Occam's Machete

Data-driven *discovery* with parsimony and causal invariance

Dion Häfner

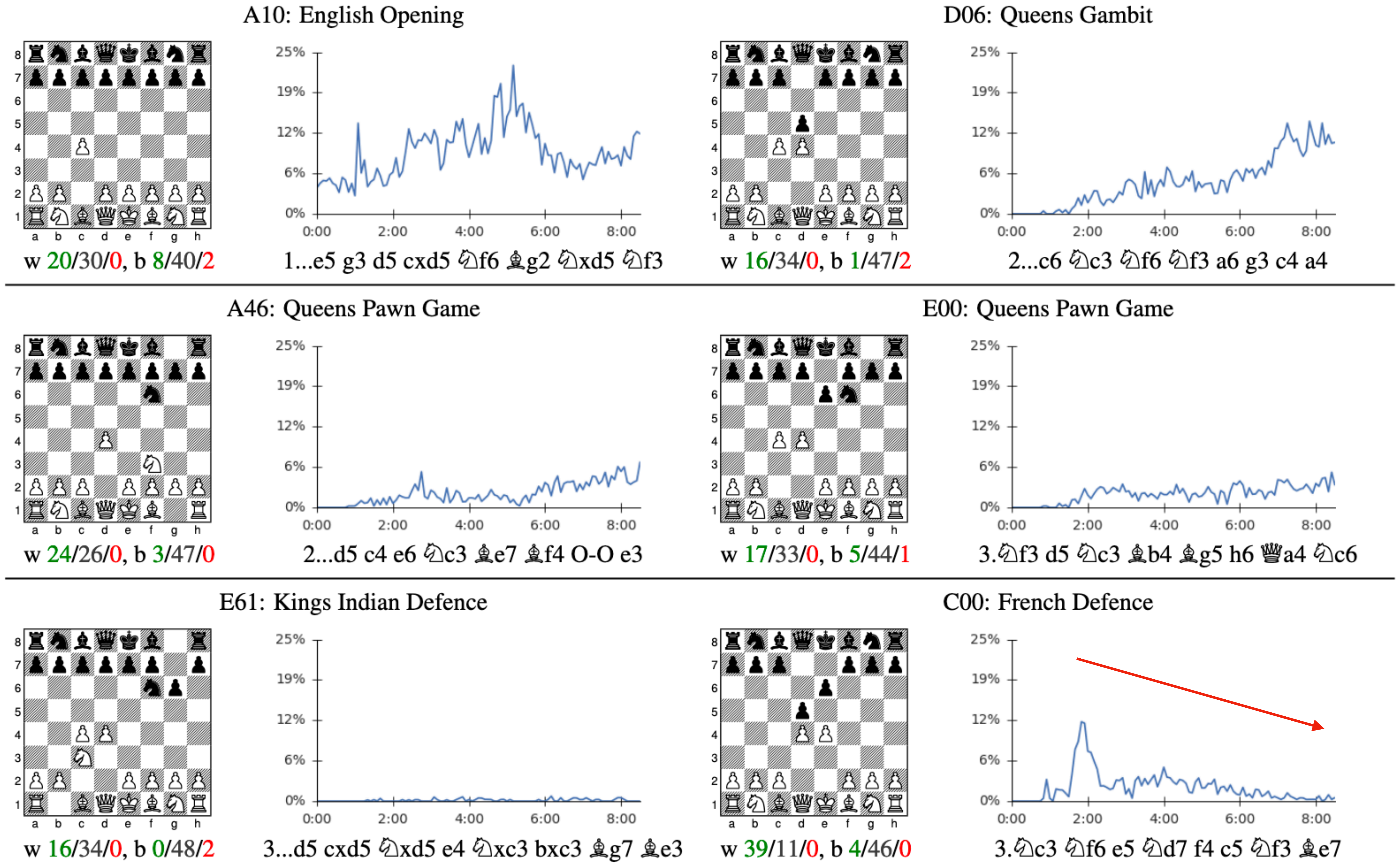
dion.haefner@simulation.science

(1) Pasteur Labs & Institute for Simulation Intelligence

(2) Niels Bohr Institute, University of Copenhagen

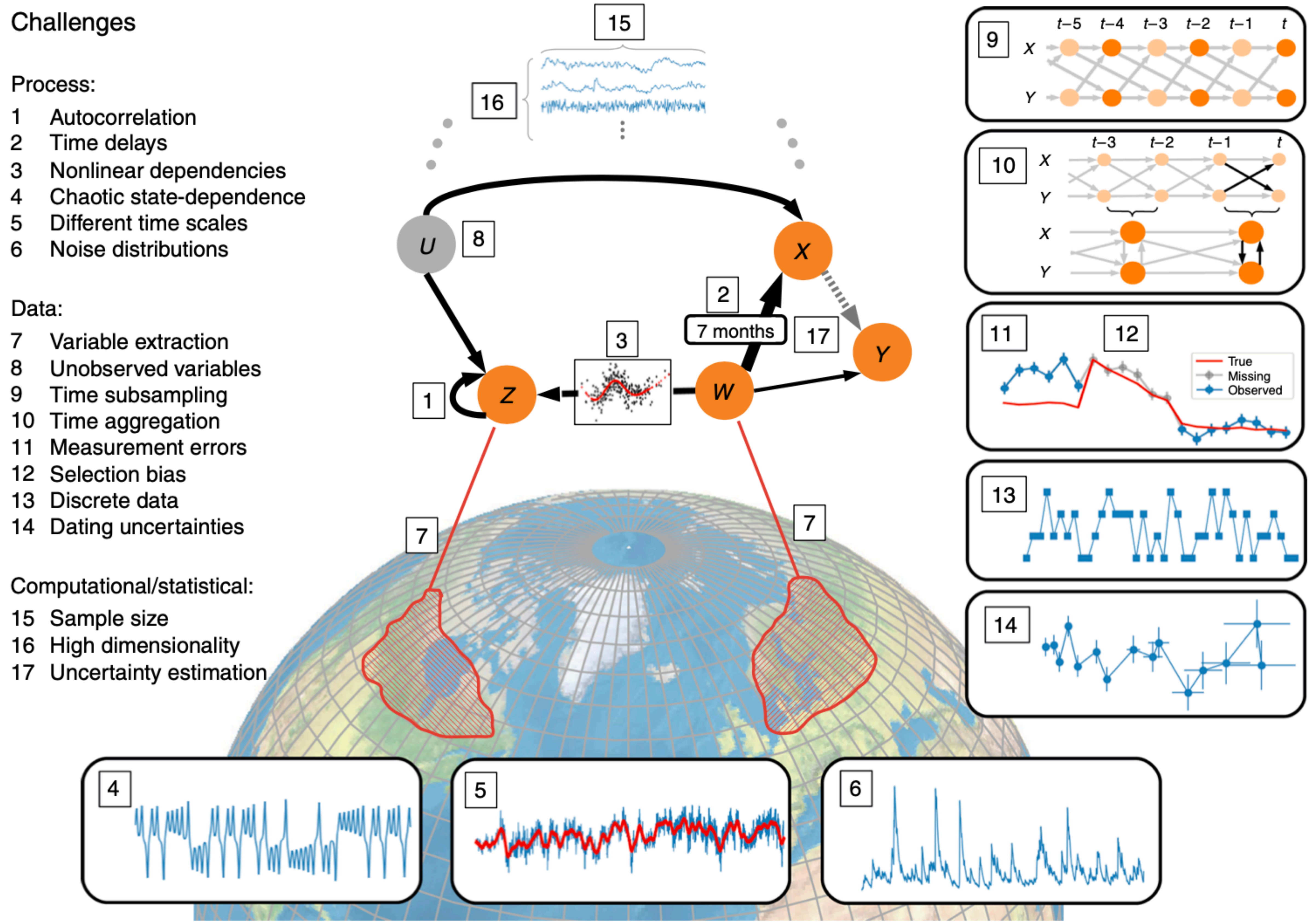
The explainability crisis

AlphaZero openings played over training time



And it gets worse

Real-world data is infinitely more difficult



[Runge et al., 2019]

Machine learning applications are often at odds with the #1 goal of science:

Machine learning applications are often at odds with the #1 goal of science:

DISCOVERY

A different guiding principle

parsimony
/ˈpɑːsɪməni/

noun
noun: **parsimony**
1. extreme unwillingness to spend money or use resources.

Fundamental in nature

$$\delta \int_{t_1}^{t_2} L(\mathbf{q}, \dot{\mathbf{q}}, t)\, \mathrm{d}t = 0$$

Principle of least action
→ Lagrangian mechanics

Pillar of the scientific method

Occam's razor

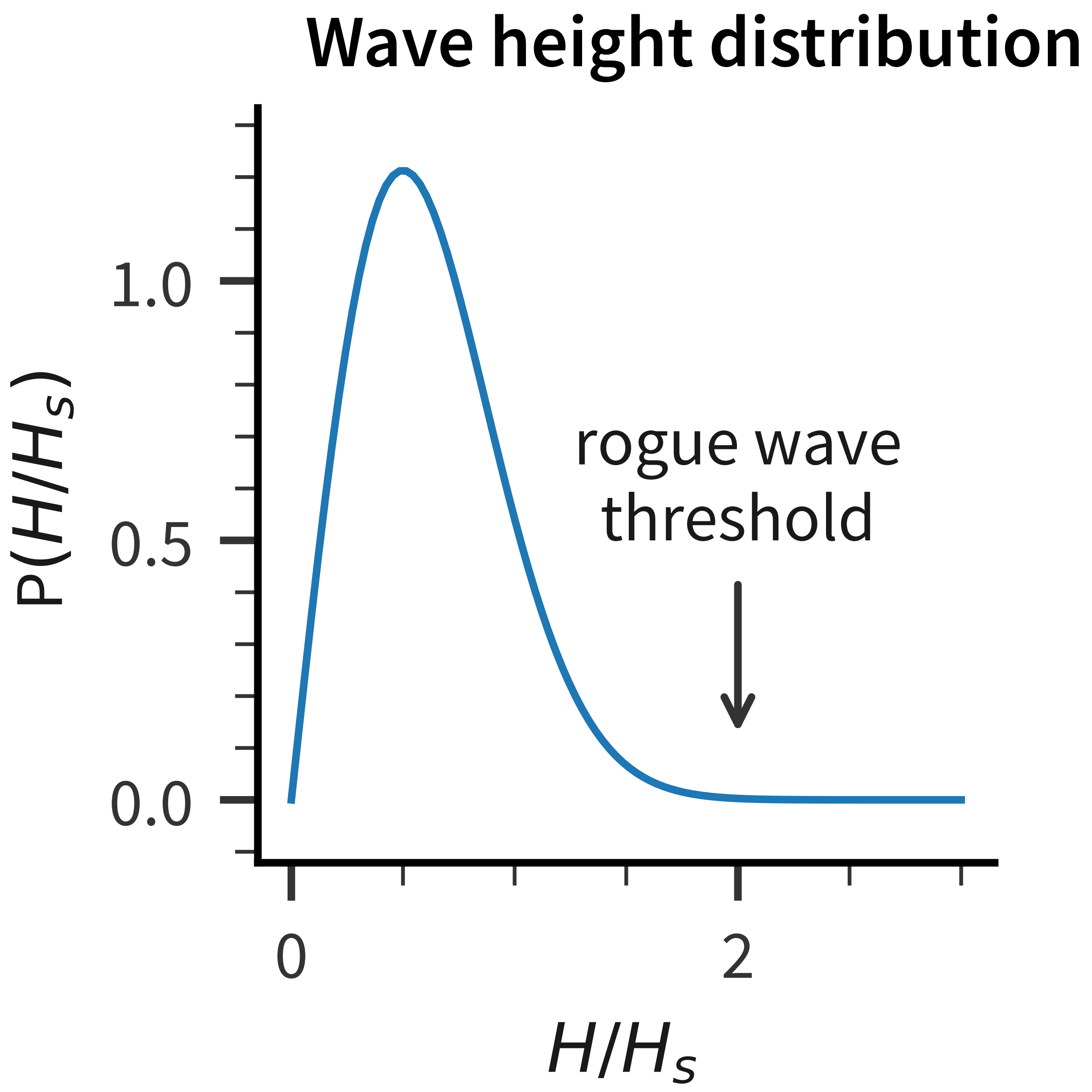
Occam's razor, Ockham's razor, or Ocham's razor, also known as the principle of *parsimony* or the law of parsimony, is the problem-solving principle that "entities should not be multiplied beyond necessity".

Let’s reboot

Can we discover something from data with parsimony-guided machine learning?

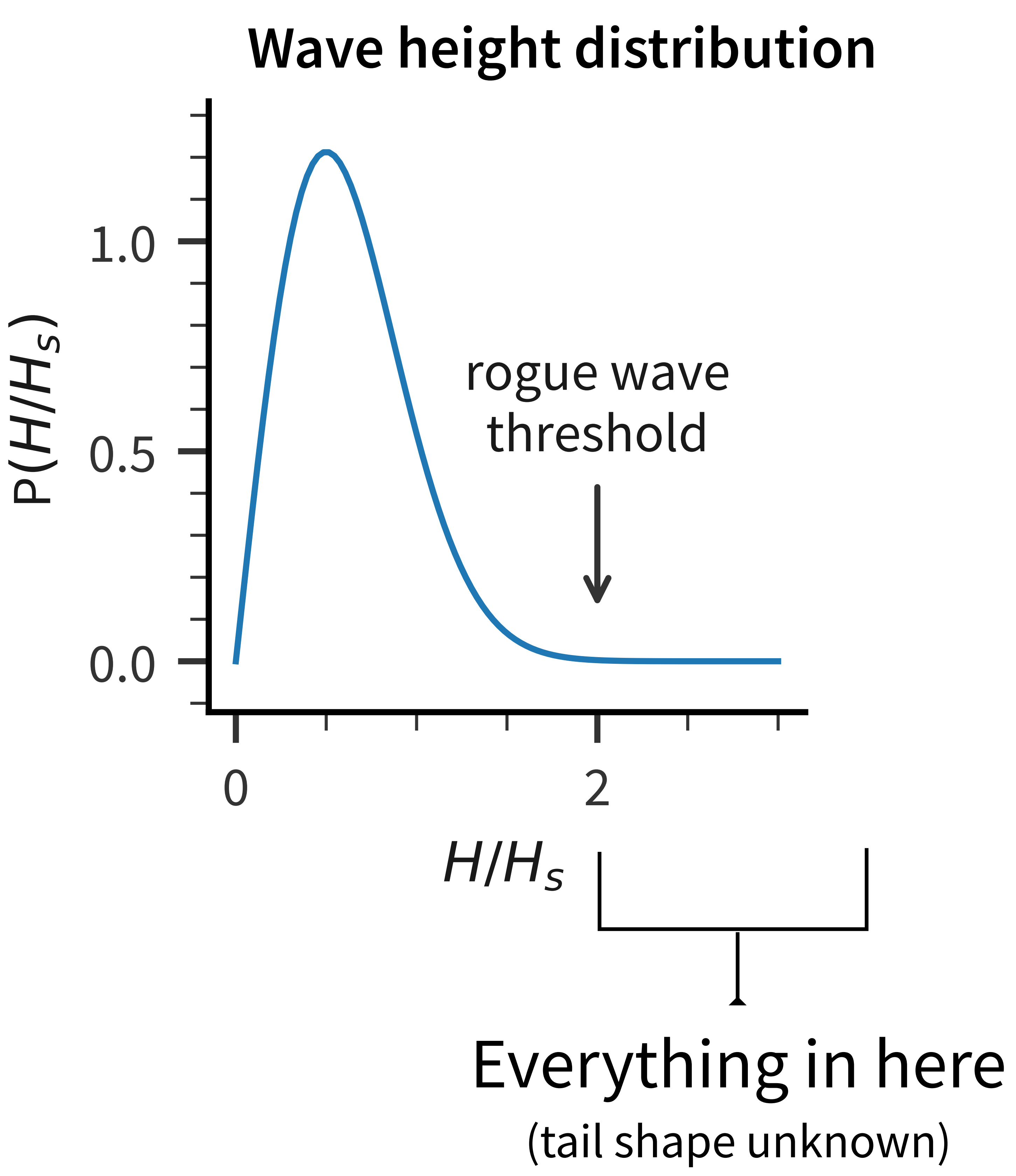
Rogue waves

Definition



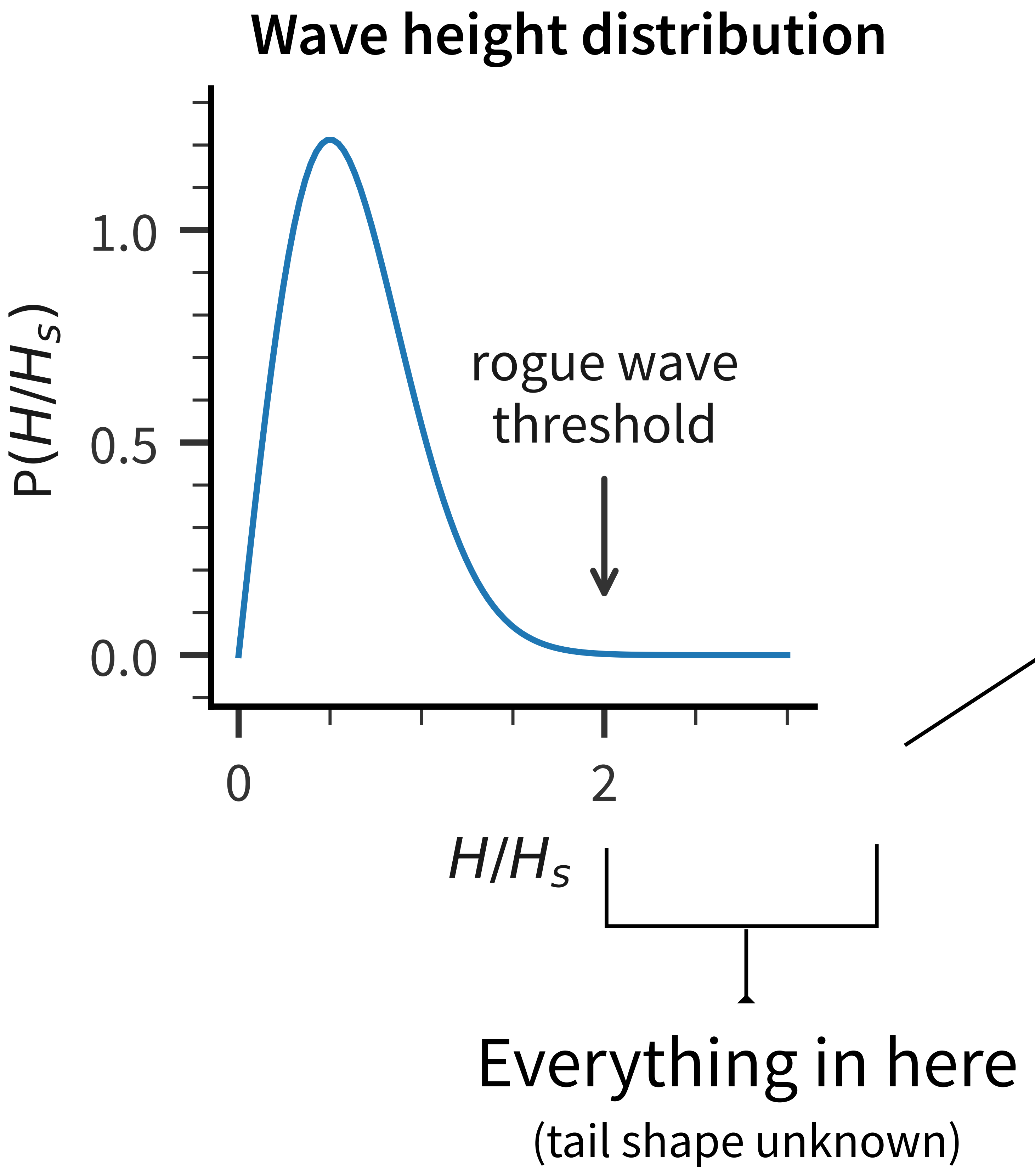
Rogue waves

Definition



Rogue waves

Definition



Binary classification

Find $P(H/H_s > 2 \mid \mathbf{x})$

Forecastable sea state parameters

Current wave theory is messy

Occurrence probability depending on sea state

Linear (bandwidth) effects

$$P(H/H_s > h) = \sqrt{\frac{1+r}{2r}}\left(1 + \frac{1-r^2}{64rh^2}\right)\exp\left(-\frac{4}{1+r}h^2\right)$$

Non-linear effects on envelope

$$p(h) = 4he^{-2h^2}\left\{1 + C_4\left(2h^4 - 4h^2 + 1\right) + C_3^2\left(4h^6 - 18h^4 + 18h^2 - 3\right)\right\}$$

depend on R, kD, eps, …

+ other effects unaccounted for by current theory

Goal

Find approximately causal predictive model

So we can…

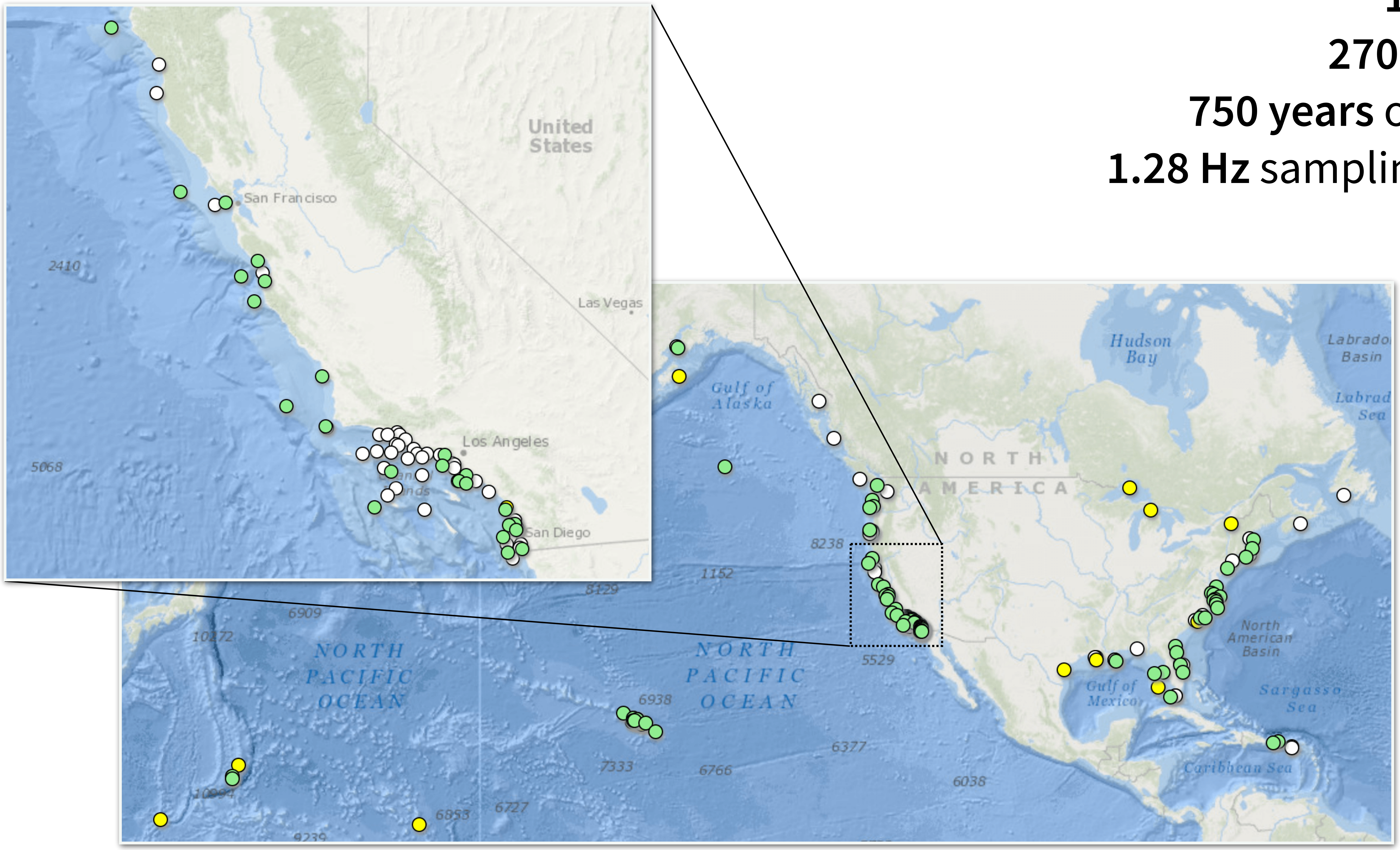
(I) Understand the generation mechanisms of real-world rogue waves

(II) Provide a better forecast

An ocean of data

Observations from CDIP buoys

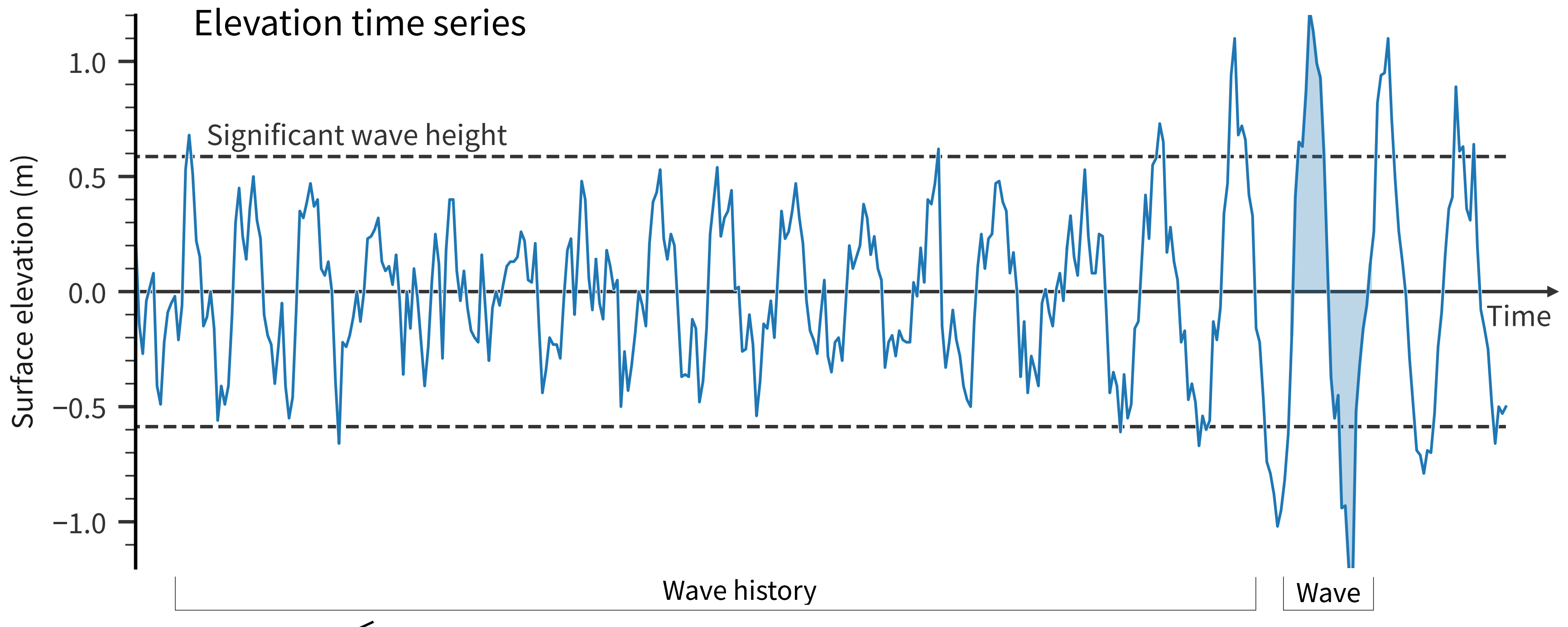
158 locations
270 GB raw data
750 years of time series
1.28 Hz sampling frequency

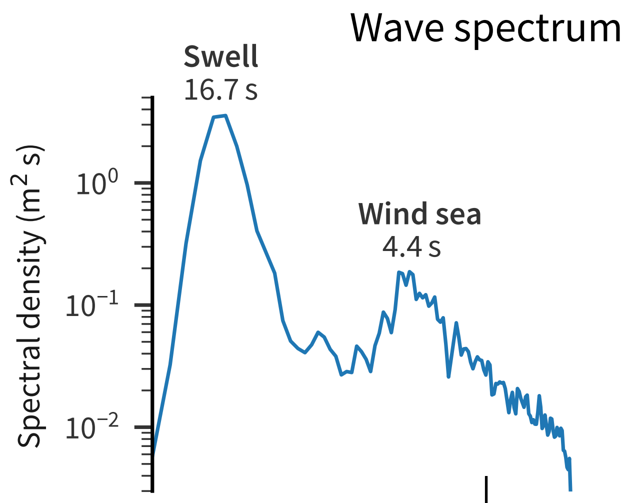


FOWD

The free ocean wave dataset

[Häfner et al., 2021]





labels

features

Sea state parameters


```
{
    "start_time": 0.0,
    "end_time": 1829.6875,
    "significant_wave_height_direct": 1.0925,
    "significant_wave_height_spectral": 1.1734,
    "mean_period_direct": 7.7227,
    "mean_period_spectral": 6.3043,
    "maximum_wave_height": 2.12,
    "rel_maximum_wave_height": 1.8067,
    "skewness": 0.0495,
    "kurtosis": 0.2568,
    "valid_data_ratio": 1.0,
    "peak_wave_period": 15.8922,
    "peak_wavelength": 393.0131,
    "steepness": 0.0066,
    "bandwidth_peakedness": 0.1962,
    "bandwidth_narrowness": 0.905,
    "benjamin_feir_index_peakedness": 0.0254,
    "benjamin_feir_index_narrowness": 0.0055,
    "crest_trough_correlation": 0.699,
    "energy_in_frequency_interval": [
        10.4343,
        633.7447,
        120.9921,
        99.3144,
        244.8272
    ],
    "rel_energy_in_frequency_interval": [
        0.0121,
        0.7331,
        0.14,
        0.1149,
        0.2832
    ]
}
```


Wave parameters

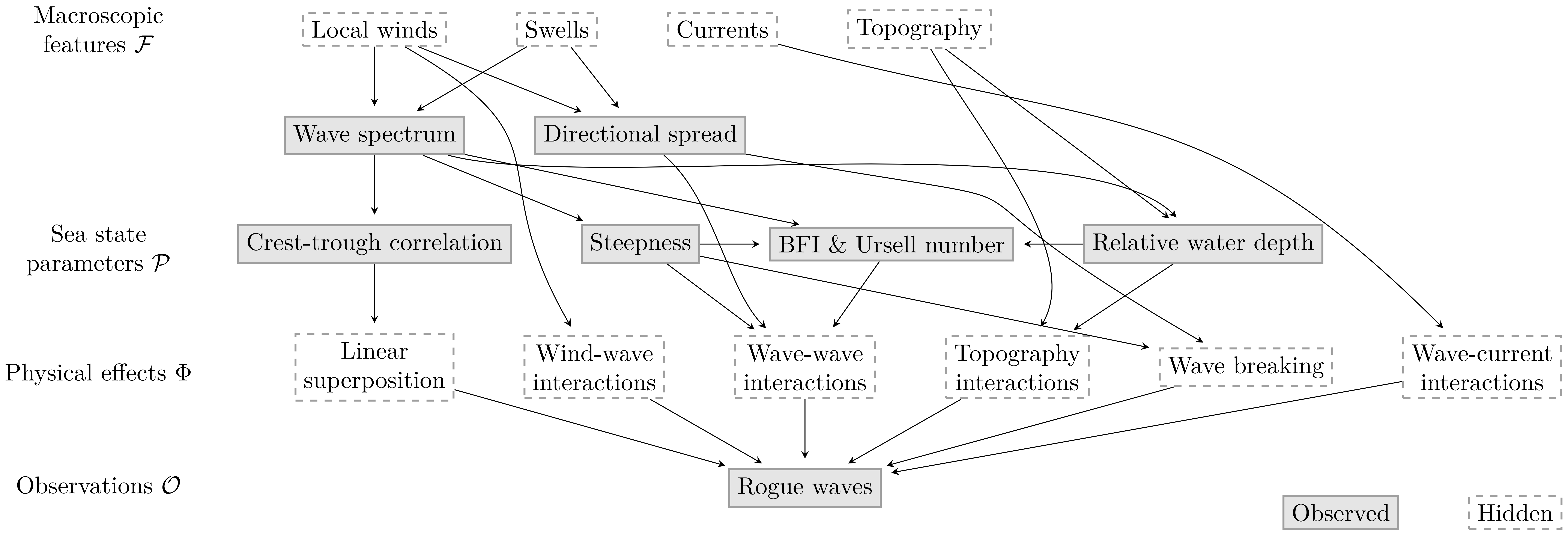

```
{
    "start_time": 1782.8125,
    "end_time": 1800.0,
    "id_local": 0.0,
    "zero_crossing_period": 16.4193,
    "zero_crossing_wavelength": 418.8372,
    "maximum_elevation_slope": 0.768,
    "crest_height": 1.22,
    "trough_depth": -1.31,
    "height": 2.53,
    "ursell_number": 0.0555,
    "raw_elevation": [
        0.41,
        0.65,
        0.63,
        0.87,
        1.22,
        1.13,
        0.99,
        0.93,
        0.6,
        0.1,
        -0.37,
        -0.55,
        -0.45,
        -0.94,
        -0.93,
        -1.2,
        -1.31,
        -0.52,
        -0.32,
        -0.16,
        -0.06
    ]
}
```


x 3 000 000 000

(1.5TB output data)

Step 1: Write down causal graph

Make assumptions explicit, reduce dimensionality



Step 2: Train neural network

On different subsets of causal features

IDEA

Parameterize rogue wave probability as

$$\log P(h > 2H_s) \sim \underbrace{f_1(r)}_{\text{linear}} + \underbrace{f_2(\text{BFI}, R)}_{\text{free waves}} + \underbrace{f_3(\varepsilon, kD)}_{\text{bound waves}} + \dots$$

where f_i are neural networks.

Step 2: Train neural network

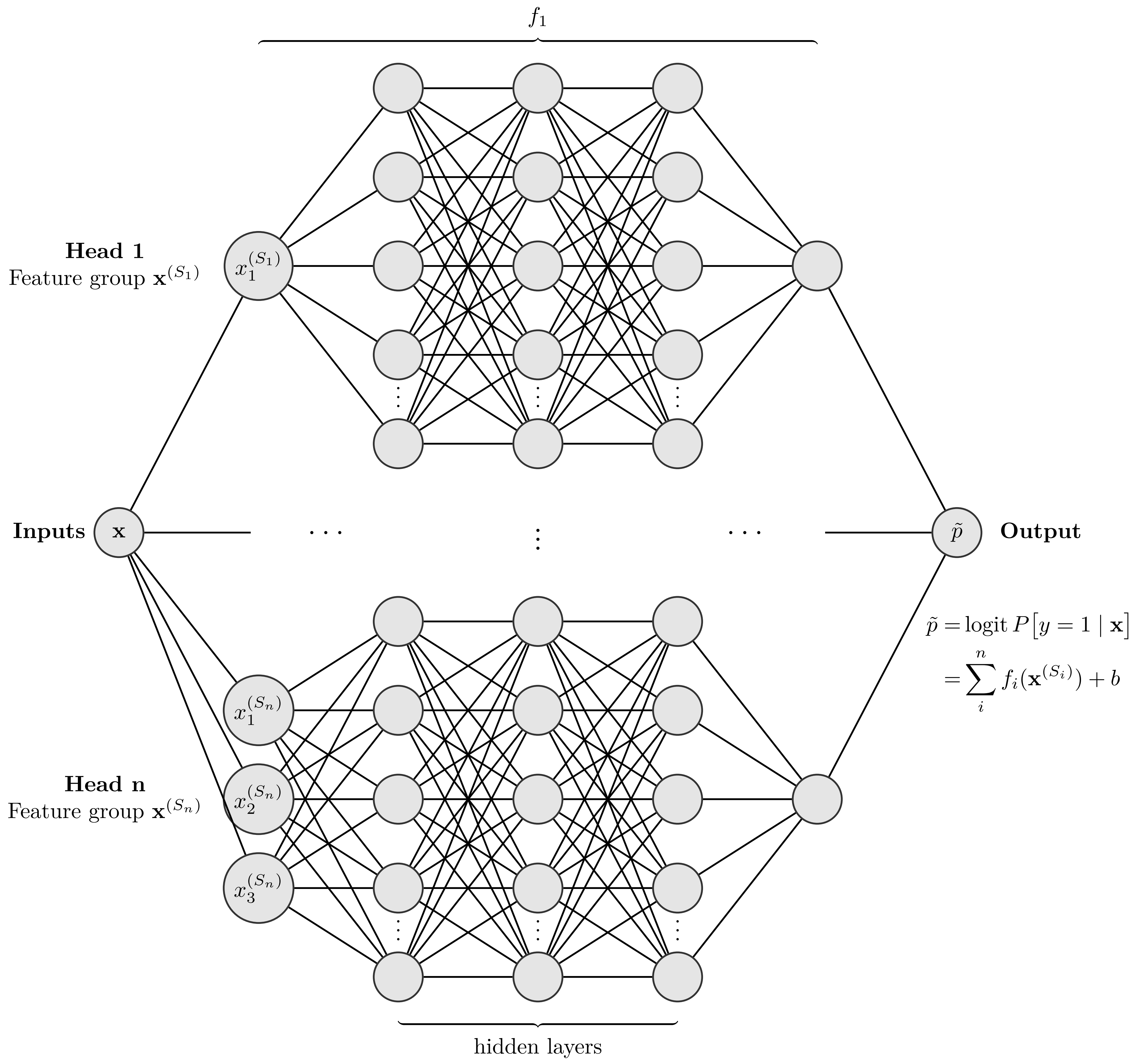
On different subsets of causal features

IDEA

Parameterize rogue wave probability as

$$\log P(h > 2H_s) \sim \underbrace{f_1(r)}_{\text{linear}} + \underbrace{f_2(\mathrm{BFI}, R)}_{\text{free waves}} + \underbrace{f_3(\varepsilon, kD)}_{\text{bound waves}} + \ldots$$

where f_i are neural networks.



Step 3: Find causally consistent network

Idea: Causal models are invariant under data shift

TABLE 1. The subsets of the validation data set used to evaluate model invariance.

Subset name	Condition	# waves
southern-california	Longitude $\in (-123.5, -117)°$, Latitude $\in (32, 38)°$	233M
deep-stations	Water depth $> 1000\,m$	33M
shallow-stations	Water depth $< 100\,m$	138M
summer	Day of year $\in (160, 220)$	44M
winter	Day of year $\in (0, 60)$	88M
Hs > 3m	$H_s > 3\,m$	55M
high-frequency	Relative swell energy < 0.15	40M
low-frequency	Relative swell energy > 0.7	42M
long-period	Mean zero-crossing period $> 9\,s$	40M
short-period	Mean zero-crossing period $< 6\,s$	90M
cnoidal	Ursell number > 8	34M
weakly-nonlinear	Steepness > 0.04	80M
spectral-narrow	Directionality index < 0.3	68M
spectral-wide	Directionality index > 1	37M
full	(all validation data)	438M

Measure how much **predictions change** after re-training on subsets

Step 3: Find causally consistent network

Idea: Causal models are invariant under data shift

TABLE 1. The subsets of the validation data set used to evaluate model invariance.

Subset name	Condition	# waves
southern-california	Longitude $\in (-123.5, -117)°$, Latitude $\in (32, 38)°$	233M
deep-stations	Water depth > 1000 m	33M
shallow-stations	Water depth < 100 m	138M
summer	Day of year $\in (160, 220)$	44M
winter	Day of year $\in (0, 60)$	88M
Hs > 3m	$H_s > 3$ m	55M
high-frequency	Relative swell energy < 0.15	40M
low-frequency	Relative swell energy > 0.7	42M
long-period	Mean zero-crossing period > 9 s	40M
short-period	Mean zero-crossing period < 6 s	90M
cnoidal	Ursell number > 8	34M
weakly-nonlinear	Steepness > 0.04	80M
spectral-narrow	Directionality index < 0.3	68M
spectral-wide	Directionality index > 1	37M
full	(all validation data)	438M

Measure how much **predictions change** after re-training on subsets

TABLE 2. Full list of experiments. $\mathcal{L}$: Prediction score (higher is better). $\mathcal{E}$: Invariance error (lower is better). $\mathcal{C}$: Calibration error (lower is better). Color coding ranges between (median − IQR, median + IQR) with inter-quartile range IQR.

	Feature groups			Scores		
ID	1	2	3	$\mathcal{L} \times 10^4$	$\mathcal{E} \times 10^2$	$\mathcal{C} \times 10^2$
1	$\{r\}$			4.62	8.52	6.90
2	$\{r, R\}$			5.05	8.58	3.86
3	$\{\varepsilon, \widetilde{D}, R\}$			0.03	22.59	6.21
4	$\{r, \widetilde{D}, R\}$			5.56	7.95	4.34
5	$\{r, \varepsilon, R\}$			5.49	8.83	3.83
6	$\{r, \varepsilon, \widetilde{D}\}$			5.35	8.89	7.05
7	$\{r, R\}$	$\{\varepsilon, \widetilde{D}\}$		5.77	9.19	4.46
8	$\{r, R, \mathrm{Ur}\}$			5.70	7.99	3.94
9	$\{r, R\}$	$\{\mathrm{Ur}, R\}$		5.64	7.49	4.31
10	$\{r, R, \mathrm{BFI}\}$			5.60	7.75	4.51
11	$\{r, R\}$	$\{\mathrm{BFI}, R\}$		5.46	8.20	4.44
12	$\{r\}$	$\{\varepsilon, \widetilde{D}, R\}$		5.67	9.24	4.67
13	$\{\sigma_f\}$	$\{\varepsilon, \widetilde{D}, R\}$		4.11	12.16	6.30
14	$\{r\}$	$\{\varepsilon, \widetilde{D}\}$	$\{\mathrm{BFI}, R\}$	5.64	9.77	6.02
15	$\{r, R\}$	$\{\varepsilon, \widetilde{D}, \sigma_\theta\}$		6.22	10.63	5.20
16	$\{r, R\}$	$\{\varepsilon, \widetilde{D}, R\}$		5.87	8.63	3.62
17	$\{r, \varepsilon, \widetilde{D}, R\}$			5.98	8.60	2.96
18	$\{r\}$	$\{\varepsilon, \widetilde{D}\}$	$\{\mathrm{BFI}, \sigma_f, \sigma_\theta\}$	6.01	11.10	8.43
19	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta\}$			5.97	9.71	6.45
20	$\{r, \varepsilon, \widetilde{D}, R, E_h\}$			6.10	9.14	5.33
21	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \nu\}$			6.31	10.04	4.00
22	$\{r, \varepsilon, \widetilde{D}, R, \mathrm{BFI}\}$			6.05	8.84	6.81
23	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \sigma_f, E_h, \mathrm{BFI}, R\}$			6.91	12.69	3.68
24	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \sigma_f, E_h, H_s, \overline{T}, \kappa, \mu, \lambda_p\}$			6.70	56.44	7.27

Symbols

r	Crest-trough correlation	ν	Spectral bandwidth (narrowness)
σ_f	Spectral bandwidth (peakedness)	σ_θ	Directional spread
ε	Peak steepness $H_s k_p$	R	Directionality index $\sigma_\theta^2/(2\nu^2)$
BFI	Benjamin-Feir index	$\widetilde{D}$	Relative peak water depth $Dk_p/(2\pi)$
E_h	Relative high-frequency energy	Ur	Ursell number
$\overline{T}$	Mean period	κ	Kurtosis
μ	Skewness	H_s	Significant wave height

Step 3: Find causally consistent network

Idea: Causal models are invariant under data shift

TABLE 1. The subsets of the validation data set used to evaluate model invariance.

Subset name	Condition	# waves
southern-california	Longitude $\in (-123.5, -117)°$, Latitude $\in (32, 38)°$	233M
deep-stations	Water depth > 1000 m	33M
shallow-stations	Water depth < 100 m	138M
summer	Day of year $\in (160, 220)$	44M
winter	Day of year $\in (0, 60)$	88M
Hs > 3m	$H_s > 3$ m	55M
high-frequency	Relative swell energy < 0.15	40M
low-frequency	Relative swell energy > 0.7	42M
long-period	Mean zero-crossing period > 9 s	40M
short-period	Mean zero-crossing period < 6 s	90M
cnoidal	Ursell number > 8	34M
weakly-nonlinear	Steepness > 0.04	80M
spectral-narrow	Directionality index < 0.3	68M
spectral-wide	Directionality index > 1	37M
full	(all validation data)	438M

Measure how much **predictions change** after re-training on subsets

TABLE 2. Full list of experiments. $\mathcal{L}$: Prediction score (higher is better). $\mathcal{E}$: Invariance error (lower is better). $\mathcal{C}$: Calibration error (lower is better). Color coding ranges between (median − IQR, median + IQR) with inter-quartile range IQR.

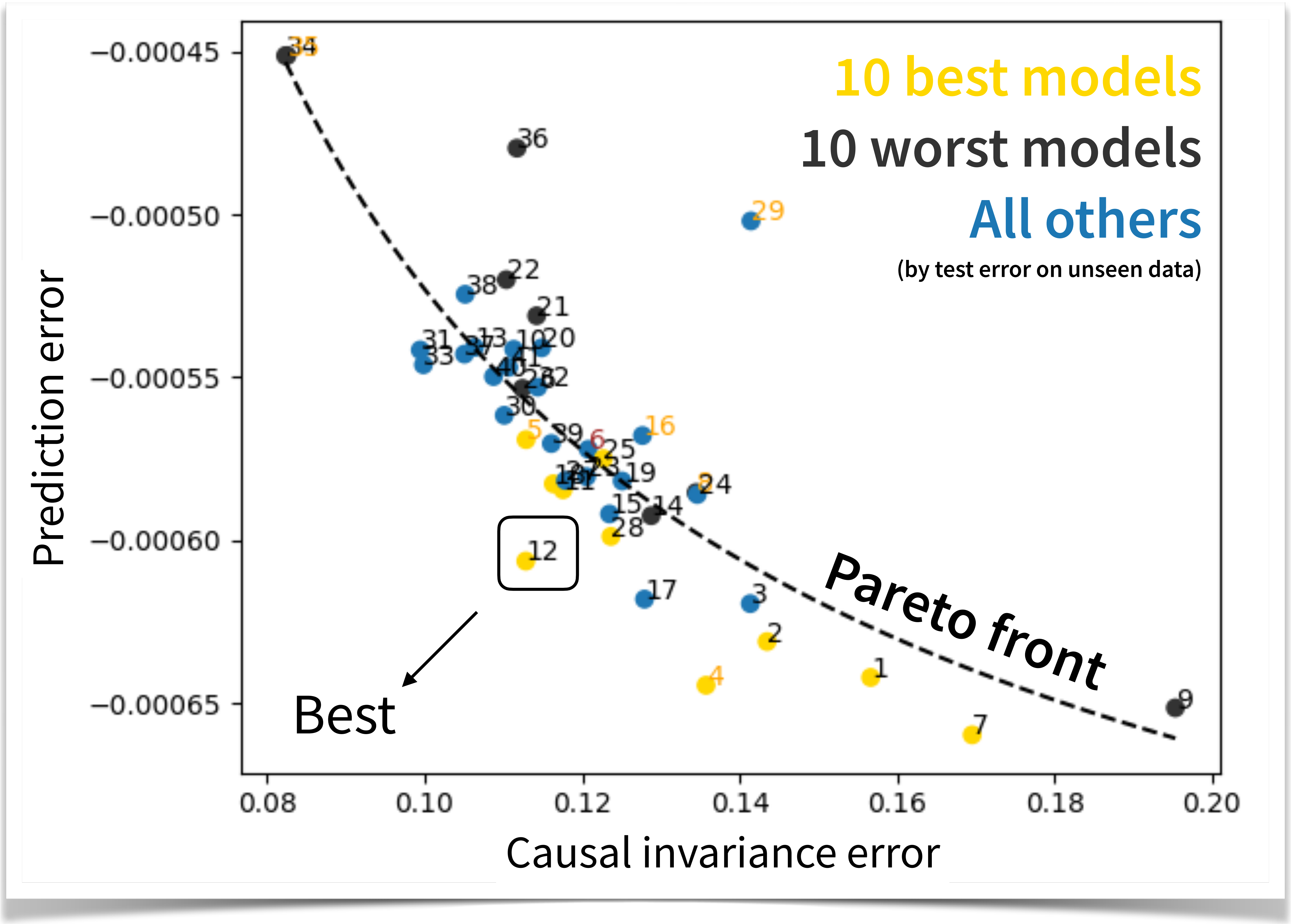
	Feature groups			Scores		
ID	1	2	3	$\mathcal{L} \times 10^4$	$\mathcal{E} \times 10^2$	$\mathcal{C} \times 10^2$
1	$\{r\}$			4.62	8.52	6.90
2	$\{r, R\}$			5.05	8.58	3.86
3	$\{\varepsilon, \widetilde{D}, R\}$			0.03	22.59	6.21
4	$\{r, \widetilde{D}, R\}$			5.56	7.95	4.34
5	$\{r, \varepsilon, R\}$			5.49	8.83	3.83
6	$\{r, \varepsilon, \widetilde{D}\}$			5.35	8.89	7.05
7	$\{r, R\}$	$\{\varepsilon, \widetilde{D}\}$		5.77	9.19	4.46
8	$\{r, R, \mathrm{Ur}\}$			5.70	7.99	3.94
9	$\{r, R\}$	$\{\mathrm{Ur}, R\}$		5.64	7.49	4.31
10	$\{r, R, \mathrm{BFI}\}$			5.60	7.75	4.51
11	$\{r, R\}$	$\{\mathrm{BFI}, R\}$		5.46	8.20	4.44
12	$\{r\}$	$\{\varepsilon, \widetilde{D}, R\}$		5.67	9.24	4.67
13	$\{\sigma_f\}$	$\{\varepsilon, \widetilde{D}, R\}$		4.11	12.16	6.30
14	$\{r\}$	$\{\varepsilon, \widetilde{D}\}$	$\{\mathrm{BFI}, R\}$	5.64	9.77	6.02
15	$\{r, R\}$	$\{\varepsilon, \widetilde{D}, \sigma_\theta\}$		[illegible]	[illegible]	[illegible]
16	$\{r, R\}$	$\{\varepsilon, \widetilde{D}, R\}$		5.87	8.63	3.62
17	$\{r, \varepsilon, \widetilde{D}, R\}$			5.98	8.60	2.96
18	$\{r\}$	$\{\varepsilon, \widetilde{D}\}$	$\{\mathrm{BFI}, \sigma_f, \sigma_\theta\}$	[illegible]	[illegible]	[illegible]
19	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta\}$			5.97	9.71	6.45
20	$\{r, \varepsilon, \widetilde{D}, R, E_h\}$			6.10	9.14	5.33
21	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \nu\}$			6.31	10.04	4.00
22	$\{r, \varepsilon, \widetilde{D}, R, \mathrm{BFI}\}$			6.05	8.84	6.81
23	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \sigma_f, E_h, \mathrm{BFI}, R\}$			6.91	12.69	3.68
24	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \sigma_f, E_h, H_s, \overline{T}, \kappa, \mu, \lambda_p\}$			6.70	56.44	7.27

Symbols

r	Crest-trough correlation	ν	Spectral bandwidth (narrowness)
σ_f	Spectral bandwidth (peakedness)	σ_θ	Directional spread
ε	Peak steepness $H_s k_p$	R	Directionality index $\sigma_\theta^2/(2\nu^2)$
BFI	Benjamin-Feir index	$\widetilde{D}$	Relative peak water depth $Dk_p/(2\pi)$
E_h	Relative high-frequency energy	Ur	Ursell number
$\overline{T}$	Mean period	κ	Kurtosis
μ	Skewness	H_s	Significant wave height

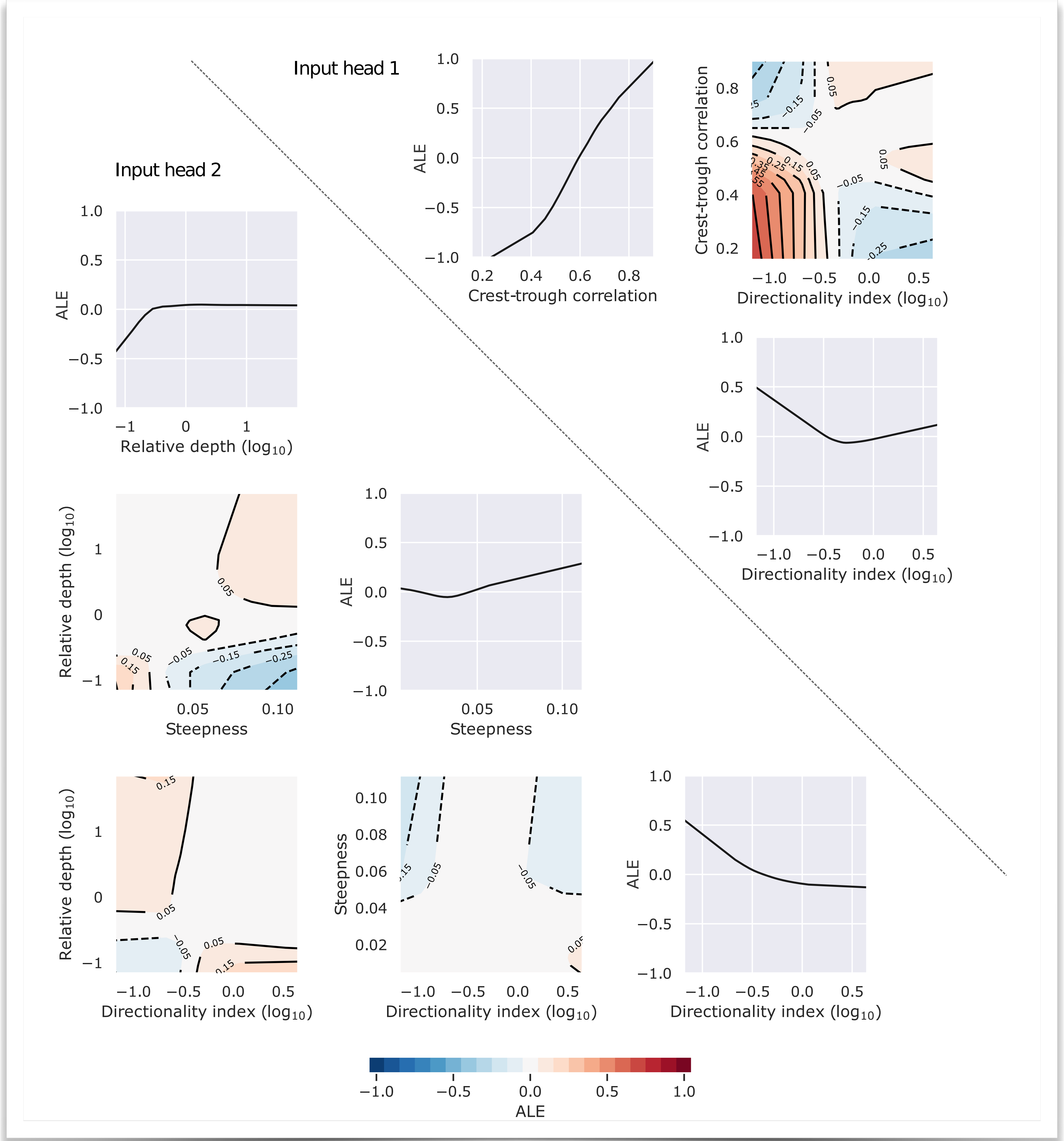
The role of parsimony (I)

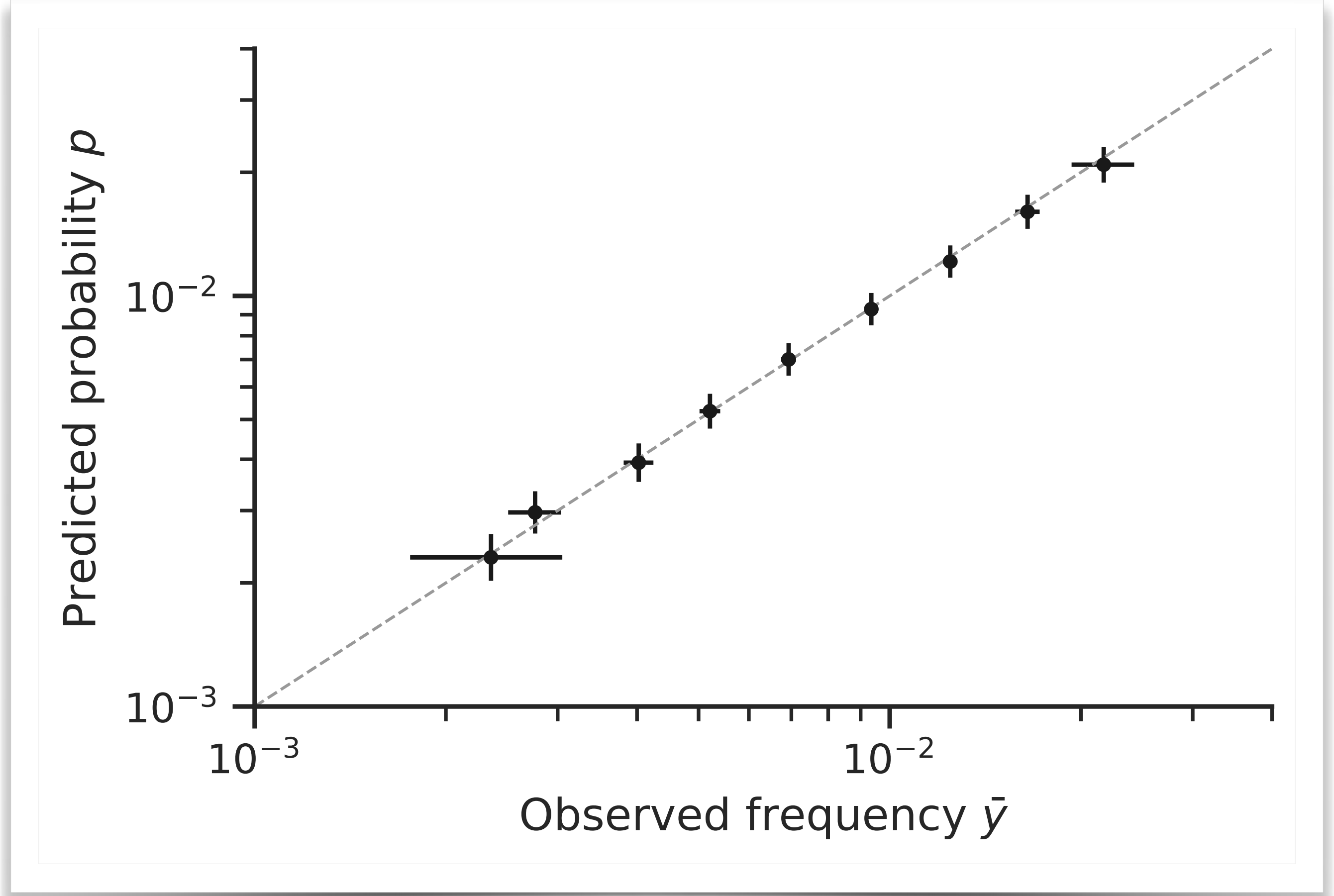
Model selection

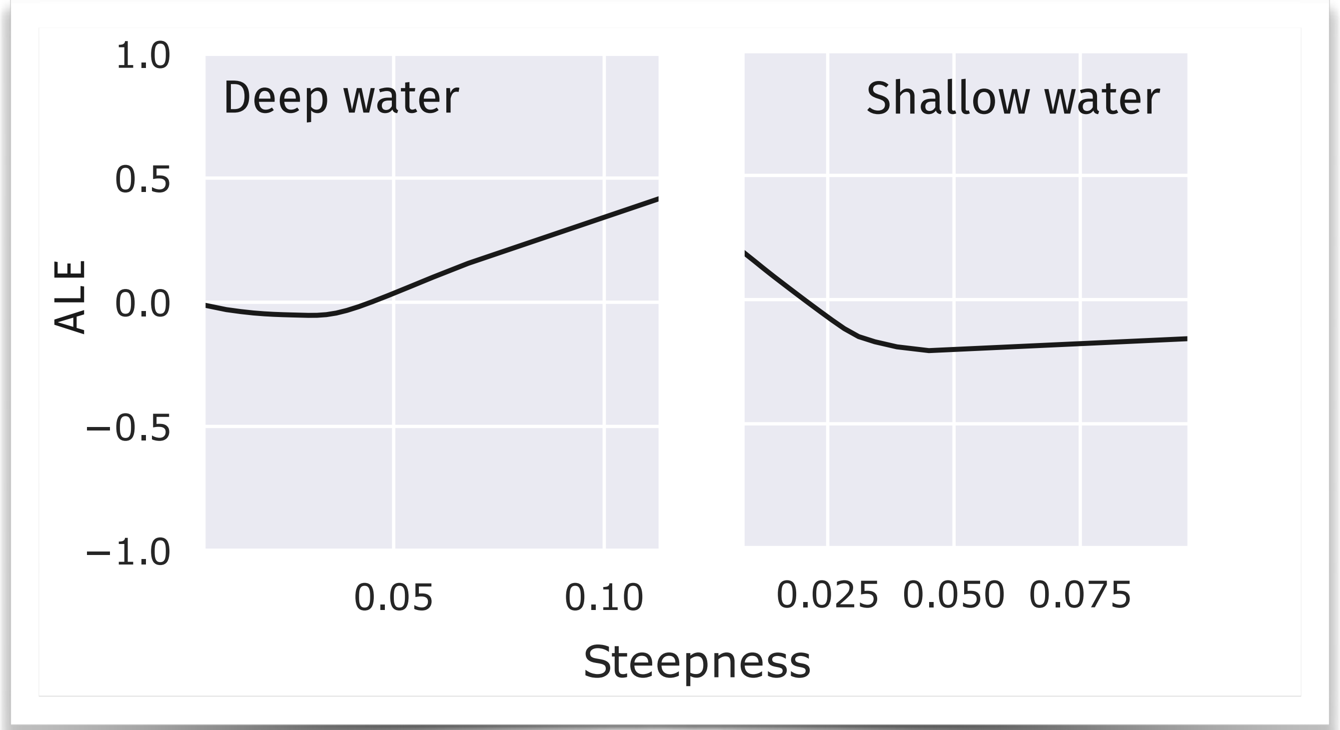


Step 4: Analyze selected model

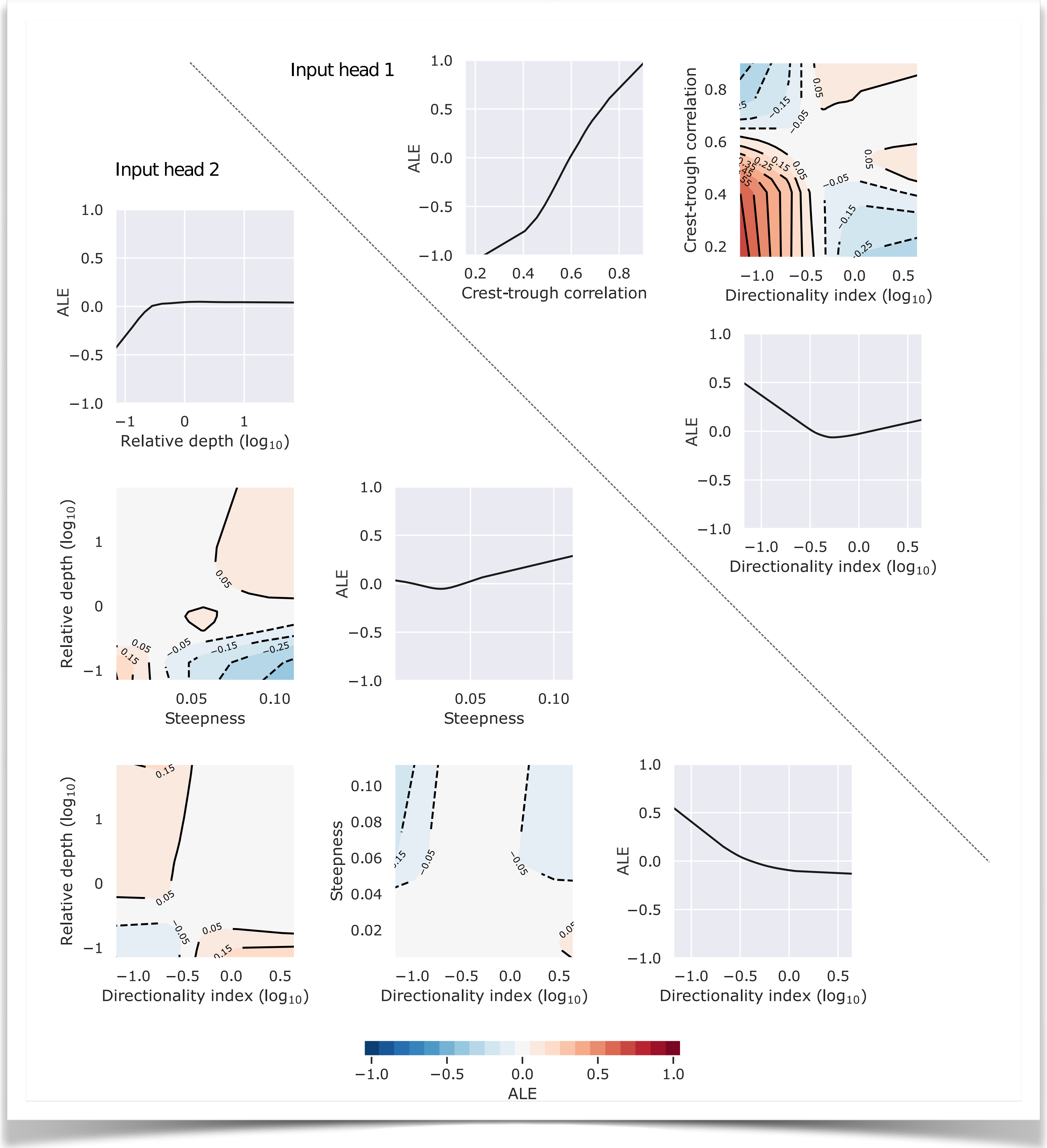
With your favorite interpretable ML methods







Next step

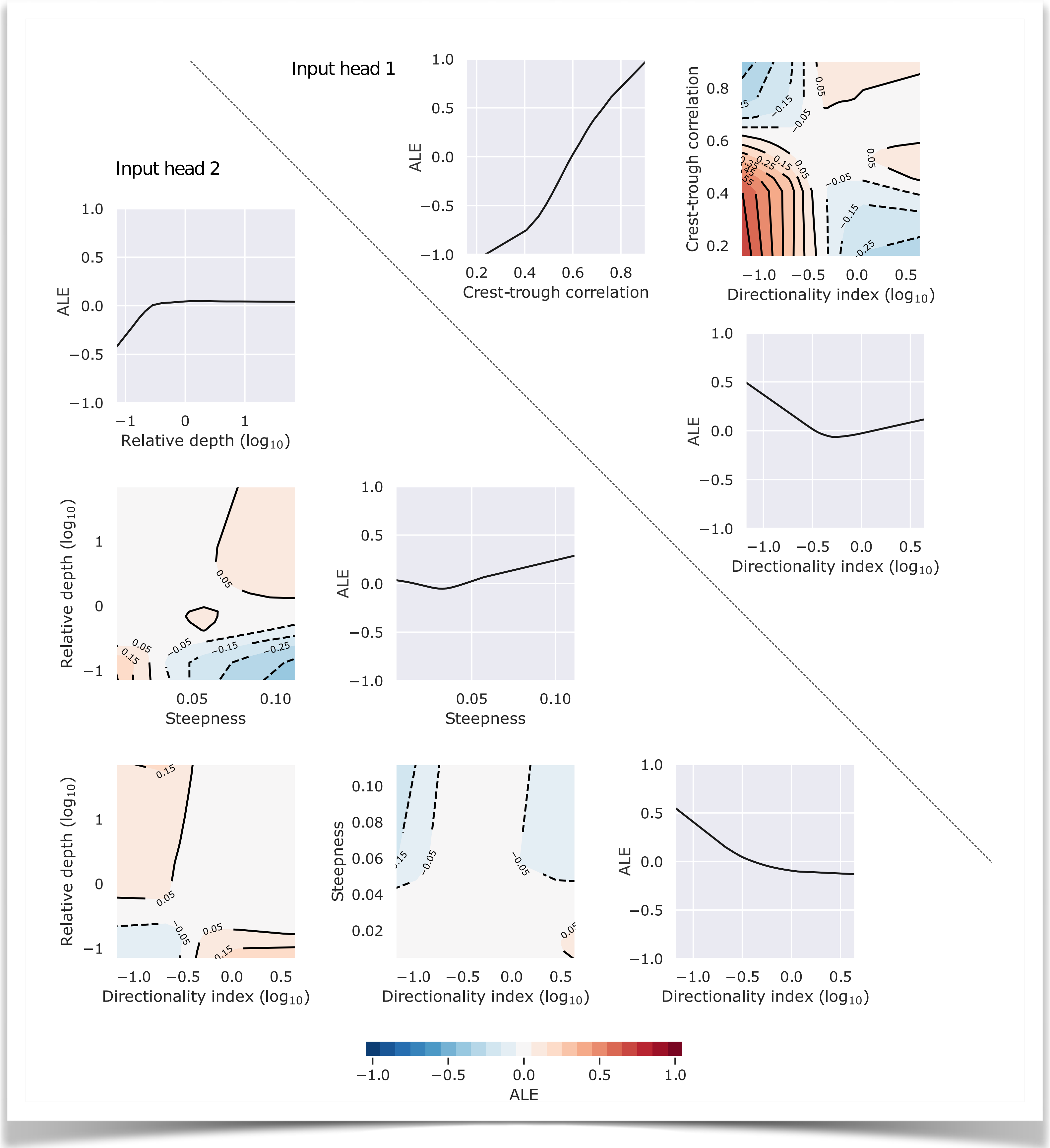


Symbolic regression

$$P(H/H_s > \kappa \mid r, \varepsilon, R, kD) = \ldots$$

Next step

DISCOVERY



Symbolic regression

$$P(H/H_s > \kappa \mid r, \varepsilon, R, kD) = \ldots$$

The role of parsimony (II)

Distillation through symbolic regression

Rediscovering orbital mechanics with machine learning

Pablo Lemos *[1,2], Niall Jeffrey †[3,2], Miles Cranmer[4], Shirley Ho[4,5,6,7], and Peter Battaglia[8]

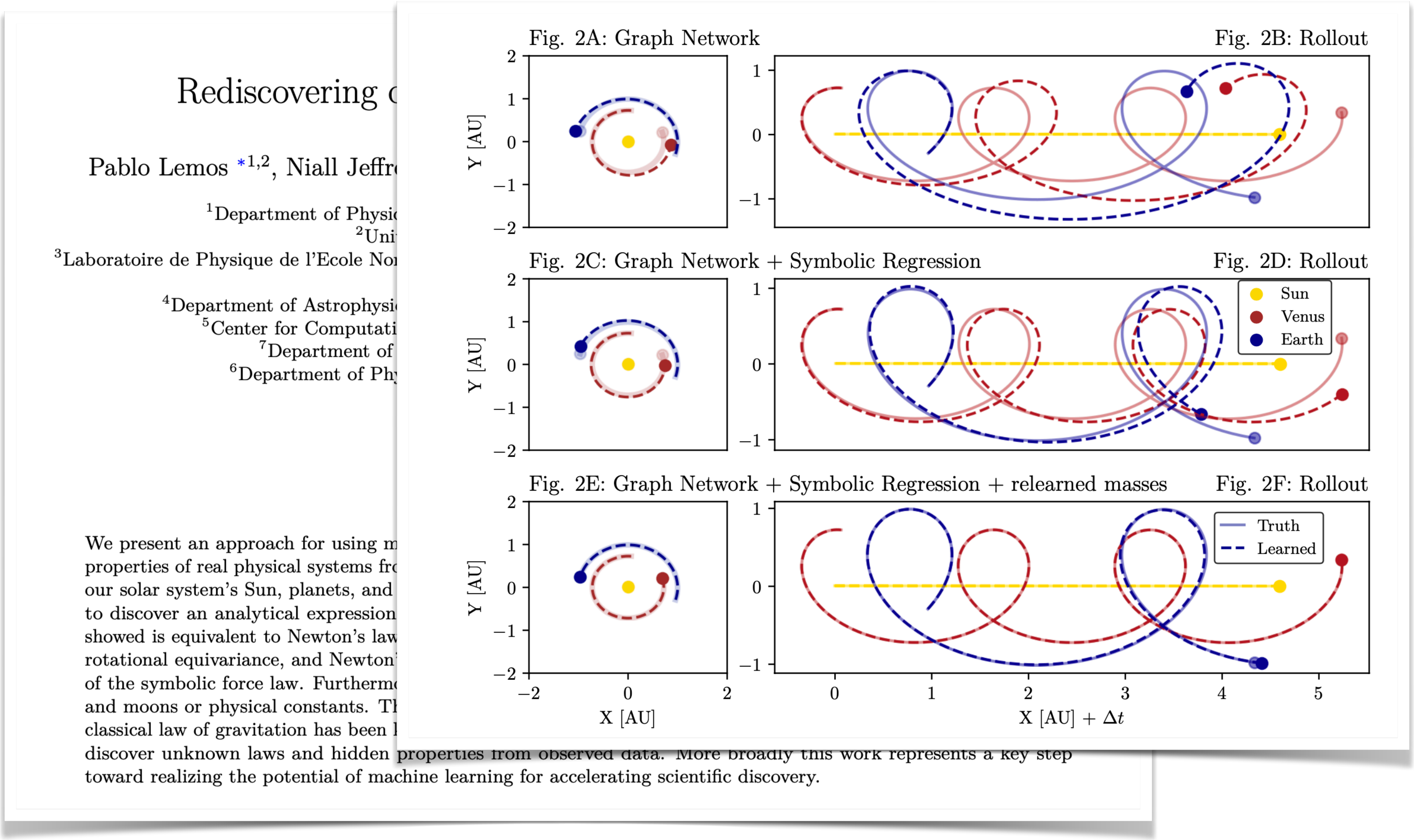
[1]Department of Physics and Astronomy, University of Sussex,Brighton, BN1 9QH, UK
[2]University College London, Gower St, London, UK
[3]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université Université de Paris, Paris, France
[4]Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey 08544, USA
[5]Center for Computational Astrophysics, Flatiron Institute, New York, NY 10010, USA
[7]Department of Physics, New York University, New York, NY 10011, USA
[6]Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15217, USA
[8]DeepMind, London, N1C 4AG, UK

Abstract

We present an approach for using machine learning to automatically discover the governing equations and hidden properties of real physical systems from observations. We train a "graph neural network" to simulate the dynamics of our solar system's Sun, planets, and large moons from 30 years of trajectory data. We then use symbolic regression to discover an analytical expression for the force law implicitly learned by the neural network, which our results showed is equivalent to Newton's law of gravitation. The key assumptions that were required were translational and rotational equivariance, and Newton's second and third laws of motion. Our approach correctly discovered the form of the symbolic force law. Furthermore, our approach did not require any assumptions about the masses of planets and moons or physical constants. They, too, were accurately inferred through our methods. Though, of course, the classical law of gravitation has been known since Isaac Newton, our result serves as a validation that our method can discover unknown laws and hidden properties from observed data. More broadly this work represents a key step toward realizing the potential of machine learning for accelerating scientific discovery.

The role of parsimony (II)

Distillation through symbolic regression


Rediscovering c

Pablo Lemos *[1,2], Niall Jeffr

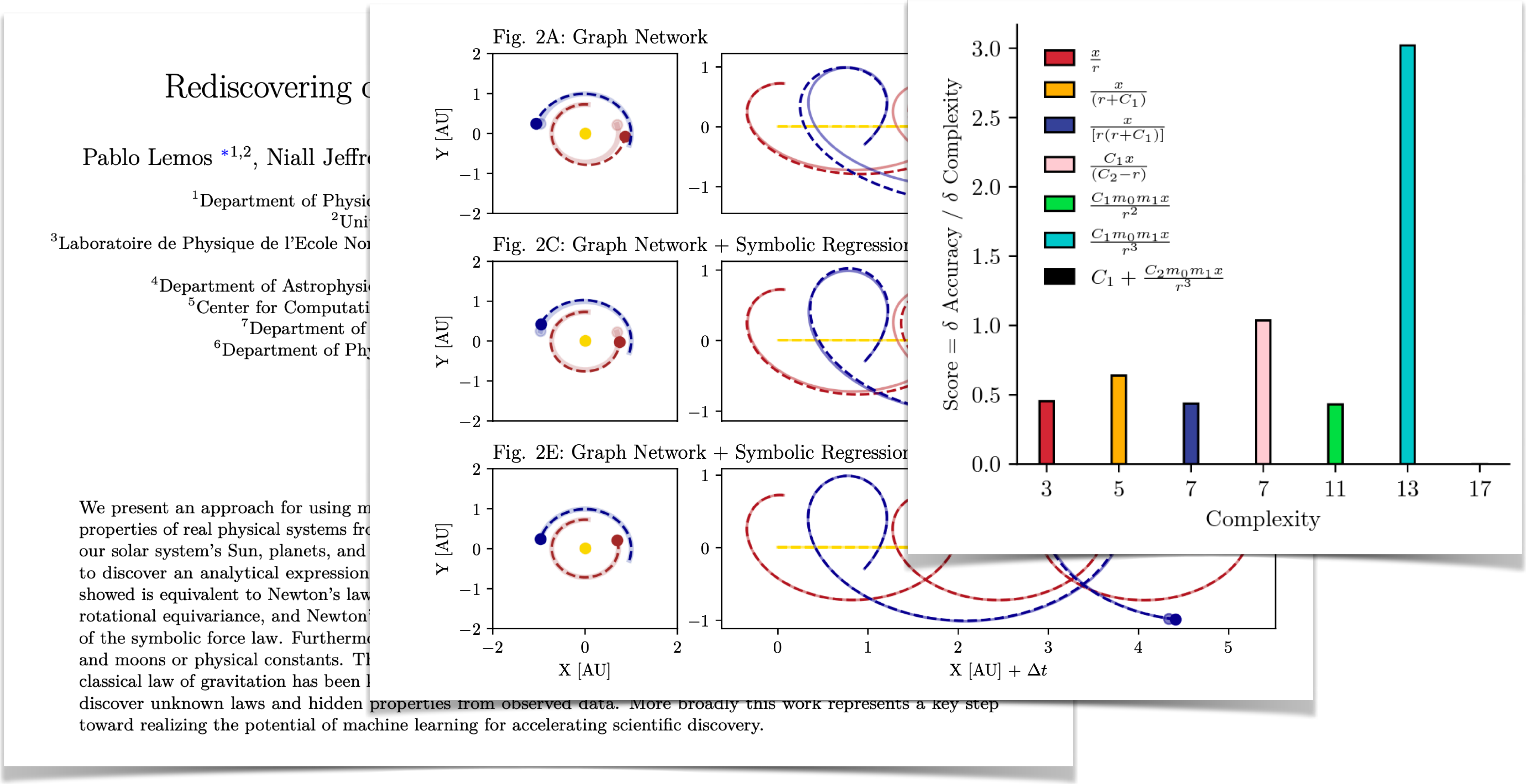
[1]Department of Physic
[2]Uni
[3]Laboratoire de Physique de l'Ecole Nor

[4]Department of Astrophysi
[5]Center for Computati
[7]Department of
[6]Department of Phy

We present an approach for using m
properties of real physical systems fr
our solar system's Sun, planets, and
to discover an analytical expression
showed is equivalent to Newton's law
rotational equivariance, and Newton
of the symbolic force law. Furtherm
and moons or physical constants. Th
classical law of gravitation has been l
discover unknown laws and hidden properties from observed data. More broadly this work represents a key step
toward realizing the potential of machine learning for accelerating scientific discovery.

The role of parsimony (II)

Distillation through symbolic regression



The role of parsimony (II)

Distillation through symbolic regression

```
0.297  0.00695565904255765

0.181  0.00386934011267578*exp(r)

0.113  0.012177238*r

0.030  0.000654510454159562*exp(4.023554*r)

0.028  0.00194348077435562*exp(3.5268505*r**2)

0.020  0.00128053886674162*exp(3.7422986*r - σ)

0.020  0.00353452980427828*exp(3.2719116*r**2 - σ)

0.018  0.00240915190512588*exp(3.2004833*r - ν - σ)

0.018  0.00173696837251109*exp(3.403016*r - ν**2 - σ)

0.017  0.00126410565365282*exp(r*(3.8113687 - ε/kD) - σ)

0.016  0.00236744728323724*exp(-r*(-3.277927 + ε/kD) - ν - σ)

0.015  0.00376661597495146*exp(3.25455527028996*r**2 - σ - 8.07518292469624*(0.30867642 - ν)**2)

0.014  0.00382349520334577*exp(3.23052130269801*r**2 - σ - 2.5516672*(0.2778663 - ν)**2/ν)

0.014  0.00382349520334577*exp(3.23052130269801*r**2 - σ - 2.43838369006115*(0.2778663 - ν)**2/(-ε + ν))

0.013  0.0030822836867589*exp(-r**4/(ν*log(kD**2/ε**2)) + 3.92468826802276*r**2 - σ)

0.011  0.00506309157021754*exp(3.399517313284*r**2 - r*(1.16199409653889*(0.927679669983524*log(ν) + 1)**2 + ε/kD)
- ν - σ)

0.011  0.00504992443082451*exp(3.399517313284*r**2 - ν - σ - (r - 0.02381676)*(1.16199409653889*(0.927679669983524*
log(ν) + 1)**2 + ε/kD))

0.011  0.00506309157021754*exp(-r**2*((log(ν) + 0.996239)**2 + 2*ε/kD) + 3.42306020292196*r**2 - ν - σ)
```

DISCOVERY

DISCOVERY

$$P(H > 2H_S) = \frac{1}{\sqrt{\sigma}} \exp\left[3.82r - 12.04 - \varepsilon \cdot \left(-65.92\varepsilon + \sqrt{1/\varepsilon} + \frac{0.23}{kD \cdot \nu}\right)\right]$$

DISCOVERY

$$P(H > 2H_S) = \frac{1}{\sqrt{\sigma}} \exp\left[3.82r - 12.04 - \varepsilon \cdot \left(-65.92\varepsilon + \sqrt{1/\varepsilon} + \frac{0.23}{kD \cdot \nu}\right)\right]$$

Nonsensical (?)
σ can be 0 in theory
observed values ~ (0.2, 1.0)
→ minor correction

Approx. first-order expansion
of Tayfun distribution

$$\sim \exp\left[-\frac{16}{1+r}\right]$$

$$= 4r - 12 + \mathcal{O}(r^2)$$

Weakly nonlinear
contribution (pos.)

Wave breaking? (neg.)

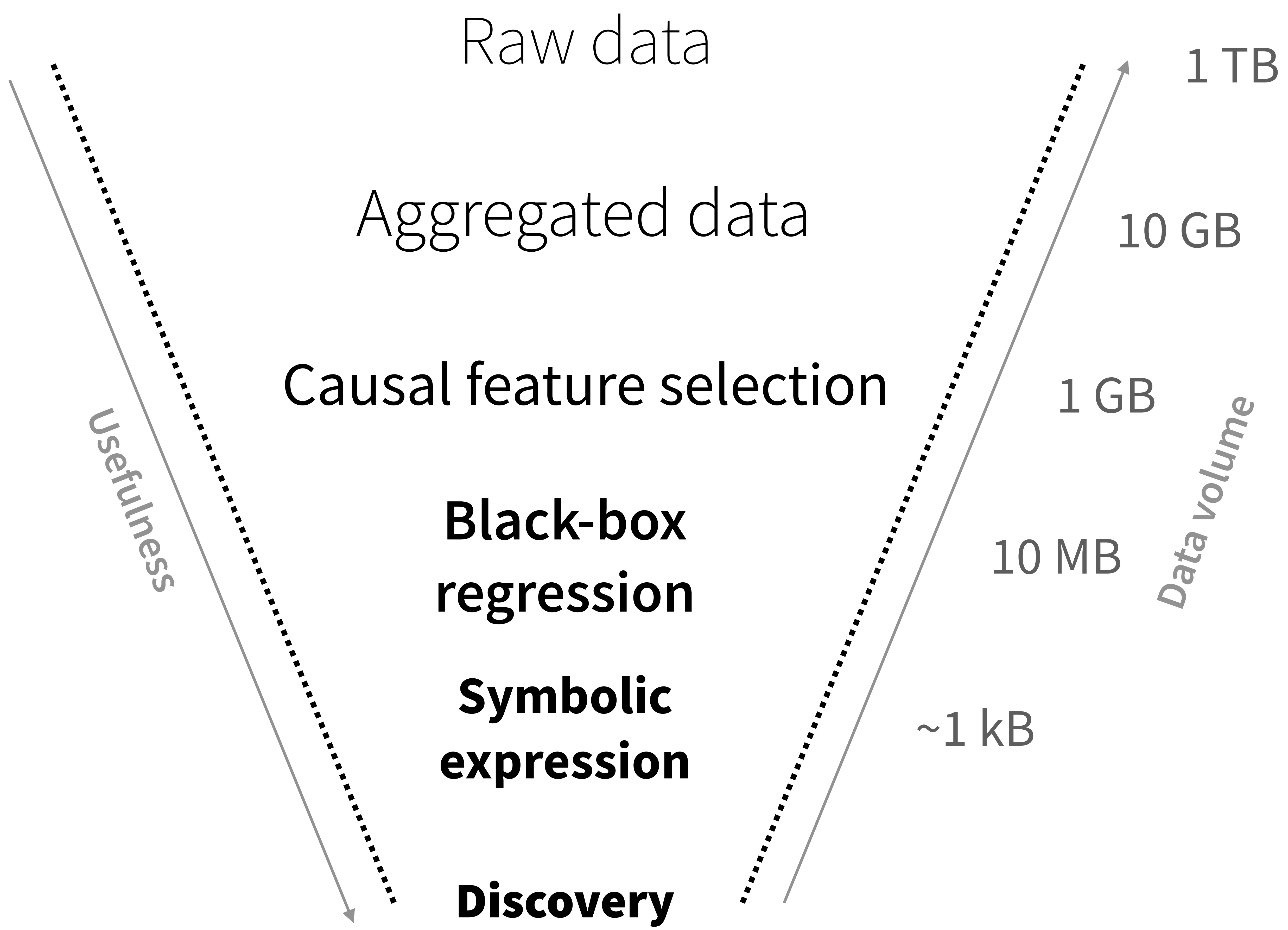
$$\varepsilon/(kD) \equiv H_S/D$$

Governing nonlinear
parameter in shallow-water
expansion.
Wave-induced current?

Some terms we understand already, some are explainable, some questionable.

The funnel

From data to science



A call to action

What we need:

(i) Incentives and best practices for **open** data

(ii) Fast, interpretable methods for **probabilistic reasoning**

(iii) Off-the-shelf **causal** methods and education on causal analysis

(iv) Prioritizing **discovery** over accuracy