Specialized Foundation Models Struggle to Beat Supervised Baselines

Zongzhe Xu*, **Ritvik Gupta***, **Wenduo Cheng, Alexander Shen, Junhong Shen**
Carnegie Mellon University
{zongzhex,ritvikgu,wenduoc,ajshen,junhongs}@andrew.cmu.edu
* denotes equal contribution; order decided by coin flip

Ameet Talwalkar
Carnegie Mellon University
talwalkar@cmu.edu

Mikhail Khodak
Princeton University
mkhodak@cs.princeton.edu

Abstract

Following its success for vision and text, the "foundation model" (FM) paradigm—pretraining large models on massive data, then fine-tuning on target tasks—has rapidly expanded to domains in the sciences, engineering, healthcare, and beyond. Has this achieved what the original FMs accomplished, i.e. the supplanting of traditional supervised learning in their domains? To answer we look at three modalities—genomics, satellite imaging, and time series—with multiple recent FMs and compare them to a standard supervised learning workflow: model development, hyperparameter tuning, and training, all using only data from the target task. Across these three specialized domains, we find that it is consistently possible to train simple supervised models—no more complicated than a lightly modified wide ResNet or UNet—that match or even outperform the latest foundation models. Our work demonstrates that the benefits of large-scale pretraining have yet to be realized in many specialized areas, reinforces the need to compare new FMs to strong, well-tuned baselines, and introduces two new, easy-to-use, open-source, and automated workflows for doing so.

1 Introduction

Recent years have witnessed a shift towards large-scale pretraining across domains like computer vision and natural language processing. This workflow generally consists of two stages: pretraining on vast amounts of domain-specific data to capture general knowledge followed by fine-tuning on target tasks (Radford and Narasimhan, 2018). This pretrain-then-finetune paradigm has been tremendously successful, enabling foundation models (Bommasani et al., 2021) to consistently outcompete traditional supervised learning methods on a wide variety of downstream tasks in the vision and language domains (Dosovitskiy et al., 2021; Liu et al., 2021; Devlin et al., 2019).

Driven by this success, the foundation model approach has been adapted to various *specialized* domains, which we define to be ML application areas—e.g. genomics, satellite imaging, and time series—whose data modalities lie outside those of classical AI tasks, i.e. natural images and text. These domains have seen the introduction of many new FMs claiming to leverage large, domain-specific pretraining datasets to achieve breakthrough performance on downstream tasks (Dalla-Torre et al., 2023; Nguyen et al., 2024; Zhou et al., 2023b; Avsec et al., 2021; Ji et al., 2021; Fuller et al., 2023; Cong et al., 2022; Mendieta et al., 2023). These claims underlie our study's motivating question:

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

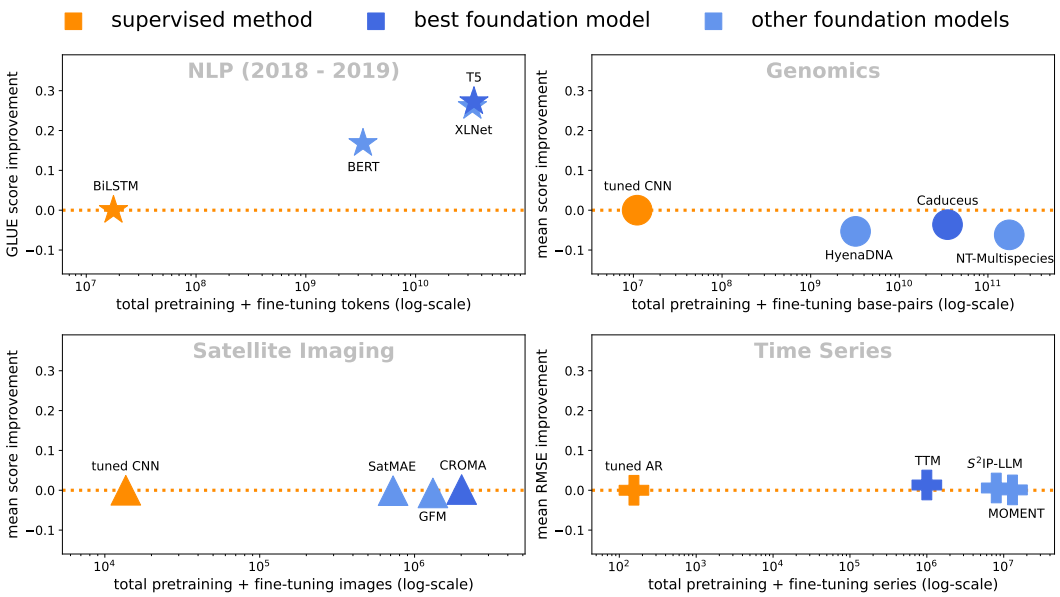

Figure 1: Across three domains—genomics, satellite imaging, and time series—specialized FMs fail to significantly improve upon tuned supervised learning despite using two-to-five orders of magnitude more data. In contrast, breakthrough FMs such as BERT dramatically outperformed supervised baselines in NLP (top left), causing the field to switch to fine-tuning as the default approach. For each domain we plot total pretraining and fine-tuning data used vs. the mean improvement across tasks over the supervised state-of-the-art. Specifics of our evaluations on the last three domains are in Section 4, while the NLP results are derived from the GLUE benchmark (Wang et al., 2019). Note that in the x-axis of the top left figure we ignore tokens used to pretrain word embeddings.

Do these new specialized FMs outperform traditional supervised learning applied to the same tasks?

Answering this question is critical because supervised workflows are usually much less expensive to implement and deploy, but FMs that allow for effective transfer learning have the potential to fundamentally transform these domains, as we have seen with language and vision processing in the past decade. However, despite ongoing efforts to promote their fair and comprehensive evaluation (Liang et al., 2022; Bommasani and Liang, 2021), many new FMs have not been adequately compared to simpler, often more efficient baselines. Indeed, we found that many works only benchmark their proposed models against other FMs, essentially creating a comparison echo chamber (Fuller et al., 2023; Mendieta et al., 2023; Nguyen et al., 2024; Zhou et al., 2023b).

We answer our motivating question by considering a reasonably representative set of three specialized domains—chosen according to the presence of multiple FMs and a standard set of evaluation tasks—and comparing their performance on those tasks with that of a traditional supervised learning workflow. As depicted in Figure 2, the latter is a model development, hyperparameter tuning, and training process in which all steps use only data from the target task, in contrast to the FM workflow, which uses vast amounts of pretraining data. By leveraging model selection tools ranging from classical information criteria to cutting-edge architecture search, we build automated pipelines that efficiently develop and train strong supervised models on over fifty tasks across three distinct domains.

Our main result is negative: we find that, despite being pretrained on massive datasets, specialized FMs struggle and very often fail to outperform models trained exclusively on downstream task data with traditional supervised learning (c.f. Figure 1). Specifically, we show that lightly adapted convolutional neural network (CNN) architectures such as wide ResNet and UNet attain state-of-the-art on the Nucleotide Transformer benchmark in genomics and match the latest pretrained satellite FMs on downstream classification. Furthermore, we show that tuned linear auto-regression (AR) matches or outperforms every open-source time series FM on a standard suite of seven forecasting tasks, despite using four or more orders of magnitude fewer parameters and data.

These results demonstrate that genomics, satellite imaging, and time series have not yet had their "BERT moment" (Devlin et al., 2019), i.e. these domains have not yet pretrained FMs that dominate traditional supervised approaches. This is despite the fact that all them have BERT-scale[1] FMs and the fact that many of them are already witnessing a shift towards not comparing with supervised

[1]Models with 100M+ parameters trained on 100x or more data than supervised tasks in the domain are given.

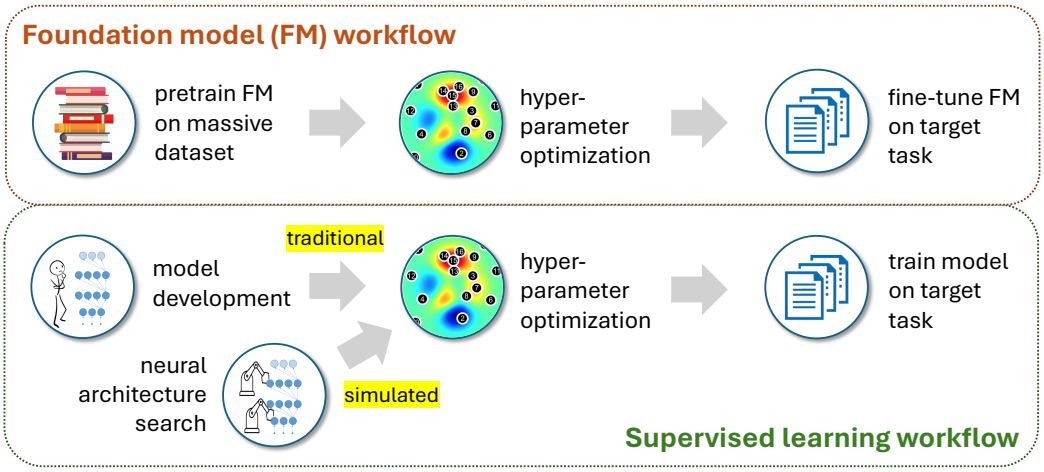

Figure 2: Our goal is to compare the pretrain-then-fine-tune paradigm (top) with a standard supervised workflow (bottom) on the tasks on which specialized FMs are evaluated. While for time series we go through a traditional process of developing and tuning a supervised model, this manual approach does not scale to many domains; as a result, in Section 3.1 we develop a way to simulate it using architecture search. Note that FM fine-tuning hyperparameters are not always tuned in practice, but we assume their creators make a best-effort attempt to present their own method in the best light.

approaches, as was seen in natural language processing (NLP) post-BERT. More broadly, since these domains are among the most high-profile areas with specialized FMs, our results challenge the prevailing assumption that pretrained models yield superior performance. They also reinforce the need for robust and well-tuned baselines, with surprising findings such as (a) simply tuning kernel sizes and dilation rates in standard CNN backbones dominates a genomics classification benchmark and (b) rescuing the century-old AR forecaster from obsolescence is as easy as considering lookback parameters larger than five and training on a GPU. To facilitate ongoing research in these and other domains, we make code associated with both our CNN-tuning pipeline (DASHA[2]) and our AR-on-GPU workflow (Auto-AR[3]) publicly available.

2 Related work

Foundation models have been trained in numerous specialized domains beyond vision and text, including genomics (Ji et al., 2021), satellite imaging (Cong et al., 2022), time series (Goswami et al., 2024), weather (Bodnar et al., 2024), pathology (Zimmermann et al., 2024), differential equation solving (Sun et al., 2024), web traffic (Zhao et al., 2023), and beyond. To get a representative sense of their success, we focus on domains that combine the following properties: (a) multiple BERT-scale FMs, (b) a standard suite of evaluation tasks, and (c) significant applied interest. These restrictions suggest looking at three domains, all of which have at least five FMs evaluated on at least nine tasks: genomics (which has some of the largest-available non-text FMs (Dalla-Torre et al., 2023)), satellite imaging (which has a large ongoing benchmarking effort (Lacoste et al., 2024)), and time series (which has already seen significant industry interest (Cohen et al., 2024)). The remainder of this section examines how different learning workflows approach problems in these domains.

2.1 Specialized foundation models

Collectively our three target domains have more than twenty-five FMs, many developed via the "lift-and-shift" approach—borrowing terminology from Rolf et al. (2024)—in which techniques from core AI areas such as vision and language processing are applied with modest tailoring to specialized domains. In particular, many methods are built on out-of-domain models such as BERT, Swin, and Hyena (Ji et al., 2021; Mendieta et al., 2023; Nguyen et al., 2024; Shen et al., 2024a), with adaptations such as specialized tokenizations, embeddings, and model modifications for handling domain-specific considerations such long-range dependencies (Dalla-Torre et al., 2023; Zhou et al., 2023b; Das et al., 2023; Cohen et al., 2024; Shen et al., 2024b) or multispectral data (Cong et al., 2022).

[2] https://github.com/ritvikgupta199/DASHA

[3] https://github.com/Zongzhe-Xu/AutoAR

While the "lift-and-shift" approach can often be useful or at least a good starting point, its widespread use underlines the need for strong *in-domain* baselines to make sure that the combination of out-of-domain tooling and massive pretraining data is actually helpful. Such comparisons are not always conducted, e.g. the satellite FM SatMAE (Cong et al., 2022) is compared to ImageNet-initialized and randomly initialized ResNet-50 (He et al., 2015), while most of the time series FMs we consider only do a full comparison to one linear baseline, DLinear (Zeng et al., 2023). While this can sometimes be justified—e.g. in the case of NLP post-BERT—our results suggest that for now, specialized FMs should still compare to in-domain supervised model development.

Lastly, we note that we are not the first to take a somewhat critical look at specialized FMs. For example, Yang et al. (2024) questioned the dominance of Transformers for protein sequence FMs by showing that convolutions could do just as well, which is related to our discovery that (supervised) CNNs were competitive with (largely Transformer-based) genomics FMs. Another study by Kedzierska et al. (2023) found that training an in-domain generative model could outperform pretraining one in single-cell biology applications, a finding generalized by our own results, which demonstrate that the underperformance of specialized FMs relative to in-domain training may be a broader trend across multiple domains. In the time series domain, Tan et al. (2024) show that the popular approach of fine-tuning text-pretrained LLMs on time series tasks often underperforms supervised models (in their case randomly initialized attention); our results generalize this to other time series FMs and use an even simpler supervised model (AR) to do so.

2.2 Specialized baselines

Both of the automated supervised learning pipelines we develop are heavily influenced by successful in-domain model development. In particular, the NAS-based pipeline we use to achieve our results in genomics and satellite imaging is inspired by the success of the human-driven specification of kernel sizes and dilation rates in successful architectures like TCN (Lea et al., 2016) and ConvNeXt (Liu et al., 2022). At the same time, for time series our approach is based upon a well-tuned GPU implementation of perhaps the most basic forecasting model, AR.

2.3 AutoML for specialized domains

While often evaluated on domains such as vision, automated techniques have long been used in specialized domains as well. An important example is Auto-ARIMA (Hyndman and Khandakar, 2008) for time series, although it has been found to underperform on the specific suite of tasks we consider (Challu et al., 2022). However, to avoid requiring significant expertise in any one domain, we also make use of AutoML methods developed specifically for diverse tasks (Roberts et al., 2021b; Shen et al., 2023), in particular the NAS method DASH (Shen et al., 2022) that can discover good kernel sizes and dilation rates for a CNN backbone faster than it can be trained from scratch.

3 Methodology

Recall that our goal is to conduct a robust comparison between traditional supervised learning and specialized FMs; the natural way to do this is to take existing benchmarks used to evaluate FMs in our three target domains and run a typical supervised workflow on the same tasks. As depicted in Figure 2, this pipeline involves three steps: (1) model development, (2) hyperparameter tuning, and (3) training. The first stage involves using both reasoning and trial-and-error to find a good architecture to tune and train on the data; for example, Lea et al. (2016) developed the temporal convolutional network (TCN) architecture with a multi-layer dilation rate pattern specifically suited to sequential data, while Liu et al. (2022) designed the breakthrough ConvNeXt architecture by methodically exploring ways to make CNNs more like Transformers without introducing attention. The second stage (hyperparameter tuning) can also be done via human-driven iteration, but there exist effective automated procedures for it as well (Li et al., 2020). Lastly, the third step of the pipeline involves simply training the selected model with the selected configuration on the data of the target task.

While it is standard to automate the last two steps of the procedure, model development is typically done by hand and so is difficult to do for fifty tasks across three domains. As a result, we settle for *approximating* the traditional supervised learning workflow by simulating the model development component using neural architecture search. To ensure fair comparison and reduce computational costs, we restrict ourselves to low-fidelity NAS methods that return an architecture in less time than it takes to train it. The results we obtain using NAS can therefore be viewed as *lower bounds* on the

Algorithm 1: Pseudocode for the DASHA workflow. Starting with a set of backbone CNNs, we use DASH (Shen et al., 2022) to set the right kernel size and dilation rate for each of its convolutional layers and then use ASHA (Li et al., 2020) to configure a training routine for the resulting architecture. Lastly, we pick the best backbone using validation data and train it.

Input: target task dataset D, candidate CNN backbone architectures A
for *CNN backbone* $a \in A$ **do**
    // set a kernel size and dilation rate for each layer of a
    $\texttt{arch}_a \leftarrow \text{DASH}(D, a)$
    // tune hyperparameters for the discovered architecture $\texttt{arch}_a$
    $\texttt{config}_a, \texttt{val\_score}_a \leftarrow \text{ASHA}(D, \texttt{arch}_a)$
// train the architecture with the highest validation score
$a \leftarrow \arg\max_{a \in A} \texttt{val\_score}_a$
Output: $\text{train}(D, \texttt{arch}_a, \texttt{config}_a)$

performance of supervised learning, as the model development might be significantly improved using less-heuristic or human-driven architecture design.

In the remainder of this section we detail how we handle the different steps of the supervised learning pipeline. Note that our NAS-dependent supervised workflow (DASHA)—which we cover in the first part of this section—yields our main results for genomics and satellite imaging but *not* for time series; in that domain we find its performance to be less competitive. There we instead focus on an even simpler approach based on linear auto-regression, whose model development and tuning we describe in the second subsection.

3.1 DASHA: Simulating the supervised workflow using NAS

To simulate model development we need a search space over architectures that is (a) efficient, (b) flexible, and (c) applicable to the types of high-dimensional unstructured data that arise in domains targeted by specialized FMs; these requirements make CNN-based search spaces a natural choice. In particular, inspired by the success of hand-tuned kernel sizes and dilation rates in traditional model development (Lea et al., 2016; Bai et al., 2018; Liu et al., 2022), we apply DASH (Shen et al., 2022), a NAS method that starts with an existing CNN backbone—e.g. a wide ResNet (Zagoruyko and Komodakis, 2017)—and uses the weight-sharing heuristic (Liu et al., 2018) to determine the right kernel size and dilation rate to use at each convolutional layer. DASH has been successfully used in AutoML competitions (Roberts et al., 2021a) and to advance the state-of-the-art on NAS benchmarks (Tu et al., 2022), making it likely to be useful beyond the domains we consider.

As described in Algorithm 1, we augment the existing DASH approach in two ways: (1) trying more than one CNN backbone (e.g. both wide ResNet and UNet (Ronneberger et al., 2015)) and (2) using the well-known hyperparameter tuner ASHA (Li et al., 2020) to configure architecture-specific training settings. This combination gives our workflow its name. Following the NAS and hyperparameter tuning stages, we train the discovered architecture with the selected configuration on the target data. Further details, including the resources given to the three steps of the pipeline and the exact search spaces used by DASH and ASHA, are provided in Appendix B.1. Note that, while our focus is on *data*-efficient baselines, we do ensure that the entire workflow is never substantially more computationally expensive than fine-tuning an FM.

3.2 Auto-AR: Making a baseline stronger by making it simpler

While DASHA can be applied to forecasting tasks, it is not competitive with state-of-the-art time series FMs. At the same time, the field of time series forecasting has long employed automated workflows, notably the Auto-ARIMA approach of Hyndman and Khandakar (2008) that uses statistical tests and information criteria to tune ARIMA's lookback and differencing parameters. Auto-ARIMA was evaluated on the time series tasks we consider by Challu et al. (2022), who found that it performed poorly compared to deep learning approaches. However, their implementation does not make use of multi-channel data and tunes up to a lookback window of at most five, which is much less data than used by time series FMs. While tuning ARIMA with larger lookback parameters is computationally costly, we find the following simplified tuning pipeline to be effective:

1. use the KPSS test (Kwiatkowski et al., 1992) to decide whether to take first differences
2. use the Bayesian Information Criterion to select the maximum lookback parameter of the auto-regressive (AR) component of ARIMA, ignoring the moving average (MA) part
3. maximize the multi-channel likelihood of AR with the chosen differencing and lookback

By dropping the MA component of the model and running the procedure on GPU, we are able to tune the lookback windows up to the maximum allowable length (usually 512); we find that longer lookbacks are critical for performance. Note that this is just a tuned version of the classic AR model.

4 Empirical results

We now present the results of applying the automated pipelines described in the previous section to our three target domains. For each domain, we provide a brief justification of the specific FMs and evaluation tasks that we consider, followed by details on how we apply our workflows; further information can be found in Appendices A and B. As there are too many separate results to present outside the appendix, in this section we mainly present aggregate statistics that summarize our findings for each domain, with detailed results relegated to Appendix C. The domains have different performance metrics, but they can all be aggregated via the following quantities: **average score**, **average rank**, and **mean / median percentage improvement over a baseline**. For each domain, we define a domain-specific baseline and measure the improvement of FMs and our approach relative to it. This standardizes comparisons across tasks of varying scales.

4.1 Genomics

We begin our investigation in the genomics domain, which has witnessed the development of numerous FMs, including the early Enformer Avsec et al. (2021), the DNABERT series (Ji et al., 2021; Zhou et al., 2023b), the HyenaDNA family (Nguyen et al., 2024), GENA-LM[4] (Fishman et al., 2024), the recent Caduceus family (Schiff et al., 2024), and the NT family (Dalla-Torre et al., 2023); The latter includes models with up to 2.5B parameters. To evaluate them, we consider the Nucleotide Transformer (NT) benchmark of Dalla-Torre et al. (2023), which contains eighteen tasks in three main categories: regulatory elements, RNA production, and histone modification. We use this benchmark because of its diversity and because it has been evaluated on by all of the aforementioned FMs, allowing us to include eight of them in the comparison.

Our numbers for these models are taken from Dalla-Torre et al. (2023, Supplementary Table 6); Following Dalla-Torre et al. (2023, Supplementary Table 5), We use F1 score and accuracy to evaluate a subset of regulatory elements and RNA production tasks, and we use Matthew's Correlation Coefficient (MCC) as the main metric for evaluation on the remaining datasets.

4.1.1 Baselines

CNNs have long been used for genomics tasks (Avsec et al., 2020; Zhou and Troyanskaya, 2015) and so constitute natural supervised baselines; in particular we include 1D variants of Wide ResNet (WRN) and UNet, which we find perform better than some domain-specific CNNs. We use these same two backbones as the candidate CNNs tuned and selected from by our DASHA workflow.

4.1.2 Results

Our genomics results are displayed in Table 1, which shows that our supervised workflow (DASHA) consistently outperforms all FMs across all aggregate metrics. As discussed in Appendix C, our strong performance is driven in large part by outstanding performance on the histone modification tasks (c.f. Table 9). The more detailed results also highlight the importance of considering diverse baselines, with Wide ResNet usually being the selected architecture but UNet performing significantly better for promoter and splice site classification tasks. Overall, DASHA arguably sets a new state-of-the-art on the NT benchmark and certainly demonstrates that supervised methods remain quite competitive in genomics, despite the availability of massive pretraining datasets.

4.2 Satellite imaging

While they do not get as large as those in genomics, numerous BERT-scale FMs have also been introduced for satellite imaging, including SeCo (Manas et al., 2021), the SatMAE family (Cong et al.,

[4]We compare to GENA-LM in Appendix C, as its reported metrics differ from the NT benchmark.

Model	Model Size	Pretraining Base-Pairs	Avg. Score ↑	Avg. Rank ↓	Mean %Imp.↑	Median %Imp.↑
Foundation Models						
Enformer	252M	4B	0.569	11.86	27.73	27.91
NT-1000G (500M)	500M	20.5T	0.625	10.52	33.48	36.74
NT-1000G (2.5B)	2.5B	20.5T	0.656	7.0	36.58	40.86
NT-Multispecies (500M)	500M	174B	0.700	3.81	40.76	45.07
NT-Multispecies (2.5B)	2.5B	174B	0.697	4.08	40.51	45.52
DNABERT-2	117M	32.5B	0.680	6.88	38.65	43.59
HyenaDNA-1K	1.6M	3.2B	0.708	6.92	41.2	43.36
HyenaDNA-32K	1.6M	3.2B	0.630	10.22	33.96	36.93
Caduceus-PS	1.9M	35B	0.689	6.69	39.08	41.38
Caduceus-PH	1.9M	35B	0.725	4.69	42.63	45.01
Supervised Methods						
Wide ResNet	2.0M	0	0.694	6.83	37.16	43.08
UNet	4.5M	0	0.68	7.78	38.67	42.69
DASHA (our workflow)	10.5M	0	**0.761**	**3.69**	**46.33**	**49.08**

Table 1: Aggregate performance on genomics tasks, showing that our supervised workflow (DASHA) attains state-of-the-art on the NT benchmark, outperforming all FMs according to most measures while using no pretraining data and oftentimes many fewer parameters. For Mean/Median %Imp., we report percentage improvement over the Raw Probe baseline from Dalla-Torre et al. (2023), and for DASHA the model size refers to the largest configuration across tasks. "-" indicates unknown quantities.

2022), the CROMA family (Fuller et al., 2023), GFM (Mendieta et al., 2023), Scale-MAE (Reed et al., 2023), Satlas (Bastani et al., 2023), Prithvi (Jakubik et al., 2023), and SkySense (Guo et al., 2024). Because our evaluation includes GeoBench (Lacoste et al., 2024), a recently introduced satellite benchmark that has not been considered by many of these FMs, we obtain all results using our own fine-tuning; therefore we only consider a restricted subset of top-performing, open-source, and compatibly-formatted models. In all cases we use the fine-tuning workflow suggested by the authors of each FM plus some automated hyperparameter tuning; note that even with the original code and extra tuning our reproductions on previous benchmarks systematically underperformed results reported in the original works. We take our tasks mainly from GeoBench's five classification tasks and then add four additional tasks—BigEarthNet (Sumbul et al., 2019), EuroSAT (Helber et al., 2019), Canadian Cropland (Jacques et al., 2023), and fMoW-Sentinel (Cong et al., 2022)—that are commonly used to evaluate other FMs.[5] As we focus on classification—sometimes with multiple labels—we report top-1 accuracy or mAP as appropriate.

Model	Model Size	Pretraining Images	Average Score ↑	Average Rank ↓	Mean %Imp.↑	Median %Imp.↑
Foundation Models						
SatMAE-Base	85.6M	700K	76.99	6.22	5.27	3.59
SatMAE-Large	303M	700K	77.75	4.5	6.52	4.62
GFM	86.8M	1.3M	77.18	5.56	5.77	4.08
SwinT-Base	86.8M	14M	76.69	5.28	4.86	1.43
CROMA-Base	90.6M	2M	77.39	4.33	5.85	4.22
CROMA-Large	312M	2M	**78.03**	**3.33**	**6.90**	**6.09**
Supervised Methods						
ResNet50	23.5M	0	73.76	8.34	0.30	00.07
Wide ResNet	17.2M	0	73.97	8.22	0.00	0.00
UNet	17.3M	0	75.73	5.89	3.01	1.07
DASHA (our workflow)	32.4M	0	77.85	**3.33**	6.67	5.16

Table 2: Aggregate performance on satellite imaging tasks, demonstrating that a supervised learning workflow (DASHA) can match the performance of state-of-the-art specialized FMs, all while using no pretraining data and having two-to-ten times fewer parameters. For Mean/Median %Imp. we report percentage improvement over a vanilla Wide ResNet, and for DASHA the model size refers to the largest configuration across tasks.

[5] In Appendix C we report results when excluding tasks where missing channels may affect performance.

Model	Model Size	Pretraining Series	Average RMSE ↓	Average Rank ↓	Mean %Imp.↑	Median %Imp.↑
Foundation Models						
GPT4TS (OFA)	87M	8M	0.555	6.70	31.29	22.71
TEST (Few Shot)	345M	8M	0.603	10.70	25.23	14.59
MOMENT	385M	13M	0.550	5.14	31.95	23.14
TTM (B)	1M	1M	0.543	2.89	32.92	**25.36**
TTM (A)	5M	1M	**0.538**	**2.21**	**33.38**	24.96
S^2IP-LLM	345M	8M	0.545	3.54	32.59	24.63
CALF	86M	8M	0.568	8.95	29.89	21.60
TEMPO (Zero Shot)	345M	8M	0.598	10.54	26.69	22.30
TimesFM (Zero Shot)	200M	5M	0.574	7.38	28.50	19.56
Supervised Methods						
DLinear	700K	0	0.567	8.13	29.68	23.18
Auto-ARIMA	10	0	0.896	12.82	0.00	0.00
AR	513	0	0.556	6.57	31.17	24.31
Auto-AR (our workflow)	513	0	0.551	5.45	31.91	**25.36**

Table 3: Aggregate performance on time series tasks across seven tasks. The latter evaluation demonstrates that simply tuning a classical AR model is competitive with state-of-the-art FMs while using no pretraining data and tens of thousands of times fewer parameters. For Mean/Median %Imp. we report percentage improvement over Auto-ARIMA, and for Auto-AR, the model size refers to the largest configuration across tasks. "-" indicates unknown quantities.

4.2.1 Baselines

Since satellite imaging resembles RGB imaging, it is common to "lift-and-shift" vision models to this domain (Rolf et al., 2024). As a result we use several CNN backbones as baselines and wide ResNet as the candidate architecture for our DASHA workflow. Lastly, we also consider the performance of fine-tuning the ImageNet-pretrained vision FM SwinT-base (Liu et al., 2021).

4.2.2 Results

Table 2 shows that our supervised workflow attains the best or second-best performance across all aggregate metrics and is only ever slightly outperformed by CROMA-large. Notably, unlike in genomics, the FMs here consistently outperform CNN backbones, likely because the associated papers compare to them as baselines. However, the frequently superior performance of DASHA suggests that domain-aware model development would yield good supervised models in this field. Another contrast with genomics is that the larger versions of the FMs consistently attain superior performance here, suggesting they are making at least somewhat effective use of the pretraining data. Nevertheless, that this improvement can also be attained by DASHA, which uses no pretraining and produces a model that is ten times smaller, suggests that there remains significant room for improvement.

4.3 Time series

Our last domain is time series, which has many FMs, including those that use the standard pretrain-then-fine-tune workflow: GPT4TS (OFA) (Zhou et al., 2023a), LLM4TS (Chang et al., 2023), MOMENT (Goswami et al., 2024), TEST (Sun et al., 2023), S^2IP-LLM (Pan et al., 2024), CALF (Liu et al., 2024), TTM (Ekambaram et al., 2024), and Time-LLM (Jin et al., 2024); and others that evaluate in a zero-shot (ZS) regime: TEMPO (Cao et al., 2024), TimesFM (Das et al., 2024), Moirai (Woo et al., 2024), and Toto (Cohen et al., 2024). As we are comparing to *supervised* baselines, our evaluation of ZS models will be in a less challenging setting than the one they report numbers for. We study the performance of these FMs and our baselines on the problem of long-horizon forecasting, which has a standard set of tasks (Goswami et al., 2024, Table 11), of which we consider seven.[6] Note that each task consists of four settings corresponding to different time horizons, so in total this yields twenty-eight tasks. Lastly, we compute aggregate metrics using RMSE, not MSE, so that performance scales linearly with prediction error; this choice has no effect on average rank.

Note that we do not include four of the above time series FMs in our main analysis. One of them, Time-LLM, reports strong results but has had difficulty being reproduced by both us and past efforts, as we detail in Appendix C.3.1. The three others—Moirai, LLM4TS, and Toto—either evaluate

[6]The two we do not consider, Exchange and ILI, are not evaluated on by most time series FMs.

on only a subset of the seven tasks or are closed-source (or both); we list their reported numbers in Table 14. As discussed in Appendix C.3.2, Moirai underperforms Auto-AR and LLM4TS performs roughly on par with TTM (A). The most recently released FM, Toto, does have strong aggregate metrics, but this is mainly due to a dominant performance on a sinle task, ETTh2 (c.f. Figure 4). As a result we do not view these excluded FMs as significantly affecting our conclusions.

4.3.1 Baselines

To baseline these FMs we use mainly linear forecasting methods, including the classical (untuned) linear auto-regression (AR), the automated workhorse Auto-ARIMA (Hyndman and Khandakar, 2008), the more recent DLinear (Zeng et al., 2023), and our own workflow Auto-AR described in Section 3.2. Lastly, we also evaluate our other approach, DASHA, on six of the tasks (c.f. Table 14).

4.3.2 Results

Table 3 shows that on the full seven-dataset evaluation our Auto-AR workflow always attains competitive performances across all aggregate metrics considered, and in particular attains the best median improvement over Auto-ARIMA. Specifically, our Auto-AR workflow achieves competitive performances with two other recent time series FMs, namely MOMENT and S^2IP-LLM, and outperforms the rest of the FMs. Although TTM surpasses all other methods across three aggregated metrics, the improvements remain relatively marginal. This observation aligns with our assertion that the substantial increase in pretraining dataset size and model scale has not yet resulted in significant advancements in model performance. Notably, the three best performing methods are *not* zero-shot, which is perhaps not surprising given the extra data. However, it does reinforce the intuition that settings with high data availability should prefer supervised methods, including simple ones like AR. Notably, even our untuned implementation of AR that uses no differencing and a large lookback window is quite effective, doing better than ZS FMs across all aggregate metrics and even MOMENT on some of them.

5 Discussion

At a high level, our results show that the foundation models in these three domains have not yet surpassed supervised learning, and thus more broadly that the latter remains a strong baseline for specialized FMs. This is a surprising and consequential finding due the paradigm's popularity and the data and compute costs associated with large-scale pretraining. In this section we discuss lessons and implications for the development of machine learning in these and other application areas.

5.1 The importance of diverse, well-tuned, and domain-specific baselines

The main lesson of our work is to select a diverse array of baselines, drawing from both "lift-and-shift" and domain-specific approaches, and then to carefully tune them. For example, in genomics the vanilla wide ResNet baseline does remarkably well, with the majority of FMs doing worse than even this "lift-and-shift" baseline on the typical task in the NT benchmark. While satellite FMs do outperform such baselines, lightly modifying these CNNs via different kernel sizes and dilation rates was enough to match state-of-the-art models there as well. Lastly, our time series results demonstrate in dramatic fashion the need to carefully tune domain-specific approaches, as we show that simply allowing the classical AR forecaster to make use of long lookback windows and GPU-based optimization leads better forecasting than all open-source FMs.

5.2 Computational efficiency considerations

While not our main focus, we nevertheless highlight that any performance gains from FMs must be balanced against their additional cost. In addition to the extensive GPU-hours used for pretraining, the resulting models are often much bigger and so lead to much more costly inference. Indeed, apart from the special case of HyenaDNA, the CNN architectures discovered and trained using our DASHA workflow are typically over ten times smaller than FMs in the case of genomics and three to ten times smaller in the case of satellite imaging. Moreover, for time series our Auto-AR approach is quick-to-train and yields simple models with less than 1K parameters—over two-thousand times smaller than any FM—while attaining performance that is often competitive even with closed-source models. In aggregate, these examples further demonstrate the efficiency of supervised approaches and the resulting high performance bar that FMs need to clear before they can be deemed useful.

5.3 The power of tuning kernel sizes and dilation rates

Our results for genomics and satellite imaging are driven by the DASHA workflow, whose crucial component is the tuning of kernel sizes and dilation rate in CNN backbones such as wide ResNet. Its success demonstrates that the procedure is an effective surrogate for human-driven model development, enabling the automated discovery of the types of diverse, domain-specific baselines stressed in Section 5.1. To understand this further, we study whether the architecture search component selects different kernel sizes and dilation rates for different tasks, and whether it does so in a consistent manner. Specifically, we run DASHA on three of the smaller datasets in the NT benchmark with fifteen different random seeds, construct eighteen-dimensional vectors of the discovered kernel sizes and dilation rates assigned to each of the nine layers, and project these to two dimensions using principal component analysis (PCA). The result in Figure 3 reveals that the architectures are clustered by task, demonstrating that the procedure selects different but consistent-within-task kernel parameters. This visualization suggests that architecture search is a useful surrogate for model development, and consequently that the DASHA workflow may also be useful for automating similar studies and baselining FMs in other domains with high-dimensional, unstructured data.

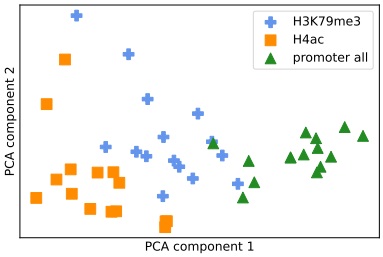

Figure 3: PCA visualization of the architectures discovered for three different tasks when the DASHA workflow is run multiples times. The spatial clustering across tasks demonstrates the within-task consistency of the workflow's architecture search component and the utility of using diverse models as baselines.

5.4 The surprising effectiveness of linear auto-regression

Perhaps our most surprising finding is the competitiveness of linear auto-regression (AR), a very old method, on long-horizon forecasting. It is likely that the lack of comparison with this baseline was driven by existing evaluations (e.g. by Challu et al. (2022)) of Auto-ARIMA (Hyndman and Khandakar, 2008), which is *perceived* to be a stronger baseline because it both combines AR with another model (MA) and tunes the lookback and differencing parameters. However, in most Auto-ARIMA packages the default maximum lookback is around five, whereas we often found much (hundred-fold) larger settings to work best. Since these implementations are also generally too slow to support such long lookbacks, the possibility of expanding the hyperparameter space was more likely to be ignored. By implementing an efficient tuning procedure over a larger space of lookback parameters, our Auto-AR workflow comprises a significant contribution to forecasting baselines.

5.5 Limitations

While our findings are significant according to measures set by past work, they should not be misinterpreted to address all possible scenarios where FMs may be useful. This is most salient for time series FMs motivated by zero-shot concerns, a setting we do not study, and to some extent for genomics FMs, which are often used for exploratory science and not supervised learning. We are also of course computationally limited and there are many other domains where FMs have been pretrained, and even in our three there are other tasks beyond classification and forecasting. Nevertheless, our evaluation is extensive—over twenty-five FMs and over fifty tasks—and so are at least strongly suggestive of the state of a field that uses benchmark performance to motivate and justify pretraining.

6 Conclusion

We conduct a thorough investigation to evaluate whether the cost of training specialized foundation models across three major domains are justified by their superior performance relative to traditional supervised learning. Our results demonstrate that FMs in these domains have not yet surpassed supervised workflows and are often outperformed by fairly simple methods, including lightly modified CNN backbones (in genomics and satellite imaging) and classical linear forecasters (for time series). As part of our study, we introduce two automated workflows—**DASHA** for simulating in-domain model development of CNNs and **Auto-AR** for tuning linear auto-regression on GPUs—that we believe will be useful tools for evaluating future work in these and other areas. The code for these pipelines and to reproduce our results is publicly available.

Acknowledgments

We thank Mononito Goswami, Esther Rolf, and Stephan Xie for useful feedback. This work was supported in part by the National Science Foundation grants IIS1705121, IIS1838017, IIS2046613, IIS2112471, a TCS Presidential Fellowship, and funding from Meta, Morgan Stanley, Amazon, Google, and Scribe. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of these funding agencies.

References

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. (2020). Base-resolution models of transcription factor binding reveal soft motif syntax. *bioRxiv*.

Avsec, , Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv.

Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., and Kembhavi, A. (2023). Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 16726–16736. IEEE.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., et al. (2024). Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L. E., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T. F., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S. P., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J. F., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y. H., Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K. P., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M. A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2021). On the opportunities and risks of foundation models. *ArXiv*.

Bommasani, R. and Liang, P. (2021). Reflections on foundation models.

Box, G. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Cao, D., Jia, F., Arik, S. O., Pfister, T., Zheng, Y., Ye, W., and Liu, Y. (2024). Tempo: Prompt-based generative pre-trained transformer for time series forecasting.

Challu, C., Olivares, K. G., Oreshkin, B. N., Garza, F., Mergenthaler, M., and Dubrawski, A. (2022). N-HiTS: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*.

Chang, C., Peng, W.-C., and Chen, T.-F. (2023). Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *arXiv preprint arXiv:2308.08469*.

Cohen, B., Khwaja, E., Wang, K., Masson, C., Ramé, E., Doubli, Y., and Abou-Amal, O. (2024). Toto: Time series optimized transformer for observability. *arXiv preprint arXiv:2407.07874*.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., and Ermon, S. (2022). SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Dalla-Torre, H., Gonzalez, L., Revilla, J. M., Carranza, N. L., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., Sirelkhatim, H., Richard, G., Skwark, M., Beguir, K., Lopez, M., and Pierrot, T. (2023). The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*.

Das, A., Kong, W., Sen, R., and Zhou, Y. (2023). A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*.

Das, A., Kong, W., Sen, R., and Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Ekambaram, V., Jati, A., Nguyen, N. H., Dayama, P., Reddy, C., Gifford, W. M., and Kalagnanam, J. (2024). Ttms: Fast multi-level tiny time mixers for improved zero-shot and few-shot forecasting of multivariate time series. *arXiv preprint arXiv:2401.03955*.

Fishman, V., Kuratov, Y., Shmelev, A., Petrov, M., Penzar, D., Shepelin, D., Chekanov, N., Kardymon, O., and Burtsev, M. (2024). Gena-lm: A family of open-source foundational dna language models for long sequences. *bioRxiv*.

Fuller, A., Millard, K., and Green, J. R. (2023). Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. (2024). Moment: A family of open time-series foundation models.

Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., et al. (2024). Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Helber, P., Bischke, B., Dengel, A., and Borth, D. (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification.

Hyndman, R. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3).

Jacques, A. A. B., Diallo, A. B., and Lord, E. (2023). The canadian cropland dataset: A new land cover dataset for multitemporal deep learning classification in agriculture.

Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S. K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D. A. B., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Li, H., Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R. K., Weldemariam, K., and Ramachandran, R. (2023). Foundation models for generalist geospatial artificial intelligence. *CoRR*, abs/2310.18660.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P., Liang, Y., Li, Y., Pan, S., and Wen, Q. (2024). Time-llm: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Kedzierska, K. Z., Crawford, L., Amini, A. P., and Lu, A. X. (2023). Assessing the limits of zero-shot foundation models in single-cell biology. bioRxiv.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178.

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E., Kerner, H., Lütjens, B., Irvin, J., Dao, D., Alemohammad, H., Drouin, A., et al. (2024). Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36.

Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (2016). Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 47–54. Springer.

Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Ben-Tzur, J., Hardt, M., Recht, B., and Talwalkar, A. (2020). A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liu, H., Simonyan, K., and Yang, Y. (2018). Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.

Liu, P., Guo, H., Dai, T., Li, N., Bao, J., Ren, X., Jiang, Y., and Xia, S.-T. (2024). Taming pre-trained llms for generalised time series forecasting via cross-modal knowledge distillation. *arXiv preprint arXiv:2403.07300*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., and Rodriguez, P. (2021). Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423.

Mendieta, M., Han, B., Shi, X., Zhu, Y., and Chen, C. (2023). Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al. (2024). Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36.

Pan, Z., Jiang, Y., Garg, S., Schneider, A., Nevmyvaka, Y., and Song, D. (2024). S^2 ip-llm: Semantic space informed prompt learning with llm for time series forecasting. *arXiv preprint arXiv:2403.05798*.

Radford, A. and Narasimhan, K. (2018). Improving language understanding by generative pre-training.

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., and Darrell, T. (2023). Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099.

Roberts, N., Guo, S., Xu, C., Talwalkar, A., Lander, D., Tao, L., Cai, L., Niu, S., Heng, J., Qin, H., Deng, M., Hog, J., Pfefferle, A., Shivakumar, S. A., Krishnakumar, A., Wang, Y., Sukthanker, R., Hutter, F., Hasanaj, E., Le, T., Khodak, M., Nevmyvaka, Y., Rasul, K., Sala, F., Schneider, A., Shen, J., and Sparks, E. R. (2021a). AutoML Decathlon: diverse tasks, modern methods, and efficiency at scale. In *Advances in Neural Information Processing Systems: Competition Track*.

Roberts, N., Khodak, M., Dao, T., Li, L., Ré, C., and Talwalkar, A. (2021b). Rethinking neural operations for diverse tasks. *Advances in Neural Information Processing Systems*, 34:15855–15869.

Rolf, E., Klemmer, K., Robinson, C., and Kerner, H. (2024). Mission critical–satellite data is a distinct modality in machine learning. *arXiv preprint arXiv:2402.01444*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

Schiff, Y., Kao, C.-H., Gokaslan, A., Dao, T., Gu, A., and Kuleshov, V. (2024). Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*.

Shen, J., Khodak, M., and Talwalkar, A. (2022). Efficient architecture search for diverse tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shen, J., Li, L., Dery, L. M., Staten, C., Khodak, M., Neubig, G., and Talwalkar, A. (2023). Cross-modal fine-tuning: Align then refine. In *International Conference on Machine Learning*, pages 31030–31056. PMLR.

Shen, J., Marwah, T., and Talwalkar, A. (2024a). Ups: Towards foundation models for pde solving via cross-modal adaptation. *arXiv preprint arXiv:2403.07187*.

Shen, J., Tenenholtz, N., Hall, J. B., Alvarez-Melis, D., and Fusi, N. (2024b). Tag-llm: Repurposing general-purpose llms for specialized domains. *arXiv preprint arXiv:2402.05140*.

Sumbul, G., Charfuelan, M., Demir, B., and Markl, V. (2019). Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE.

Sun, C., Li, H., Li, Y., and Hong, S. (2023). Test: Text prototype aligned embedding to activate llm's ability for time series. *arXiv preprint arXiv:2308.08241*.

Sun, J., Liu, Y., Zhang, Z., and Schaeffer, H. (2024). Towards a foundation model for partial differential equation: Multi-operator learning and extrapolation. *arXiv preprint arXiv:2404.12355*.

Tan, M., Merrill, M. A., Vinayak Gupta, T. A., and Hartvigsen, T. (2024). Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems*.

Tu, R., Roberts, N., Khodak, M., Shen, J., Sala, F., and Talwalkar, A. (2022). NAS-Bench-360: Benchmarking diverse tasks for neural architecture search. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 7th International Conference on Learning Representations*.

Wang, Y., Wu, H., Dong, J., Liu, Y., Long, M., and Wang, J. (2024). Deep time series models: A comprehensive survey and benchmark.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. (2024). Unified training of universal time series forecasting transformers.

Yang, K. K., Fusi, N., and Lu, A. X. (2024). Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294.

Zagoruyko, S. and Komodakis, N. (2017). Wide residual networks.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. (2023). Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128.

Zhao, R., Zhan, M., Deng, X., Wang, Y., Wang, Y., Gui, G., and Xue, Z. (2023). Yet another traffic classifier: A masked autoencoder based traffic transformer with multi-level flow representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5420–5427.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934.

Zhou, T., Niu, P., Wang, X., Sun, L., and Jin, R. (2023a). One fits all:power general time series analysis by pretrained lm.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. (2023b). Dnabert-2: Efficient foundation model and benchmark for multi-species genome.

Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., Fuchs, T., Fusi, N., Liu, S., and Severson, K. (2024). Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv.

A Tasks

A.1 Genomics

For the Genomics domain, we use the eighteen classification tasks from the Nucleotide Transformer benchmark (Dalla-Torre et al., 2023) that has widely been used for other genomics FMs. The benchmark datasets consist of nucleotide base sequences ranging from 200 to 600 bases in length. It provides a realistic and biological meaningful benchmark across four main categories: promoter (human/mouse), enhancer (human), splice site (SS; human/multispecies) and histone modification (yeast). Within the benchmark, the `enhancers_types` and `splice_sites_all` tasks are classification tasks with three classes each, while the remaining tasks are binary classification tasks.

Dataset	# of classes	# of samples		Maximum sequence length	Metric
		train	test		
enhancers	2	14968	400	200	MCC
enhancers_types	3	14968	400	200	MCC
promoter_all	2	53276	5920	300	F1
promoter_no_tata	2	47767	5299	300	F1
promoter_tata	2	5509	621	300	F1
splice_sites_acceptors	2	19961	2218	600	F1
splice_sites_all	3	27000	3000	400	Accuracy
splice_sites_donors	2	19775	2198	600	F1
H3	2	13468	1497	500	MCC
H3K14ac	2	29743	3305	500	MCC
H3K36me3	2	31392	3488	500	MCC
H3K4me1	2	28509	3168	500	MCC
H3K4me2	2	27614	3069	500	MCC
H3K4me3	2	25953	2884	500	MCC
H3K79me3	2	25953	2884	500	MCC
H3K9ac	2	25003	2779	500	MCC
H4	2	13140	1461	500	MCC
H4ac	2	30685	3410	500	MCC

Table 4: Statistics for Genomics datasets

A.2 Satellite imaging

In the satellite imaging domain, we aim to conduct evaluations with real-world relevance to Earth science. To achieve this, we include a variety of data from different sources to cover a diverse range of tasks, such as brick kiln identification, deforestation prediction, and photovoltaic monitoring. We utilize five classification tasks provided by the GeoBench dataset (Lacoste et al., 2024), a recently developed benchmark that offers a clean and carefully curated collection of tasks specifically designed for satellite imaging. In addition to GeoBench, we evaluate our model on three additional datasets (Helber et al., 2019; Jacques et al., 2023; Sumbul et al., 2019) commonly used in the literature as benchmarks for this domain . This brings the total to eight datasets, encompassing a wide range of features. These tasks vary in complexity, with single-class classification ranging from binary to 62-class problems, as well as two multilabel classification tasks. The datasets are further characterized by diverse input channels, ranging from 3 RGB channels to 18 channels that integrate data from both Sentinel-1 and Sentinel-2 formats.

For Geo-Bench datasets, we do not use any `mixup` and `cutmix` augmentations. For other datasets, we universally use `mixup = 0.8`, `cutmix = 1.0`, and a switch probability of 0.5. Following Fuller et al. (2023), we use only 10% of training set from BigEarthNet and fMoW-Sentinel while using the full evaluation set for validation.

Dataset	Image Size	# of classes	# of samples			# of channels
			train	val	test	
m-bigearthnet	120×120	43	20000	1000	1000	12
m-brickkiln	64×64	2	15063	999	999	13
m-so2sat	32×32	17	19992	986	986	18
m-forestnet	332×332	12	6464	989	993	6
m-pv4ger	320×320	2	11814	999	999	3
BigEarthNet	120×120	19	31166	103944	103728	12
EuroSAT	64×64	13	16200	10800	5400	13
Canadian Cropland	120×120	10	53884	11414	11674	12
fMoW-Sentinel	96×96	62	71287	84939	84966	13

Table 5: Statistics for Satellite datasets

A.3 Time series

In the time series domain, we focus on the long horizon forecasting task. We use a subset of the common benchmark datasets for evaluating models across different domains (ETT, Electricity, Weather, Illness, Traffic, Exchange Rate) (Wang et al., 2024), specifically, the ETT, Weather, Electricity, Illness (ILI), and Traffic datasets. Note that the ETT dataset is actually a collection of four series: ETTh1, ETTh2, ETTm1, and ETTm2; we follow the rest of the literature in treating each series as a separate dataset. Each dataset contains measurements of one or more channels at evenly spaced time steps.

Dataset	# of channels	# of samples train	val	test
ETTh1	7	8033	2785	2785
ETTh2	7	8033	2785	2785
ETTm1	7	33953	11425	11425
ETTm2	7	33953	11425	11425
Weather	21	36280	5175	10444
Electricity	321	17805	2537	5165
ILI	7	69	2	98
Traffic	862	11673	1661	3413

Table 6: Statistics for Time Series datasets

B Implementation details

B.1 DASHA

Following the architecture search, we perform hyperparameter tuning using ASHA. The hyperparameter search space includes learning rate, weight decay, momentum, drop rate, and random seed for model initialization. We define a continuous search space, with further specific details provided in Table 7. Using ASHA, we evaluate 200 sample configurations over a maximum of 20 epochs, using a reduction factor of 2. The low-performing configurations are pruned based on their validation scores.

Before retraining the final model, we load the model checkpoint corresponding to the optimal hyperparameter configuration. The model is then trained for 200 epochs on the training data, with the best-performing checkpoint selected based on validation performance. This process is repeated for each backbone architecture, and the best-performing backbone is selected using the validation score. Finally, the checkpoint for the selected backbone is evaluated on the test set to obtain the final score.

Hyperparameter	Search Space	Type of Search Space
`random_seed`	$[0, 500]$	Integer
`lr`	$[10^{-5}, 5 \times 10^{-1}]$	Log Uniform
`drop_rate`	$\{0, 0.05, 0.1\}$	Discrete
`weight_decay`	$[5 \times 10^{-7}, 5 \times 10^{-3}]$	Log Uniform
`momentum`	$[0.9, 1]$	Uniform

Table 7: Hyperparameter Search Space

B.2 Auto-AR

A fairly complete description is provided in Section 3.2. Here we note only that, because we minimize the total maximum likelihood across (independent) channels, to determine the amount of differencing used for each task we run the KPSS test separately on each channel and use the differencing needed by the majority of the channels. Notably, this results in a differencing of one for each task.

C Detailed results

C.1 Genomics

We include all FMs listed in Dalla-Torre et al. (2023, Supplementary Table 6) with addition of the two recently released models, Caduceus (Schiff et al., 2024) and Gena-LM (Fishman et al., 2024); note that the last row of tasks in the NT paper(promoter and splice sites) are mislabeled, but we infer an order in combination with previous information obtained from a (now-deleted) Huggingface leaderboard.[7] In alignment with the leaderboard, we apply a 0.1 validation split for DASHA during our evaluation. Additionally, we use an architecture set that includes both Wide ResNet and UNet for the search with DASHA on these datasets. We use batch size= 128 for all datasets, and cross entropy loss for all the training and finetuning. Individual scores for each task in the benchmark are provided in Tables 9 and 8.

Model	Regulatory Elements					RNA Production		
	enhancers	enhancers types	promoter all	promoter no_tata	promoter tata	splice_sites acceptors	splice_sites all	splice_sites donors
Enformer	0.454	0.312	0.955	0.955	0.959	0.915	0.847	0.906
NT-1000G (500M)	0.509	0.395	0.951	0.951	0.936	0.965	0.968	0.971
NT-1000G (2.5B)	0.546	0.432	0.965	0.967	0.957	0.98	0.976	0.979
NT-Multispecies (500M)	0.559	0.438	0.976	0.976	0.965	0.981	0.984	0.987
NT-Multispecies (2.5B)	0.545	0.444	0.975	0.977	0.959	0.986	0.982	0.987
DNABERT-1	0.495	0.367	0.961	0.962	0.956	–	0.975	–
DNABERT-2	0.525	0.423	0.972	0.972	0.955	0.975	0.939	0.963
HyenaDNA-1K	0.52	0.403	0.959	0.959	0.944	0.959	0.956	0.947
HyenaDNA-32K	0.489	0.352	0.956	0.954	0.939	0.96	0.962	0.957
Caduceus-PS	0.491	0.416	0.967	0.968	0.957	0.936	0.927	0.874
Caduceus-PH	0.546	0.439	0.97	0.969	0.953	0.937	0.94	0.948
Wide ResNet	0.525	0.416	0.952	0.946	0.93	0.821	0.457	0.815
UNet	0.49	0.366	0.956	0.954	0.95	0.956	0.955	0.968
DASHA	0.527	0.432	0.958	0.962	0.957	0.978	0.979	0.978

Table 8: Regulatory Elements and RNA Production Downstream Tasks

Model	H3	H3K14ac	H3K36me3	H3K4me1	H3K4me2	H3K4me3	H3K79me3	H3K9ac	H4	H4ac
Enformer	0.724	0.284	0.345	0.291	0.207	0.156	0.498	0.415	0.735	0.275
NT-1000G (500M)	0.736	0.381	0.468	0.38	0.26	0.235	0.562	0.479	0.755	0.342
NT-1000G (2.5B)	0.754	0.453	0.53	0.418	0.278	0.311	0.574	0.491	0.787	0.408
NT-Multispecies (500M)	0.786	0.549	0.624	0.55	0.32	0.406	0.63	0.567	0.799	0.496
NT-Multispecies (2.5B)	0.793	0.538	0.618	0.541	0.324	0.408	0.623	0.547	0.808	0.492
DNABERT-1	0.763	0.403	0.474	0.396	0.282	0.258	0.578	0.505	0.784	0.359
DNABERT-2	0.785	0.515	0.591	0.512	0.333	0.353	0.615	0.545	0.797	0.465
HyenaDNA-1K	0.781	0.608	0.614	0.512	0.455	0.55	0.669	0.586	0.763	0.564
HyenaDNA-32K	0.747	0.405	0.479	0.387	0.276	0.291	0.567	0.472	0.761	0.379
Caduceus-PS	0.799	0.541	0.609	0.488	0.388	0.44	0.679	0.604	0.789	0.525
Caduceus-PH	0.815	0.631	0.601	0.523	0.487	0.544	0.697	0.622	0.811	0.621
Wide ResNet	0.798	0.667	0.670	0.554	0.541	0.660	0.706	0.620	0.754	0.657
UNet	0.797	0.647	0.482	0.541	0.553	0.292	0.562	0.624	0.760	0.389
DASHA	0.790	0.683	0.630	0.528	0.640	0.714	0.721	0.709	0.776	0.744

Table 9: Histone Modification Downstream Tasks

The performance of the models on individual tasks is detailed in Tables 8 and 9. In the regulatory elements domain (c.f. Table 8), DASHA performs slightly behind the largest models like NT-Multispecies (Dalla-Torre et al., 2023) on the enhancers tasks but consistently outperforms models such as DNABERT-1 (Ji et al., 2021), Enformer (Avsec et al., 2021), and HyenaDNA (Nguyen et al., 2024); in the promoters task in generally performs worse than all reported FMs. In the RNA production domain, DASHA performs near the middle of the FMs. However, where DASHA truly excels is in the histone modification tasks (in Table 9), where it not only competes with, but often outperforms, the other FMs, consistently achieving top scores in nearly all tasks.

Note that because GENA-LM reports using MCC on all datasets, which is different than the metrics used by NT and Caduceus paper, we compare its results with DASHA in a separate table, where all datasets are evaluated through MCC. As shown in Table 10 and Table 11, DASHA also outperform GENA-LM by a large margin in terms of average MCC.

[7] `https://huggingface.co/spaces/InstaDeepAI/nucleotide_transformer_benchmark`

Model	H3	H3K14ac	H3K36me3	H3K4me1	H3K4me2	H3K4me3	H3K79me3	H3K9ac	H4	H4ac
GENA-LM	0.79	0.6	0.61	0.53	0.46	0.55	0.67	0.61	0.78	0.59
DASHA	0.79	0.68	0.63	0.53	0.64	0.71	0.72	0.71	0.78	0.74

Table 10: Histone Modification for GENA-LM (all scores are obtained using MCC)

Model	enhancers	enhancers types	promoter all	promoter no_tata	promoter tata	splice_sites acceptors	splice_sites all	splice_sites donors	Average
GENA-LM	0.55	0.45	0.94	0.94	0.91	0.92	0.91	0.91	0.707
DASHA	0.53	0.43	0.92	0.92	0.90	0.97	0.96	0.96	0.755

Table 11: Regulatory Elements and RNA Production for GENA-LM. (all scores are obtained using MCC) The last column shows the average MCC across all 18 datasets.

C.2 Satellite imaging

Training on satellite datasets requires relatively large computational resources due to the high number of channels and the size of the datasets. To ensure a fair comparison, we fine-tuned all the foundation models ourselves by sweeping across a fixed set of base learning rates $[5e-3, 2e-3, 2e-4, 4e-5]$. We then calculate the actual learning rate from base learning rate following previous work by $lr = base\_lr \cdot \frac{batch\_size}{256}$. This approach ensures that approximately the same amount of resources were used as during the DASHA tuning process, allowing for a balanced evaluation of model performance.

We closely followed the reported evaluation processes from previous studies on FMs (Cong et al., 2022; Fuller et al., 2023; Mendieta et al., 2023). These models do not employ a validation set for hyperparameter tuning or model selection, and we adhered to this same approach when fine-tuning the FMs. However, for DASHA, since we performed extensive hyperparameter optimization over a large search space, we used a validation set to ensure fair and accurate comparisons between DASHA and the FMs. This is a less favorable setting for DASHA, as it relies on extensive hyperparameter tuning, but we demonstrate that, even under these conditions, DASHA matches the performance of the FMs.

It is also important to note that SatMAE only accepts 3-channel and 12-channel inputs, while CROMA is limited to 12-channel inputs. GeoBench, however, includes a wide range of tasks with varying numbers of input channels, ranging from 3 to 18. Despite these differences, we include all datasets in our evaluation because they are valuable benchmarks in the satellite image domain, and it is crucial for FMs in this field to generalize across diverse datasets. For datasets where the input size does not match the model requirements, we pad missing channels with zeros and prune any extra channels. However, to ensure a fair comparison, in addition to reporting the average scores across all datasets, we also provide average scores excluding m-pv4ger and m-forestnet, where missing channels may affect the performance of the FMs. The aggregate scores excluding m-pv4ger and m-forestnet are presented in Figure 12.

Model	Average Score ↑	Average Rank ↓	Mean %Imp.↑	Median %Imp.↑
Foundation Model				
SatMAE-Base	77.71	6.00	6.74	4.56
SatMAE-Large	78.44	4.07	7.87	**6.75**
GFM	76.95	5.57	5.40	4.08
SwinT-Base	76.12	6.50	3.98	1.43
CROMA-Base	77.98	3.57	6.95	5.30
CROMA-Large	**79.06**	**2.14**	**8.84**	6.26
Supervised Methods				
ResNet50	73.25	9.00	-0.27	-0.38
Wide ResNet	73.90	8.00	0.00	0.00
UNet	75.29	6.57	2.33	0.87
DASHA	78.00	3.57	7.02	5.16

Table 12: Aggregated metrics on Satellite imaging tasks excluding the m-pv4ger and m-forestnet.

For training and finetuning, we universally use `batch_size = 16` and loss function as cross entropy with 0.1 label smoothing for single label classification, and multi-label soft margin loss for multilabel classification. Individual scores for each task are provided in Table 13.

Model	Average	m-bigearthnet	m-brickkiln	m-so2sat	m-forestnet	m-pv4ger	BigEarth Net	EuroSAT	Canadian Cropland	fMoW Sentinel
SatMAE-Base	76.99	72.3	98.22	54.56	51.89	97.0	86.04	98.69	74.64	59.55
SatMAE-Large	77.75	73.82	98.6	55.79	53.7	96.92	86.75	98.86	75.38	59.89
GFM	77.18	71.97	98.35	57.52	59.38	96.54	85.93	99.02	72.03	53.84
SwinT-Base	76.69	70.14	98.81	56.49	59.78	97.54	85.91	98.99	70.15	52.37
CROMA-Base	77.39	72.07	98.99	60.04	54.07	96.6	86.94	98.81	75.87	53.14
CROMA-Large	78.03	73.36	99.01	59.22	51.96	96.9	87.98	98.98	76.56	58.32
ResNet50	73.76	60.31	98.4	51.83	54.29	96.79	79.16	98.23	72	52.84
Wide ResNet	73.97	69.15	98.95	49.04	52.14	96.34	80.48	98.6	72.05	49.00
UNet	75.73	69.89	98.6	56.87	57.18	97.39	83.9	98.99	72.68	46.11
DASHA	77.85	72.72	98.92	56.28	57.24	97.4	86.09	99.07	75.69	57.20

Table 13: Satellite Imaging Tasks

C.3 Time series

The long horizon forecasting task for a time series can be summarized as follows: at every timestep t, take the L historical observations at times $(t - L + 1, ..., t)$ for each channel and predict the next H observations $(t + 1, ..., t + H)$ for every channel. Following the literature, we evaluate models on $H \in \{24, 36, 48, 60\}$ for the Illness dataset and $H \in \{96, 192, 336, 720\}$. For most methods, we report results when $L = 512$.

Model	ETTh1				ETTh2				ETTm1				ETTm2			
Horizon	96	192	336	720	96	192	336	720	96	192	336	720	96	192	336	720
GPT4TS (OFA)	0.376	0.416	0.442	0.477	0.285	0.354	0.373	0.406	0.292	0.332	0.366	0.417	0.173	0.229	0.286	0.378
TEST (Few shot)	0.455	0.572	0.611	0.723	0.332	0.401	0.408	0.459	0.392	0.423	0.471	0.552	0.233	0.303	0.359	0.452
LLM4TS	0.371	0.403	0.42	0.422	0.269	0.328	0.353	0.383	0.285	0.324	0.353	0.408	0.165	0.22	0.268	0.35
MOMENT	0.387	0.41	0.422	0.454	0.288	0.349	0.369	0.403	0.293	0.326	0.352	0.405	0.17	0.227	0.275	0.363
TTM (B)	0.36	0.392	0.401	0.436	0.269	0.336	0.359	0.39	0.291	0.325	0.363	0.419	0.164	0.219	0.277	0.35
TTM (A)	0.363	0.392	0.413	0.442	0.262	0.324	0.351	0.392	0.283	0.332	0.353	0.393	0.158	0.213	0.269	0.369
S^2IP-LLM	0.366	0.401	0.412	0.44	0.278	0.346	0.367	0.4	0.288	0.323	0.359	0.403	0.165	0.222	0.277	0.363
CALF	0.369	0.427	0.456	0.479	0.279	0.353	0.362	0.404	0.323	0.374	0.409	0.477	0.178	0.242	0.307	0.397
TEMPO (Zero Shot)	0.4	0.426	0.441	0.443	0.301	0.355	0.379	0.409	0.438	0.461	0.515	0.591	0.185	0.243	0.309	0.386
TimesFM (Zero Shot)	0.421	0.472	0.51	0.514	0.326	0.399	0.434	0.451	0.357	0.411	0.441	0.507	0.205	0.294	0.367	0.473
Moirai (Zero Shot)	0.375	0.399	0.412	0.413	0.281	0.34	0.362	0.38	0.404	0.435	0.462	0.49	0.205	0.261	0.319	0.415
Toto (Zero Shot)	0.307	0.329	0.396	0.419	0.093	0.135	0.16	0.294	0.306	0.328	0.39	0.463	0.2	0.269	0.264	0.354
ARIMA	0.646	0.704	0.732	0.738	0.324	0.411	0.456	0.462	1.131	1.172	1.197	1.231	0.225	0.298	0.37	0.478
AR (d=0)	0.358	0.39	0.41	0.424	0.271	0.334	0.361	0.395	0.299	0.336	0.368	0.426	0.163	0.218	0.271	0.366
DLinear	0.375	0.405	0.439	0.472	0.289	0.383	0.448	0.605	0.299	0.335	0.369	0.425	0.167	0.224	0.281	0.397
Auto-AR	0.357	0.39	0.41	0.422	0.269	0.332	0.359	0.394	0.299	0.336	0.368	0.426	0.163	0.218	0.271	0.367
DASHA	0.369	0.401	0.430	0.478	0.284	0.377	0.396	0.745	0.305	0.335	0.367	0.418	0.169	0.224	0.290	0.378

Model	Weather				Electricity				ILI				Traffic			
Horizon	96	192	336	720	96	192	336	720	96	192	336	720	96	192	336	720
GPT4TS (OFA)	0.162	0.204	0.254	0.326	0.139	0.153	0.169	0.206	2.063	1.868	1.79	1.979	0.388	0.407	0.412	0.45
TEST (Few shot)	0.163	0.23	0.278	0.301	0.143	0.158	0.176	0.23	-	-	-	-	0.415	0.425	0.436	0.489
LLM4TS	0.147	0.191	0.241	0.313	0.128	0.146	0.163	0.2	–	–	–	–	0.372	0.391	0.405	0.437
MOMENT	0.154	0.197	0.246	0.315	0.136	0.152	0.167	0.205	2.728	2.669	2.728	2.883	0.391	0.404	0.414	0.45
TTM (B)	0.146	0.19	0.242	0.323	0.129	0.149	0.163	0.2	-	-	-	-	0.368	0.403	0.395	0.431
TTM (A)	0.149	0.192	0.24	0.318	0.128	0.144	0.162	0.191	-	-	-	-	0.352	0.359	0.375	0.419
S^2IP-LLM	0.145	0.19	0.243	0.312	0.135	0.149	0.167	0.2	-	-	-	-	0.379	0.397	0.407	0.44
CALF	0.164	0.214	0.269	0.355	0.145	0.161	0.175	0.222	-	-	-	-	0.407	0.43	0.444	0.477
TEMPO (Zero Shot)	0.211	0.254	0.292	0.37	0.178	0.198	0.209	0.279	3.0	2.956	2.651	2.701	0.476	0.496	0.503	0.538
TimesFM (Zero Shot)	0.122	0.169	0.242	0.391	0.119	0.137	0.158	0.206	2.595	2.984	3.34	3.227	0.327	0.353	0.378	0.42
Moirai (Zero Shot)	0.173	0.216	0.26	0.32	0.205	0.22	0.236	0.27	–	–	–	–	–	–	–	–
Toto (Zero Shot)	0.18	0.235	0.252	0.356	0.124	0.138	0.155	0.211	–	–	–	–	–	–	–	–
ARIMA	0.217	0.263	0.33	0.425	1.22	1.264	1.311	1.364	5.554	6.94	7.192	6.648	1.997	2.044	2.096	2.138
AR (d=0)	0.171	0.215	0.263	0.332	0.138	0.153	0.17	0.212	2.084	2.04	2.004	2.011	0.398	0.413	0.426	0.464
DLinear	0.176	0.22	0.265	0.323	0.14	0.153	0.169	0.203	2.215	1.963	2.13	2.368	0.41	0.423	0.436	0.466
Auto-AR	0.172	0.215	0.263	0.332	0.138	0.153	0.17	0.212	2.084	2.04	2.004	2.011	0.398	0.413	0.426	0.464
DASHA	0.163	0.205	0.251	0.314	0.136	0.151	0.165	0.200	–	–	–	–	–	–	–	–

Table 14: Time Series Forecasting Tasks

In addition to the results for DASHA, Auto-AR, DLinear, and FMs, we evaluate the performance of two simple baselines

1. Vanilla Autoregressive Model (Box and Jenkins, 1976): This model predicts the (scalar) value of a time series at $t + 1$ as a linear combination of the last L timesteps and a constant,

i.e. $\hat{x}_{t+1} = \alpha_0 + \alpha_1 x_t + \alpha_2 x_{t-1} + ... + \alpha_L x_{t-L+1}$ for learnable parameters $\alpha_0, ..., \alpha_L$. We fit these parameters using standard maximum likelihood techniques.

2. ARIMA is a statistical method used for time series forecasting that combines three components: AutoRegressive (AR), Integrated (I), and Moving Average (MA). The AR component models the relationship between an observation and its lagged (past) values, assuming that past values have a linear influence on future ones. The Integrated component applies differencing to the data to remove trends or seasonality, making the time series stationary by stabilizing its mean over time. The MA component models the relationship between an observation and the residual errors from a moving average model applied to previous observations. ARIMA is characterized by three parameters: p (the number of lag observations), d (the number of differencing steps to achieve stationarity), and q (the number of lagged forecast errors). This model is particularly effective for univariate time series forecasting where patterns like trends or seasonality are present.

All results are reported on a 70/10/20 train/validation/test split for each datasets, except for the ETT datasets which have predefined splits. MSE is reported after all datasets have been scaled by the mean and variance of the training data. Both autoregressive models have only one tunable hyperparameter (number of lags). Similarly, the linear model has only one tunable hyperparameter (number of training epochs).

The baselines as described can handle only univariate time series, while all of the benchmark datasets are multivariate (multiple channels). These baselines are trained under channel independence: each channel of a time series is treated independently. While channel independence fails to take into account cross-channel dependencies, we note that developing methods that leverage cross-channel dependencies for a variable number of channels remains an open problem.

C.3.1 Time-LLM

Due to differences in the results reported for Time-LLM between the original paper and reproductions in several works (Goswami et al., 2024; Ekambaram et al., 2024; Pan et al., 2024; Tan et al., 2024), we to attempt our own partial reproduction in order to determine whether to include their numbers. To do so, we used code provided by the authors and looked at ETTh1 with input lengths from $96, 192$ and ILI with input lengths from $24, 36, 48, 90$ (going beyond this was infeasible due to the resources required to finetune LLaMA-7B).

Our reproduced MSE for ETTh1 with an input length of 96 was 0.405 as compared to that of 0.362 reported by Jin et al. (2024), while for input length of 192 it was 0.431 vs. 0.398. On ILI the average increase in MSE across time horizons from their reported results to our reproduction was more than 0.4. Due to these large discrepancies we chose to exclude their model from our analysis.

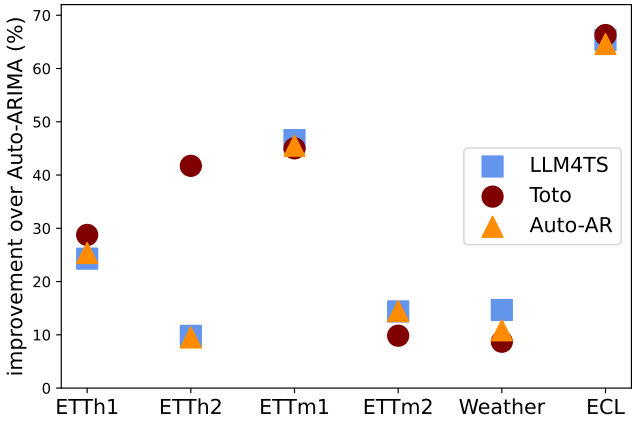

Figure 4: Scatterplot of improvement in RMSE (averaged across time horizons) of two closed-source FMs, LLM4TS and Toto, on a subset of datasets in Table 14. It shows that both FMs usually perform similarly to the Auto-AR baseline, with Toto attaining strong aggregate metrics due to a dominant performance on ETTh2.

C.3.2 Closed-source models

In Table 14 we consider three additional FMs that either do not report complete set of results (Moirai) on all 7 datasets or are closed-source (Toto and LLM4TS). While the performance of Auto-AR is comparably to that of LLM4TS, Toto dominates the three aggregated metrics we considered in Table 15; impressively, it does this despite being zero-shot. However, a look at Figure 4 reveals that Toto is not superior across all six tasks, with the aggregated metrics being strongly boosted by its dramatically better performance on one of them (ETTh2). Thus, while its ZS performance is quite good, it is unclear whether this result would be maintained with additional tasks.

Model	Model Size	Pretraining Series	**Partial setting (6 datasets / 24 tasks)** Avg. RMSE ↓	Avg. %Imp.↑	Median %Imp.↑
Moirai (Zero Shot)	14M	6M	0.566	24.53	18.52
LLM4TS	60M	8M	0.526	29.25	20.96
TTM (A)	5M	1M	0.525	29.38	19.87
Toto (Zero Shot)	103M	-	**0.505**	**32.40**	**31.35**
Auto-AR	513	0	0.534	28.12	19.63
DASHA	480K	0	0.549	25.94	16.78

Table 15: Additional time series results on the 6 datasets setting