How do we build a general intelligence?

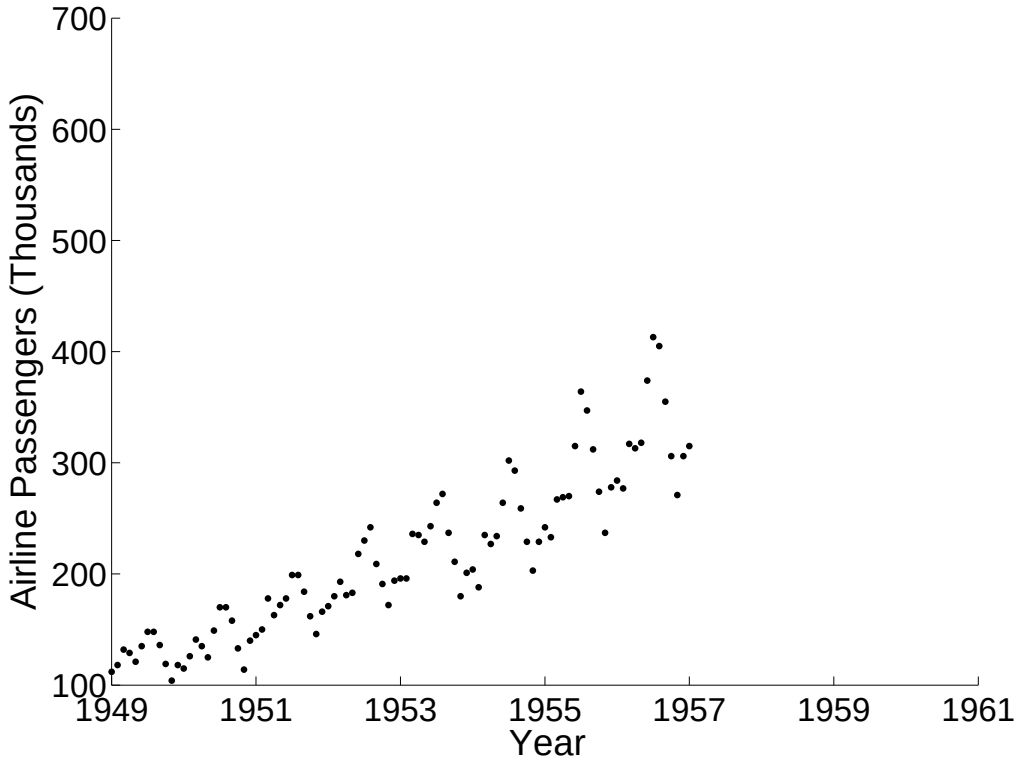
Andrew Gordon Wilson

https://cims.nyu.edu/~andrewgw
Courant Institute of Mathematical Sciences
Center for Data Science
New York University

How far are we from AGI?
May 11, 2024

Special thanks: Marc Finzi, Micah Goldblum, Pavel Izmailov, Sanae Lotfi, Shikai Qiu, Nate Gruver, Polina Kirichenko, Andres Potapczynski

Model Selection



Which model should we choose?

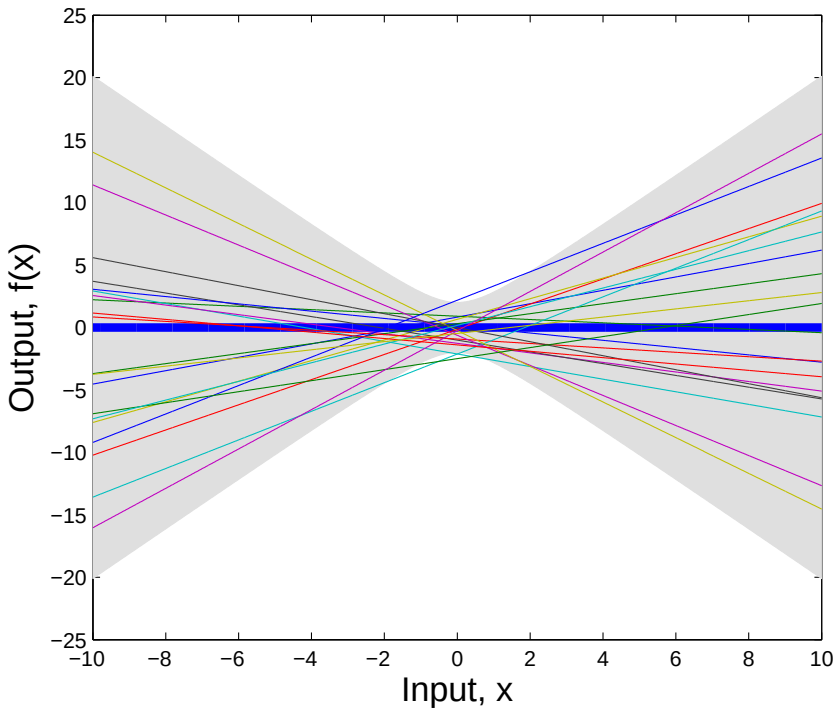
(1): $f_1(x) = a_0 + a_1 x$ **(2):** $f_2(x) = \sum_{j=0}^{3} a_j x^j$ **(3):** $f_3(x) = \sum_{j=0}^{10^4} a_j x^j$

A Function-Space View

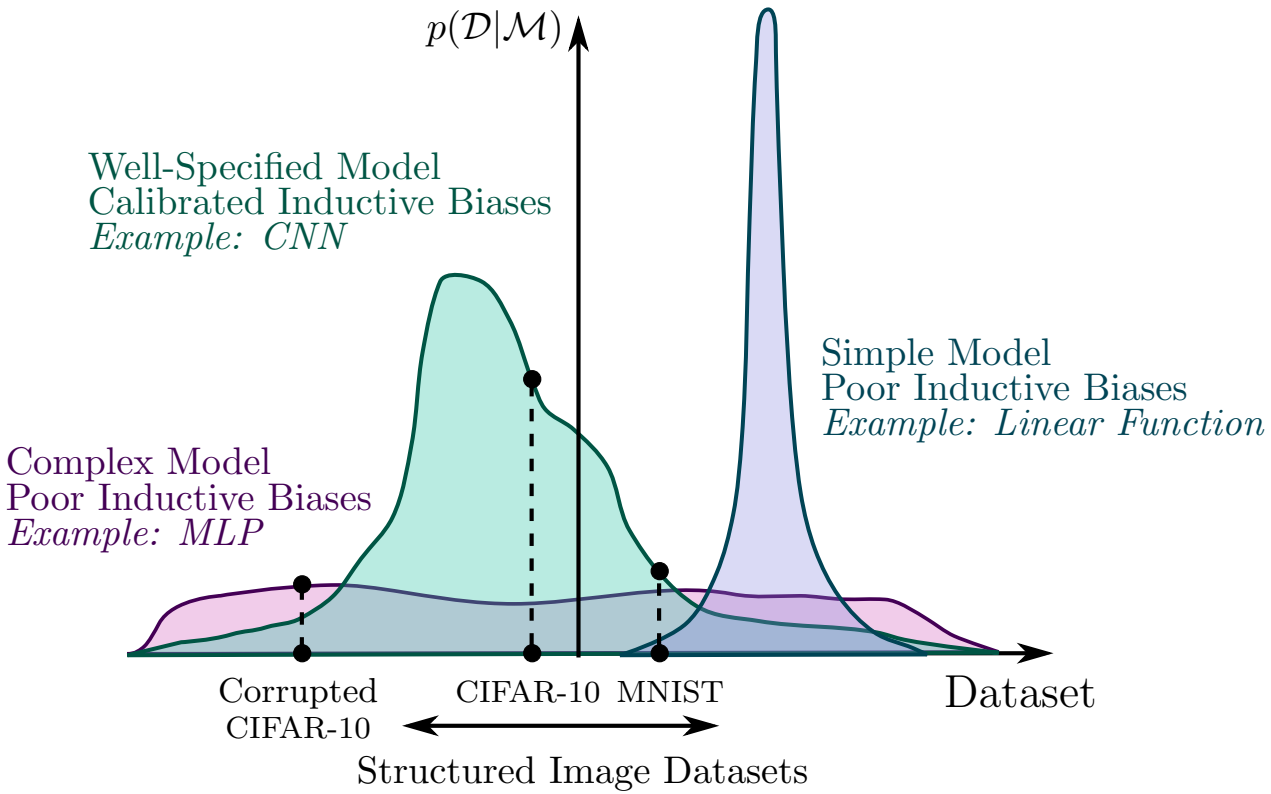
Consider the simple linear model,

$$f(x) = w_0 + w_1 x\,, \tag{1}$$

$$w_0, w_1 \sim \mathcal{N}(0, 1)\,. \tag{2}$$

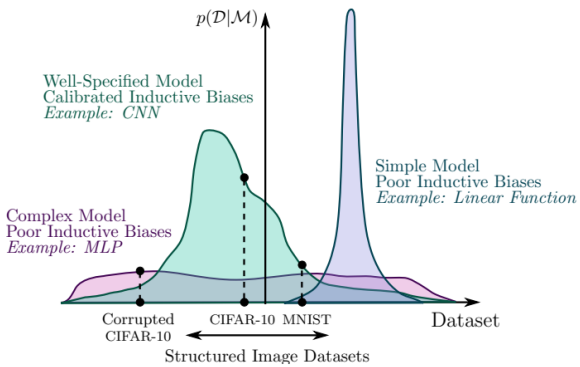


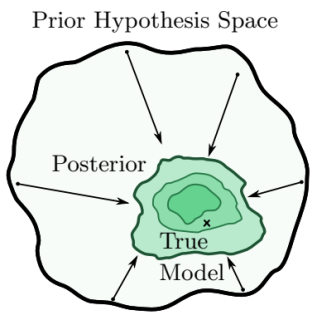
Model Construction and Generalization

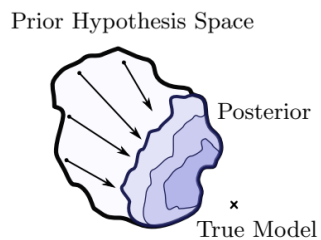


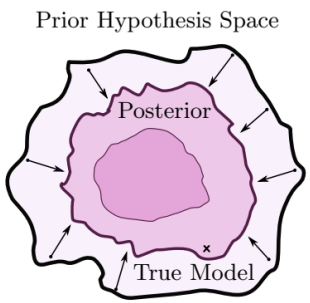
How do we learn?

- The ability for a system to learn is determined by its *support* (which solutions are a priori possible) and *inductive biases* (which solutions are a priori likely).
- We should not conflate *flexibility* and *complexity*.
- An influx of new *massive* datasets provide great opportunities to automatically learn rich statistical structure, leading to new scientific discoveries.





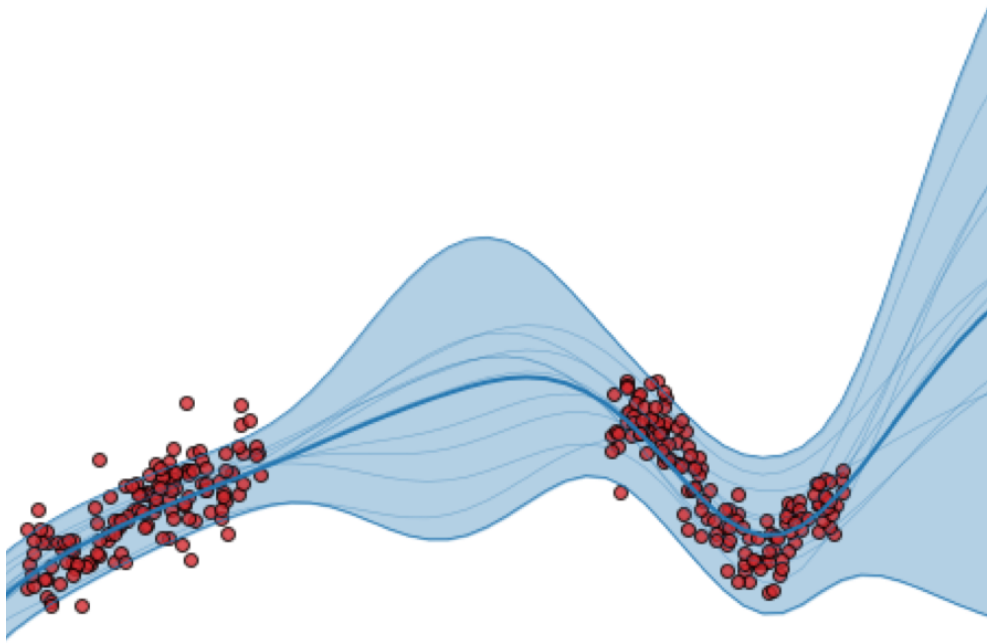




Bayesian Deep Learning and a Probabilistic Perspective of Generalization
Wilson and Izmailov, NeurIPS 2020

What is Bayesian learning?

- The key distinguishing property of a Bayesian approach is **marginalization** instead of optimization.
- Rather than use a single setting of parameters **w**, use all settings weighted by their posterior probabilities in a *Bayesian model average*.



Bayesian Marginalization

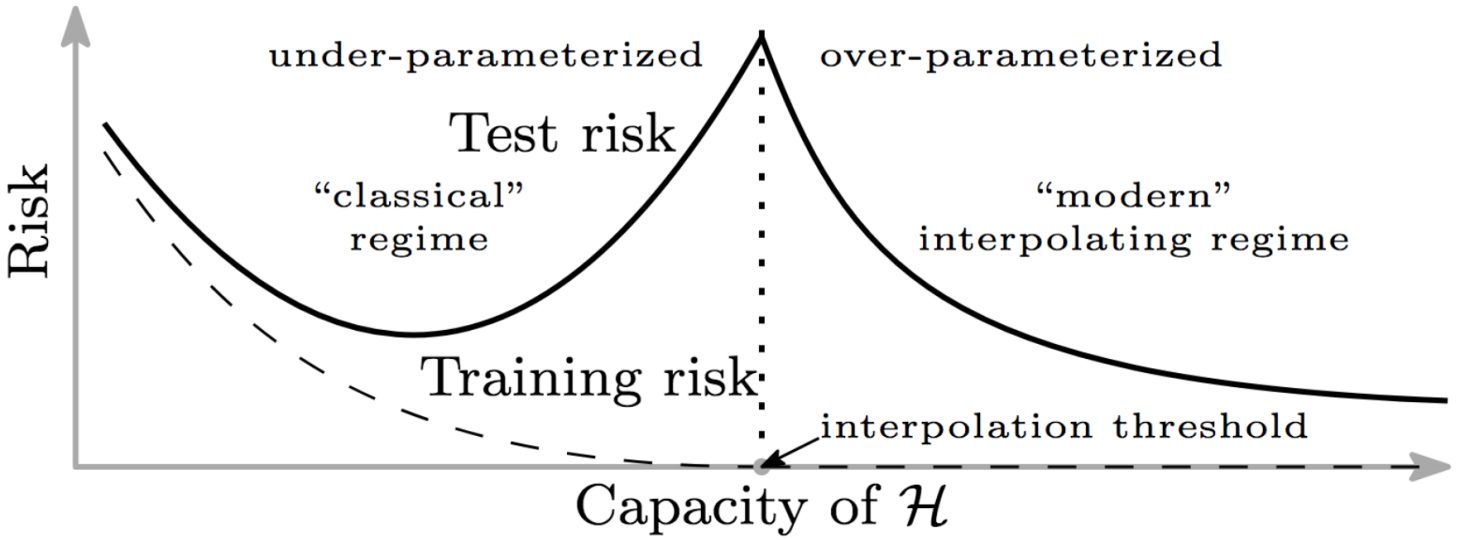
Sum rule: $p(x) = \sum_x p(x, y)$. **Product rule:** $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$.

$$p(y|x_*, \mathbf{y}, X) = \int p(y|x_*, \mathbf{w})p(\mathbf{w}|\mathbf{y}, X)d\mathbf{w}\,. \tag{3}$$

- Think of each setting of $\mathbf{w}$ as a different model. Eq. (3) is a *Bayesian model average*, an average of infinitely many models weighted by their posterior probabilities.
- Automatically calibrated complexity even with highly flexible models.
- Can view classical training as using an approximate posterior $q(\mathbf{w}|\mathbf{y}, X) = \delta(w = w_{\text{MAP}})$.
- Typically more interested in the induced distribution over **functions** than in parameters $\mathbf{w}$. Can be hard to have intuitions for priors on $p(\mathbf{w})$.

Can We Understand Deep Learning With Probability?

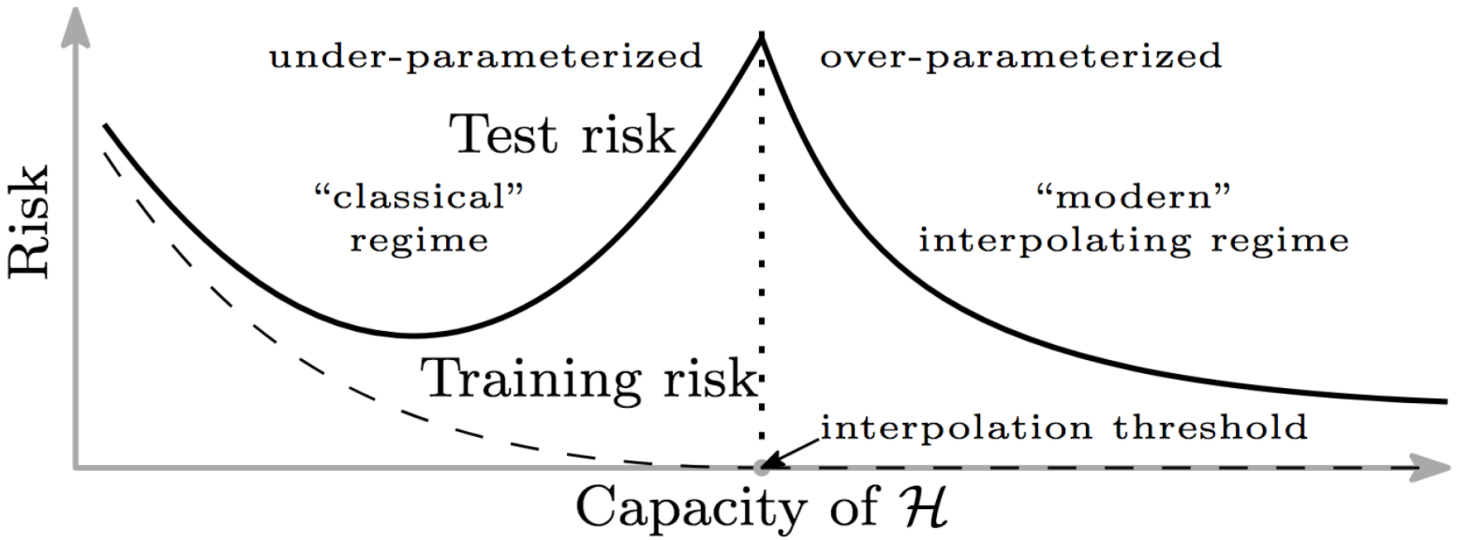
Double Descent



Belkin et. al (2018)

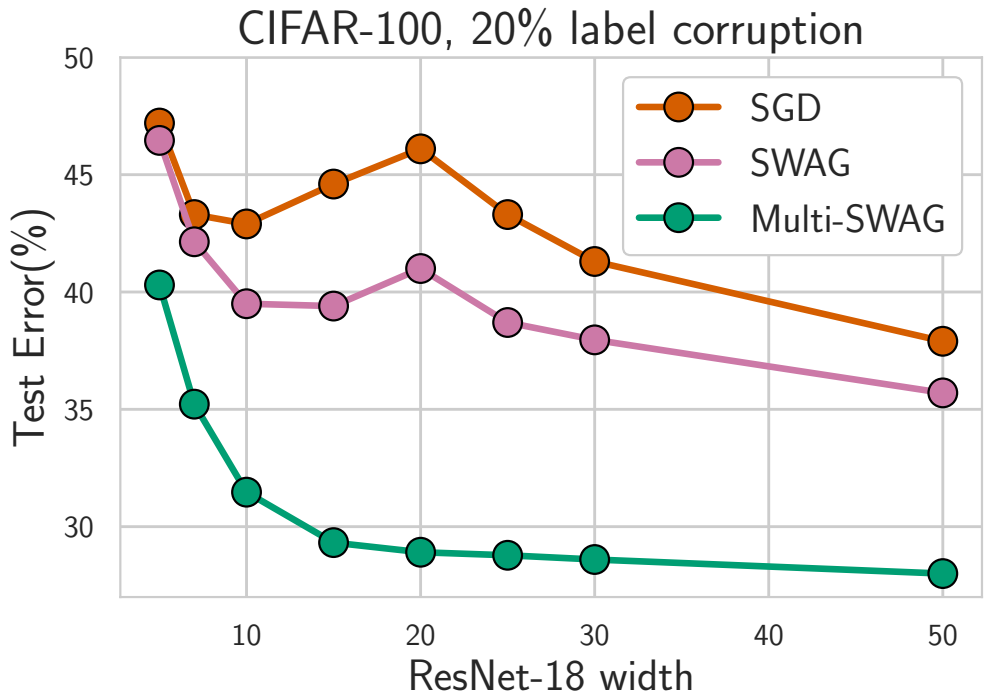
Reconciling modern machine learning practice and the bias-variance trade-off. Belkin et. al, 2018

Double Descent

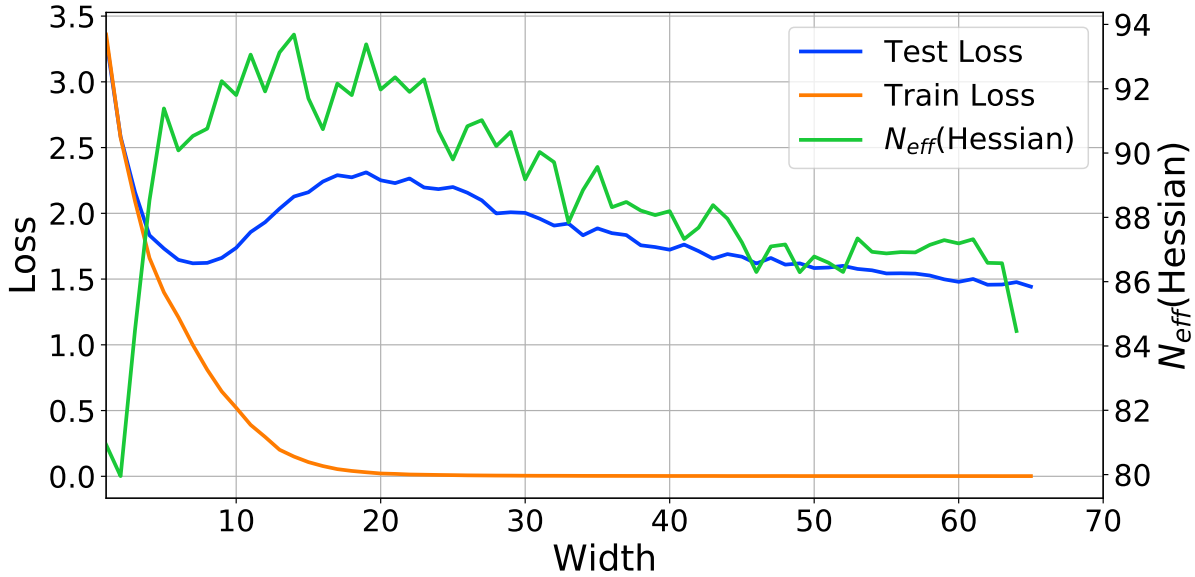


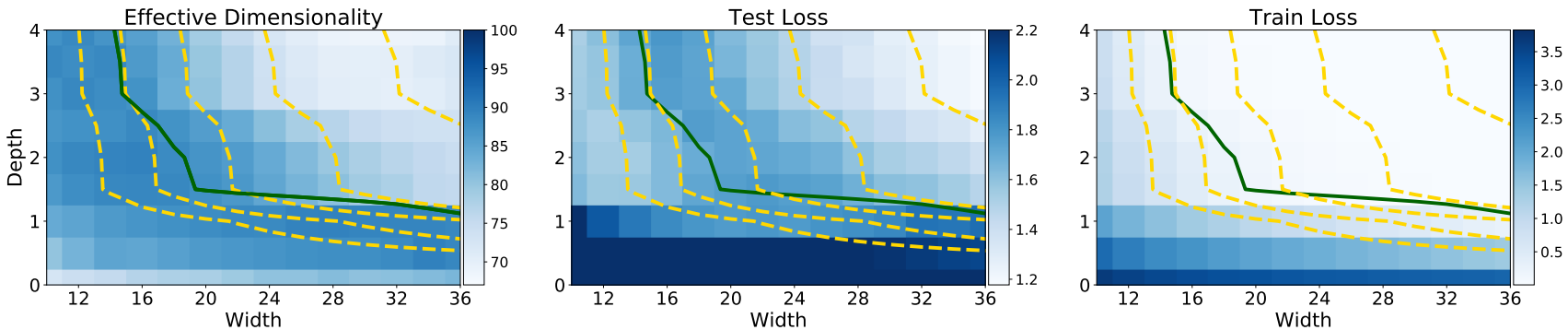
Should a Bayesian model experience double descent?

Bayesian Model Averaging Alleviates Double Descent



Double Descent Explained by Occam's Razor





$$N_{\text{eff}}(H) = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}$$

Rethinking Parameter Counting in Deep Models: Effective Dimensionality Revisited.
W. Maddox, G. Benton, A.G. Wilson, 2020.

Gaussian processes: a function space view

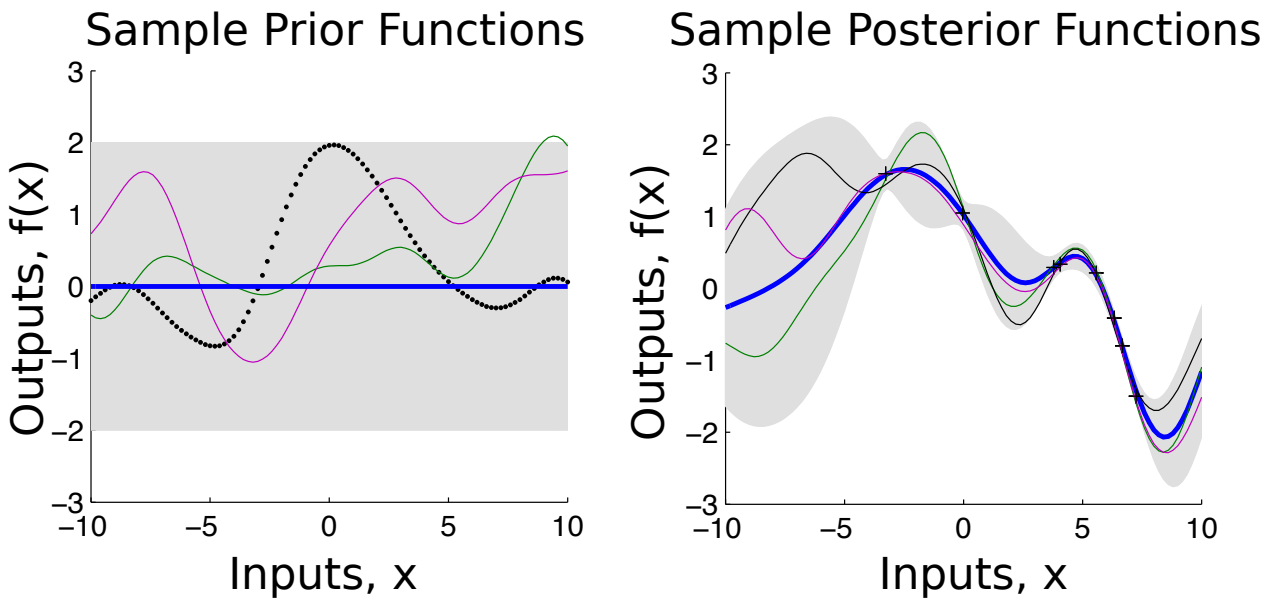
Definition

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. **Gaussian processes assign priors directly in function-space**.

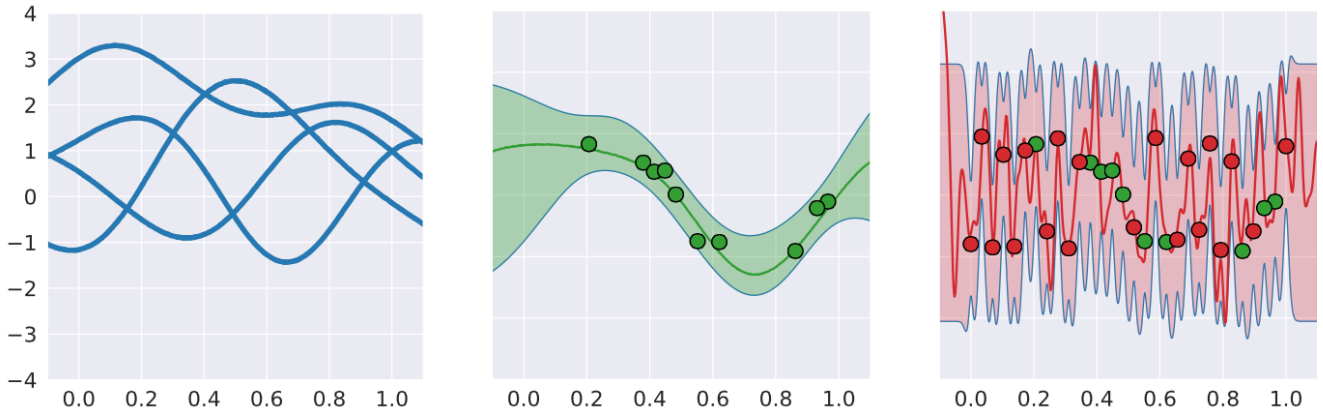
Nonparametric Regression Model

- Prior: $f(x) \sim \mathcal{GP}(m(x), k(x,x'))$, meaning $(f(x_1), \ldots, f(x_N)) \sim \mathcal{N}(\boldsymbol{\mu}, K)$, with $\boldsymbol{\mu}_i = m(x_i)$ and $K_{ij} = \mathrm{cov}(f(x_i), f(x_j)) = k(x_i, x_j)$.

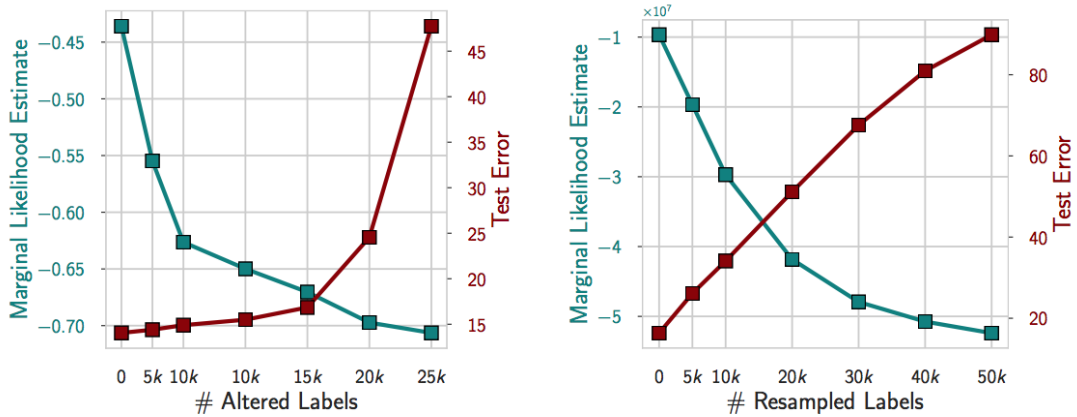
$$\overbrace{p(f(x)|\mathcal{D})}^{\text{GP posterior}} \propto \overbrace{p(\mathcal{D}|f(x))}^{\text{Likelihood}} \overbrace{p(f(x))}^{\text{GP prior}}$$



Rethinking Generalization



(a) Prior Draws (b) True Labels (c) Corrupted Labels



(d) Gaussian Process (e) PreResNet-20

Bayesian Deep Learning and a Probabilistic Perspective of Generalization.
A.G. Wilson, P. Izmailov, NeurIPS 2020

Can we build generalist models?

Can we actually build "AGI"? Models that are simultaneously good on many real-world problems?

The no free lunch theorems are sometimes used to argue that we can't.

The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning
Micah Goldblum, Marc Finzi, Keefer Rowan, Andrew Gordon Wilson
March 2023
arXiv:2304.05366
Appearing at ICML 2024

No Free Lunch Theorems

- Every learner is equally good in expectation over all datasets sampled uniformly (Wolpert 1996; Wolpert & Macready 1997).
- No single learner can achieve high accuracy on every problem (Shalev-Shwartz & Ben-David, 2014).
- Many others.

Suggests we may need to build highly specialized learners for particular tasks...

Do the no free lunch theorems preclude AGI?

A Polymath Model

In practice, we see the opposite trend...

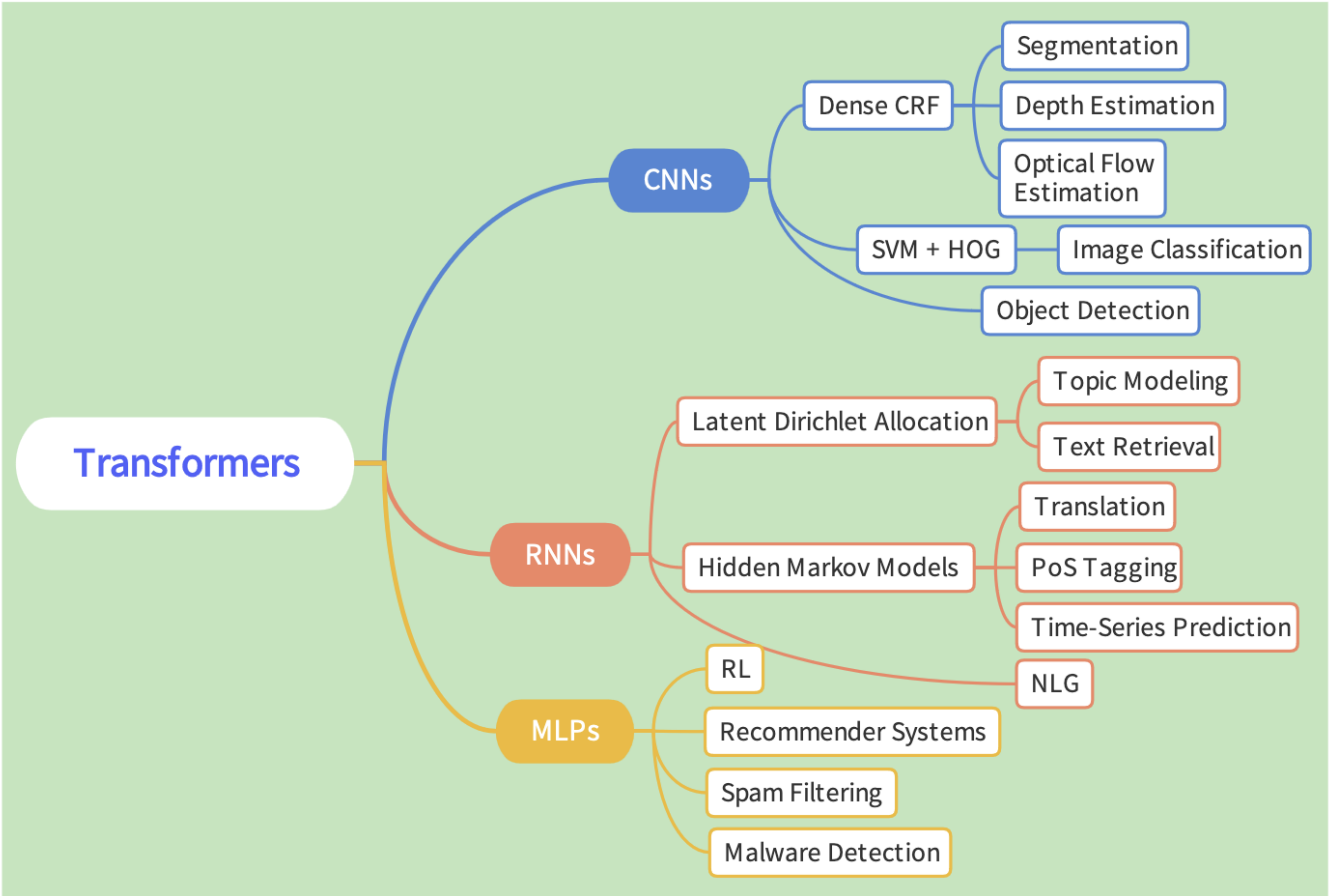


Figure 1: Over time, tasks that were performed by domain-specialized ML systems are increasingly performed by unified neural network architectures.

How can we have generalist models?

- Naturally occurring problems could involve highly structured data.
- Aspects of this structure could be largely shared across problems.
- We can explore the alignment between structure in real-world data and machine learning models through the lens of *Kolmogorov complexity*.
- **Our contention: a single low-complexity biased prior can suffice on a wide variety of problems due to the low Kolmogorov complexity of data.**

Kolmogorov Complexity and Generalization Bounds

- $K(y|x)$: the length of the shortest program (in bits) that inputs x and outputs y.
- Consider the universal prior that assigns higher probability to compressible hypotheses: $P(h) = 2^{-K_p(h)}/Z$ where $K_p(h) \leq K(h) + 2\log_2 K(h)$.

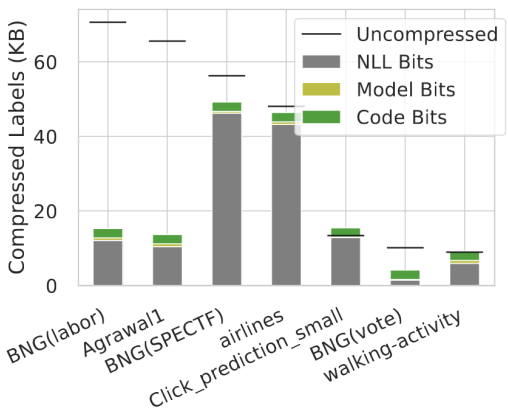
$$R(h) \leq \hat{R}(h) + \sqrt{\frac{K_p(h)\log 2 + \log 1/\delta}{2n}}. \tag{4}$$

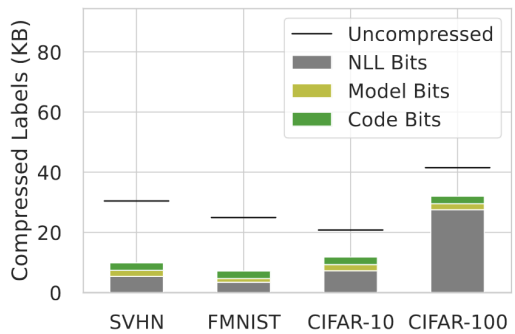
Even under an arbitrarily large hypothesis space, generalization is possible if we assign prior mass disproportionately to the highly structured data that typically occurs.

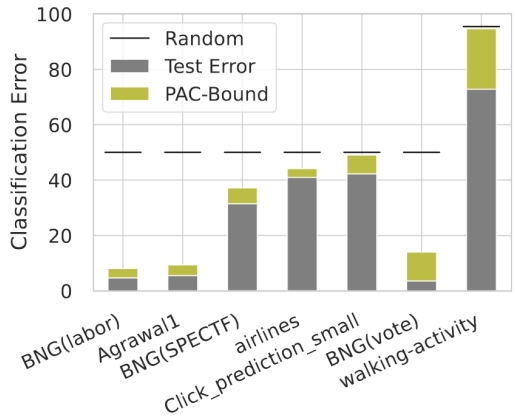
Key Findings

1. Models can significantly compress the labels of data.
2. CNNs provably generalize well on completely different modalities like tabular data due to their simplicity bias!
3. GPT-3 generates low-complexity sequences with exponentially higher probability than complex ones (and bigger models do more compression!)
4. Even randomly initialized GPT models have a low-complexity bias.
5. Worries about overfitting to benchmark test sets is overblown (see paper).
6. We can design models that work well in small and large data regimes, by embracing a flexible hypothesis space combined with a strong simplicity bias.

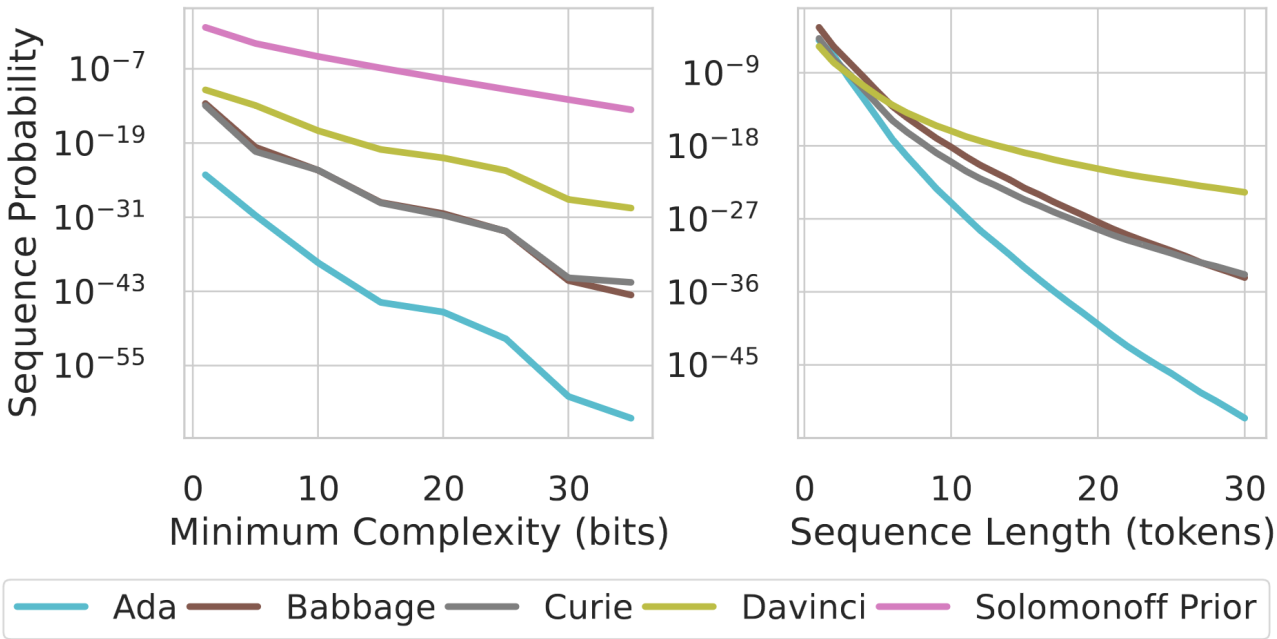
Cross-Domain Generalization Bounds







Simplicity Bias



Good in All Regimes?

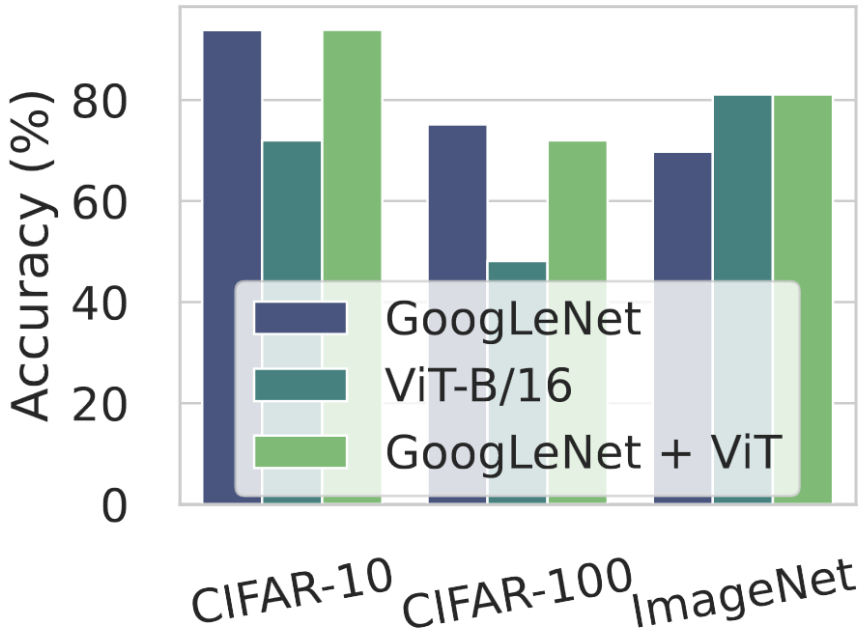
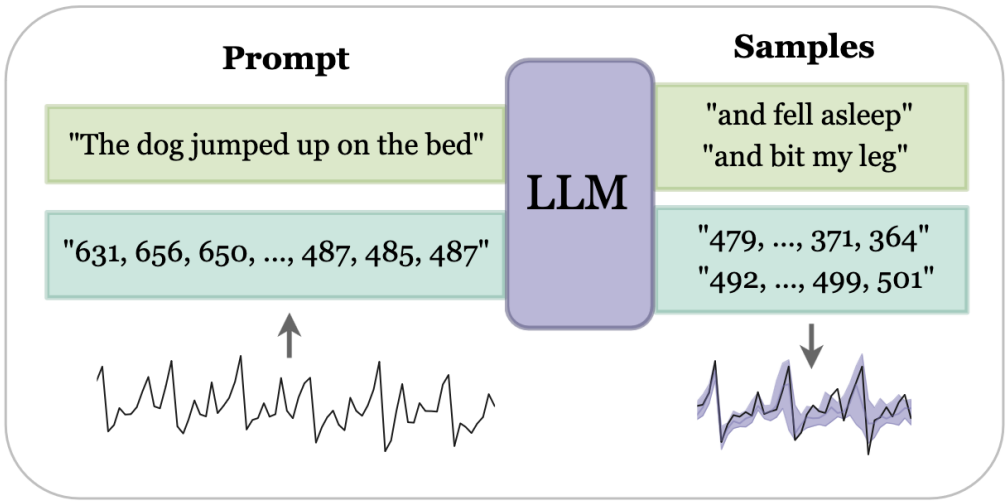
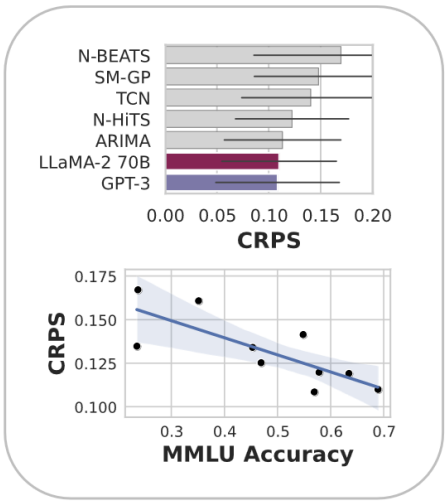


Figure 4: A single learner, which is more expressive than a ViT but also prefers simple solutions representable by a GoogLeNet, can simultaneously solve small and large scale problems.

The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning
Micah Goldblum, Marc Finzi, Keefer Rowan, Andrew Gordon Wilson
March 2023
arXiv:2304.05366
Appearing at ICML 2024

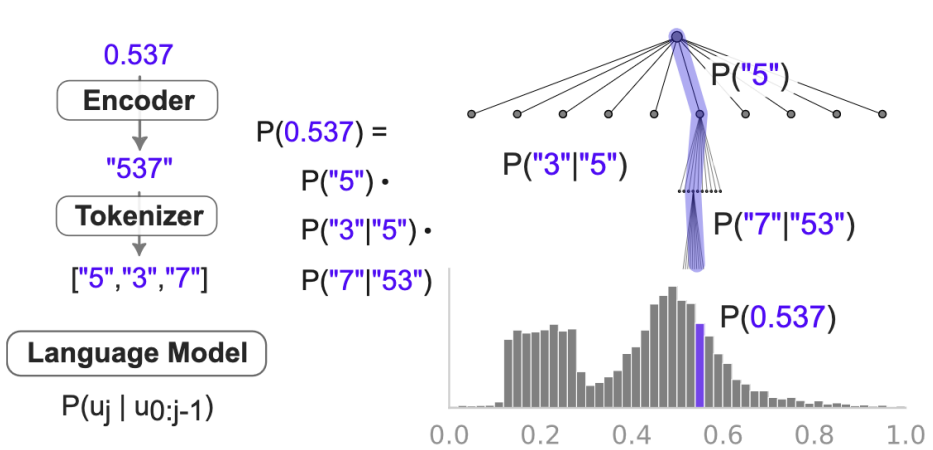
Large Language Models for Time Series Forecasting

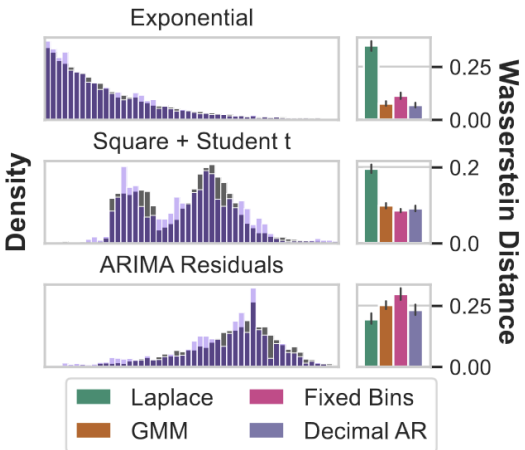




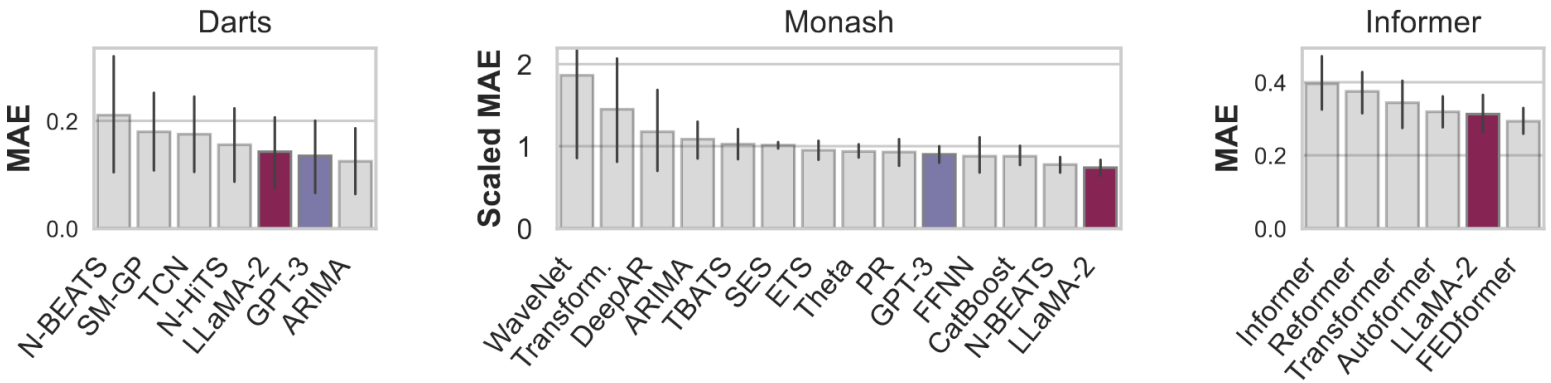
Large Language Models are Zero-Shot Time Series Forecasters
N. Gruver, M. Finzi, S. Qiu, A.G. Wilson
NeurIPS 2023

Uncertainty Representation in LLMTime

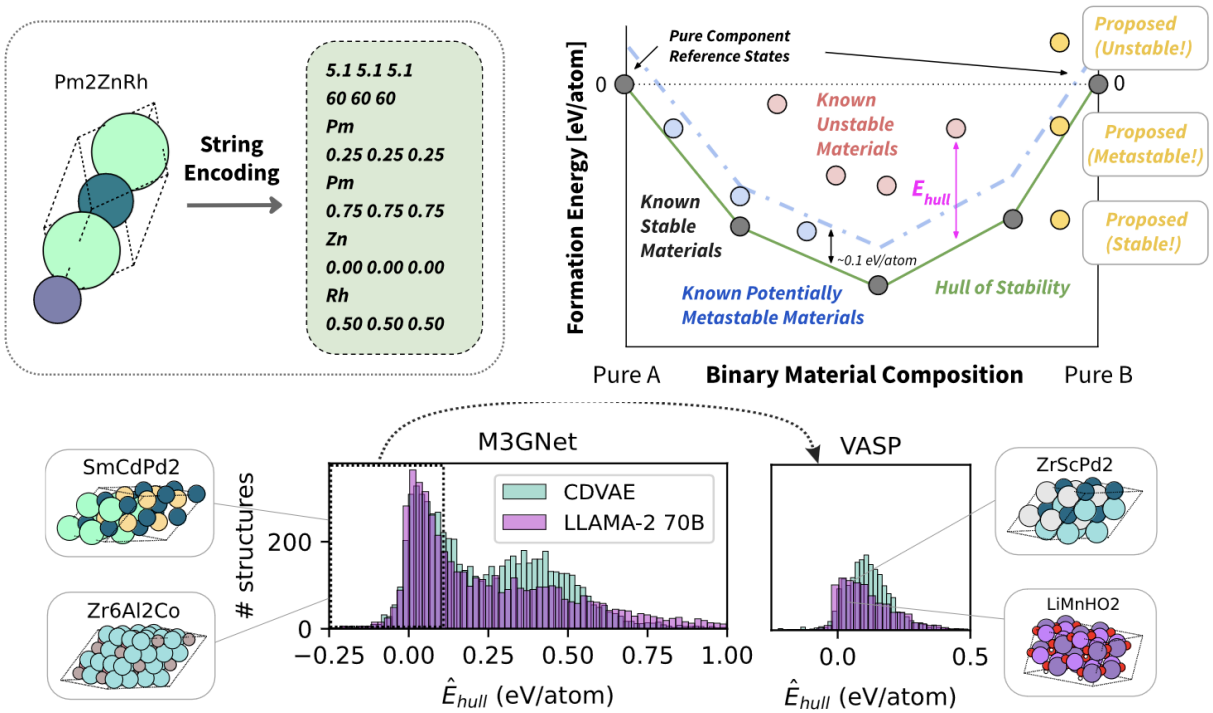




LLMTime Results

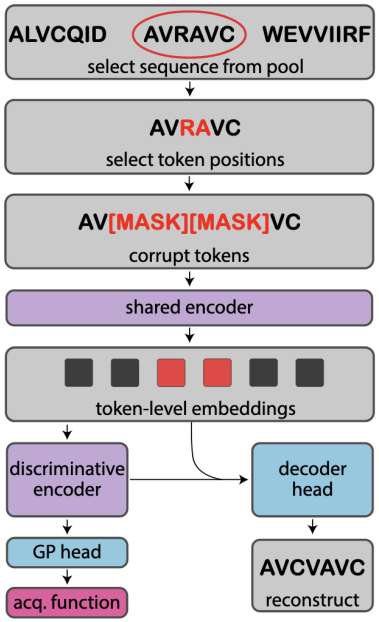


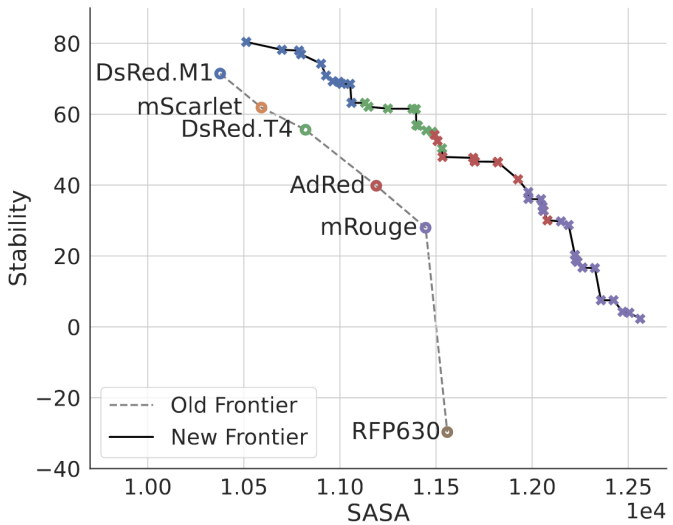
LLMs for Materials



Fine-Tuned Language Models Generate Stable Inorganic Materials as Text.
N. Gruver, A. Sriram, A. Madotto, A. G. Wilson, C. L. Zitnick, Z. Ulissi
ICLR 2014

Bayesian Optimization for Protein Design





- Biological sequence design is a high-dimensional discrete optimization problem, where querying the objective is expensive
- We develop state-of-the-art approaches based on deep kernel learning with de-noising autoencoders and discrete diffusion models.
- Drug discovery is one of the most potentially impactful and up-and-coming applications of machine learning

Accelerating Bayesian Optimization for Biological Sequence Design with Denoising Autoencoders.
S. Stanton, W. Maddox, N. Gruver, P. Mafetonne, E. Delaney, P. Greenside, A.G. Wilson, ICML 2022.

Protein Design with Guided Discrete Diffusion.
N. Gruver, S. Stanton, N. Frey, T. Rudner, I. Hotzel, J. Lafrance-Vanasse, A. Rajpal, K. Cho, A. G. Wilson, NeurIPS 2023.

Numerical Linear Algebra is a Foundation for AI

Exploiting Compositional Structure for Automatic and Efficient Numerical Linear Algebra.
A. Potapczynski, M. Finzi, G. Pleiss, A. G. Wilson, NeurIPS 2023.

Conclusions

- We can use probability theory to develop a *prescriptive understanding* of model construction and generalization, resolving otherwise mysterious behaviour.
- We should not conflate *flexibility* with *complexity*, or do *parameter counting*.
- **Universal learning (general intelligence) in the real world should be possible.**
- Neural networks represent many compelling solutions to a given problem, which is perfect for **Bayesian model averaging**.
- Large language models combine **expressiveness** with a **strong simplicity bias** for effective zero-shot and few-shot performance in many domains, including time-series **forecasting**.

What is >100 years away? Discoveries scientific theories. How do we develop algorithm that will propose a theory like general relativity?