

A DESCRIPTION OF BOHB

Algorithm 1 BOHB

Input: b_{min} , b_{max} and η \triangleright b stands for budget, while η stands for the downsampling rate
Initialization: $s_{max} = \log_{\eta} \frac{b_{max}}{b_{min}}$
for s in $(s_{max}, s_{max}-1, \dots, 0)$ **do**
 $n = \frac{s_{max}+1}{s+1}$ \triangleright Number of configurations
 Call SH routine with $\eta^{-s} \cdot b_{max}$ as an initial budget \triangleright SH = Successing Halving routine
end for
Output: Return hyperparameter configuration with the smallest loss

BOHB (Falkner et al., 2018a) is a hyperparameter optimization algorithm that extends Hyperband (Li et al., 2017) 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, by default the criteria is set to equal $S + 1$, where S is the dimensionality of the search space.

BOHB achieves strong anytime results by combining Random Search and Bayesian optimization and helps deep neural networks in achieving faster convergence compared with traditional Bayesian Optimisation methods.

B CONFIGURATION SPACES

B.1 METHOD IMPLICIT SEARCH SPACE

Category	Hyperparameter	Type	Range	Conditionality
Cosine Annealing	Iterations multiplier	Continuous	{2.0}	Scheduler = COS
	Max. iterations	Integer	{15}	Scheduler = COS
Network	Activation	Nominal	{ReLU}	—
	Bias initialization	Nominal	{Yes}	—
	Blocks in a group	Integer	{2}	—
	Embeddings	Nominal	{None}	—
	Number of groups	Integer	{2}	—
	Resnet shape	Nominal	{Brick}	Type = Shaped-Resnet
	Type	Nominal	{Shaped-Resnet}	—
	Units in a layer	Integer	{512}	—
Preprocessing	Preprocessor	Nominal	{None}	—
Resampling	Target size	Nominal	{Median, Upsample}	—
	Under sampling	Nominal	{Random, None}	
Training	Batch size	Integer	{128}	—
	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 2: The configuration space of the training and model architecture hyperparameters.

Table 2 presents the implicit search space used in all our experiments. The implicit search space is shared between all the individual regularizers and the regularization cocktail.

B.2 AUTO-SKLEARN: GRADIENT BOOSTED DECISION TREE SEARCH SPACE

For Experiment 3, we set up the search space of Auto-Sklearn as follows:

Hyperparameter	Type	Range	Conditionality
Early Stopping	Nominal	{Off, Train, Valid}	Estimator = Gradient Boosting
L_2 Regularization	Continuous	$[1e - 10, 1]$	Estimator = Gradient Boosting
Learning Rate	Continuous	$[0.01, 1]$	Estimator = Gradient Boosting
Max Leaf Nodes	Integer	$[3, 2047]$	Estimator = Gradient Boosting
Min Samples Leaf	Integer	$[1, 200]$	Estimator = Gradient Boosting
# Iterations No Change	Integer	$[1, 20]$	Estimator = Gradient Boosting
Validation Fraction	Continuous	$[0.01, 0.4]$	Estimator = Gradient Boosting

Table 3: The search space of the training and model hyperparameters for the gradient boosting estimator of the Auto-Sklearn tool.

Furthermore, the estimator for Auto-Sklearn 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 also our regularization cocktails 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 features one classifier and not multiple ones. The seed is set to 11 as it was set in the experiments with the regularization cocktail, so we can have 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 workers in parallel is set to 10 to match the parallel resources that were given to the experiment with the regularization cocktails.

C PLOTS

C.1 EXPERIMENT 1: REGULARIZATION COCKTAIL PERFORMANCE

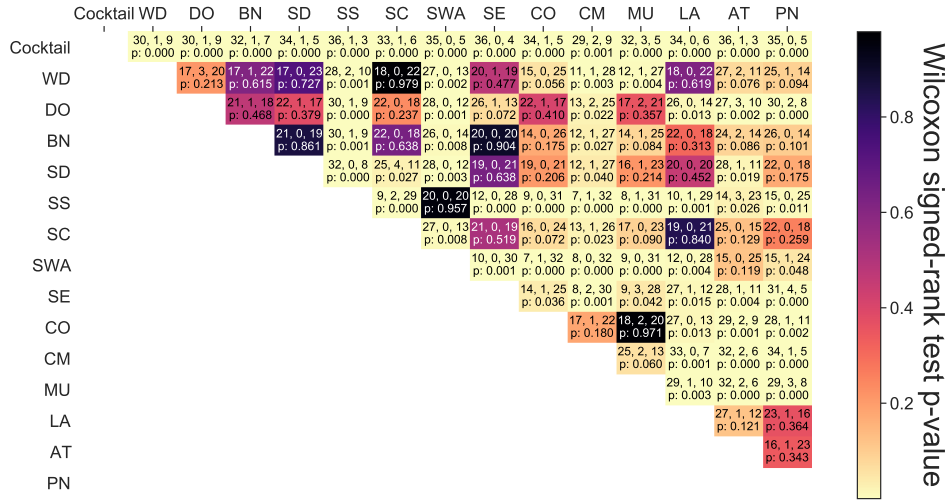


Figure 6: **Pairwise statistical significance and comparison.** For every entry, the first row shows the wins, draws and losses of the horizontal method with the vertical method on all datasets, calculated on the test set; the second row presents the p-value for the statistical significance test.

In Figure 6, we present the results of each pairwise comparison. The results 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 the Wilcoxon signed rank test. As can be seen from the results, the regularization cocktail is the only method that has statistically significant results compared to all the other methods.

C.2 EXPERIMENT 2: DATASET-DEPENDENT OPTIMAL COCKTAILS

In Figure 7, 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.

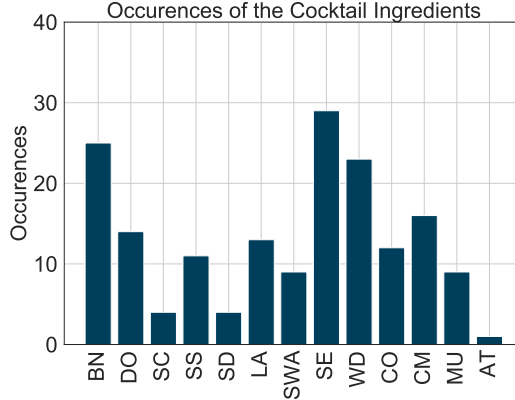


Figure 7: **Frequency of the regularization techniques.** The occurrences of the individual regularization techniques in the best hyperparameter configurations found by the cocktail across 42 datasets.

D TABLES

In this section, at Table 4 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.

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	numera128.6	96320	22	50.52	49.48
233121	Devnagari-Script	92000	1025	2.17	2.17
233122	helena	65196	28	6.14	0.17
233123	jannis	83733	55	46.01	2.01
233124	volkert	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 4: **Datasets.** The collection of datasets used in our experiments, combined with detailed information for each dataset.

Moreover, Table 5 shows the results for the comparison between the Regularization Cocktail and the Top-5 cocktail variants as described in Experiment 2. The results are calculated on the test set for all datasets, after retraining on the best dataset-specific hyperparameter configuration.

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 5: **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.

At Table 6 we provide the results of all our experiments for the 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 the balanced accuracy.

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

Lastly, at Table 7 we present the results of Experiment 3 where we compare our method with GBDT. The results describe the balanced accuracy calculated on the test set after retraining on both methods on the best hyperparameter configuration found within the given budget.

Task Id	GBDT	Cockt.	Task Id	GBDT	Cockt.	Task Id	GBDT	Cockt.	Task Id	GBDT	Cockt.
233090	90.000	89.270	233091	99.850	99.850	233092	46.850	61.461	233093	97.500	98.000
233088	71.191	74.643	233094	80.165	82.576	233096	63.353	74.381	233099	79.830	82.443
233102	62.764	63.923	233103	88.341	86.619	233104	99.967	99.953	233106	68.947	68.107
233107	97.217	96.826	233108	93.519	95.833	233109	64.985	67.617	233110	72.283	85.993
233112	72.645	80.073	233113	98.571	99.948	233114	72.926	73.546	233115	88.589	87.088
233116	100.000	99.587	233117	93.074	93.723	233118	90.457	91.950	233119	83.070	97.471
233120	52.421	52.668	233121	77.897	98.370	233122	21.144	27.70	233123	55.593	65.287
233124	63.428	71.667	233126	94.137	94.015	233130	91.797	92.535	233131	74.447	74.262
233132	98.704	99.049	233133	70.120	69.183	233134	78.878	79.217	233135	95.119	94.045
233137	74.620	94.01	233142	13.534	97.17	233143	92.514	92.531	233146	58.201	64.280

Table 7: **Results of Experiment 3** The performances of the Regularization Cocktail and the GBDT algorithm over the different datasets.