
Well-tuned Simple Nets Excel on Tabular Datasets

Arlind Kadra

Department of Computer Science
University of Freiburg
kadraa@cs.uni-freiburg.de

Marius Lindauer

Institute for Information Processing
Leibniz University Hannover
lindauer@tnt.uni-hannover.de

Frank Hutter

Department of Computer Science
University of Freiburg &
Bosch Center for Artificial Intelligence
fh@cs.uni-freiburg.de

Josif Grabocka

Department of Computer Science
University of Freiburg
grabocka@informatik.uni-freiburg.de

Abstract

Tabular datasets are the last “unconquered castle” for deep learning, with traditional ML methods like Gradient-Boosted Decision Trees still performing strongly even against recent specialized neural architectures. In this paper, we hypothesize that the key to boosting the performance of neural networks lies in rethinking the joint and simultaneous application of a large set of modern regularization techniques. As a result, we propose regularizing plain Multilayer Perceptron (MLP) networks by searching for the optimal combination/cocktail of 13 regularization techniques for each dataset using a joint optimization over the decision on which regularizers to apply and their subsidiary hyperparameters.

We empirically assess the impact of these *regularization cocktails* for MLPs in a large-scale empirical study comprising 40 tabular datasets and demonstrate that (i) well-regularized plain MLPs significantly outperform recent state-of-the-art specialized neural network architectures, and (ii) they even outperform strong traditional ML methods, such as XGBoost.

1 Introduction

In contrast to the mainstream in deep learning (DL), in this paper, we focus on tabular data, a domain that we feel is understudied in DL. Nevertheless, it is of great relevance for many practical applications, such as climate science, medicine, manufacturing, finance, recommender systems, etc. During the last decade, traditional machine learning methods, such as Gradient-Boosted Decision Trees (GBDT) [5], dominated tabular data applications due to their superior performance, and the success story DL has had for raw data (e.g., images, speech, and text) stopped short of tabular data.

Even in recent years, the existing literature still gives mixed messages on the state-of-the-art status of deep learning for tabular data. While some recent neural network methods [1, 46] claim to outperform GBDT, others confirm that GBDT are still the most accurate method on tabular data [48, 26]. The extensive experiments on 40 datasets we report indeed confirm that recent neural networks [1, 46, 11] do not outperform GBDT when the hyperparameters of all methods are thoroughly tuned.

We hypothesize that the key to improving the performance of neural networks on tabular data lies in exploiting the recent DL advances on regularization techniques (reviewed in Section 3), such as data augmentation, decoupled weight decay, residual blocks and model averaging (e.g., dropout or snapshot ensembles), or on learning dynamics (e.g., look-ahead optimizer or stochastic weight averaging).

Indeed, we find that even plain Multilayer Perceptrons (MLPs) achieve state-of-the-art results when regularized by multiple modern regularization techniques applied jointly and simultaneously.

Applying multiple regularizers jointly is already a common standard for practitioners, who routinely mix regularization techniques (e.g. Dropout with early stopping and weight decay). However, the deeper question of “Which subset of regularizers gives the largest generalization performance on a particular dataset among dozens of available methods?” remains unanswered, as practitioners currently combine regularizers via inefficient trial-and-error procedures. In this paper, we provide a simple, yet principled answer to that question, by posing the selection of the optimal subset of regularization techniques and their inherent hyperparameters, as a joint search for the best combination of MLP regularizers for each dataset among a pool of 13 modern regularization techniques and their subsidiary hyperparameters (Section 4).

From an empirical perspective, this paper is the first to provide compelling evidence that well-regularized neural networks (even simple MLPs!) indeed surpass the current state-of-the-art models in tabular datasets, including recent neural network architectures and GBDT (Section 6). In fact, the performance improvements are quite pronounced and highly significant.¹ We believe this finding to potentially have far-reaching implications, and to open up a garden of delights of new applications on tabular datasets for DL.

Our contributions are as follows:

1. We demonstrate that modern DL regularizers (developed for DL applications on raw data, such as images, speech, or text) also substantially improve the performance of deep multi-layer perceptrons on tabular data.
2. We propose a simple, yet principled, paradigm for selecting the optimal subset of regularization techniques and their subsidiary hyperparameters (so-called *regularization cocktails*).
3. We demonstrate that these regularization cocktails enable even simple MLPs to outperform both recent neural network architectures, as well as traditional strong ML methods, such as GBDT, on tabular data. Specifically, we are the first to show neural networks to significantly (and substantially) outperform XGBoost in a fair, large-scale experimental study.

2 Related Work on Deep Learning for Tabular Data

Recently, various neural architectures have been proposed for improving the performance of neural networks on tabular data. TabNet [1] introduced a sequential attention mechanism for capturing salient features. Neural oblivious decision ensembles (NODE [46]) blend the concept of hierarchical decisions into neural networks. Self-normalizing neural networks [29] have neuron activations that converge to zero mean and unit variance, which in turn, induces strong regularization and allows for high-level representations. Regularization learning networks train a regularization strength on every neural weight by posing the problem as a large-scale hyperparameter tuning scheme [48]. The recent NET-DNF technique introduces a novel inductive bias in the neural structure corresponding to logical Boolean formulas in disjunctive normal forms [26]. An approach that is often mistaken as deep learning for tabular data is AutoGluon Tabular [11], which builds ensembles of basic neural networks together with other traditional ML techniques, with its key contribution being a strong stacking approach. We emphasize that some of these publications claim to outperform Gradient Boosted Decision Trees (GBDT) [1, 46], while other papers explicitly stress that the neural networks tested do not outperform GBDT on tabular datasets [48, 26]. In contrast, we do not propose a new kind of neural architecture, but a novel paradigm for learning a combination of regularization methods.

3 An Overview of Regularization Methods for Deep Learning

Weight decay: The most classical approaches of regularization focused on minimizing the norms of the parameter values, e.g., either the L1 [51], the L2 [52], or a combination of L1 and L2 known as the Elastic Net [63]. A recent work fixes the malpractice of adding the decay penalty term before

¹In that sense, this paper adds to the growing body of literature on demonstrating the sensitivity of modern ML methods to hyperparameter settings [19], demonstrating that proper hyperparameter tuning can yield substantial improvements of modern ML methods [6], and demonstrating that even simple architectures can obtain state-of-the-art performance with proper hyperparameter settings [38].

momentum-based adaptive learning rate steps (e.g., in common implementations of Adam [27]), by decoupling the regularization from the loss and applying it after the learning rate computation [36].

Data Augmentation: Among the augmentation regularizers, Cut-Out [10] proposes to mask a subset of input features (e.g., pixel patches for images) for ensuring that the predictions remain invariant to distortions in the input space. Along similar lines, Mix-Up [60] generates new instances as a linear span of pairs of training examples, while Cut-Mix [58] suggests super-positions of instance pairs with mutually-exclusive pixel masks. A recent technique, called Aug-Mix [20], generates instances by sampling chains of augmentation operations. On the other hand, the direction of reinforcement learning (RL) for augmentation policies was elaborated by Auto-Augment [7], followed by a technique that speeds up the training of the RL policy [34]. Recently, these complex and expensive methods were superseded by simple and cheap methods that yield similar performance (RandAugment [8]) or even improve on it (TrivialAugment [41]). Last but not least, adversarial attack strategies (e.g., FGSM [17]) generate synthetic examples with minimal perturbations, which are employed in training robust models [37].

Ensemble methods: Ensembled machine learning models have been shown to reduce variance and act as regularizers [45]. A popular ensemble neural network with shared weights among its base models is Dropout [49], which was extended to a variational version with a Gaussian posterior of the model parameters [28]. As a follow-up, Mix-Out [32] extends Dropout by statistically fusing the parameters of two base models. Furthermore, so-called *snapshot ensembles* [21] can be created using models from intermediate convergence points of stochastic gradient descent with restarts [35]. In addition to these efficient ensembling approaches, ensembling independent classifiers trained in separate training runs can yield strong performance (especially for uncertainty quantification), be it based on independent training runs only differing in random seeds (*deep ensembles* [31]), training runs differing in hyperparameter settings (*hyperdeep ensembles*, [55]), or training runs with different neural architectures (*neural ensemble search* [59]).

Structural and Linearization: In terms of structural regularization, ResNet adds skip connections across layers [18], while the Inception model computes latent representations by aggregating diverse convolutional filter sizes [50]. A recent trend adds a dosage of *linearization* to deep models, where skip connections transfer embeddings from previous less non-linear layers [18, 22]. Along similar lines, the Shake-Shake regularization deploys skip connections in parallel convolutional blocks and aggregates the parallel representations through affine combinations [15], while Shake-Drop extends this mechanism to a larger number of CNN architectures [56].

Implicit: The last family of regularizers broadly encapsulates methods that do not directly propose novel regularization techniques but have an *implicit* regularization effect as a virtue of their ‘modus operandi’ [2]. The simplest such implicit regularization is Early Stopping [57], which limits overfitting by tracking validation performance over time and stopping training when validation performance no longer improves. Another implicit regularization method is Batch Normalization, which improves generalization by reducing internal covariate shift [24]. The scaled exponential linear units (SELU) represent an alternative to batch-normalization through self-normalizing activation functions [30]. On the other hand, stabilizing the convergence of the training routine is another implicit regularization, for instance by introducing learning rate scheduling schemes [35]. The recent strategy of stochastic weight averaging relies on averaging parameter values from the local optima encountered along the sequence of optimization steps [25], while another approach conducts updates in the direction of a few ‘lookahead’ steps [61].

4 Regularization Cocktails for Multilayer Perceptrons

4.1 Problem Definition

A training set is composed of features $\mathbf{X}^{(\text{Train})}$ and targets $\mathbf{y}^{(\text{Train})}$, while the test dataset is denoted by $\mathbf{X}^{(\text{Test})}, \mathbf{y}^{(\text{Test})}$. A parametrized function f , i.e., a neural network, approximates the targets as $\hat{\mathbf{y}} = f(\mathbf{X}; \boldsymbol{\theta})$, where the parameters $\boldsymbol{\theta}$ are trained to minimize a differentiable loss function \mathcal{L} as $\arg \min_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{y}^{(\text{Train})}, f(\mathbf{X}^{(\text{Train})}, \boldsymbol{\theta}))$. To generalize into minimizing $\mathcal{L}(\mathbf{y}^{(\text{Test})}, f(\mathbf{X}^{(\text{Test})}, \boldsymbol{\theta}))$, the parameters of f are controlled with a regularization technique Ω that avoids overfitting to the peculiarities of the training data. With a slight abuse of notation we denote $f(\mathbf{X}; \Omega(\boldsymbol{\theta}; \boldsymbol{\lambda}))$ to be the predictions of the model f whose parameters $\boldsymbol{\theta}$ are optimized under the regime of the regularization

method $\Omega(\cdot; \lambda)$, where $\lambda \in \Lambda$ represents the hyperparameters of Ω . The training data is further divided into two subsets as training and validation splits, the later denoted by $\mathbf{X}^{(\text{Val})}, \mathbf{y}^{(\text{Val})}$, such that λ can be tuned on the validation loss via the following hyperparameter optimization objective:

$$\begin{aligned} \lambda^* \in \arg \min_{\lambda \in \Lambda} \mathcal{L}(\mathbf{y}^{(\text{Val})}, f(\mathbf{X}^{(\text{Val})}; \theta_\lambda^*)), \\ \text{s.t. } \theta_\lambda^* \in \arg \min_{\theta} \mathcal{L}(\mathbf{y}^{(\text{Train})}, f(\mathbf{X}^{(\text{Train})}; \Omega(\theta; \lambda))). \end{aligned} \quad (1)$$

After finding the optimal (or in practice at least a well-performing) configuration λ^* , we re-fit θ on the entire training dataset, i.e., $\mathbf{X}^{(\text{Train})} \cup \mathbf{X}^{(\text{Val})}$ and $\mathbf{y}^{(\text{Train})} \cup \mathbf{y}^{(\text{Val})}$.

While the search for optimal hyperparameters λ is an active field of research in the realm of AutoML [23], still the choice of the regularizer Ω mostly remains an ad-hoc practice, where practitioners select a few combinations among popular regularizers (Dropout, L2, Batch Normalization, etc.). In contrast to prior studies, we hypothesize that the optimal regularizer is a cocktail mixture of a large set of regularization methods, all being simultaneously applied with different strengths (i.e., dataset-specific hyperparameters). Given a set of K regularizers $\{\Omega^{(k)}(\cdot; \lambda^{(k)})\}_{k=1}^K := \{\Omega^{(1)}(\cdot; \lambda^{(1)}), \dots, \Omega^{(K)}(\cdot; \lambda^{(K)})\}$, each with its own hyperparameters $\lambda^{(k)} \in \Lambda^{(k)}, \forall k \in \{1, \dots, K\}$, the problem of finding the optimal cocktail of regularizers is:

$$\begin{aligned} \lambda^* \in \arg \min_{\lambda := (\lambda^{(1)}, \dots, \lambda^{(K)}) \in (\Lambda^{(1)}, \dots, \Lambda^{(K)})} \mathcal{L}(\mathbf{y}^{(\text{Val})}, f(\mathbf{X}^{(\text{Val})}; \theta_\lambda^*)) \\ \text{s.t. } \theta_\lambda^* \in \arg \min_{\theta} \mathcal{L}\left(\mathbf{y}^{(\text{Train})}, f\left(\mathbf{X}^{(\text{Train})}; \left\{\Omega^{(k)}(\theta, \lambda^{(k)})\right\}_{k=1}^K\right)\right) \end{aligned} \quad (2)$$

The intuitive interpretation of Equation 2 is searching for the optimal hyperparameters λ (i.e., strengths) of the cocktail’s regularizers using the validation set, given that the optimal prediction model parameters θ are trained under the regime of all the regularizers being applied jointly. We stress that, for each regularizer, the hyperparameters $\lambda^{(k)}$ include a conditional hyperparameter controlling whether the k -th regularizer is applied or skipped. The best cocktail might comprise only a subset of regularizers.

4.2 Cocktail Search Space

To build our regularization cocktails we combine the 13 regularization methods listed in Table 1, which represent the categories of regularizers covered in Section 3. The regularization cocktail’s search space with the exact ranges for the selected regularizers’ hyperparameters is given in the same table. In total, the optimal cocktail is searched in a space of 19 hyperparameters.

While we can in principle use any hyperparameter optimization method, we decided to use the multi-fidelity Bayesian optimization method BOHB [12] since it achieves strong performance across a wide range of computing budgets by combining Hyperband [33] and Bayesian Optimization [40], and since BOHB can deal with the categorical hyperparameters we use for enabling or disabling regularization techniques and the corresponding conditional structures. Appendix A describes the implementation details for the deployed HPO method. Some of the regularization methods cannot be combined, and we, therefore, introduce the following constraints to the proposed search space: (i) Shake-Shake and Shake-Drop are not simultaneously active since the latter builds on the former; (ii) Only one data augmentation technique out of Mix-Up, Cut-Mix, Cut-Out, and FGSM adversarial learning can be active at once due to a technical limitation of the base library we use [62].

5 Experimental Protocol

5.1 Experimental Setup and Datasets

We use a large collection of 40 tabular datasets (listed in Table 9 of Appendix D). This includes 31 datasets from the recent open-source OpenML AutoML Benchmark [16]². In addition, we added

²The remaining 8 datasets from that benchmark were too large to run effectively on our cluster.

Group	Regularizer	Hyperparameter	Type	Range	Conditionality
Implicit	BN	BN-active	Boolean	{True, False}	-
	SWA	SWA-active	Boolean	{True, False}	-
	LA	LA-active Step size Num. steps	Boolean Continuous Integer	{True, False} [0.5, 0.8] [5, 10]	- LA-active LA-active
W. Decay	WD	WD-active Decay factor	Boolean Continuous	{True, False} [10 ⁻⁵ , 0.1]	- WD-active
Ensemble	DO	DO-active	Boolean	{True, False}	-
		Dropout shape	Nominal	{funnel, long funnel, diamond, hexagon, brick, triangle, stairs}	DO-active
		Drop rate	Continuous	[0.0, 0.8]	DO-active
	SE	SE-active	Boolean	{True, False}	-
Structural	SC	SC-active MB choice	Boolean Nominal	{True, False} {SS, SD, Standard}	- SC-active
	SD	Max. probability	Continuous	[0.0, 1.0]	SC-active \wedge MB choice = SD
	SS	-	-	-	SC-active \wedge MB choice = SS
Augmentation	-	Augment	Nominal	{MU, CM, CO, AT, None}	-
	MU	Mix. magnitude	Continuous	[0.0, 1.0]	Augment = MU
	CM	Probability	Continuous	[0.0, 1.0]	Augment = CM
	CO	Probability	Continuous	[0.0, 1.0]	Augment = CO
		Patch ratio	Continuous	[0.0, 1.0]	Augment = CO
	AT	-	-	-	Augment = AT

Table 1: The configuration space for the regularization cocktail regarding the **explicit regularization hyperparameters** of the methods and the conditional constraints enabling or disabling them. (BN: Batch Normalization, SWA: Stochastic Weight Averaging, LA: Lookahead Optimizer, WD: Weight Decay, DO: Dropout, SE: Snapshot Ensembles, SC: Skip Connection, MB: Multi-branch choice, SD: Shake-Drop, SS: Shake-Shake, MU: Mix-Up, CM: Cut-Mix, CO: Cut-Out, and AT: FGSM Adversarial Learning)

9 popular datasets from UCI [3] and Kaggle that contain roughly 100K+ instances. Our resulting benchmark of 40 datasets includes tabular datasets that represent diverse classification problems, containing between 452 and 416 188 data points, and between 4 and 2 001 features, varying in terms of the number of numerical and categorical features. The datasets are retrieved from the OpenML repository [54] using the OpenML-Python connector [14] and split as 60% training, 20% validation, and 20% testing sets. The data is standardized to have zero mean and unit variance where the statistics for the standardization are calculated on the training split.

We ran all experiments on a CPU cluster, each node of which contains two Intel Xeon E5-2630v4 CPUs with 20 CPU cores each, running at 2.2GHz and a total memory of 128GB. We chose the PyTorch library [43] as a deep learning framework and extended the AutoDL-framework Auto-Pytorch [39, 62] with our implementations for the regularizers of Table 1. We provide the code for our implementation at the following link: <https://github.com/releaunifreiburg/WellTunedSimpleNets>.

To optimally utilize resources, we ran BOHB with 10 workers in parallel, where each worker had access to 2 CPU cores and 12GB of memory, executing one configuration at a time. Taking into account the dimensions D of the considered configuration spaces, we ran BOHB for at most 4 days, or at most $40 \times D$ hyperparameter configurations, whichever came first. During the training phase, each configuration was run for 105 epochs, in accordance with the cosine learning rate annealing with restarts (described in the following subsection). For the sake of studying the effect on more datasets, we only evaluated a single train-val-test split. After the training phase is completed, we report the results of the best hyperparameter configuration found, retrained on the joint train and validation set.

5.2 Fixed Architecture and Optimization Hyperparameters

In order to focus exclusively on investigating the effect of regularization we fix the neural architecture to a simple multilayer perceptron (MLP) and also fix some hyperparameters of the general training

procedure. These fixed hyperparameter values, as specified in Table 4 of Appendix B.1, have been tuned for maximizing the performance of an unregularized neural network on our dataset collection (see Table 9 in Appendix D). We use a 9-layer feed-forward neural network with 512 units for each layer, a choice motivated by previous work [42].

Moreover, we set a low learning rate of 10^{-3} after performing a grid search for finding the best value across datasets. We use AdamW [36], which implements decoupled weight decay, and cosine annealing with restarts [35] as a learning rate scheduler. Using a learning rate scheduler with restarts helps in our case because we keep a fixed initial learning rate. For the restarts, we use an initial budget of 15 epochs, with a budget multiplier of 2, following published practices [62]. Additionally, since our benchmark includes imbalanced datasets, we use a weighted version of categorical cross-entropy and balanced accuracy [4] as the evaluation metric.

5.3 Research Hypotheses and Associated Experiments

Hypothesis 1: Regularization cocktails outperform state-of-the-art deep learning architectures on tabular datasets.

Experiment 1: We compare our well-regularized MLPs against the recently proposed deep learning architectures Node [46] and TabNet [1]. Additionally, we compare against two versions of AutoGluon Tabular [11], a version that features stacking and a version that additionally includes hyperparameter optimization. Moreover, we add an unregularized version of our MLP for reference, as well as a version of our MLP regularized with Dropout (where the dropout hyperparameters are tuned on every dataset). Lastly, we also compare against self-normalizing neural networks [29] by using the same MLP backbone as with our regularization cocktails.

Hypothesis 2: Regularization cocktails outperform Gradient-Boosted Decision Trees (GBDTs), the most commonly used traditional ML method and de-facto state-of-the-art for tabular data.

Experiment 2: We compare against three different implementations of GBDT: an implementation from scikit-learn [44] and optimized by Auto-sklearn [13], the popular XGBoost [5], and lastly, the recently proposed CatBoost [47].

Hypothesis 3: Regularization cocktails are time-efficient and achieve strong anytime results.

Experiment 3: We compare our regularization cocktails against XGBoost over time.

5.4 Experimental Setup for the Baselines

All baselines use the same train, validation, and test splits, the same seed, and the same HPO resources and constraints as for our automatically-constructed regularization cocktails (4 days on 20 CPU cores with 128GB of memory). After finding the best incumbent configuration, the baselines are refitted on the union of the training and validation sets and evaluated on the test set. The baselines consist of two recent neural architectures, two versions of AutoGluon Tabular with neural networks, and three implementations of GBDT, as follows:

TabNet: This library does not provide an HPO algorithm by default; therefore, we also used BOHB for this search space, with the hyperparameter value ranges recommended by the authors [1].

Node: This library does not offer an HPO algorithm by default. We performed a grid search among the hyperparameter value ranges as proposed by the authors [46]; however, we faced multiple memory and runtime issues in running the code. To overcome these issues we used the default hyperparameters the authors used in their public implementation.

AutoGluon Tabular: This library constructs stacked ensembles with bagging among diverse neural network architectures having various kinds of regularization [11]. The training of the stacking ensemble of neural networks and its hyperparameter tuning are integrated into the library. Hyperparameter optimization (HPO) is deactivated by default to give more resources to stacking, but here we study AutoGluon based on either stacking or HPO (and HPO actually performs somewhat better). While AutoGluon Tabular by default uses a broad range of traditional ML techniques, here, in order to study it as a “pure” deep learning method, we restrict it to only use neural networks as base learners.

ASK-GBDT: The GBDT implementation of scikit-learn offered by Auto-sklearn [13] uses SMAC for HPO, and we used the default hyperparameter search space given by the library.

XGBoost: The original library [5] does not incorporate an HPO algorithm by default, so we used BOHB for its HPO. We defined a search space for XGBoost’s hyperparameters following the best practices by the community; we describe this in the Appendix B.2.

CatBoost: Like for XGBoost, the original library [47] does not incorporate an HPO algorithm, so we used BOHB for its HPO, with the hyperparameter search space recommended by the authors.

For in-depth details about the different baseline configurations with the exact hyperparameter search spaces, please refer to Appendix B.2.

Dataset	#Ins./#Feat.	MLP	MLP+D	XGB.	ASK-G.	TabN.	Node	AutoGL. S	MLP+C
anneal	898 / 39	84.131	86.916	85.416	90.000	84.248	20.000	80.000	89.270
kr-vs-kp	3196 / 37	99.701	99.850	99.850	99.850	93.250	97.264	99.687	99.850
arrhythmia	452 / 280	37.991	38.704	48.779	46.850	43.562	N/A	48.934	61.461
mfeat.	2000 / 217	97.750	98.000	98.000	97.500	97.250	97.250	98.000	98.000
credit-g	1000 / 21	69.405	68.095	68.929	71.191	61.190	73.095	69.643	74.643
vehicle	846 / 19	83.766	82.603	74.973	80.165	79.654	75.541	83.793	82.576
kc1	2109 / 22	70.274	72.980	66.846	63.353	52.517	55.803	67.270	74.381
adult	48842 / 15	76.893	78.520	79.824	79.830	77.155	78.168	80.557	82.443
walking.	149332 / 5	60.997	63.754	61.616	62.764	56.801	N/A	60.800	63.923
phoneme	5404 / 6	87.514	88.387	87.972	88.341	86.824	82.720	83.943	86.619
skin-seg.	245057 / 4	99.971	99.962	99.968	99.967	99.961	N/A	99.973	99.953
ldpa	164860 / 8	62.831	67.035	99.008	68.947	54.815	N/A	53.023	68.107
nomao	34465 / 119	95.917	96.232	96.872	97.217	95.425	96.217	96.420	96.826
cnae	1080 / 857	87.500	90.741	94.907	93.519	89.352	96.759	92.593	95.833
blood.	748 / 5	67.836	68.421	62.281	64.985	64.327	50.000	67.251	67.617
bank.	45211 / 17	78.076	83.145	72.658	72.283	70.639	74.607	79.483	85.993
connect.	67557 / 43	73.627	76.345	72.374	72.645	72.045	N/A	75.622	80.073
shuttle	58000 / 10	99.475	99.892	98.563	98.571	88.017	42.805	83.433	99.948
higgs	98050 / 29	67.752	66.873	72.944	72.926	72.036	N/A	73.798	73.546
australian	690 / 15	86.268	86.268	89.717	88.589	85.278	83.468	88.248	87.088
car	1728 / 7	97.442	99.690	92.376	100.000	98.701	46.119	99.675	99.587
segment	2310 / 20	94.805	94.589	93.723	93.074	91.775	90.043	91.991	93.723
fashion.	70000 / 785	90.464	90.507	91.243	90.457	89.793	N/A	91.336	91.950
jungle.	44819 / 7	97.061	97.237	87.325	83.070	73.425	N/A	93.017	97.471
numera1	96320 / 22	50.262	50.301	52.363	52.421	51.599	52.364	51.706	52.668
devnagari	92000 / 1025	96.125	97.000	93.310	77.897	94.179	N/A	97.734	98.370
helena	65196 / 28	16.836	23.983	21.994	21.144	19.032	N/A	27.115	27.701
jannis	83733 / 55	51.505	55.118	55.225	55.593	56.214	N/A	58.526	65.287
volkert	58310 / 181	65.081	66.996	64.170	63.428	59.409	N/A	70.195	71.667
miniboone	130064 / 51	90.639	94.099	94.024	94.137	62.173	N/A	94.978	94.015
apsfailure	76000 / 171	87.759	91.194	88.825	91.797	51.444	N/A	88.890	92.535
christine	5418 / 1637	70.941	70.756	74.815	74.447	69.649	73.247	74.170	74.262
dilbert	10000 / 2001	96.930	96.733	99.106	98.704	97.608	N/A	98.758	99.049
fabert	8237 / 801	63.707	64.814	70.098	70.120	62.277	66.097	68.142	69.183
jasmine	2984 / 145	78.048	76.211	80.546	78.878	76.690	80.053	80.046	79.217
sylvine	5124 / 21	93.070	93.363	95.509	95.119	83.595	93.852	93.753	94.045
dionis	416188 / 61	91.905	92.724	91.222	74.620	83.960	N/A	94.127	94.010
aloi	108000 / 129	92.331	93.852	95.338	13.534	93.589	N/A	97.423	97.175
ccfraud	284807 / 31	50.000	50.000	90.303	92.514	85.705	N/A	91.831	92.531
clickpred.	399482 / 12	63.125	64.367	58.361	58.201	50.163	N/A	54.410	64.280
Wins/Losses/Ties	MLP+C vs ...	35/5/0	30/8/2	26/12/2	29/11/0	38/2/0	19/2/0	30/9/1	-
Wilcoxon p-value	MLP+C vs ...	5.3×10^{-7}	8.9×10^{-6}	6×10^{-4}	2.8×10^{-4}	4.5×10^{-8}	8.2×10^{-8}	4×10^{-5}	-

Table 2: Comparison of well-regularized MLPs vs. other methods in terms of balanced accuracy. N/A values indicate a failure due to exceeding the cluster’s memory (24GB per process) or runtime limits (4 days). The acronyms stand for MLP+D: MLP with Dropout, XGB.: XGBoost, ASK-G.: GBDT by Auto-sklearn, AutoGL. S: Autogluon with stacking enabled, TabN.: TabNet and MLP+C: our MLP regularized by cocktails.

6 Experimental Results

We present the comparative results of our MLPs regularized with the proposed regularization cocktails (MLP+C) against ten baselines (descriptions in Section 5.4): (a) two state-of-the-art architectures (NODE, TabN.); (b) two AutoGluon Tabular variants with neural networks that features stacking (AutoGL. S) and additionally HPO (AutoGL. HPO); (c) three Gradient-Boosted Decision Tree

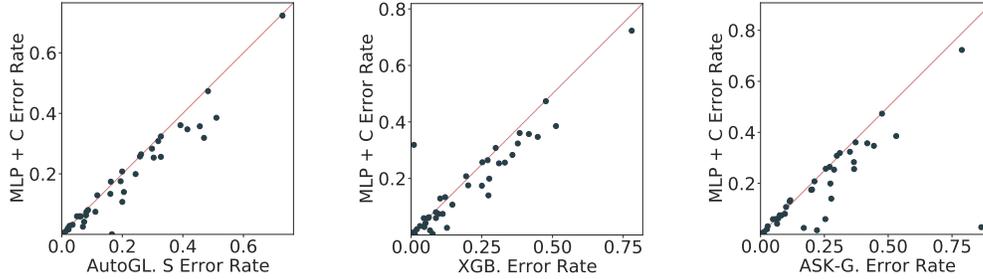


Figure 1: Comparison of our proposed dataset-specific cocktail (MLP+C) against the top three baselines. Each dot in the plot represents a dataset, the y-axis our method’s errors and the x-axis the baselines’ errors.

(GBDT) implementations (XGB., ASK-G., and CatBoost); (d) as well as three reference MLPs (unregularized (MLP), and regularized with Dropout (MLP+D) [49] or SELU (MLP+SELU) [30]).

Table 2 shows the comparison against a subset of the baselines, while the full detailed results involving all the remaining baselines are located in Table 13 in the appendix. It is worth re-emphasizing that the hyperparameters of all the presented baselines (except the unregularized MLP, which has no hyperparameters and AutoGL. S) are carefully tuned on a validation set as detailed in Section 5 and the appendices referenced therein. The table entries represent the test sets’ balanced accuracies achieved over the described large-scale collection of 40 datasets. Figure 1 visualizes the results showing substantial improvements for our method.

To assess the statistical significance, we analyze the ranks of the classification accuracies across the 40 datasets. We use the Critical Difference (CD) diagram of the ranks based on the Wilcoxon significance test, a standard metric for comparing classifiers across multiple datasets [9]. The overall empirical comparison of the elaborated methods is given in Figure 2. The analysis of neural network baselines in Subplot 2a reveals a clear statistical significance of the regularization cocktails against the other methods. Apart from AutoGluon (both versions), the other neural architectures are not competitive even against an MLP regularized only with Dropout and optimized with our standard, fixed training pipeline of Adam with cosine annealing. To be even fairer to the weaker baselines (TabNet and Node) we tried boosting them by adding early stopping (indicated with "+ES"), but their rank did not improve. Overall, the large-scale experimental analysis shows that Hypothesis 1 in Section 5.3 is validated: **well-regularized simple deep MLPs outperform specialized neural architectures.**

Next, we analyze the empirical significance of our well-regularized MLPs against the GBDT implementations in Figure 2b. The results show that our MLPs outperform all three GBDT variants (XGBoost, auto-sklearn, and CatBoost) with a statistically significant margin. We added early stopping ("+ES") to XGBoost, but it did not improve its performance. Among the GBDT implementations, XGBoost without early stopping has a non-significant margin over the GBDT version of auto-sklearn

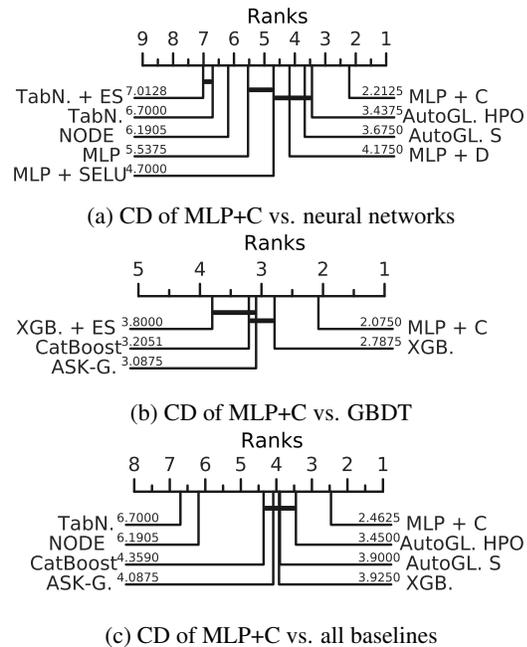


Figure 2: **Critical difference diagrams** with a Wilcoxon significance analysis on 40 datasets. Connected ranks via a bold bar indicate that performances are not significantly different ($p > 0.05$).

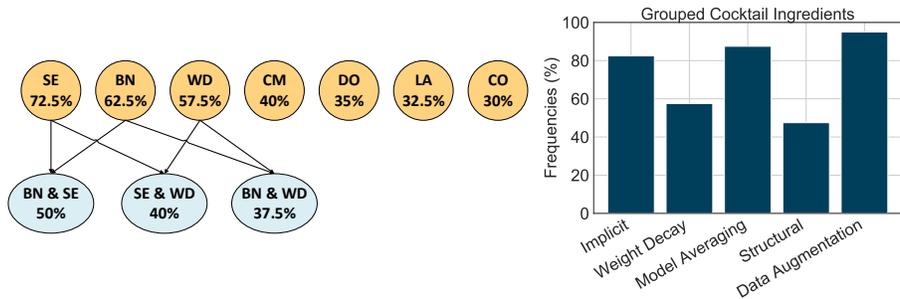


Figure 3: **Left:** Cocktail ingredients occurring in at least 30% of the datasets. **Right:** Clustered histogram (union of member occurrences) with the acronyms from Table 1. Implicit: {BN, LA, SWA}, M. Averaging: {DO, SE}, Structural: {SC, SS, SD}, D. Augmentation: {MU, CM, CO, AT}.

as well as CatBoost. We conclude that **well-regularized simple deep MLPs outperform GBDT**, which validates Hypothesis 2 in Section 5.3.

The final cumulative comparison in Figure 2c provides a further result: none of the specialized previous deep learning methods (TabNet, NODE, AutoGluon Tabular) outperforms GBDT significantly. To the best of our awareness, this paper is therefore the first to demonstrate that neural networks beat GBDT with a statistically significant margin over a large-scale experimental protocol that conducts a thorough hyperparameter optimization for all methods.

Figure 3 provides a further analysis on the most prominent regularizers of the MLP cocktails, based on the frequency with which our HPO procedure selected the various regularization methods for each dataset’s cocktail. In the left plot, we show the frequent individual regularizers, while in the right plot the frequencies are grouped by types of regularizers. The grouping reveals that a cocktail for each dataset often has at least one ingredient from every regularization family (detailed in Section 3), highlighting the need for jointly applying diverse regularization methods.

Lastly, Table 3 shows the efficiency of our regularization cocktails compared to XGBoost over increasing HPO budgets. The descriptive statistics are calculated from the hyperparameter configurations with the best validation performance for all datasets during the HPO search, however, taking their respective test performances for comparison. A dataset is considered in the comparison only if the HPO procedure has managed to evaluate at least one hyperparameter configuration for the cocktail or baseline. As the table shows, our regularization cocktails achieve a better performance in only 15 minutes for the majority of datasets. After 30 minutes of HPO time, regularization cocktails are statistically significantly better than XGBoost. As more time is invested, the performance gap with XGBoost increases, and the results get even more significant; this is further visualized in the ranking plot over time in Figure 4. Based on these results, we conclude that **regularization cocktails are time-efficient and achieve strong anytime results**, which validates Hypothesis 3 in Section 5.3.

Time (Hours)	Wins	Ties	Losses	p-value
0.25	21	1	17	0.8561
0.5	25	1	13	0.0145
1	24	1	14	0.0120
2	27	1	11	0.0006
4	28	1	11	0.0006
8	28	1	11	0.0004
16	28	1	11	0.0005
32	29	1	10	0.0003
64	30	1	9	0.0002
96	30	1	9	0.0002

Table 3: Comparing the cocktails and XGBoost over different HPO budgets. The statistics are based on the test performance of the incumbent configurations over all the benchmark datasets.

7 Conclusion

Summary. Focusing on the important domain of tabular datasets, this paper studied improvements to deep learning (DL) by better regularization techniques. We presented *regularization cocktails*, per-dataset-optimized combinations of many regularization techniques, and demonstrated that these improve the performance of even simple neural networks enough to substantially and significantly

surpass XGBoost, the current state-of-the-art method for tabular datasets. We conducted a large-scale experiment involving 13 regularization methods and 40 datasets and empirically showed that (i) modern DL regularization methods developed in the context of raw data (e.g., vision, speech, text) substantially improve the performance of deep neural networks on tabular data; (ii) regularization cocktails significantly outperform recent neural networks architectures, and most importantly iii) regularization cocktails outperform GBDT on tabular datasets.

Limitations. To comprehensively study basic principles, we have chosen an empirical evaluation that has many limitations. We only studied classification, not regression. We only used somewhat balanced datasets (the ratio of the minority class and the majority class is above 0.05). We did not study the regimes of extremely few or extremely many data points (our smallest data set contained 452 data points, our largest 416 188 data points). We also did not study datasets with extreme outliers, missing labels, semi-supervised data, streaming data, and many more modalities in which tabular data arises. An important point worth noticing is that the recent neural network architectures (Section 5.4) could also benefit from our regularization cocktails, but integrating the regularizers into these baseline libraries requires considerable coding efforts.

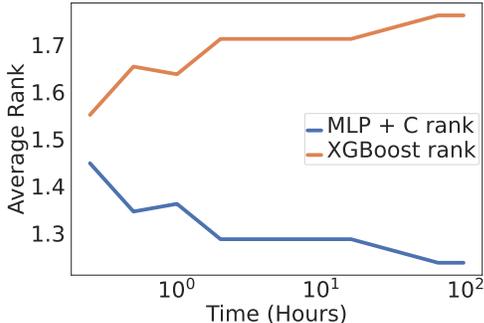


Figure 4: Ranking plot comparing XGBoost and regularization cocktails over time.

Future Work. This work opens up the door for a wealth of exciting follow-up research. Firstly, the per-dataset optimization of regularization cocktails may be substantially sped up by using meta-learning across datasets [53]. Secondly, as we have used a fixed neural architecture, our method’s performance may be further improved by using joint architecture and hyperparameter optimization. Thirdly, regularization cocktails should also be tested under all the data modalities under “Limitations” above. In addition, it would be interesting to validate the gain of integrating our well-regularized MLPs into modern AutoML libraries, by combining them with enhanced feature preprocessing and ensembling.

Take-away. *Even simple neural networks can achieve competitive classification accuracies on tabular datasets when they are well regularized, using dataset-specific regularization cocktails found via standard hyperparameter optimization.*

Societal Implications

Enabling neural networks to advance the state-of-the-art on tabular datasets may open up a garden of delights in many crucial applications, such as climate science, medicine, manufacturing, and recommender systems. In addition, the proposed networks can serve as a backbone for applications of data science for social good, such as the realm of fair machine learning where the associated data are naturally in a tabular form. However, there are also potential disadvantages in advancing deep learning for tabular data. In particular, even though complex GBDT ensembles are also hard to interpret, simpler traditional ML methods are much more interpretable than deep neural networks; we therefore encourage research on interpretable deep learning on tabular data.

Acknowledgements

We acknowledge funding by the Robert Bosch GmbH, by the Eva Mayr-Stihl Foundation, the MWK of the German state of Baden-Württemberg, the BrainLinks-BrainTools CoE, the European Research Council (ERC) under the European Union’s Horizon 2020 programme, grant no. 716721, and the German Federal Ministry of Education and Research (BMBF, grant RenormalizedFlows 01IS19077C). The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 39/963-1 FUGG.

References

- [1] S. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning. In *AAAI Conference on Artificial Intelligence*, 2021.
- [2] S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alche Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7413–7424. Curran Associates, Inc., 2019.
- [3] A. Asuncion and D. Newman. Uci machine learning repository, 2007.
- [4] K. Henning Brodersen, C., and E. Klaas Stephan J. M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE, 2010.
- [5] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.
- [6] Yutian Chen, Aja Huang, Ziyu Wang, Ioannis Antonoglou, Julian Schrittwieser, David Silver, and Nando de Freitas. Bayesian optimization in alphago. *CoRR*, abs/1812.06855, 2018.
- [7] E. Cubuk, B. Zoph, D. Mane, and V. Vasudevan and Q. Le. Autoaugment: Learning augmentation strategies from data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [8] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [9] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006.
- [10] T. Devries and G. Taylor. Improved regularization of convolutional neural networks with cutout. *ArXiv*, abs/1708.04552, 2017.
- [11] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola. Autogluontabular: Robust and accurate automl for structured data. *CoRR*, abs/2003.06505, 2020.
- [12] S. Falkner, A. Klein, and F. Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pages 1436–1445, July 2018.
- [13] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, page 2755–2763. MIT Press, 2015.
- [14] M. Feurer, JN. van Rijn, A. Kadra, P. Gijsbers, N. Mallik, S. Ravi, A. Müller, J. Vanschoren, and F. Hutter. Openml-python: an extensible python api for openml. *Journal of Machine Learning Research*, 22(100):1–5, 2021.
- [15] X. Gastaldi. Shake-shake regularization of 3-branch residual networks. In *5th International Conference on Learning Representations, ICLR*. OpenReview.net, 2017.
- [16] P. Gijsbers, E. LeDell, S. Poirier, J. Thomas, B. Bischl, and J. Vanschoren. An open source automl benchmark. *arXiv preprint arXiv:1907.00909 [cs.LG]*, 2019. Accepted at AutoML Workshop at ICML 2019.
- [17] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR*, 2015.
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

- [19] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. In S. McIlraith and K. Weinberger, editors, *Proceedings of the Conference on Artificial Intelligence (AAAI'18)*. AAAI Press, 2018.
- [20] D. Hendrycks, N. Mu, E. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. Augmix: A simple method to improve robustness and uncertainty under data shift. In *International Conference on Learning Representations*, 2020.
- [21] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. Hopcroft, and K. Weinberger. Snapshot Ensembles: Train 1, Get M for Free. *International Conference on Learning Representations*, November 2017.
- [22] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [23] F. Hutter, L. Kotthoff, and J. Vanschoren, editors. *Automated Machine Learning: Methods, Systems, Challenges*. Springer, 2019. In press, available at <http://automl.org/book>.
- [24] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456. PMLR, 07–09 Jul 2015.
- [25] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, and A. Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI*, pages 876–885. AUAI Press, 2018.
- [26] L. Katzir, G. Elidan, and R. El-Yaniv. Net-{dnf}: Effective deep modeling of tabular data. In *International Conference on Learning Representations*, 2021.
- [27] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [28] D. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15*, page 2575–2583. MIT Press, 2015.
- [29] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. In *Proceedings of the 31st international conference on neural information processing systems*, pages 972–981, 2017.
- [30] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [31] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [32] C. Lee, K. Cho, and W. Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations*, 2020.
- [33] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.*, 18(1):6765–6816, January 2017.
- [34] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim. Fast autoaugment. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6665–6675. Curran Associates, Inc., 2019.
- [35] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR) 2017 Conference Track*, April 2017.

- [36] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- [38] Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*, 2018.
- [39] H. Mendoza, A. Klein, M. Feurer, J. Tobias Springenberg, M. Urban, M. Burkart, M. Dippel, M. Lindauer, and F. Hutter. Towards automatically-tuned deep neural networks. In F. Hutter, L. Kotthoff, and J. Vanschoren, editors, *AutoML: Methods, Systems, Challenges*, chapter 7, pages 141–156. Springer, December 2019.
- [40] J. Mockus. Application of bayesian approach to numerical methods of global and stochastic optimization. *J. Glob. Optim.*, 4(4):347–365, 1994.
- [41] Samuel G. Müller and Frank Hutter. Trivialaugument: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 774–782, October 2021.
- [42] A. Emin Orhan and X. Pitkow. Skip connections eliminate singularities. *arXiv preprint arXiv:1701.09175*, 2017.
- [43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.
- [44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [45] R. Polikar. Ensemble Learning. In C. Zhang and Y. Ma, editors, *Ensemble Machine Learning: Methods and Applications*, pages 1–34. Springer US, 2012.
- [46] S. Popov, S. Morozov, and A. Babenko. Neural oblivious decision ensembles for deep learning on tabular data. In *International Conference on Learning Representations*, 2020.
- [47] L. Prokhorenkova, G. Gusev, A. Vorobev, AV. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 2018.
- [48] I. Shavitt and E. Segal. Regularization learning networks: Deep learning for tabular datasets. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 1386–1396. Curran Associates Inc., 2018.
- [49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.
- [50] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.
- [51] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.
- [52] A. Tikhonov. On the stability of inverse problems. In *Doklady Akademii Nauk SSSR*, 1943.
- [53] J. Vanschoren. Meta-learning. In F. Hutter, L. Kotthoff, and J. Vanschoren, editors, *Automated Machine Learning - Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pages 35–61. Springer, 2019.

- [54] J. Vanschoren, J. Van Rijn, B. Bischl, and L. Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.
- [55] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6514–6527. Curran Associates, Inc., 2020.
- [56] Y. Yamada, M. Iwamura, and K. Kise. Shakedrop regularization, 2018.
- [57] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, August 2007.
- [58] S. Yun, D. Han, S. Joon, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019.
- [59] Sheheryar Zaidi, Arber Zela, Thomas Elsken, Chris Holmes, Frank Hutter, and Yee Whye Teh. Neural ensemble search for uncertainty estimation and dataset shift. In *Advances in Neural Information Processing Systems*, 2021.
- [60] H. Zhang, M. Cisse, Y. Dauphin, and D. Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.
- [61] M. Zhang, J. Lucas, J. Ba, and G. Hinton. Lookahead optimizer: k steps forward, 1 step back. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 9593–9604, 2019.
- [62] L. Zimmer, M. Lindauer, and F. Hutter. Auto-pytorch tabular: Multi-fidelity metalearning for efficient and robust autodl. *IEEE TPAMI*, 2021. IEEE Early Access.
- [63] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
 - (b) Did you describe the limitations of your work? [Yes] See Section 7.
 - (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the previous section.
 - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? [N/A]
 - (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 1 and Section 5.
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and Section 4.
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Yes, we report the statistical significance of the performance difference between our method and the baselines in Section 6
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? [Yes]
 - (b) Did you mention the license of the assets? [Yes]
 - (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 1.
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Description of BOHB

BOHB [12] is a hyperparameter optimization algorithm that extends Hyperband [33] by sampling from a model instead of sampling randomly from the hyperparameter search space.

Initially, BOHB performs random search and favors exploration. As it iterates and gets more observations, it builds models over different fidelities and trades off exploration with exploitation to avoid converging in bad regions of the search space. BOHB samples from the model of the highest fidelity with a probability p and with $1 - p$ from random. A model is built for a fidelity only when enough observations exist for that fidelity; by default, this limit is set to equal $S + 1$ observations, where S is the dimensionality of the search space.

We used the original implementation of BOHB in HPBandSter³ [12]. For simplicity, we only used a single fidelity level for BOHB, effectively using it as a blackbox optimization algorithm; we believe that future work should revisit this choice.

B Configuration Spaces

B.1 Method implicit search space

Category	Hyperparameter	Type	Range
Cosine Annealing	Iterations multiplier	Continuous	{2.0}
	Max. iterations	Integer	{15}
Network	Activation	Nominal	{ReLU}
	Bias initialization	Nominal	{Yes}
	Blocks in a group	Integer	{2}
	Embeddings	Nominal	{One-Hot encoding}
	Number of groups	Integer	{2}
	Resnet shape	Nominal	{Brick}
	Type	Nominal	{Shaped-Resnet}
Preprocessing	Preprocessor	Nominal	{None}
	Batch size	Integer	{128}
Training	Imputation	Nominal	{Median}
	Initialization method	Nominal	{Default}
	Learning rate	Continuous	{ 10^{-3} }
	Loss module	Nominal	{Weighted Cross-Entropy}
	Normalization strategy	Nominal	{Standardize}
	Optimizer	Nominal	{AdamW}
	Scheduler	Nominal	{COS}
	Seed	Integer	{11}

Table 4: The configuration space of the training and model architecture hyperparameters. All these hyperparameters only have one value in their range, meaning they are fixed.

Table 4 presents the network architecture and the training pipeline choices used in all our experiments for the individual regularizers and for the regularization cocktails.

B.2 Benchmark search space

For the experiments conducted in our work, we set up the search space and the individual configurations of the state-of-the-art competitors used for the comparison as follows:

Auto-Sklearn. The estimator is restricted to only include GBDT, for the sake of fully comparing against the algorithm as a baseline. We do not activate any preprocessing since our regularization cocktails also do not make use of preprocessing algorithms in the pipeline. The time left is always selected based on the time it took BOHB to find the hyperparameter with the best validation accuracy from the start of the hyperparameter optimization phase. The ensemble size is kept to 1 since our method only uses models from one training run, not multiple ones. The seed is set to 11 as it was set in the experiments with the regularization cocktail, to obtain the same data splits. To keep the comparison fair, there is no warm start for the initial configurations with meta-learning, since our method also does not make use of meta-learning. Lastly, the number of parallel workers is set to 10, to match the parallel resources that were given to the experiment with the regularization cocktails. The search space of the hyperparameters is left to the default search space offered by Auto-Sklearn which is shown in Table 5. We use version 0.10.0 of the library.

³<https://github.com/automl/HPBandSter>

Hyperparameter	Type	Range
<i>early_stopping</i>	Nominal	{Off, Train, Valid}
<i>l2_regularization</i>	Continuous	[1e - 10, 1]
<i>learning_rate</i>	Continuous	[0.01, 1]
<i>max_leaf_nodes</i>	Integer	[3, 2047]
<i>min_samples_leaf</i>	Integer	[1, 200]
<i>nr_iterations_no_change</i>	Integer	[1, 20]
<i>validation_fraction</i>	Continuous	[0.01, 0.4]

Table 5: The search space of the training and model hyperparameters for the gradient boosting estimator of the Auto-Sklearn tool.

XGBoost. To have a well-performing configuration space for XGBoost we augmented the default configuration spaces previously used in Auto-Sklearn⁴ with further recommended hyperparameters and ranges from Amazon⁵. In Table 6, we present a refined version of the configuration space that achieves a better performance on the benchmark. We would like to note that we did not apply One-Hot encoding to the categorical features for the experiment, since we observed better overall results when the categorical features were ordinal encoded. Nevertheless, in Table 13 we ablate the choice of encoding (+ENC) as a categorical hyperparameter (one-hot vs. ordinal) and also early stopping (+ES). When early stopping is activated, the *num_rounds* is increased to 4000. All the results can be found in Appendix D.

Hyperparameter	Type	Range	Log scale
<i>eta</i>	Continuous	[0.001, 1]	✓
<i>lambda</i>	Continuous	[1e - 10, 1]	✓
<i>alpha</i>	Continuous	[1e - 10, 1]	✓
<i>num_round</i>	Integer	[1, 1000]	-
<i>gamma</i>	Continuous	[0.1, 1]	✓
<i>colsample_bylevel</i>	Continuous	[0.1, 1]	-
<i>colsample_bynode</i>	Continuous	[0.1, 1]	-
<i>colsample_bytree</i>	Continuous	[0.5, 1]	-
<i>max_depth</i>	Integer	[1, 20]	-
<i>max_delta_step</i>	Integer	[0, 10]	-
<i>min_child_weight</i>	Continuous	[0.1, 20]	✓
<i>subsample</i>	Continuous	[0.01, 1]	-

Table 6: The hyperparameter search space for the XGBoost library.

TabNet. For the search space of the TabNet model, we used the default hyperparameter ranges suggested by the authors which were found to perform best in their experiments.

For our experiments with the TabNet and XGBoost models, we also used BOHB for hyperparameter tuning, using the same parallel resources and limiting conditions as for our regularization cocktail. In the above search spaces for the experiments with the XGBoost and TabNet models, we did not include early stopping; however, we did actually run experiments with early stopping for both models,

⁴https://github.com/automl/auto-sklearn/blob/v.0.4.2/autosklearn/pipeline/components/classification/xgradient_boosting.py

⁵<https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-tuning.html> We reduced the size of some ranges since the ranges given at this website were too broad and resulted in poor performance.

Hyperparameter	Type	Values
n_a	Integer	{8, 16, 24, 32, 64, 128}
$learning_rate$	Continuous	{0.005, 0.01, 0.02, 0.025}
$gamma$	Continuous	{1.0, 1.2, 1.5, 2.0}
n_steps	Integer	{3, 4, 5, 6, 7, 8, 9, 10}
λ_{sparse}	Continuous	{0, 0.000001, 0.0001, 0.001, 0.01, 0.1}
$batch_size$	Integer	{256, 512, 1024, 2048, 4096, 8192, 16384, 32768}
$virtual_batch_size$	Integer	{256, 512, 1024, 2048, 4096}
$decay_rate$	Continuous	{0.4, 0.8, 0.9, 0.95}
$decay_iterations$	Integer	{500, 2000, 8000, 10000, 20000}
$momentum$	Continuous	{0.6, 0.7, 0.8, 0.9, 0.95, 0.98}

Table 7: The hyperparameter search space for TabNet.

and results did not improve. Lastly, for both experiments, we imputed missing values with the most frequent strategy (the implementation we used did not accept the median strategy for categorical value imputation).

AutoGluon. The library is configured to construct stacked ensembles with bagging among diverse neural network architectures having various kinds of regularization with the *preset* = **'Best Quality'** to achieve the best predictive accuracy. Furthermore, we used the same seed as for our MLPs with regularization cocktails to obtain the same dataset splits. We allowed AutoGluon to make use of early stopping and additionally, we allowed feature preprocessing since different feature preprocessing techniques are embedded in different model types, to allow for better overall performance. For all the other training and hyperparameter settings, we used the library’s default⁶ following the explicit recommendation of the authors on the efficacy of their proposed stacking without needing any HPO [11].

We also investigate using HPO with AutoGluon and compare the results against the version that uses stacking; the full detailed results can be found in Appendix D.

NODE. For our experiments with NODE we used the official implementation⁷. In our initial experiment iterations, we used the search space that was proposed by the authors [46]. However, evaluating the search space proposed is infeasible, since the memory and run-time requirements of the experiments are very high and cannot be satisfied within our cluster constraints. The high run-time and memory issues are also noted by the authors in the official implementation.

To alleviate these problems, we used the default configuration suggested by the authors in the examples, where $num_layers = 2$, $total_tree_count = 1024$ and $tree_depth = 6$. Lastly, we use the same seed as for our experiment with the regularization cocktails to obtain the same data splits.

CatBoost. We use version 0.26 of the official library and we use the hyperparameter search space that is recommended by the authors [47], provided in Table 8.

C Plots

C.1 Regularization Cocktail Performance

To investigate the performance of our formulation, we compare plain MLPs regularized with only one individual regularization technique at a time against the dataset-specific regularization cocktails. The hyperparameters for all methods are tuned on the validation set and the best configuration is refitted on the full training set. In Figure 5, we present the results of each pairwise comparison. The results

⁶We used version 0.2.0 of the AutoGluon library

⁷<https://github.com/Qwicen/node>

Hyperparameter	Type	Range	Log scale
<i>learning_rate</i>	Continuous	$[e^{-7}, 1]$	✓
<i>random_strength</i>	Integer	$[1, 20]$	-
<i>one_hot_max_size</i>	Integer	$[0, 25]$	-
<i>num_round</i>	Integer	$[1, 4000]$	-
<i>l2_leaf_reg</i>	Continuous	$[1, 10]$	✓
<i>bagging_temperature</i>	Continuous	$[0, 1]$	-
<i>gradient_iterations</i>	Integer	$[1, 10]$	-

Table 8: The hyperparameter search space for the CatBoost library.

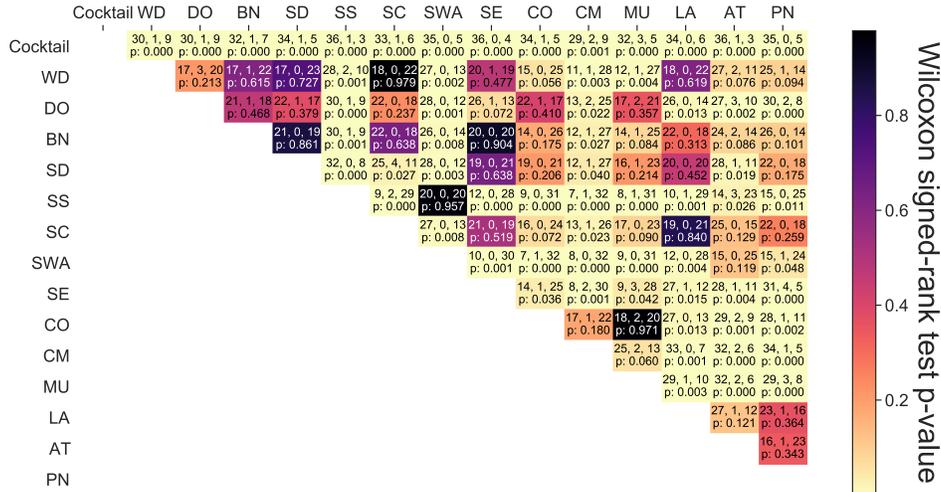


Figure 5: **Pairwise statistical significance and comparison.** For every entry, **the first row** shows the wins, draws and losses of the horizontal method with the vertical method on all datasets, calculated on the test set; **the second row** presents the p-value for the statistical significance test.

presented are calculated on the test set after the refit phase is completed on the best hyperparameter configuration. The p-value is generated by performing a Wilcoxon signed-rank test. As can be seen from the results, the regularization cocktail is the only method that has statistically significant improvements compared to all other methods (with a p-value ≤ 0.001 in all cases). The detailed results for all methods on every dataset are shown in Table 11.

C.2 Dataset-dependent optimal cocktails

To verify the necessity for dataset-specific regularization cocktails, we initially investigate the best-found hyperparameter configurations to observe the occurrences of individual regularization techniques. In Figure 6, we present the occurrences of every regularization method over all datasets. The occurrences are calculated by analyzing the best-found hyperparameter configuration for each dataset and observing the number of times the regularization method was chosen to be activated by BOHB. As can be seen from Figure 6, there is no regularization method or combination that is always chosen for every dataset.

Additionally, we compare our regularization cocktails against the top-5 frequently chosen regularization techniques and the top-5 best performing regularization techniques. For the top-5 baselines, the regularization techniques are activated and their hyperparameters are tuned on the validation set. The results of the comparison as shown in Table 10 show that the cocktail outperforms both top-5 variants, indicating the need for dataset-specific regularization cocktails.

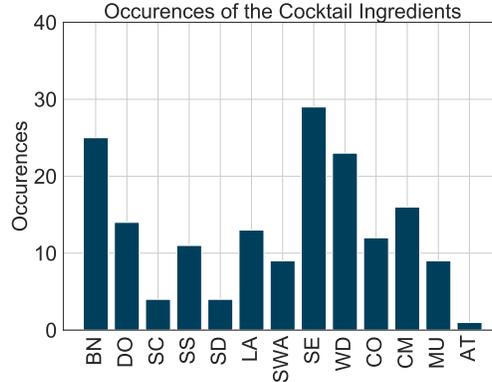


Figure 6: **Frequency of the regularization techniques.** The occurrences of the individual regularization techniques in the best hyperparameter configurations found by the cocktail across 40 datasets.

C.3 Learning rate as a hyperparameter

In the majority of our experiments, we keep a fixed initial learning rate to investigate in detail the effect of the individual regularization techniques and the regularization cocktails. The learning rate is set to a fixed value that achieves the best results across the chosen benchmark of datasets. To investigate the role and importance of the learning rate in the regularization cocktail performance, we perform an additional experiment, where, the learning rate is an additional hyperparameter that is optimized individually for every dataset. The results as shown in Table 12, indicate that regularization cocktails with a dynamic learning rate outperform the regularization cocktails with a fixed learning rate in 21 out of 40 datasets, tie in 1 and lose in 18. However, the results are not statistically significant with a p -value of 0.7 and do not indicate a clear region where the dynamic learning rate helps.

D Tables

In Table 9, we provide information about the datasets that are considered in our experiments. Concretely, we provide descriptive statistics and the identifiers for every dataset. The identifier (the task id) can be used to download the datasets from OpenML (<http://www.openml.org>) by using the OpenML-Python connector [14].

Table 10 shows the results for the comparison between the Regularization Cocktail and the Top-5 cocktail variants. The results are calculated on the test set for all datasets, after retraining on the best dataset-specific hyperparameter configuration.

Table 11 provides the results of all our experiments for the plain MLP baseline, the individual regularization methods, and the regularization cocktail. All the results are calculated on the test set, after retraining on the best-found hyperparameter configurations. The evaluation metric used for the performance is balanced accuracy.

Additionally, in Table 12, we provide the results of the regularization cocktails with a fixed learning rate and with the learning rate being a hyperparameter optimized for every dataset.

Lastly, in Table 13, we present the remaining baselines from our experiments/ablations and their final performances on every dataset. The results show the test set performances after the incumbent configuration is refit.

Task Id	Dataset Name	Number of Instances	Number of Features	Majority Class Percentage	Minority Class Percentage
233090	anneal	898	39	76.17	0.89
233091	kr-vs-kp	3196	37	52.22	47.78
233092	arrhythmia	452	280	54.20	0.44
233093	mfeat-factors	2000	217	10.00	10.00
233088	credit-g	1000	21	70.00	30.00
233094	vehicle	846	19	25.77	23.52
233096	kc1	2109	22	84.54	15.46
233099	adult	48842	15	76.07	23.93
233102	walking-activity	149332	5	14.73	0.61
233103	phoneme	5404	6	70.65	29.35
233104	skin-segmentation	245057	4	79.25	20.75
233106	ldpa	164860	8	33.05	0.84
233107	nomao	34465	119	71.44	28.56
233108	cnae-9	1080	857	11.11	11.11
233109	blood-transfusion	748	5	76.20	23.80
233110	bank-marketing	45211	17	88.30	11.70
233112	connect-4	67557	43	65.83	9.55
233113	shuttle	58000	10	78.60	0.02
233114	higgs	98050	29	52.86	47.14
233115	Australian	690	15	55.51	44.49
233116	car	1728	7	70.02	3.76
233117	segment	2310	20	14.29	14.29
233118	Fashion-MNIST	70000	785	10.00	10.00
233119	Jungle-Chess-2pcs	44819	7	51.46	9.67
233120	numerai28.6	96320	22	50.52	49.48
233121	Devnagari-Script	92000	1025	2.17	2.17
233122	helen	65196	28	6.14	0.17
233123	jannis	83733	55	46.01	2.01
233124	voklet	58310	181	21.96	2.33
233126	MiniBooNE	130064	51	71.94	28.06
233130	APSFailure	76000	171	98.19	1.81
233131	christine	5418	1637	50.00	50.00
233132	dilbert	10000	2001	20.49	19.13
233133	fabert	8237	801	23.39	6.09
233134	jasmine	2984	145	50.00	50.00
233135	sylvine	5124	21	50.00	50.00
233137	dionis	416188	61	0.59	0.21
233142	aloi	108000	129	0.10	0.10
233143	C.C.FraudD.	284807	31	99.83	0.17
233146	Click prediction	399482	12	83.21	16.79

Table 9: **Datasets.** The collection of datasets used in our experiments, combined with detailed information for each dataset.

Task Id	Cockt.	Top-5 F	Top-5 R	Task Id	Cockt.	Top-5 F	Top-5 R	Task Id	Cockt.	Top-5 F	Top-5 R
233090	89.27	89.71	88.54	233091	99.85	99.85	98.20	233092	61.46	59.94	57.21
233093	98.00	98.75	98.75	233088	74.64	71.43	74.76	233094	82.58	82.01	80.33
233096	74.38	78.03	73.96	233099	82.44	82.35	82.24	233102	63.92	62.21	54.10
233103	86.62	85.90	82.33	233104	99.95	99.96	99.85	233106	68.11	68.81	55.45
233107	96.83	96.67	96.59	233108	95.83	95.83	95.83	233109	67.62	67.32	68.20
233110	85.99	86.35	86.06	233112	80.07	79.57	77.49	233113	99.95	97.95	85.34
233114	73.55	73.25	72.06	233115	87.09	88.11	87.60	233116	99.59	100.00	98.20
233117	93.72	93.94	90.69	233118	91.95	91.83	91.59	233119	97.47	92.66	85.53
233120	52.67	52.49	51.70	233121	98.37	98.41	96.93	233122	27.70	28.82	28.09
233123	65.29	65.13	62.11	233124	71.67	70.87	66.06	233126	94.02	88.13	93.16
233130	92.53	96.24	95.89	233131	74.26	71.86	74.63	233132	99.05	98.95	98.55
233133	69.18	68.75	69.03	233134	79.22	78.21	77.71	233135	94.05	94.43	93.95
233137	94.01	94.33	92.43	233142	97.17	97.06	96.06				
233146	64.28	64.53	63.28	233143	92.53	92.13	92.59				

Table 10: **Top-5 baselines.** The test set performance for the Regularization Cocktail against the Top-5 Most Frequent (Top-5 F) and the Top-5 Highest Ranks (Top-5 R) baselines.

Task Id	PN	BN	LA	SE	SWA	SC	AT	SS	SD	MU	CO	CM	WD	DO	Cocktail
233090	84.13	86.78	83.99	86.48	87.96	87.21	86.92	84.28	87.21	89.27	85.60	86.77	87.06	86.92	89.27
233091	99.70	99.85	99.70	99.70	99.55	100.00	99.85	99.85	99.69	99.85	99.55	99.85	99.85	99.85	99.85
233092	37.99	41.91	36.14	37.31	25.94	53.42	38.79	55.61	53.26	42.19	32.48	42.22	35.76	38.70	61.46
233093	97.75	98.50	96.00	97.75	69.25	98.25	97.25	97.25	98.25	98.00	98.00	97.75	98.00	98.00	98.00
233088	69.40	68.69	70.83	69.76	69.40	66.43	69.29	66.43	67.14	70.00	70.36	64.29	69.29	68.10	74.64
233094	83.77	83.17	84.36	84.39	83.36	80.82	83.17	83.20	81.98	83.77	81.47	78.65	83.20	82.60	82.58
233096	70.27	66.56	71.95	76.43	75.44	77.40	71.95	65.31	78.31	72.43	76.84	74.94	67.33	72.98	74.38
233099	76.89	77.92	75.95	78.23	76.38	78.38	76.75	75.56	78.61	78.67	82.56	82.23	76.99	78.52	82.44
233102	61.00	62.89	61.32	63.57	56.67	60.79	59.99	43.04	60.77	61.95	63.30	63.49	64.03	63.75	63.92
233103	87.51	87.02	88.25	87.03	87.22	85.90	87.99	87.64	85.90	87.12	87.26	86.59	86.74	88.39	86.62
233104	99.97	99.96	99.96	99.94	2.57	99.97	99.95	92.77	99.97	99.95	99.96	99.97	99.96	99.96	99.95
233106	62.83	68.90	62.46	65.70	62.16	61.85	61.89	44.63	62.05	66.29	65.43	64.99	66.50	67.04	68.11
233107	95.92	95.93	96.01	96.36	95.23	95.76	95.77	95.37	96.22	96.52	96.10	96.55	95.98	96.23	96.83
233108	87.50	91.20	85.65	87.96	50.00	93.98	92.59	94.91	94.44	94.44	93.06	95.37	91.67	90.74	95.83
233109	67.84	73.68	66.52	68.20	66.45	65.20	66.89	66.74	67.03	68.64	67.32	70.18	66.23	68.42	67.62
233110	78.08	72.58	72.70	83.40	66.93	72.74	74.12	70.16	74.76	74.09	85.71	85.76	72.34	83.14	85.99
233112	73.63	74.68	73.37	74.33	77.36	73.86	72.91	72.06	74.35	72.08	76.23	75.74	72.48	76.35	80.07
233113	99.47	99.89	99.92	99.87	55.86	98.11	99.46	90.60	98.11	99.94	99.92	99.91	99.88	99.89	99.95
233114	67.75	68.90	68.81	69.11	67.36	68.08	67.44	67.70	68.56	68.59	71.93	73.13	67.80	66.87	73.55
233115	86.27	85.79	88.73	86.44	87.26	87.74	88.39	87.74	88.39	88.73	88.25	88.90	87.91	86.27	87.09
233116	97.44	100.00	96.79	97.44	87.35	99.47	99.14	97.46	99.69	99.37	97.64	99.04	97.44	99.69	99.59
233117	94.81	92.86	93.51	93.51	90.48	93.72	92.86	92.64	93.72	93.51	93.07	93.72	93.94	94.59	93.72
233118	90.46	90.86	90.73	90.75	81.72	89.91	90.69	86.69	90.06	91.11	91.09	91.88	90.70	90.51	91.95
233119	97.06	93.76	97.79	96.08	92.15	87.83	97.16	87.08	87.68	98.14	96.50	97.51	97.33	97.24	97.47
233120	50.26	50.95	51.29	50.50	51.63	50.92	50.17	50.23	51.00	50.72	52.35	52.10	50.41	50.30	52.67
233121	96.12	97.83	96.45	96.74	92.40	95.31	96.34	91.38	95.15	97.52	97.88	97.80	96.88	97.00	98.37
233122	16.84	22.26	17.20	19.65	20.90	24.53	16.77	18.71	24.35	23.62	23.43	24.10	17.52	23.98	27.70
233123	51.51	51.74	50.86	53.16	56.11	53.58	49.65	49.88	51.94	51.22	60.98	61.67	51.13	55.12	65.29
233124	65.08	66.82	65.57	66.56	66.15	57.71	65.26	64.97	58.04	67.24	70.03	68.84	66.86	67.00	71.67
233126	90.64	58.17	90.42	92.94	92.60	93.99	90.45	88.55	93.98	93.58	93.86	93.87	92.97	94.10	94.02
233130	87.76	87.81	88.98	88.99	70.72	87.99	50.00	85.25	88.35	92.43	50.00	95.81	94.92	91.19	92.53
233131	70.94	69.28	71.59	70.94	71.31	72.14	71.59	71.59	72.32	70.94	72.69	72.42	70.76	70.76	74.26
233132	96.93	98.62	97.52	97.14	94.58	96.85	97.00	97.27	96.90	98.66	98.14	99.15	96.81	96.73	99.05
233133	63.71	65.11	65.00	66.05	64.57	66.21	62.82	64.33	65.98	68.75	66.58	66.28	64.36	64.81	69.18
233134	78.05	75.87	79.05	78.22	80.38	78.38	76.88	78.38	78.38	76.88	77.38	76.54	76.88	76.21	79.22
233135	93.07	92.49	92.10	93.17	93.17	92.10	93.17	93.27	92.10	92.58	92.68	94.53	93.75	93.36	94.05
233137	91.91	93.71	92.16	92.56	90.38	91.58	91.36	88.09	91.60	92.72	92.48	92.39	92.95	92.72	94.01
233142	92.33	96.70	92.90	92.35	63.59	95.47	91.43	93.60	95.56	93.47	93.81	93.25	92.60	93.85	97.17
233143	50.00	92.30	92.76	50.00	70.81	90.28	50.00	50.31	89.26	50.00	50.00	50.00	92.26	50.00	92.53
233146	63.12	60.06	62.79	64.16	63.39	64.42	63.52	54.64	64.21	64.26	64.05	64.57	64.41	64.37	64.28

Table 11: **Detailed Table of Results.** The test set performance for the plain network, individual regularization methods and for the regularization cocktails.

Task Id	Fixed LR Cocktail	Dynamic LR Cocktail
233090	89.270	90.000
233091	99.850	99.850
233092	61.461	56.518
233093	98.000	98.250
233088	74.643	64.881
233094	82.576	79.654
233096	74.381	70.058
233099	82.443	82.551
233102	63.923	63.884
233103	86.619	87.854
233104	99.953	99.967
233106	68.107	69.081
233107	96.826	96.446
233108	95.833	95.370
233109	67.617	67.836
233110	85.993	86.596
233112	80.073	78.985
233113	99.948	83.263
233114	73.546	73.276
233115	87.088	88.077
233116	99.587	99.690
233117	93.723	93.939
233118	91.950	91.964
233119	97.471	98.039
233120	52.668	52.204
233121	98.370	98.522
233122	27.701	28.008
233123	65.287	63.293
233124	71.667	72.243
233126	94.015	93.930
233130	92.535	94.894
233131	74.262	72.140
233132	99.049	99.404
233133	69.183	68.877
233134	79.217	78.887
233135	94.045	94.435
233137	94.010	93.961
233142	97.175	97.106
233143	92.531	92.592
233146	64.280	64.362

Table 12: The test set performances of the regularization cocktails with a fixed initial learning rate value and a dynamic learning rate chosen by BOHB.

Task Id	MLP	MLP + D	MLP + S	XGB. + ES	XGB. + ES + ENC	XGB. + ENC	CatBoost	TabN. + ES	AutoGL. + HPO	MLP + C
233090	84.131	86.916	87.354	90.000	89.000	89.000	90.000	66.678	78.854	89.270
233091	99.701	99.850	99.687	99.701	99.687	99.701	99.197	99.687	99.238	99.850
233092	37.991	38.704	51.321	50.631	48.841	48.779	51.882	30.447	51.943	61.461
233093	97.750	98.000	97.000	96.750	98.000	98.250	97.250	97.250	97.500	98.000
233088	69.405	68.095	66.429	66.786	65.595	68.214	67.976	63.571	77.738	74.643
233094	83.766	82.603	85.444	75.081	73.241	74.405	79.732	66.061	82.624	82.576
233096	70.274	72.980	79.429	50.000	62.713	65.027	65.660	60.065	72.219	74.381
233099	76.893	78.520	81.354	78.790	78.917	79.529	81.045	76.715	82.768	82.443
233102	60.997	63.754	62.709	60.312	61.734	60.703	61.001	54.327	61.119	63.923
233103	87.514	88.387	87.886	87.079	87.841	87.553	87.579	79.027	87.975	86.619
233104	99.971	99.962	99.962	99.962	99.977	99.967	99.968	99.958	99.971	99.953
233106	62.831	67.035	63.560	98.668	98.850	98.721	69.889	53.774	56.662	68.107
233107	95.917	96.232	96.273	96.460	97.263	97.049	96.232	96.212	96.826	96.826
233108	87.500	90.741	93.519	95.370	93.519	93.519	95.370	85.648	94.907	95.833
233109	67.836	68.421	69.956	57.237	57.456	62.427	59.576	64.547	69.152	67.617
233110	78.076	83.145	86.137	73.329	72.087	72.252	74.152	67.893	83.097	85.993
233112	73.627	76.345	75.809	72.849	72.875	73.730	71.578	60.440	76.312	80.073
233113	99.475	99.892	99.494	97.143	98.571	98.571	98.571	63.273	83.851	99.948
233114	67.752	66.873	67.985	72.781	72.779	72.800	72.809	72.431	73.167	73.546
233115	86.268	86.268	83.159	91.186	50.000	89.376	91.016	87.428	86.300	87.088
233116	97.442	99.690	97.757	96.085	93.512	97.870	100.000	88.195	97.565	99.587
233117	94.805	94.589	93.723	93.723	91.991	93.290	92.424	91.126	91.126	93.723
233118	90.464	90.507	89.007	91.064	91.136	91.093	90.893	89.171	90.964	91.950
233119	97.061	97.237	90.997	90.938	90.085	90.748	82.150	88.403	99.270	97.471
233120	50.262	50.301	51.917	51.708	52.257	52.190	52.083	51.340	52.452	52.668
233121	96.125	97.000	2.174	91.967	93.082	92.533	95.777	NaN	97.299	98.370
233122	16.836	23.983	18.618	19.004	22.969	23.190	23.213	20.945	26.466	27.701
233123	51.505	55.118	56.543	54.858	54.705	55.295	55.453	55.212	60.209	65.287
233124	65.081	66.996	63.064	61.342	63.472	64.717	62.660	63.458	67.986	71.667
233126	90.639	94.099	93.701	93.880	94.055	93.979	93.932	49.790	95.172	94.015
233130	87.759	91.194	93.720	89.546	87.304	86.607	87.910	90.012	91.776	92.535
233131	70.941	70.756	69.926	74.815	74.170	75.738	73.708	69.649	73.801	74.262
233132	96.930	96.733	96.462	97.107	96.467	96.519	99.259	97.266	98.856	99.049
233133	63.707	64.814	64.877	68.952	68.911	70.331	71.708	63.811	66.074	69.183
233134	78.048	76.211	75.539	77.867	77.197	78.370	80.052	75.527	80.211	79.217
233135	93.070	93.363	95.217	95.119	94.827	94.924	95.119	92.681	93.364	94.045
233137	91.905	92.724	91.513	89.141	90.793	89.026	NaN	91.334	93.551	94.010
233142	92.331	93.852	92.707	95.019	93.870	94.047	85.379	89.384	96.207	97.175
233143	50.000	50.000	50.000	88.263	87.749	89.791	92.345	85.701	92.829	92.531
233146	63.125	64.367	50.000	57.866	58.143	58.324	56.563	50.340	59.203	64.280

Table 13: The results for the remaining baselines used in our experiments. Each performance represents the test accuracy of the incumbent configuration after being refit.