Hypothesis-driven interpretable neural network for interactions between genes

Modelling and predicting genotype-fitness maps

Shuhui Wang, Alexandre Allauzen, Philippe Nghe, Vaitea Opuu

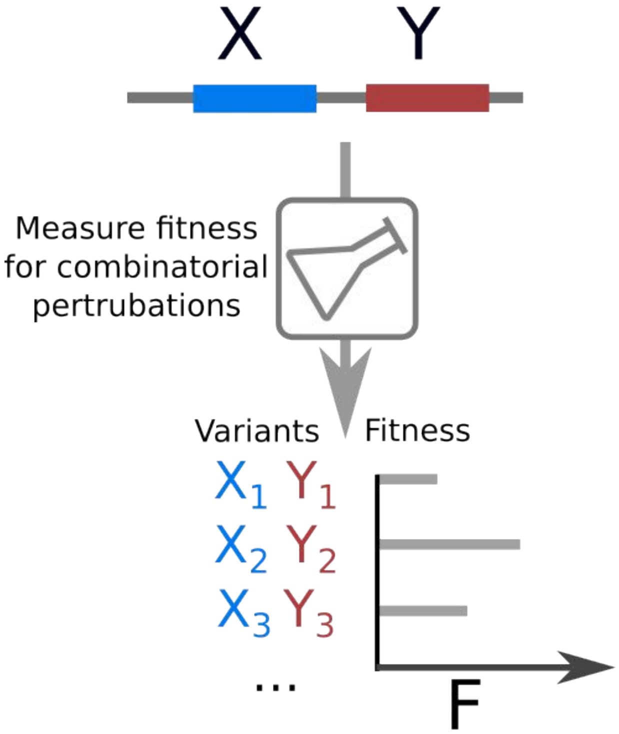
ESPCI PARIS | PSL

cnrs

Laboratoire
Biophysique
et Évolution

Modelling genotype-fitness maps

- Collection of **mutation-fitness**
- Predictive genotype-fitness model
- Interpretation to build hypotheses
- Biological system engineering



SOTA

- **Mechanistic:**

Explicit modelling of the biological system

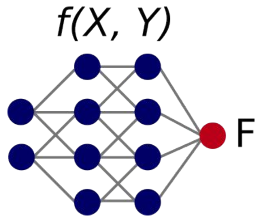
$$F(X,Y) = \left(w + \mu\varphi - \frac{\nu}{1/\eta - \varphi}\right)(1 - \theta_X X - \theta_Y Y)\,,$$

highly interpretable but not streamlined & not scalable

- **Machine learning:**

Statistical modelling of the data

easy modelling & high accuracy but low interpretability



Hypothesis-Driven Modelling

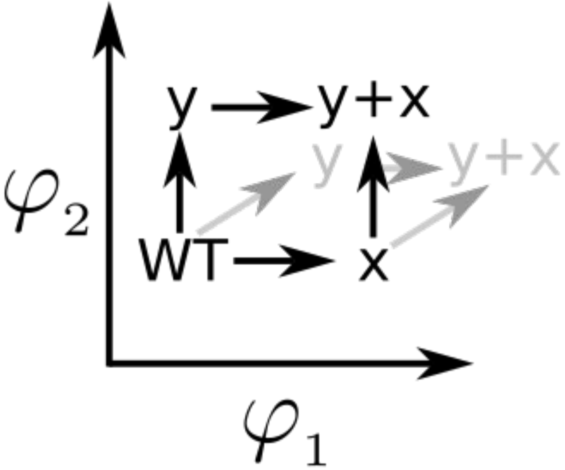
ML ~ Mechanistic

- Phenotype inference
- Identify genetic trade-offs
- Extrapolate outside of the data domain

Shuhui Wang

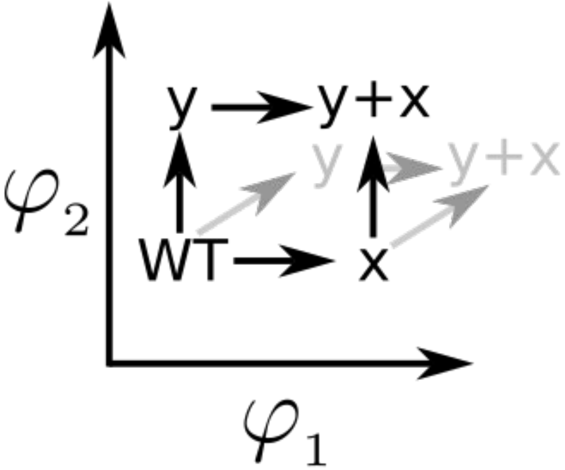
One phenotype — one latent variable

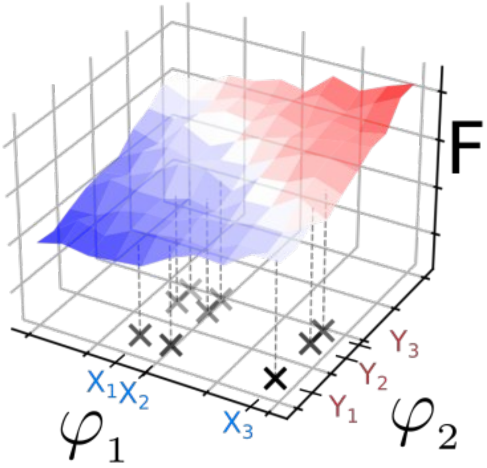
- 1 **gene** ↔ 1 **phenotype** ↔ 1 **latent variable**

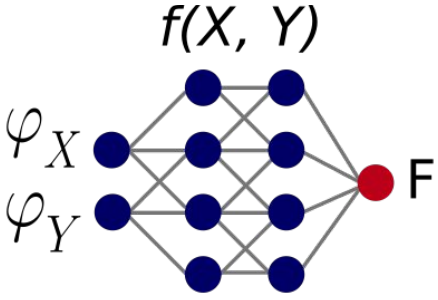


One phenotype — one latent variable

- 1 **gene** ↔ 1 **phenotype** ↔ 1 **latent variable**
- **Fitness** = **nonlinear** function **combining phenotypes**

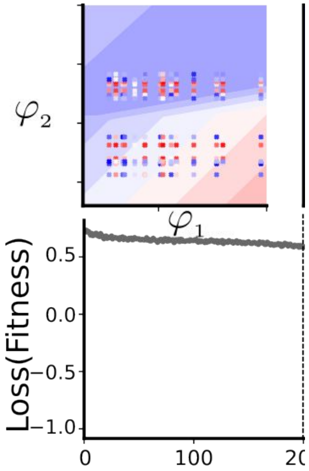






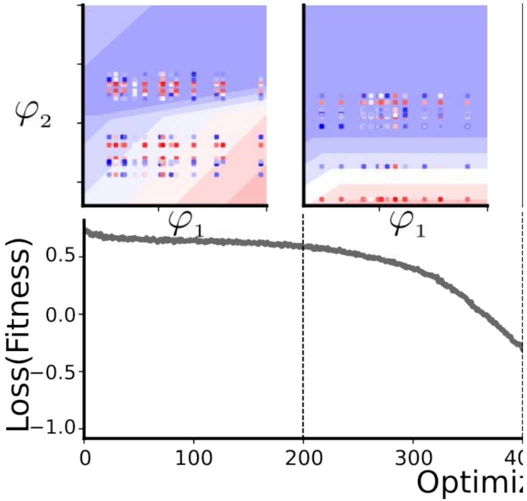
One phenotype — one latent variable

- 1 **gene** ↔ 1 **phenotype** ↔ 1 **latent variable**
- **Fitness** = **nonlinear** function **combining phenotypes**



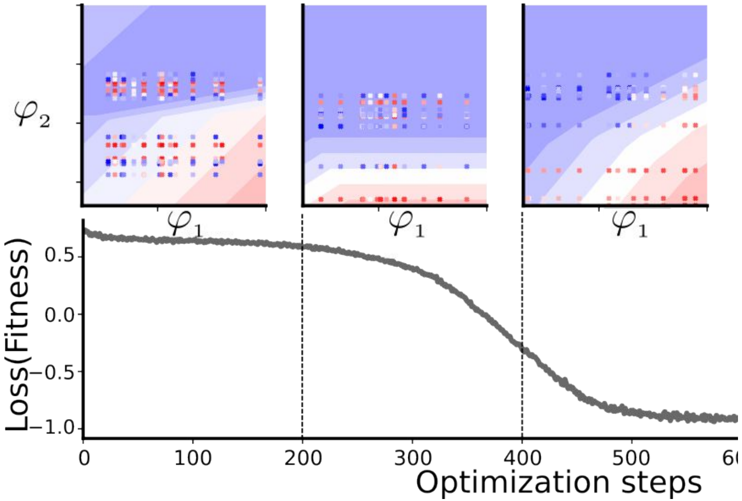
One phenotype — one latent variable

- 1 **gene** ↔ 1 **phenotype** ↔ 1 **latent variable**
- **Fitness** = **nonlinear** function **combining phenotypes**



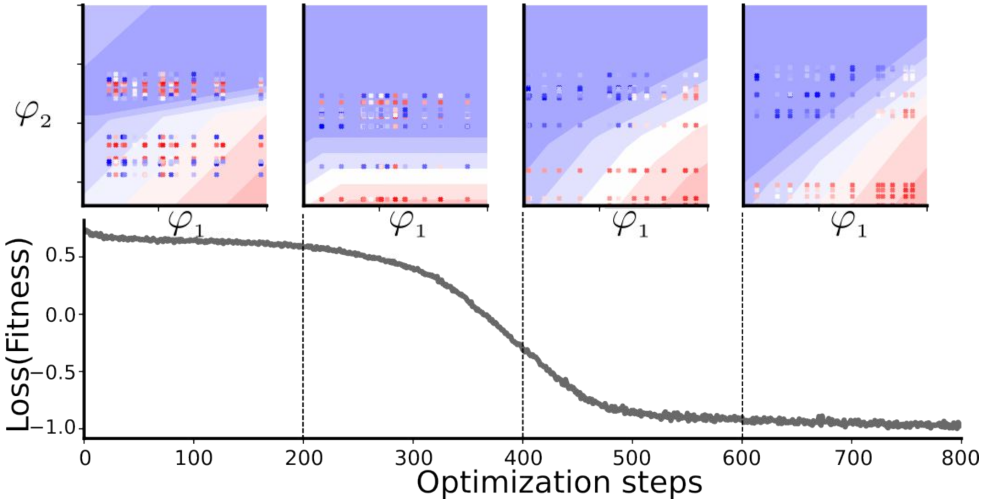
One phenotype — one latent variable

- 1 **gene** ↔ 1 **phenotype** ↔ 1 **latent variable**
- **Fitness** = **nonlinear** function **combining phenotypes**

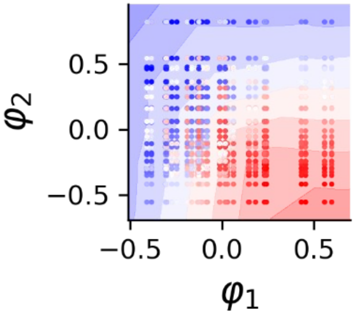


One phenotype — one latent variable

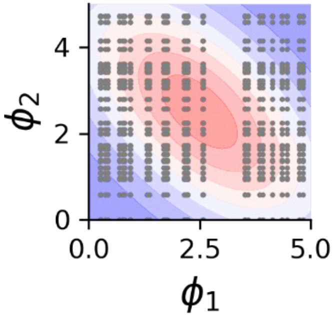
- 1 **gene** ↔ 1 **phenotype** ↔ 1 **latent variable**
- **Fitness** = **nonlinear** function **combining phenotypes**



The challenge of non-monotonous landscapes



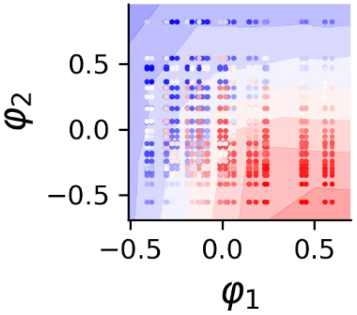
Naive model



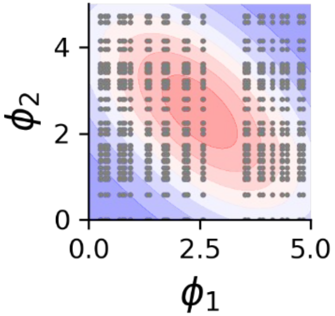
Artificial data

The challenge of non-monotonous landscapes

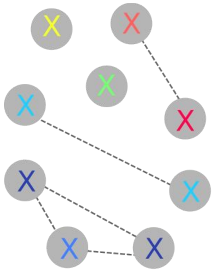
- Construct a graph of mutations
- Spectral initialization (Laplacian)



Naive model



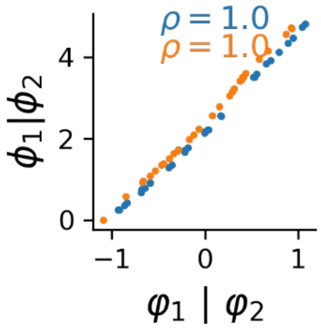
Artificial data

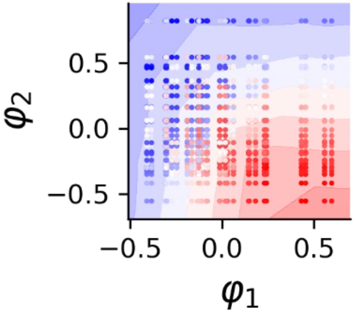


Similarity graph

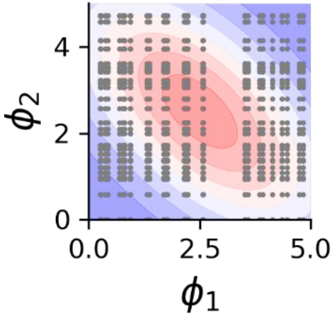
The challenge of non-monotonous landscapes

- Construct a graph of mutations
- Spectral initialization (Laplacian)

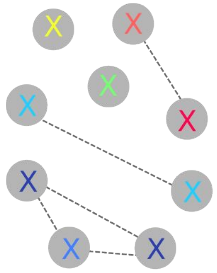




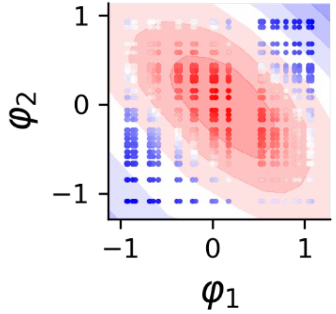
Naive model



Artificial data



Similarity graph



Spectral initialization

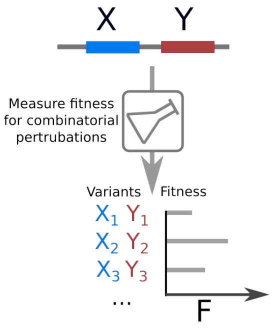
Phenotype inference

Latent variation = phenotypic variation

- Artificial fitness landscape:

$$F(X,Y) = \left(w + \mu\varphi - \frac{\nu}{1/\eta - \varphi}\right)(1 - \theta_X \boxed{X} - \theta_Y \boxed{Y}) \,,$$

Kemble *et al* 2020



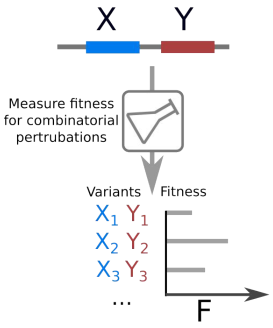
- Assign numerical **phenotypic values**

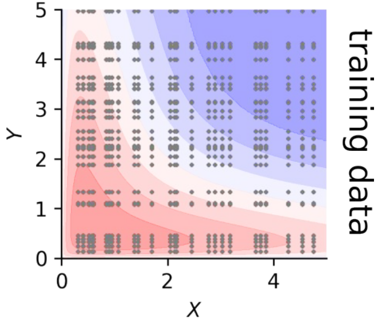
Latent variation = phenotypic variation

- Artificial fitness landscape:

$$F(X,Y) = \left(w + \mu\varphi - \frac{\nu}{1/\eta - \varphi}\right)\left(1 - \theta_X \boxed{X} - \theta_Y \boxed{Y}\right),$$

Kemble *et al* 2020





- Assign numerical **phenotypic values**

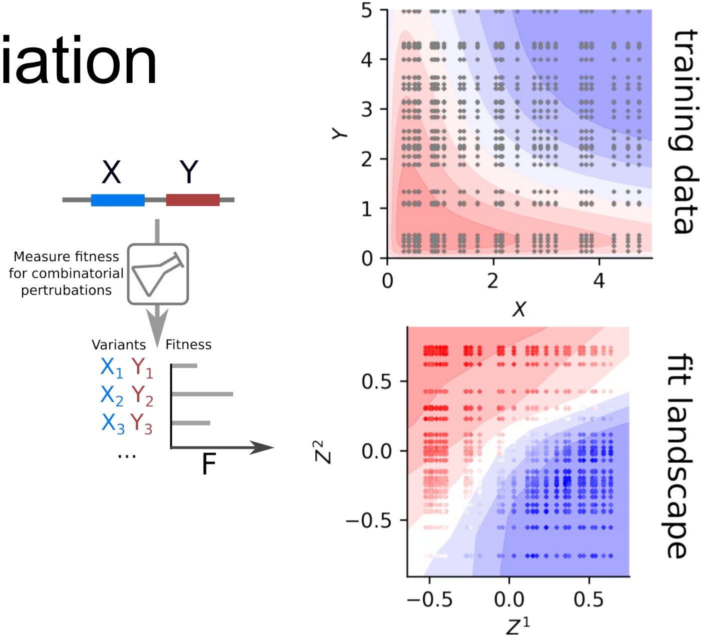
Latent variation = phenotypic variation

- Artificial fitness landscape:

$$F(X,Y) = \left(w + \mu\varphi - \frac{\nu}{1/\eta - \varphi}\right)(1 - \theta_X \boxed{X} - \theta_Y \boxed{Y}),$$

Kemble *et al* 2020

- Assign numerical **phenotypic values**



Measure fitness for combinatorial pertrubations

Latent variation = phenotypic variation

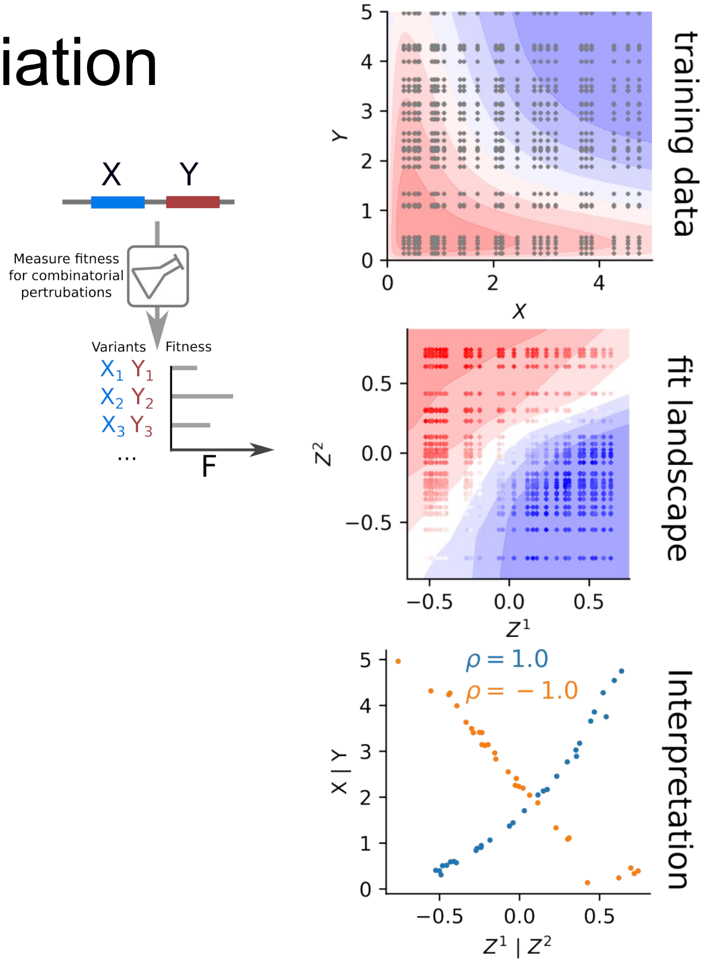
- Artificial fitness landscape:

$$F(X,Y)=\left(w+\mu\varphi-\frac{\nu}{1/\eta-\varphi}\right)(1-\theta_X\boxed{X}-\theta_Y\boxed{Y})\,,$$

Kemble *et al* 2020

- Assign numerical **phenotypic values**

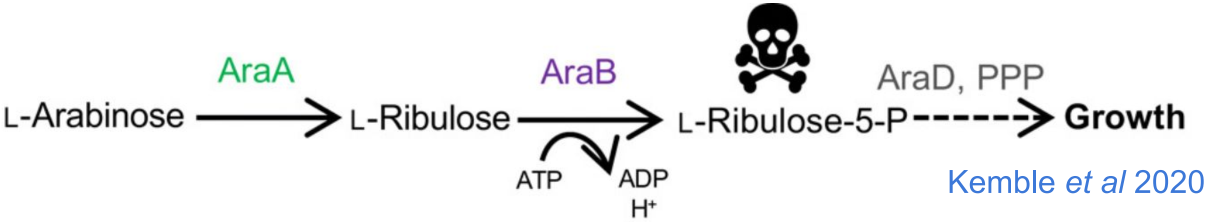
- **Latent variables ∝ phenotype**



Measure fitness for combinatorial pertrubations

Identify trade-offs

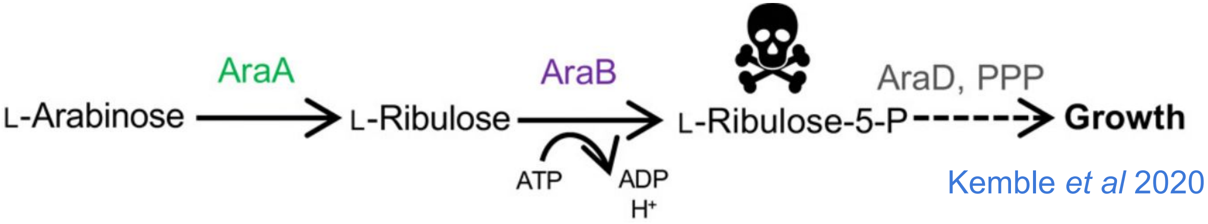
Identify a genetic trade-off



Kemble *et al* 2020

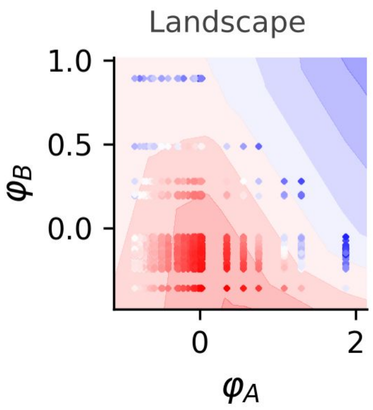
- **2 genes** influencing growth
- **Fitness** measured: **growth** (sequencing/barcoding)

Identify a genetic trade-off

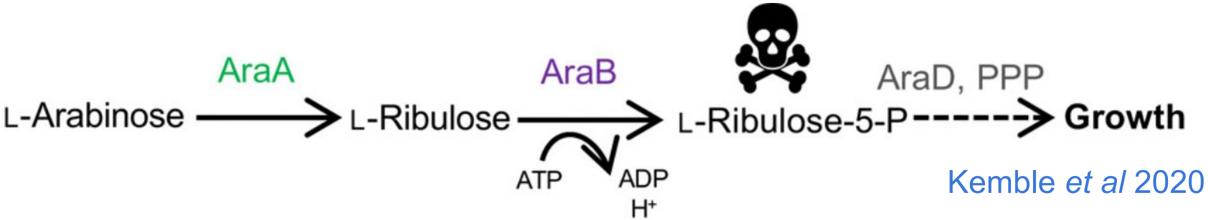


Kemble *et al* 2020

- **2 genes** influencing growth
- **Fitness** measured: **growth** (sequencing/barcoding)

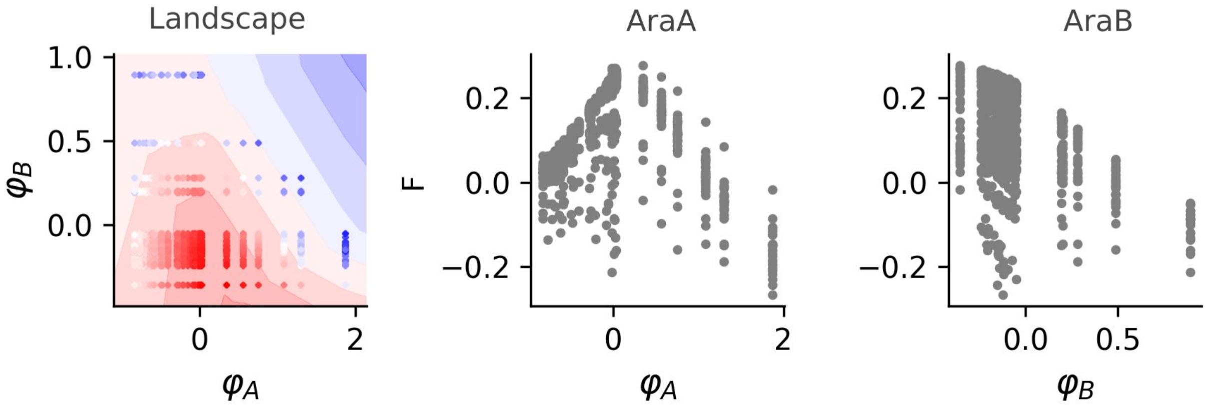


Identify a genetic trade-off

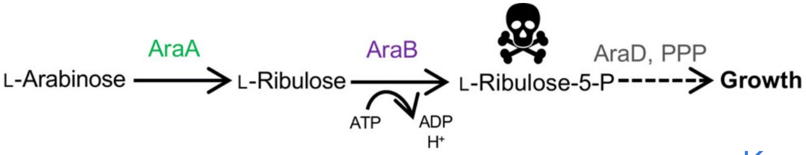


Kemble *et al* 2020

- **2 genes** influencing growth
- **Fitness** measured: **growth** (sequencing/barcoding)



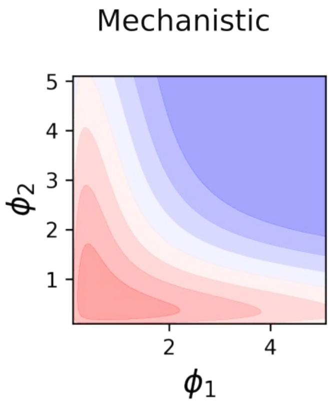
Recovery of biophysical/mechanistic insights



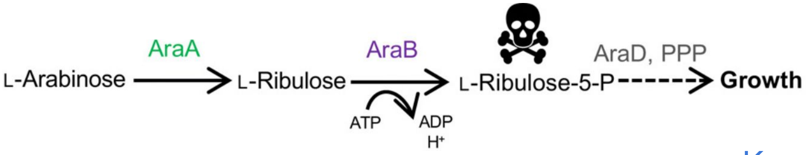
$$F(X,Y) = \left(w + \mu\varphi - \frac{\nu}{1/\eta - \varphi}\right)(1 - \theta_X X - \theta_Y Y)\,,$$

Kemble *et al* 2020

- **Previous biophysical modelling**
- **Latent variable correlated to earlier hypotheses**


Mechanistic

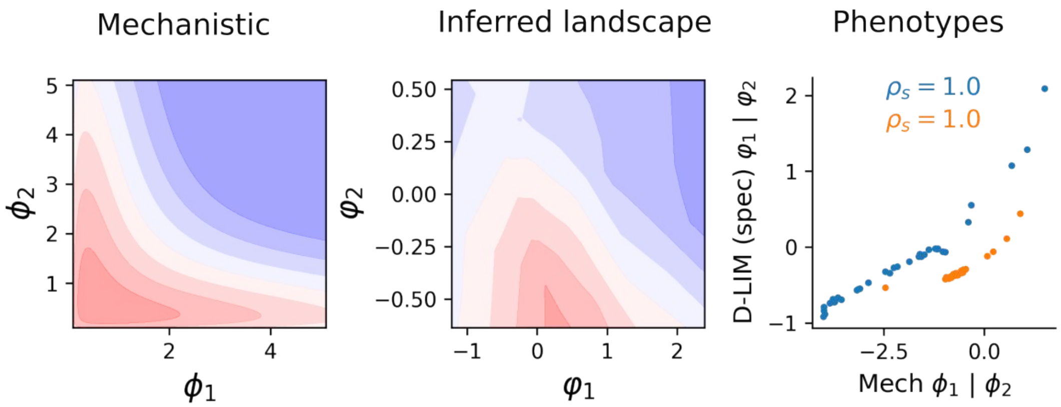
Recovery of biophysical/mechanistic insights



$$F(X,Y) = \left(w + \mu\varphi - \frac{\nu}{1/\eta - \varphi}\right)(1 - \theta_X X - \theta_Y Y)\,,$$

Kemble *et al* 2020

- **Previous biophysical modelling**
- **Latent variable correlated to earlier hypotheses**

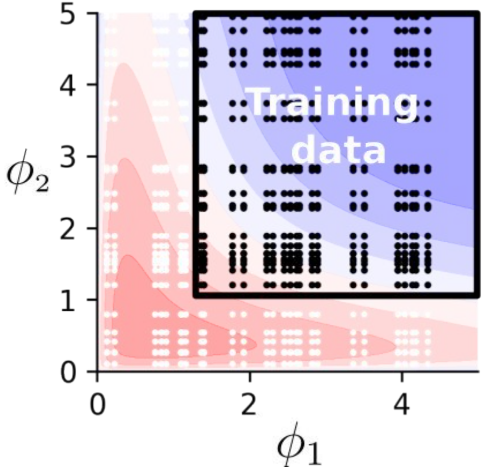


Extrapolate beyond training data

Latent ↔Phenotype map

- No data for high fitness

- Measure phenotypic values
  (Single measures ↔combinatorial measures)

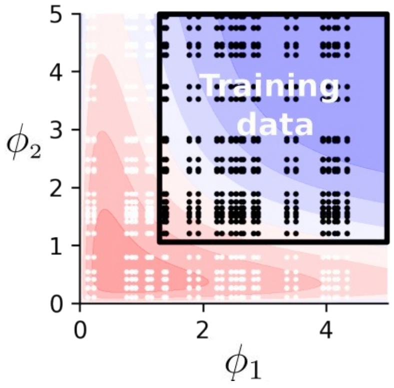


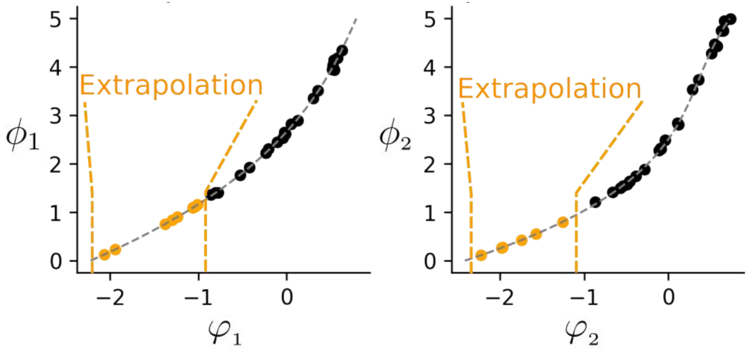
Latent ↔Phenotype map

- No data for high fitness

- Measure phenotypic values
  (Single measures ↔combinatorial measures)

- Fit phenotypic values with latent variables





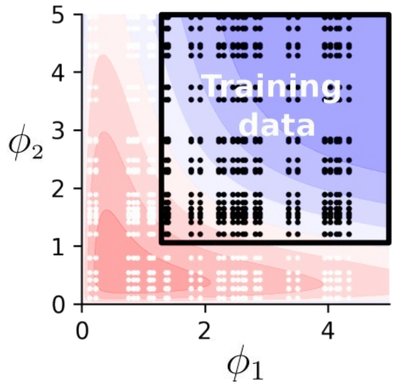
Latent ↔Phenotype map

- No data for high fitness

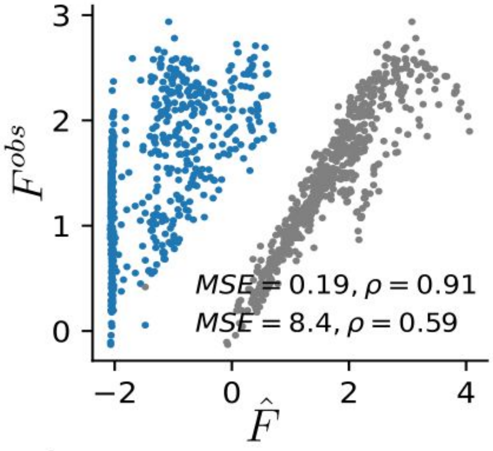
- Measure phenotypic values
  (Single measures ↔combinatorial measures)

- Fit phenotypic values with latent variables

- Infer fitness



Fitting quality



Thank you !

Shuhui Wang, Alexandre Allauzen, Philippe Nghe, Vaitea Opuu

Laboratoire Biophysique et Évolution

cnrs

ESPCI PARIS | PSL