A COMPARISON OF NAS WITH EXPERT ARCHTECTURES

We create a more challenging baseline for NAS by evaluating hand-designed architectures for each specific task. Hand-crafted networks are selected according to best-effort search. The full evaluation results of NAS methods vs. non-NAS baselines can be found in Table 5. Figure 4 illustrates a comparison between best-performing NAS methods vs. best non-NAS methods. Surprisingly, GAEA PC-DARTS beats all the baselines on a portion of the tasks.

Here is a brief summary of these expert models and their citations:

1. DenseNet-BC (CIFAR-100): a more sophisticated version of ResNet, achieving state-of-the-art performance on vision classification (Huang et al., 2017).
2. S2CNN (Spherical): a spherical CNN contains special operations designed for spherical signals, state-of-the-art on spherically-projected MNIST (Cohen et al., 2018).
3. Fourier Neural Operator (FNO) Network (Darcy Flow): via parametrization in Fourier space, FNO can efficiently learn a family of partial differential equations and their mapping to solutions (Li et al., 2021b).
4. DEEPCON (PSICOV): a dilated-convolution neural network combined with dropout to optimize for protein distance prediction (Adhikari, 2020a).
5. deepCR-mask (Cosmic): a modified version of UNet retaining data dimension to keep pixels at the borders to suit astronomy applications, state-of-the-art on this task (Zhang & Bloom, 2020).
6. Attention-based model (NinaPro): a lightweight feed-forward neural network adopting attention modules in place of convolutions (Josephs et al., 2020).
7. VGG-like (FSD50K): a smaller VGG network with output features combining both global max pooling and average pooling for audio (Fonseca et al., 2020).
8. ResNet-1D (ECG): ResNet with 1D convolution, using a larger kernel size of 16 and a stride of 2 for all convolutions. The architecture is state-of-the-art on several time-series prediction tasks in medicine (Hong et al., 2020).
9. ROCKET (Satellite): a simple linear classifier with random convolution kernel as a feature extractor, achieving state-of-the-art performance on UCR time-series prediction tasks (Dempster et al., 2020).
10. DeepSEA model (DeepSEA): the original 1D convolution model accompanying the dataset, state-of-the-art on DeepSEA itself (Zhou & Troyanskaya, 2015).

B EXPERIMENT DETAILS

B.1 HYPERPARAMETER TUNING AND BACKBONE

We use a wide residual network with 16 layers and a widening factor of 4 (WRN-16-4) for all tasks.

For tuning hyperparameters, we use ASHA's default elimination schedule and search over 7 to 256 randomly sampled hyperparameter configurations matching GAEA PC-DART's runtime. The maximum epochs that a single configuration could be trained is equal to that of Wide ResNet's default, 200.

We have selected the following hyperparameter ranges for tuning the Wide ResNet backbone:

- $\log_{10}$(learning rate): Unif[-4, -1]
- momentum: Unif{0.0, 0.3, 0.6, 0.9}
- $\log_{10}$(weight decay): Unif[-5, -2]
- dropout: Unif{0.0, 0.3, 0.6}
- batch size: 128 (all point tasks except FSD50K), 4 (Darcy Flow), 8 (PSICOV, Cosmic), 256(FSD50K, ECG, Deepsea), 4096 (Satellite)

Table 4: Experiment training runtimes of NAS-Perf-360 (GPU hours)

Task	GAEA	DenseNAS	WRN	AMBER / Auto-DeepLab
CIFAR-100	9.5	2.5	2	n/a
Spherical	16.5	2.5	2	n/a
Darcy Flow	6.5	0.5	0.5	5.5
PSICOV	18	24	18.5	19
Cosmic	21.5	2.5	4	17.5
NinaPro	0.5	0.2	0.2	n/a
FSD50K	37	4.5	4	n/a
ECG	140	6.5	5	27
Satellite	28	3	4.5	26
DeepSEA	39.5	2	1.5	28

B.2 Reference Runtimes

Using a Nvidia V100 GPU, we have recorded the following runtimes for each experiment in this benchmark in Table 4. Overall, GAEA PC-DARTS is more costly than backbone with hyperparameter optimization, which is more costly than DenseNAS. The protein tasks requires heavy computation since the data is not static but generated during training.

B.3 Model sizes and FLOPS statistics

Full information of model parameter counts and FLOPs can be found in Table reftable-6 and Table 7.

B.4 Adjustments for Dense Prediction Tasks

On the wide ResNet backbone, we add an adaptive averaging pooling operation to upsample the features back to their original dimensions before output. On the DARTS space, we prevent downsampling and keep spatial dimensions unchanged by disabling reduction cells and replacing them with normal cells. On DenseNAS, we configure the super-network to contain only blocks with the original spatial dimensions.

B.5 Adjustments for 1D Prediction Tasks

The WRN-1D does not have a convolution stem and uses larger kernel sizes of 8,5,3 in each convolution block. We substitute 2D operations with 1D operations within the DARTS and DenseNAS search spaces.

B.6 Random Seeds

For main experiments, we fix the random seed to be 0,1,2 for each of the 3 trials respectively.

For AMBER experiments, we completed three trials as the package did not offer the option of setting random seeds.

B.7 Correlation between Performance and Model Size

We plot performances of 30 random architectures from the DenseNAS search space across three tasks in Figure 3. From our random search experiment, larger models searched by NAS are not always better-performing. We study the Pearson correlation coefficient between test performance vs. model size in number of parameters for three tasks: FSD50K, Cosmic, and ECG. On Cosmic and ECG, the

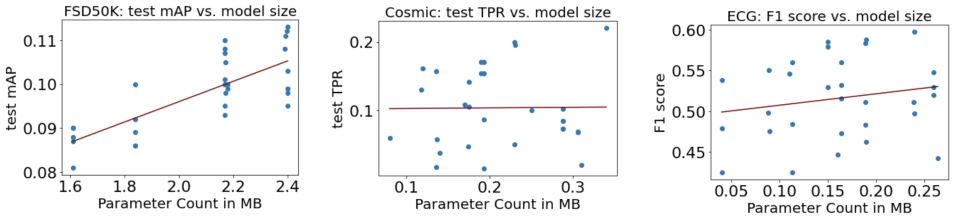

Figure 3: Performances v. Model sizes for three sample tasks.

correlation is very weak ($r = 0.01$ and $r = 0.19$ respectively). On FSD50K, a stronger correlation ($r = 0.79$) is observed but performance varies significantly even for architectures of the same size.

C Supplementary Materials

C.1 Data License

- CIFAR-100: CC BY 4.0 (on https://www.tensorflow.org/datasets/catalog/cifar100)
- Spherical CIFAR-100: CC BY-SA
- NinaPro: CC BY-ND
- FSD50k: CC BY 4.0
- Darcy Flow: MIT
- DeepCov, PSICOV: GPL
- Cosmic: Open License (https://registry.opendata.aws/hst/)
- ECG: ODC-BY 1.0
- Satellite: GPL 3.0
- Deepsea: CC BY 4.0

C.2 Data Preprocessing Details

CIFAR-100: while the 10,000 testing images are kept aside only for evaluating architectures, the 50,000 training images are randomly partitioned into 40,000 for architecture search and 10,000 for validation. On all of the 50,000 training images, we apply standard CIFAR augmentations including random crops and horizontal flipping, and finally normalize them using a pre-calculated mean and standard deviation of this set. On the 10,000 testing images, we only apply normalization with the same constants.

Spherical: with the same split ratios CIFAR-100, the generated spherical image data is directly used for training and evaluation without data augmentation and pre-processing.

NinaPro: Containing less than 4,000 samples, the data is comprised of single-channel signals with an irregular shape of 16*52 pixels. This task also differs from CIFAR for its class imbalance, as over 65% of all gestures are the neutral position. We split the data using the same ratio as CIFAR, resulting in 2638 samples for training, and 659 samples for validation and testing each. No additional pre-processing is performed.

FSD50K: The raw sound files are first resampled at a frequency of 22,050 Hz and transformed into 96-band, log-mel spectrograms, which is a representation of the sound's power spectrum. Small overlapping audio chunks of 1 second are obtained from these larger clips, resulting in an input size of 101*96 (101 frames of 96-band spectrograms). As data augmentation, background noise of the same frequency is also mixed into 75% of the training data. We split 4,170 clips into the validation set and 10,231 clips into the test set following the original paper. During training, we train on one randomly-sampled chunk, instead of all chunks, from each clip.

Darcy Flow: we use scripts provided by (Li et al., 2021b) to generate the PDEs and their solutions, for a total of 900 data points for training, 100 for validation, and 100 for testing. All input data is normalized with constants calculated on the training set before fed into the neural network and de-normalized following an encode-decode scheme. The solutions, or labels, for the training set are also encoded and decoded this way. The test labels are not processed.

PSICOV: we adopt the chosen subset of DeepCov proteins in (Adhikari, 2020b), consisting of 3,456 proteins each with 128*128 feature maps across 57 channels. 100 proteins from this set are used for validation and the rest for training. Test data for final evaluation is gathered from another set of 150 proteins, PSICOV. Since these produce feature maps that are larger (512*512), we run the prediction network over all of its non-overlapping 128*128 patches.

Cosmic: we use data from a specific filter, F435W, of the space telescope, representing the 3605–4882 Å spectral range. Image stamps of 256*256 pixels are taken from large images. The dataset contains 4,347 stamps for training, and 420 for test, and 483 for validation to match the test set size.

ECG: from the sliding window approach, 12,186 single lead recordings are converted into more than 330,000 recording segments comprised of 261,740 for training, 33,281 for validation, and 33,494 for test. Each segment is of the shape 1*1,000, representing one channel of 1,000-long temporal sequence.

Satellite: each satellite time-series is single-channel of length 46 (1*46). After applying standard normalization, we divide the one million entries to 800,000 for training, 100,000 for validation, and 100,000 for test. Zero-padding to 48-length sequences is required for DenseNAS' downsampling network.

DeepSEA: the genome sequences are 1,000-base pair (bp) long and represented as a 1000×4 binary matrix, as each bp is represented as an one-hot encoding corresponding to either A,C,T,G at that location. Total training set size is 71,753. Validation and test sizes that are not subsampled are 2,490 and 149,400 respectively.

C.3 NAS VS. NON-NAS

C.4 MODEL SIZE TABLE

C.5 MODEL FLOPS TABLE

Table 5: Performance of NAS vs. non-NAS baselines across the tasks of NAS-Perf-360. All results are averages of three random seeds.

Search space	Search algorithm	CIFAR-100 0-1 error ^{l}	Spherical 0-1 error ^{l}	Darcy Flow relative ℓ_2 ^{l}	PSICOV MAE_8 ^{l}	Cosmic FNR ^{l}
DenseNAS	random	25.49±0.41	71.23±1.65	0.071±0.006	3.70±0.06	70.42±6.07
DenseNAS	original	25.98±0.38	72.99±0.95	0.10±0.01	3.84±0.15	79.52±2.20
DARTS	GAEA	24.02±1.92	**48.23±2.87**	0.026±0.001	**2.94±0.13**	31.15±3.48
Auto-DL	DARTS	n/a	n/a	0.049±0.005	6.73±0.73	99.79±0.02
WRN	default	23.35±0.05	85.77±0.71	0.073±0.001	3.84±0.053	51.76±2.09
WRN	ASHA	23.39±0.01	75.46±0.40	0.066±0.00	3.84±0.05	37.53±10.16
Expert	default	**19.39±0.20**	67.41±0.76	**0.008±0.001**	3.35±0.14	**25.29±1.44**

Search space	Search algorithm	NinaPro 0-1 error ^{l}	FSD50K mAP ^{h}	ECG F1 score ^{h}	Satellite 0-1 error ^{l}	DeepSEA AUROC ^{h}
DenseNAS	random	8.45±0.56	**0.40±0.001**	0.58±0.01	13.91±0.13	0.60±0.001
DenseNAS	original	10.17±1.31	0.36±0.002	0.60±0.01	13.81±0.69	0.60±0.001
DARTS	GAEA	17.67±1.39	0.06±0.02	0.66±0.01	**12.51±0.24**	0.64±0.02
AMBER	ENAS	n/a	n/a	0.67±0.015	12.97±0.07	0.68±0.01
WRN	default	**6.78±0.26**	0.08±0.001	0.57±0.01	15.49±0.03	0.60±0.001
WRN	ASHA	7.34±0.76	0.09±0.03	0.57±0.01	15.84±0.52	0.59±0.002
Expert	default	8.73±0.90	0.38±0.004	**0.72±0.00**	19.80±0.00	**0.70±0.024**

$^{h\,/\,l}$ a higher / lower value of the metric indicates better performance.

Table 6: Paramater counts of searched and baseline models for all tasks of NAS-Perf-360. Searched model sizes are reported as mean±standard deviation of three random seeds. Results are reported in millions (M). Architectures with the best performance are bolded.

Search space	Search algorithm	CIFAR-100	Spherical	Darcy Flow	PSICOV	Cosmic
DenseNAS	random	1.74±0.12	2.23±0.47	1.00±0.18	1.21±0.16	0.25±0.06
DenseNAS	original	2.03±0.53	1.84±0.15	0.38±0.13	0.93±0.36	0.15±0.16
DARTS	GAEA	4.92±0.28	**1.67±0.14**	0.63±0.08	**0.53±0.05**	0.43±0.15
Auto-DL	DARTS	n/a	n/a	22.98±3.49	6.50±1.84	7.61±2.14
WRN	default	2.77	2.77	2.75	2.76	2.75
Expert	default	**3.08**	0.16	**1.19**	0.60	**0.10**

Search space	Search algorithm	NinaPro	FSD50K	ECG	Satellite	DeepSEA
DenseNAS	random	6.80±0.46	**2.40±0.00**	0.18±0.05	0.79±0.16	0.25±0.04
DenseNAS	original	6.69±0.53	1.45±0.00	0.11±0.05	1.08±0.63	0.19±0.00
DARTS	GAEA	3.35±0.48	0.81±0.11	3.31±0.07	**3.35±0.35**	2.91±0.47
AMBER	ENAS	n/a	n/a	6.61±0.33	6.22±1.36	8.44±1.47
WRN	default	**2.75**	2.80	0.50	0.51	0.51
Expert	default	1.36	0.35	**16.5**	0.48	**60.9**

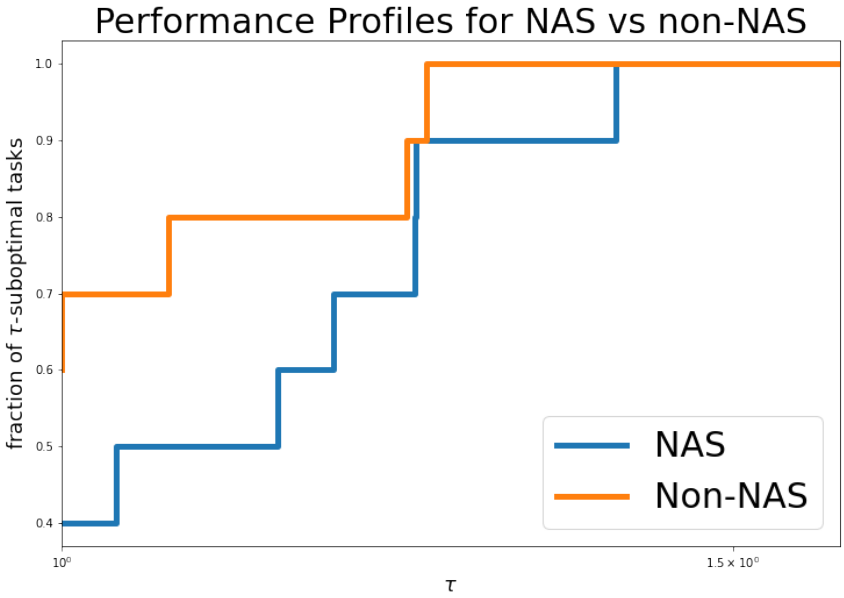

Figure 4: Performance profiles on all tasks for best-performing NAS vs. Non-NAS. The y-value indicates the fraction of tasks on which a plotted method's error is within a multiplicative factor τ of the lowest error achieved by all plotted methods..

Table 7: FLOPS of searched and baseline models for all tasks of NAS-Perf-360. Searched model FLOPS are reported as mean±standard deviation of three random seeds. Results are reported in GFLOPS. Architectures with the best performance are bolded.

Search space	Search algorithm	CIFAR-100	Spherical	Darcy Flow	PSICOV	Cosmic
DenseNAS	random	0.46±0.07	0.91±0.07	14.42±2.58	39.80±5.09	8.42±2.11
DenseNAS	original	0.44±0.53	1.84±0.15	5.43±1.82	30.51±11.90	5.00±5.30
DARTS	GAEA	1.42±0.09	**1.91±0.65**	9.33±1.13	**17.74±1.68**	14.27±4.90
Auto-DL	DARTS	n/a	n/a	2.54±1.20	3.43±1.27	2.44±0.26
WRN	default	0.78	2.78	39.72	90.58	90.06
Expert	default	**1.18**	n/a	**n/a**	0.01	**1.96**
Search space	Search algorithm	NinaPro	FSD50K	ECG	Satellite	DeepSEA
DenseNAS	random	1.02±0.06	**0.40±0.00**	0.11±0.02	0.02±0.01	0.15±0.02
DenseNAS	original	0.97±0.14	0.80±0.00	0.16±0.03	0.02±0.01	0.10±0.00
DARTS	GAEA	0.89±0.12	2.57±0.47	2.28±0.05	**0.11±0.07**	2.01±0.33
AMBER	ENAS	n/a	n/a	0.03±0.01	0.03±0.01	0.04±0.01
WRN	default	**0.64**	7.56	1.02	0.04	1.02
Expert	default	0.02	0.66	**0.70**	0.01	**0.12**

*some expert models contain non-standard modules without FLOPS count.