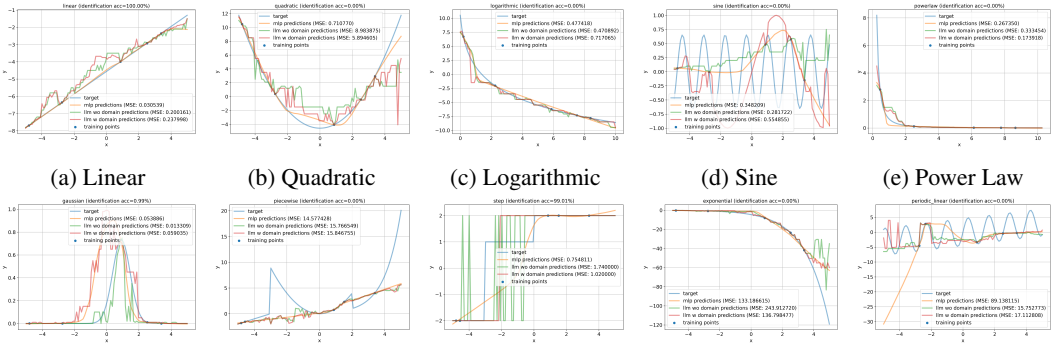
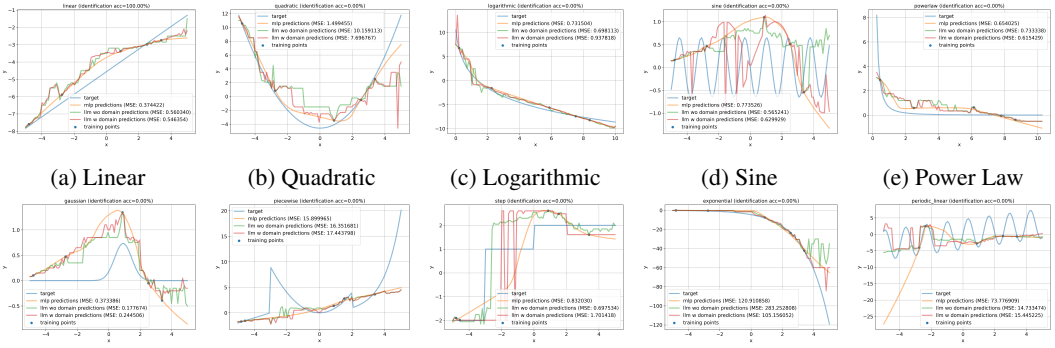


(a) Linear (b) Quadratic (c) Logarithmic (d) Sine (e) Power Law

(f) Gaussian (g) Piecewise (h) Step (i) Exponential (j) Periodic Linear

Figure 6: Basic evaluations of function modeling using 5 training points w/o any noise.



(a) Linear (b) Quadratic (c) Logarithmic (d) Sine (e) Power Law

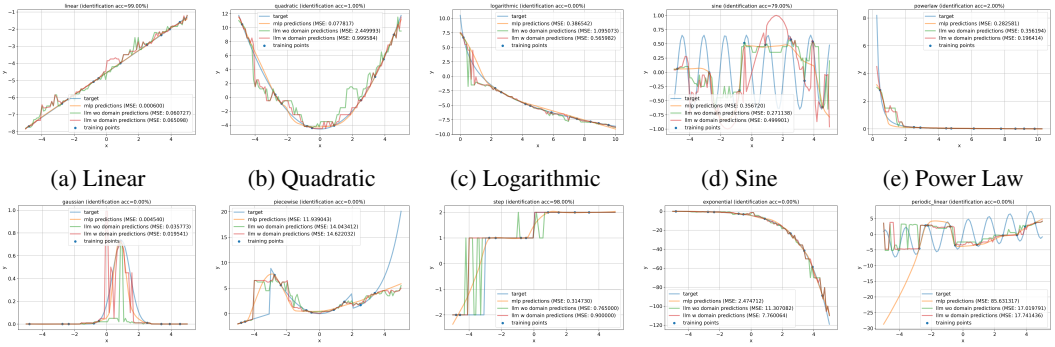
(f) Gaussian (g) Piecewise (h) Step (i) Exponential (j) Periodic Linear

Figure 7: Basic evaluations of function modeling using 5 training points after adding Gaussian noise using standard deviation of 0.5.

A SYNTHETIC FUNCTION MODELING

We visualize the results for synthetic function modeling using just 5 training examples in Fig. 6. The results after additive Gaussian noise with $\sigma = 0.5$ are presented in Fig. 7.

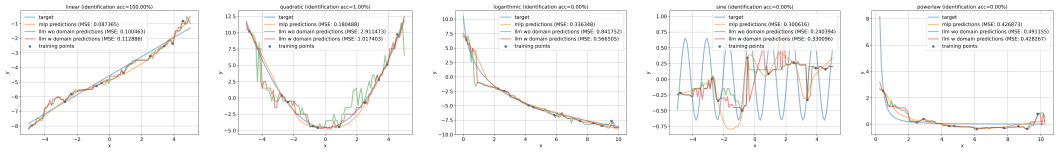
The results with 10 training examples without and with additive Gaussian noise are presented in Fig. 8 and Fig. 9 respectively.



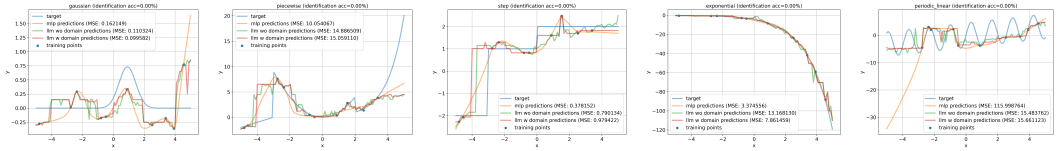
(a) Linear (b) Quadratic (c) Logarithmic (d) Sine (e) Power Law

(f) Gaussian (g) Piecewise (h) Step (i) Exponential (j) Periodic Linear

Figure 8: Basic evaluations of function modeling using 10 training points w/o any noise.

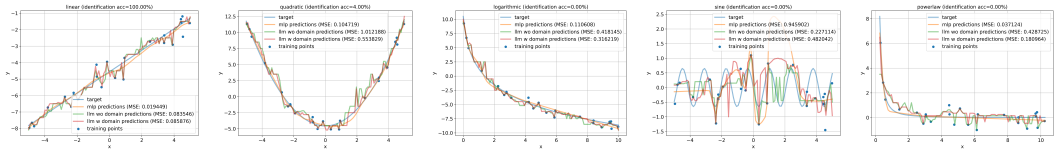


(a) Linear (b) Quadratic (c) Logarithmic (d) Sine (e) Power Law

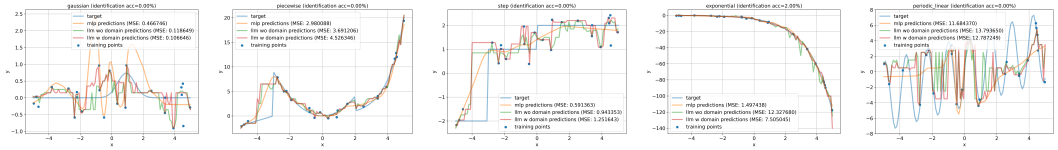


(f) Gaussian (g) Piecewise (h) Step (i) Exponential (j) Periodic Linear

Figure 9: Basic evaluations of function modeling using 10 training points after adding Gaussian noise using standard deviation of 0.5.



(a) Linear (b) Quadratic (c) Logarithmic (d) Sine (e) Power Law



(f) Gaussian (g) Piecewise (h) Step (i) Exponential (j) Periodic Linear

Figure 10: Basic evaluations of function modeling using 25 training points after adding Gaussian noise using standard deviation of 0.5.

Finally, the results with 25 training examples and additive Gaussian noise are presented in Fig. 10 (see Fig. 3 in the main text for the no noise counterpart).

B OVERVIEW OF DATA AND TASKS

Table 7: A summary of the real-world task considered in this work.

	CO_2	**Income**	**Currencies**
Input	Univariate	Multivariate	Multivariate
Output	Univariate	Univariate	Multivariate
Domain	climate science	socio-economics	finance
ML model	GP	NN	GP
Eval. method	{direct prediction, kernel design}	{direct prediction, feature selection}	{kernel design}

C PROMPTS

C.1 SYNTHETIC DATA

Task and setup description
In the following task, I am going to give you some data x, y, and ask you to tell the possible relationship between x and y.

During this process, you need to directly output your answer without resorting to external softwares or libraries. Your answer needs not to be very accurate; however, providing code example or general data analysis guidance will not be acceptable.

Do you understand the setup? You are welcome to ask me any question.

Pattern inference - 1
Below is the data:

x= ...
y= ...

Can you tell me what is the possible relationship between x and y? A rough/intuitive estimate is enough.

Pattern inference - 2
Can you suggest two or three important characteristics of the function behind this data?

Domain inference
What do you think this dataset come from (provided that it is collected in real-world settings)?

C.2 INCOME

Task and setup description
I am going to let you judge if an individual in the US has high income, using some features describing that individual as input. The features are:

['age', 'workclass', 'rep. weight', 'degree', 'marital status', 'occupation', 'relationship', 'ethnicity', 'gender', 'capital net gain', 'hours per day', 'native-country']

Below are some examples:
id:1, features: [38.8909, ' State-gov', 77516, ' Bachelors', 'not married', ' Adm-clerical', ' Not-in-family', ' White', ' Male', 2174, 8.0, ' United-States'] , high income? False
id:2, ...

Do you understand the setup? You are welcome to ask me questions if there is any uncertainty.

In-context learning
Below are some additional samples you can use to learn. When you learn from these samples, please combine what you learn from the data and any domain knowledge you have about this task. The samples are:

id:1, features: [38.8909, ' State-gov', 77516, ' Bachelors', 'not married', ' Adm-clerical', ' Not-in-family', ' White', ' Male', 2174, 8.0, ' United-States'] , high income? False
id:2, ...

Let me know when you already finish learning from these samples. At the same time, feel free to ask for more samples for learning if necessary.

Direct prediction
Now let us proceed to make some predictions. During this process, you are required to directly answer yes or no for each sample. Your answer needs not to be very accurate; an intuitive estimate is enough. However, providing code or general data analysis guidance will not be acceptable.

Below are the 25 samples I would like you to predict:

id:10000, features: [22.4297, ' Private', 176486, ' Some-college', 'not married', ' Other-service', ' Other-relative', ' White', ' Female', 0, 5.0, ' United-States'],
id: 10001, ...

Can you make a reasonable guess on whether these people are of high income or not, combining what you learned from the data above and any domain knowledge?

Feature importance ranking
If you were to rank the importance of each feature using a score between 0 to 1, what would you do? Again, please jointly consider the data provided as well as prior knowledge.

Feature selection
Now instead of giving a score for each feature, can you give the top 5 features whose combination are the most predictive?

Feature selection vs feature importance
I realise there is a difference between the top 5 features selected and the top 5 features with highest scores. Can you explain this difference?

Asking for detailed reason for a specific feature being not selected
(right after the last question)
Is it why 'workclass' is dropped out of the top 5? Can you explain it in more details?

C.3 CO2

Task and setup description
I am going to ask you to predict the CO2 concentration level at some time based on some historical records. These historical data was collected at an observatory.

You should directly answer an estimate of the concentration level rather than resorting to external libraries (e.g. scipy). Also, showing me general guidance or code is unacceptable.

During this process, you are encouraged to take into account both the patterns in the data and the domain knowledge.

Do you understand the setup?

Hinting to combine domain knowledge and data patterns
Here you only need to provide a rough estimate rather than an accurate prediction comparable to a machine learning model. This is just to test your intuition about (a) the patterns in the data and (b) the domain knowledge.

Are you ready? Feel free to ask me any question before I show you the data.

In-context learning
Below are the data collected during the years 1958 to 1975. Here, x is the time of the data collected (where 1958.25 means 0.25 years after the year of 1958 i.e. March 1958. Others figures are similar), and y is the level of atmospheric CO2 concentration measured at the corresponding time.

x= ...
y= ...

Please learn from these data. You can truncate the dataset if you feel it too difficult to learn from the whole dataset. Let me know when you have finished learning.

Asking about general prediction rule
(after in-context learning)

Before doing prediction, let us discuss the patterns you observe from the data in more details. Please name two or three major patterns you have seen from the data, and ideally relates it to some domain knowledge.

Perform direct prediction - predicting the yearly average
Now please make predictions for the following dates respectively:

x = ...

Perform direct prediction - detailed prediction
The estimate you gave above corresponded to the average for the year of 1990. I still want a more detailed prediction for each of the dates mentioned. Can you do this for me, or do you think this detailed prediction too difficult? Again a rough estimate is enough.

Hinting to account for the increasing changing rate
As you said, there may be going up trend for the CO2 concentration level. Can you try to calibrate your prediction using this fact?

Hinting to make more reasonable modelling regarding the changing rate
I realise that you update the increasing rate to 1.4 from 1.2 in 1992. Would you think this change too steep when considering the whole picture?

C.4 CURRENCY EXCHANGE RATES

Task and setup description
I want to test you in modelling functions. Especially, your goal is to formulate a model utilizing Gaussian Process to predict the missing part in international currencies(CAD, JPY, and AUD).

Here are the details of the dataset. 'Exchange rates data set' consists of the daily exchange rate w.r.t. USD of the top ten international currencies (CAD, EUR, JPY, GBP, CHF, AUD, HKD, NZD, KRW, and MXN) and three precious metals (gold, silver, and platinum) in the unknown year.

Unfortunately, some parts in between are missing in the three currencies, CAD, JPY, and AUD. The missing parts are different for each currency. Furthermore, there are also missing parts in three precious metals. The task is to predict these missing parts of three currencies from the given whole dataset.

Before we dive into modelling GP, let's do step by step. I cannot use normal GP directly since there are missing values. Can you suggest a good way to predict accurate missing values by exploiting the given data? You can ask a few questions to me.

Ask Kernels from domain knowledge
As we have a method to deal with such missing values, we need to devise proper kernels. To do so, I want you to incorporate your knowledge of currency w.r.t USD in the unknown year. What do you know about this? And how can you include insight from this domain knowledge into designing kernels?

Ask Kernels from both domain knowledge and data
Here are the real data, CAD, downsampled every 5 (y[::5]). What can you capture from this real data? I want you to incorporate your domain knowledge of currency in an unknown year and the intuition from the real data for designing kernels.

Accounting for the output dependence in function modelling
Remark that what we would like to predict are three currencies (CAD, JPY, AUD) rather than only one. And we would like to make use of all possible data/information we have. Do you have any suggestion about how we can do better for this multivariate output prediction task?

Hinting to use autoregressive modelling
If I believe that multi-output GP is too weak in modeling the correlation/dependence between the outputs, what should we do?

D HUMAN EXPERT KERNEL DESIGN ON THE CO_2 DATASET

In this section, we showcase the kernel formulation crafted by human experts. Their design encompasses three distinct types of kernels:

1. A combination of RBF and exponential sine squared kernels tailored to capture seasonality.
2. The rational quadratic kernel employed to grasp irregularities.
3. An additional combination of RBF and white noise kernels aimed at modeling noise.

```
from sklearn.gaussian_process.kernels import RBF
from sklearn.gaussian_process.kernels import ExpSineSquared
from sklearn.gaussian_process.kernels import RationalQuadratic
from sklearn.gaussian_process.kernels import WhiteKernel

long_term_trend_kernel = 50.0**2 * RBF(length_scale=50.0)

```

```
seasonal_kernel = (
    2.0**2
    * RBF(length_scale=100.0)
    * ExpSineSquared(length_scale=1.0, periodicity=1.0,
    periodicity_bounds="fixed")
)

irregularities_kernel = 0.5**2 * RationalQuadratic(length_scale=1.0,
    alpha=1.0)

noise_kernel = 0.1**2 * RBF(length_scale=0.1) + WhiteKernel(
    noise_level=0.1**2, noise_level_bounds=(1e-5, 1e5)
)

co2_kernel = (
    long_term_trend_kernel + seasonal_kernel + irregularities_kernel +
    noise_kernel
)
```

Listing 1: Python code for human expert kenel design on CO_2 data