wenzel, F., Dittadi, A., Gehler, P.V., Simon-Gabriel, C.J., Horn, M., Zietlow, D., Kernert, D., Russell, C., Brox, T., Schiele, B., et al.: Assaying out-of-distribution generalization in transfer learning. arXiv preprint arXiv:2207.09239 (2022)

Appendices

A Additional results

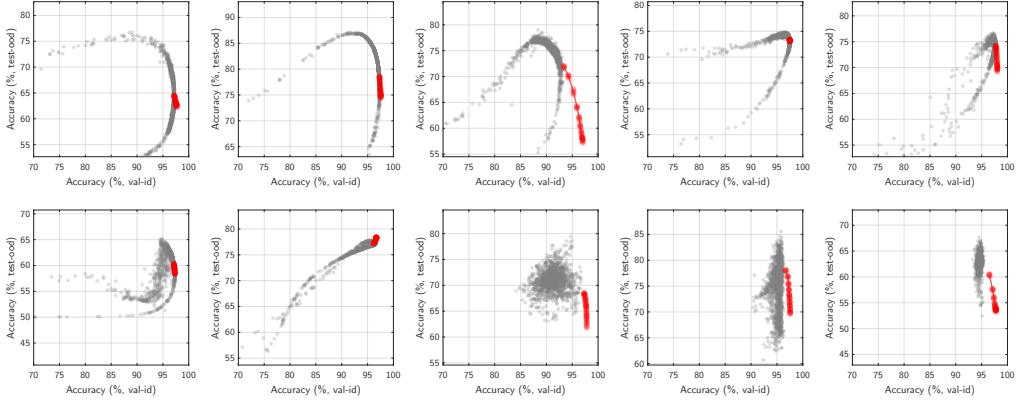


Figure 5: As in Figure 3, we show that higher OOD accuracy can be sometimes be traded off for a lower ID accuracy. Each panel shows results from a different pretrained model (i.e. pretrained with a different random seed). Each dot represents a linear classifier retrained on features from this pretrained model with standard ERM (red dots ●) or with a diversity regularizer [19] (gray dots ○). The latter produces models with clearly higher OOD accuracy but lower ID accuracy.

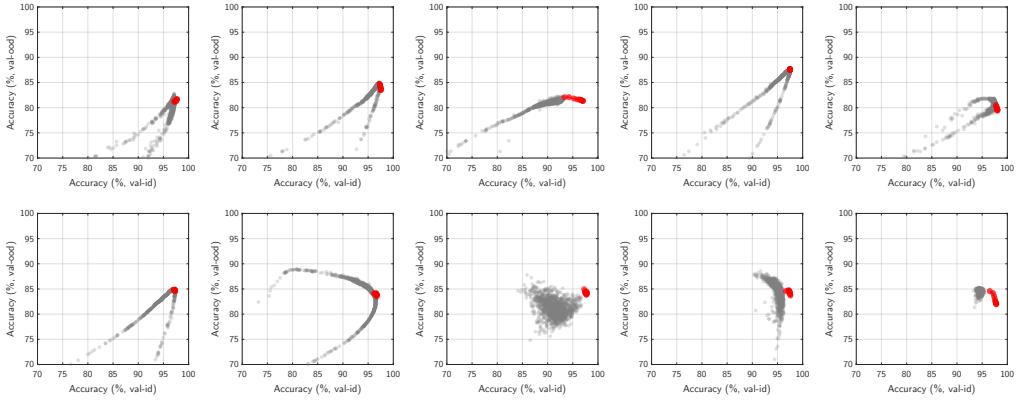


Figure 6: Same as in Figure 5, but using `val-ood` (instead of `test-ood`) as the OOD evaluation set.

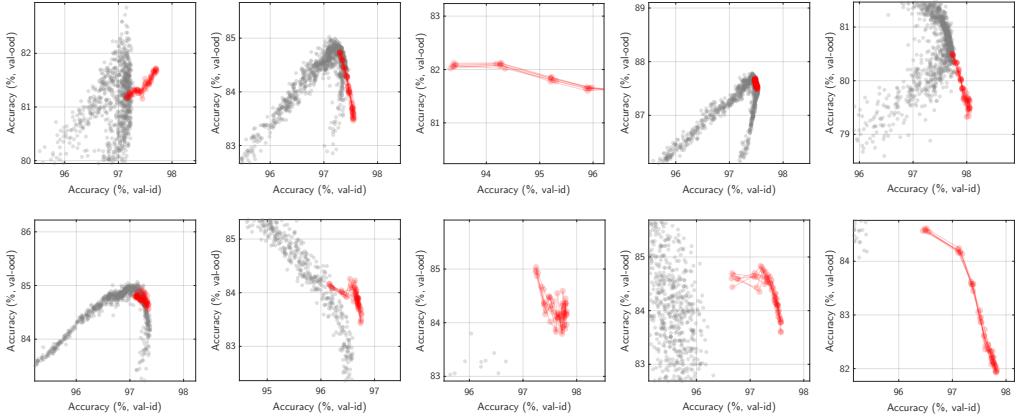


Figure 7: Same as in Figure 6, zoomed-in on ERM-trained models (red dots ●).