Kermut: Composite kernel regression for protein variant effects

Peter Mørch Groth*, **Mads** Herbert Kerrn*,
Lars Olsen, Jesper Salomon, and Wouter Boomsma

* Equal contributions

KØBENHAVNS UNIVERSITET
UNIVERSITY OF COPENHAGEN

novonesis

M · L · L · S

PIONEER
CENTRE
FOR AI

Variant effect prediction

- Variant effects are measurable **changes in protein function caused by mutations** in the amino acid sequence.
- Predicting potentially **beneficial** or **deleterious** mutations is crucial for engineering and optimizing proteins, e.g., to increase activity and stability.

Reference	V	F	A	H	P	E	T	L	1.0
Variant 1	V	F	A	H	P	W	T	L	0.2
Variant 2	V	F	A	H	A	E	T	L	1.2
Variant 3	V	E	A	H	P	E	T	L	1.5

- Central question:

How can we predict variant effects given a reference protein and experimental data for a number of variants?

Desiderata

- **Supervised** model
  - Not all protein properties correlate with zero-shot fitness estimates
  - We want to learn from our data to guide exploration for protein engineering

- **Uncertainty quantification**
  - Valuable to quantify predictive uncertainties
  - Uncertainties should be **well-calibrated**

- **Leverage pre-trained models**
  - We're often working with few labeled sequences

Model of choice: **Gaussian processes**

- Explicitly uses similarities between datapoints to reason about the function of interest
- Provides predictive uncertainties

- Fully specified by **two** components:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

  - A mean function

$$m(\mathbf{x})$$

  - And a covariance function (also known as kernel function)

$$k(\mathbf{x}, \mathbf{x}') = \mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}'))$$

Figure source: https://doi.org/10.1103/PhysRevD.98.063511

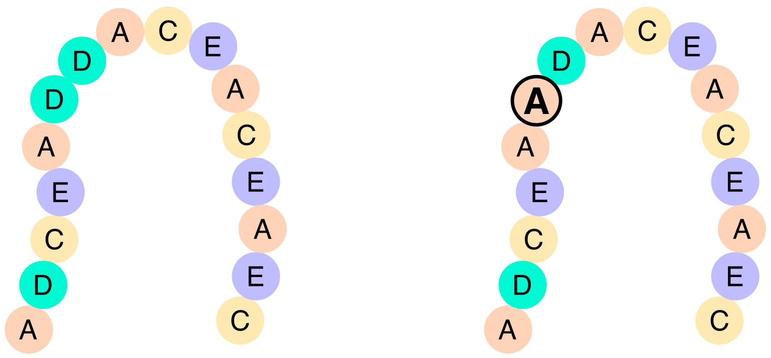
When are variants similar?

- Similarity given **local environments**

- Use an inverse-folding model to obtain **structure-conditioned amino acid distributions** at all sites

1. Local environments should be similar
2. Individual mutations should be similar
3. Mutates sites should be physically close

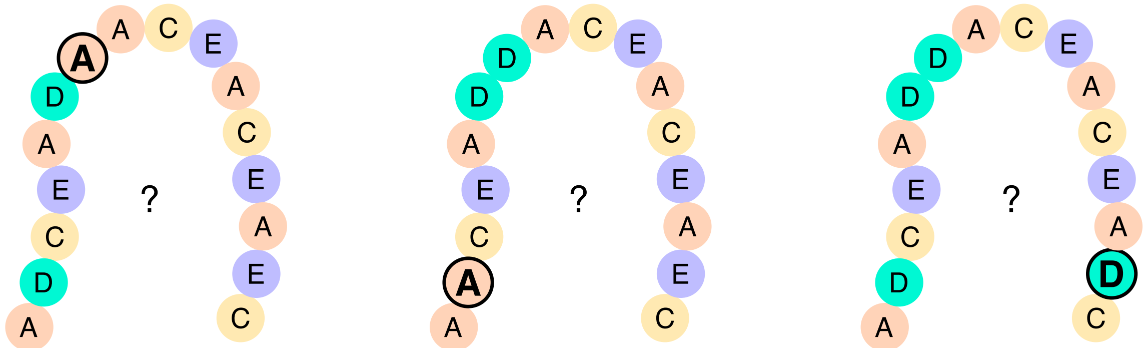
$y_{\mathrm{WT}} = 1.0$ $y = 2.0$


A D C E A D D A C E A C E A E C
A D C E A A D A C E A C E A E C

Wild type

Observed variant

Given the wild type and observed variant, what can we say about the following variants?


A D C E A D A A C E A C E A E C
?
A A C E A D D A C E A C E A E C
?
A D C E A D D A C E A C E A D C
?

Test variant 1 $y \approx 2.0$

Test variant 2 $y \approx 1.5$

Test variant 3 $y \approx 1.0$

Kermut: a kernel for modeling mutation similarity

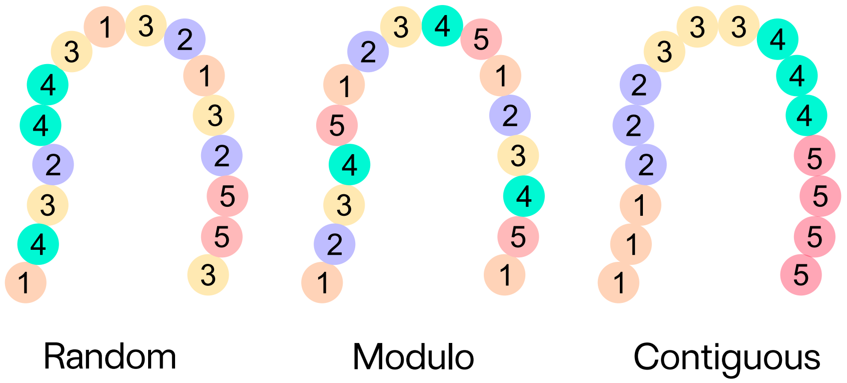
$$k(\mathbf{x}, \mathbf{x}') = \pi k_{\text{struct}}(\mathbf{x}, \mathbf{x}') + (1 - \pi) k_{\text{seq}}(\mathbf{x}, \mathbf{x}')$$

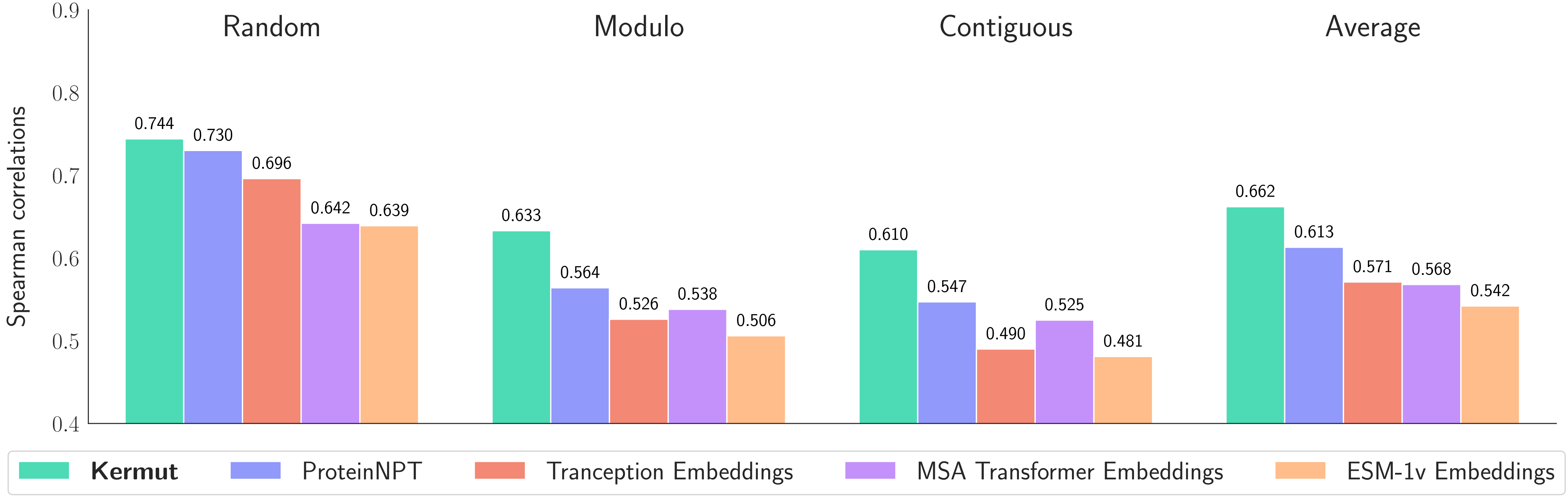
$$k_{\text{struct}}(\mathbf{x}, \mathbf{x}') = \underbrace{\lambda k_H(\mathbf{x}, \mathbf{x}')}_{\text{Site comparison}} \cdot \underbrace{k_p(\mathbf{x}, \mathbf{x}')}_{\text{Mutation comparison}} \cdot \underbrace{k_d(\mathbf{x}, \mathbf{x}')}_{\text{Distance comparison}}$$

$$k_{\text{seq}}(\mathbf{x}, \mathbf{x}') = k_{\text{SE}}(f_1(\mathbf{x}), f_1(\mathbf{x}')) = k_{\text{SE}}(\mathbf{z}, \mathbf{z}') = \underbrace{\exp\left(-\frac{||\mathbf{z} - \mathbf{z}'||_2^2}{2\sigma^2}\right)}_{\text{Embedding comparison}}$$

Results

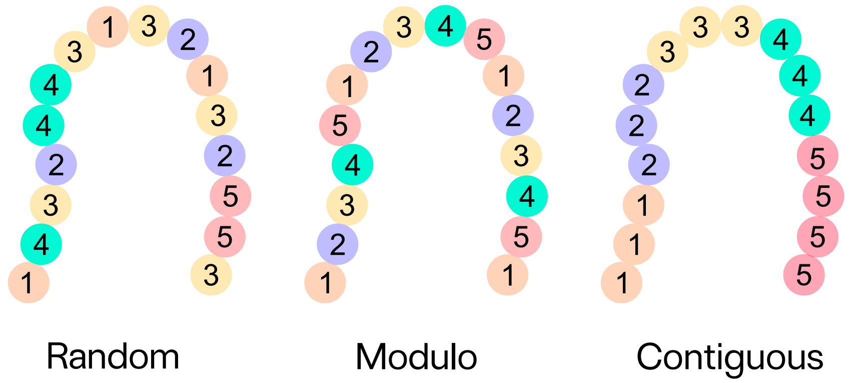
- Supervised ProteinGym benchmark
  - 217 DMS substitutions assays
  - 3 split schemes are defined for each assay with 5-fold CV in each

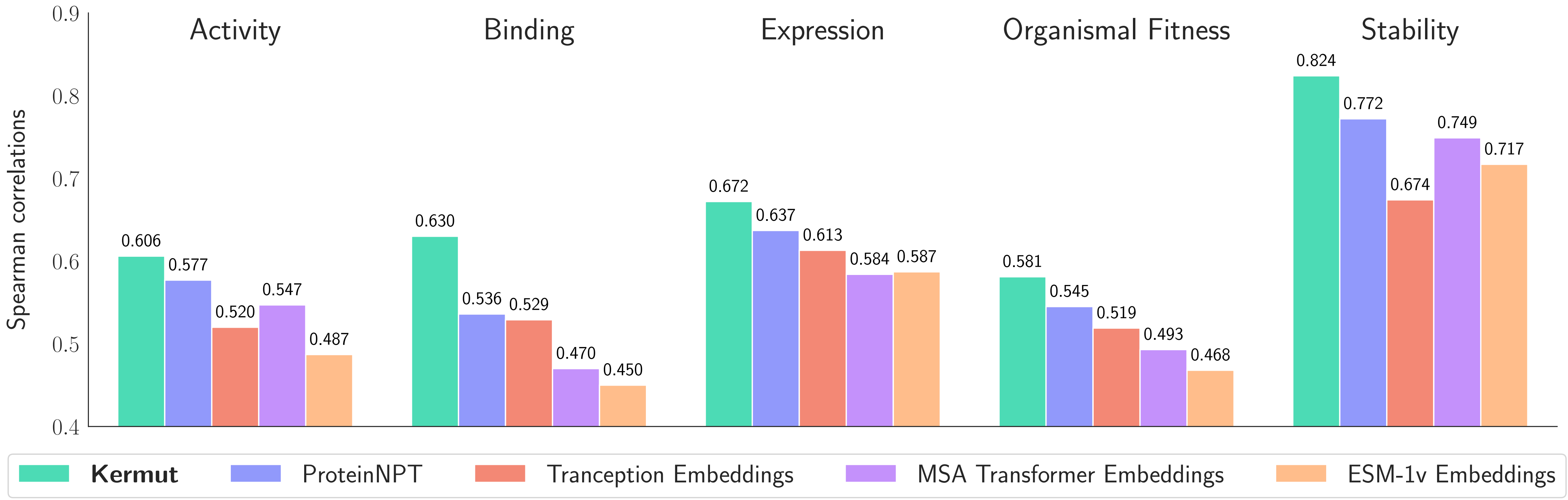




Results per functional category

- Supervised ProteinGym benchmark
  - 217 DMS substitutions assays
  - 3 split schemes are defined for each assay with 5-fold CV in each





Speed

Dataset	Kermut	ProteinNPT	N	L
BLAT_ECOLX	111s	$\approx$ 32h	4996	286
PA_I34A1	45s	$\approx$ 52h	1820	716
TCRG1_MOUSE	19s	$\approx$ 22h	621	37
OPSD_HUMAN	14s	$\approx$ 40h	165	348

Conclusion

- Kermut achieves **state-of-the-art performance** for supervised variant effect prediction
- Provides **well-calibrated uncertainties** out-of-the-box
- Can be trained and evaluated orders of magnitude **faster** than competing methods
- Can easily be adapted for new pre-trained models

Limitations

- Does not support insertions and deletions
- Due to scaling, GPs scale cubically with number of datapoints*
- Structure kernel models multi-mutants linearly – only epistasis via sequence embeddings
- Extrapolation to higher order mutations is difficult and needs further analysis

*: Not a practical concern in most protein engineering campaigns.