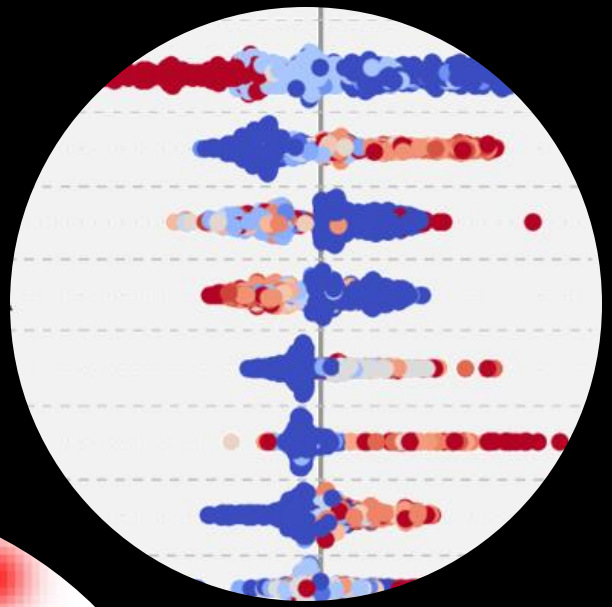
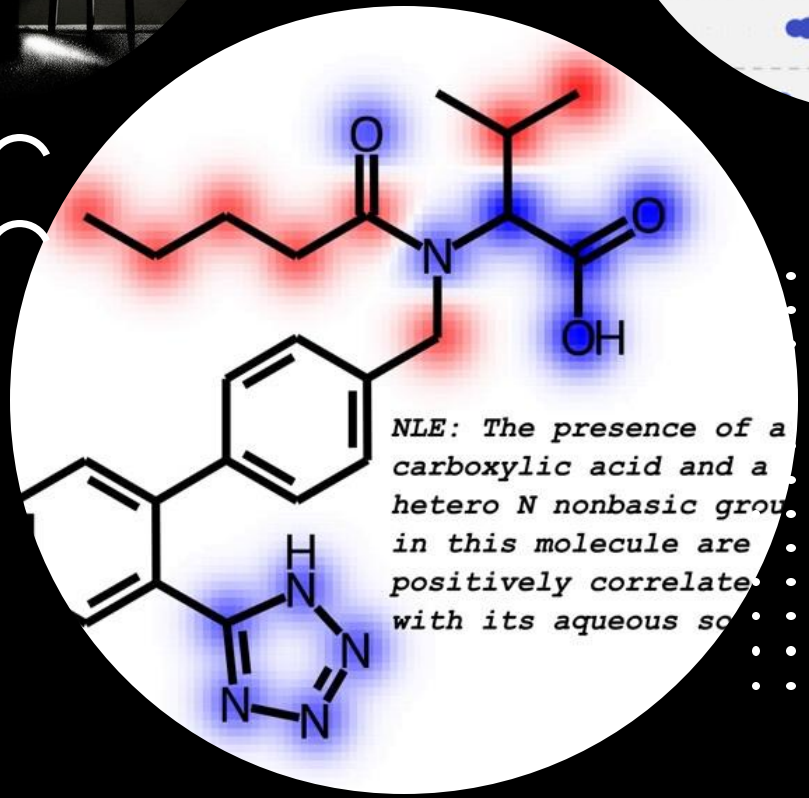
Explainable Artificial Intelligence

how to interrogate a black box?





The what and why's of explainable AI

Explainable AI (XAI) encompasses a broad set of topics that deal with the problem of trying to rationalize, understand, and/or distill meaning from ML models

This is important and under-leveraged:

- XAI makes ML models more practical and practicable
  we are more likely to use models if we can understand why they work
- XAI is important from a regulatory or compliance perspective
  dissemination and action depends on credibility and trust;
  XAI is now becoming a part of guidelines and legislation
- XAI can help prevent misuse
  if we understand why something works, we may anticipate when it won't

In physical science and engineering

- we are often tasked with *understanding*; it is part of our job
- there is bias against the use of black-box models and enhanced scrutiny
- we are probably leaving fruit on the tree; there's juice still to squeeze; etc.

Interpretation vs. Justification vs. Explanation

There are many useful routes in XAI; however, we should be clear on what it is that we are actually explaining!

- **Interpretation:** why has the model made the prediction it has made? Here, we are ascertaining an underlying cause for the decision. This explains the model but not the phenomena.
- **Justification:** why should the prediction be trusted? Here, we are considering whether there are valid reasons to support the model's decision.
- **Explanation:** how is the prediction actually linked to the input? Why are things the way that they are. This is more distilling cause and effect at the meta level. Such explanations usually afford an understanding that is actionable.

Explanations via post hoc analysis

ML models and their parameters can themselves be difficult to interpret; often, “explanations” are formulated via *post hoc* analysis

- **Training data importance:** what data is is most important for formulating a given prediction? This provides context for a prediction.
- **Feature Importance:** what factors (usually of the input) were most influential in making the prediction? Analysis usually reveals important trends of the data.
- **Counterfactual analysis:** how can predictions/outcomes minimally differ and what changes? That’s the reason! right?

Training Data Importance

ML models are the way they are in response to the training data.
Therefore, we can posit that the prediction being the way it is should be traceable to some subset of the training data.

Basic idea: evaluate predictions in the presence/absence of training with certain data

$$\mathcal{I}(\boldsymbol{x}_i, \boldsymbol{x}) = \hat{f}_{-\boldsymbol{x}_i}(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})$$

"influence" of $\boldsymbol{x}_i$

*This possesses conceptual similarities to **ablation analysis***
(or is perhaps a specific manifestation thereof);
ablation analysis is usually used to identify important factors in performance

Feature Importance Assessment

Feature importance is currently the most straightforward and common way of interpreting the outcomes of ML; there are both global and local variants

Basic formulations:

- *How responsive is the output to the inputs (gradient-esque evaluation)?*

  consider a linear function: $\hat{y} = \boldsymbol{\theta}^T \boldsymbol{x} + \boldsymbol{b}$

  non-liner conceptual extrapolation: $\hat{f}(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x}') + \nabla \hat{f}(\boldsymbol{x}')(\boldsymbol{x} - \boldsymbol{x}')$

  can use automatic differentiation methods or packages to implement; direct gradients may not be useful!

- *How is a prediction formulated in terms of relative contributions?*

  Shapley Additive Explanations is a very common variant

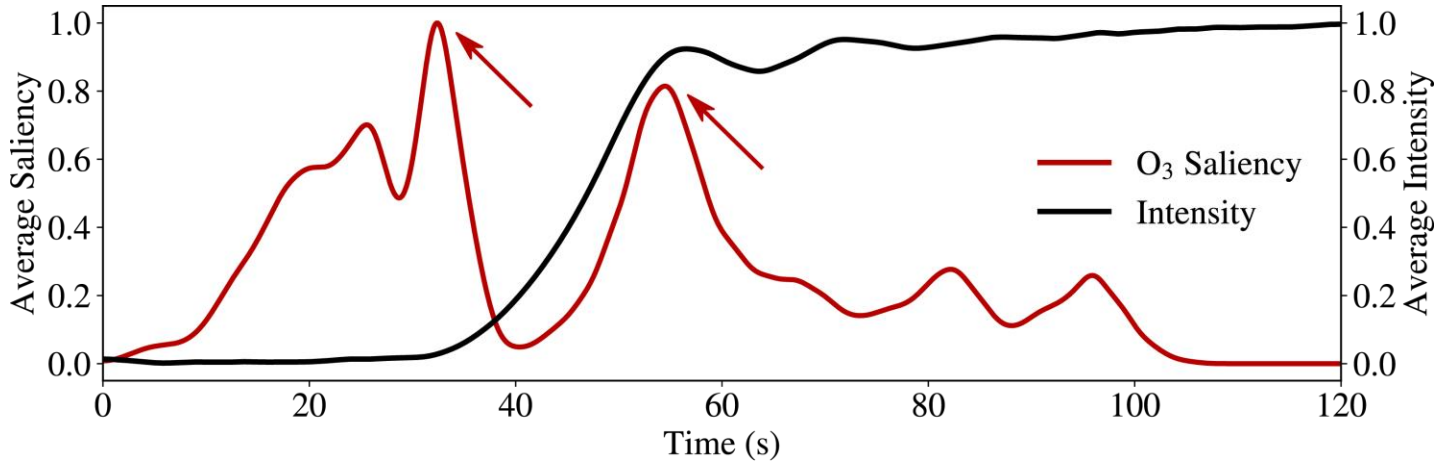
$$\hat{f}(\boldsymbol{x}) = \sum_i \phi_i(\boldsymbol{x}) \qquad \phi_i(\boldsymbol{x}) = \frac{1}{Z} \sum_{S \in N - \{x_i\}} v(S \cup x_i) - v(S)$$

SHAP analysis is model-agnostic, which is nice!

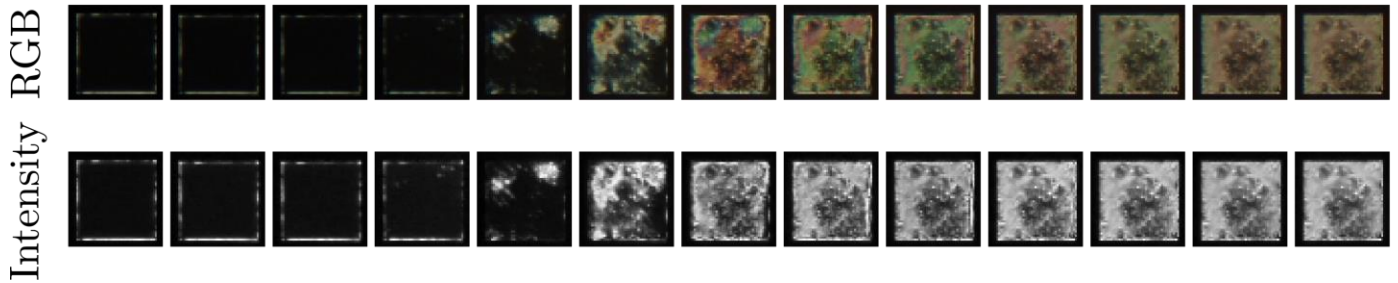
$$\phi_i(\boldsymbol{x}) \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{f}(\boldsymbol{z}_{+i}(\sim \boldsymbol{x}_m)) - \hat{f}(\boldsymbol{z}_{-i}(\sim \boldsymbol{x}_m)))$$

Saliency/Gradient Analysis: What's important?

$$S = \frac{\partial F(x)}{\partial x} = \frac{\partial \text{ Model output}}{\partial \text{ Model input}}$$



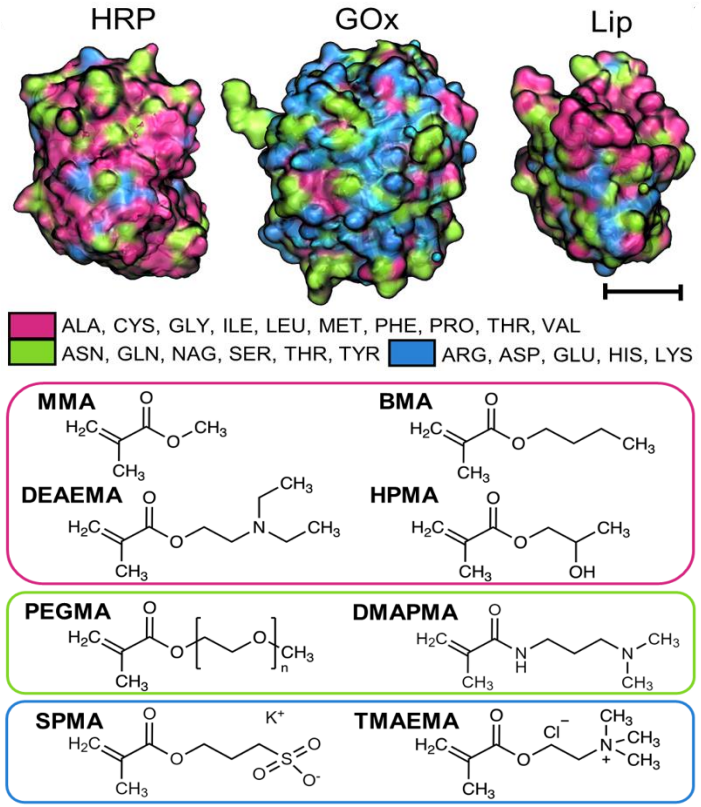
optical response of LC array

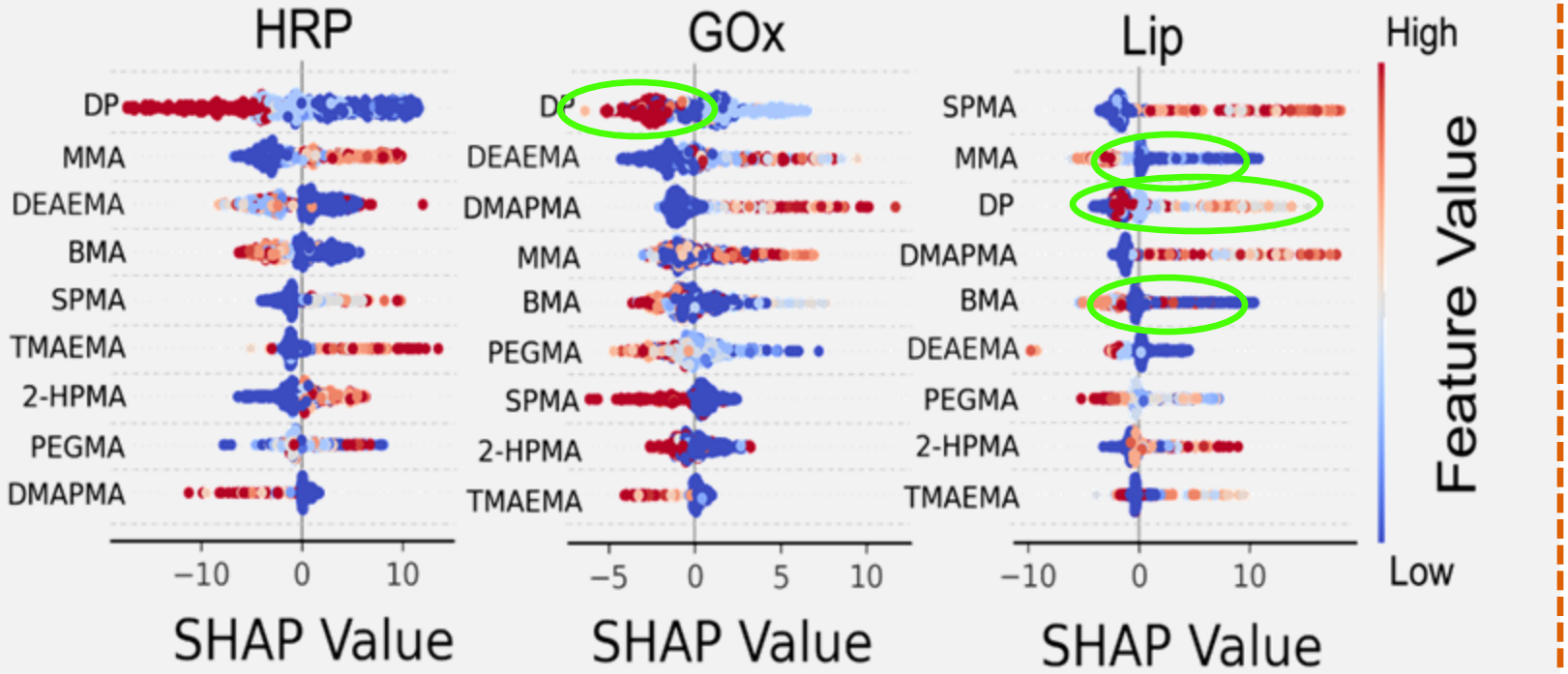


SHAP Analysis Plots

- enzymes are exposed to ***thermal stress*** that would denature enzyme alone
- GPR to predict ***retained enzyme activity*** (output) from ***copolymer features*** (input)

$$\hat{f}(\boldsymbol{x}) = \sum_i \phi_i(\boldsymbol{x}) \qquad \phi_i(\boldsymbol{x}) = \frac{1}{Z} \sum_{S \in N - \{x_i\}} v(S \cup x_i) - v(S)$$





Impact on REA prediction

$$\phi_i(\boldsymbol{x}) \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{f}(\boldsymbol{z}_{+i}(\sim \boldsymbol{x}_m)) - \hat{f}(\boldsymbol{z}_{-i}(\sim \boldsymbol{x}_m)))$$

Tamasi†, Patel†, Kosuri†, Borca† et al. "Machine Learning on a Robotic Platform for the Design of Protein-Polymer Hybrids" *Advanced Materials* **(2022)**

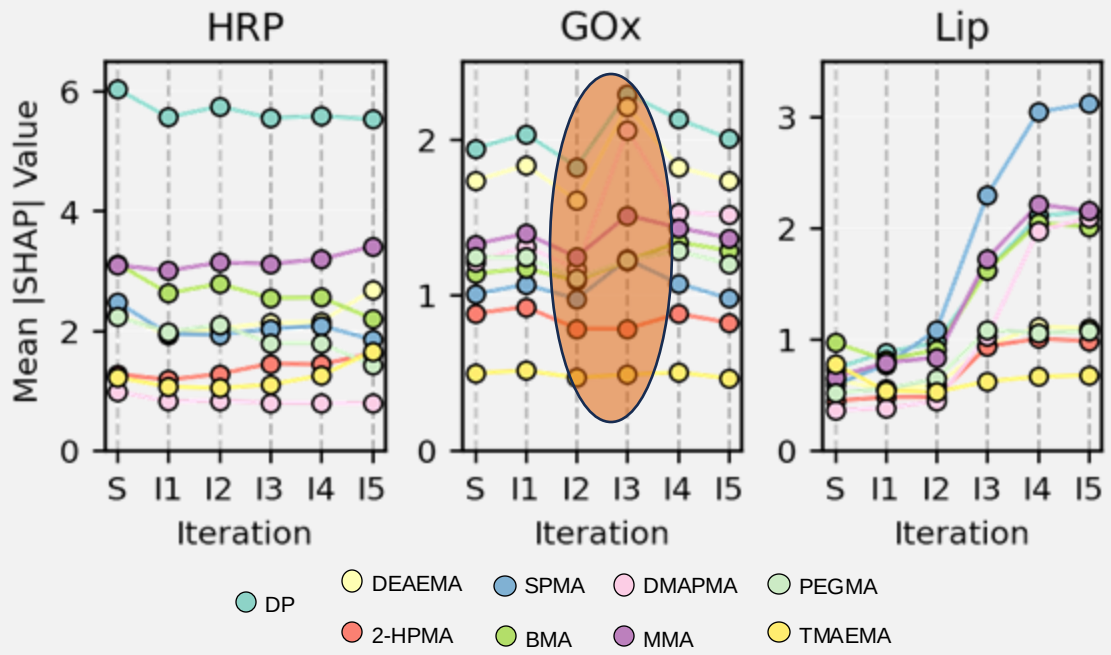
SHAP Analysis Plots

Feature Importance by SHapley Additive exPlanations

Relative Importance of Polymer Length

Each enzyme displays unique and intriguing preferences on the degree of polymerization

Evolution of Importance during Active Learning



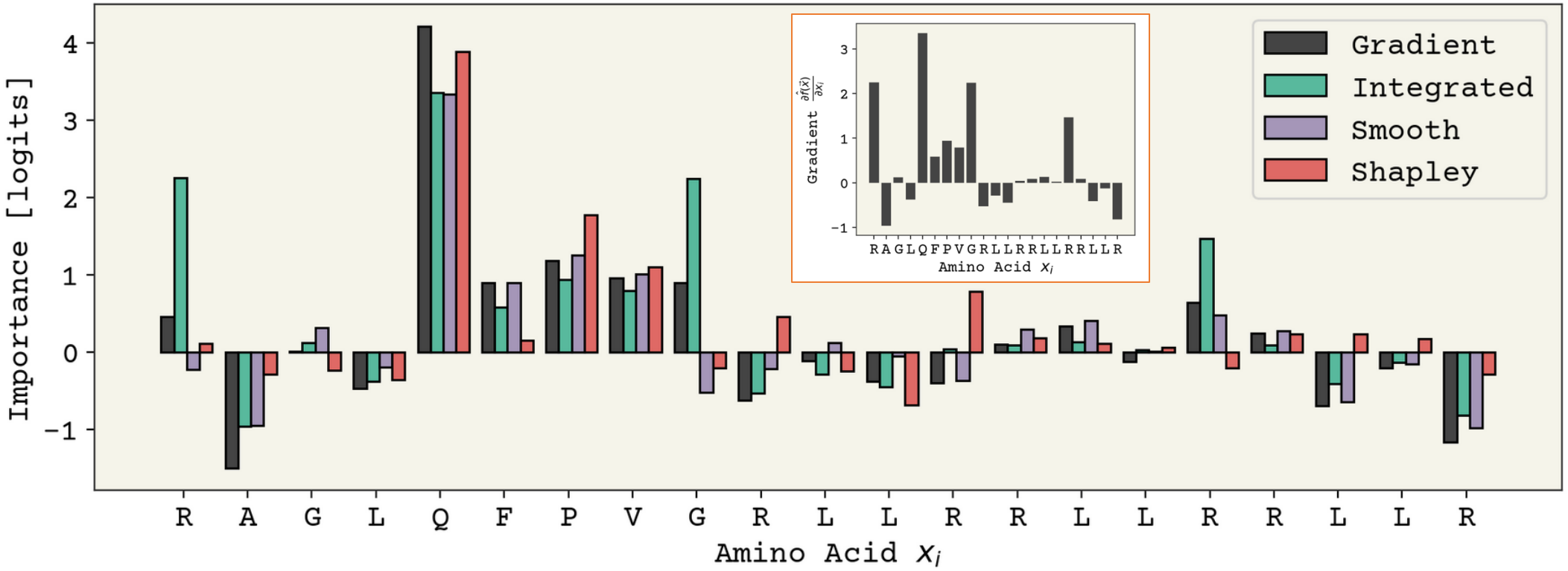
Discovery of high-performing systems can be a steady march or be a challenging excursion

Tamasi†, Patel†, Kosuri†, Borca† et al. “Machine Learning on a Robotic Platform for the Design of Protein-Polymer Hybrids” *Advanced Materials* **(2022)**

Comparison of Feature Importance Methods

Task: is a peptide homolytic (i.e., will it kill red blood cells) ?

deep learning for
molecules & materials



Counterfactual Analysis

Counterfactual analysis aims to provide an explanation by comparison. Given a particular example/behavior, the idea is to find something that is close in all respects except the behavior to be explained.

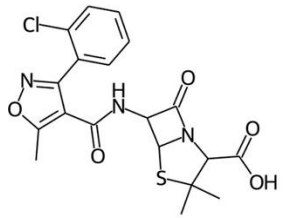
$$\min d(\boldsymbol{x}, \boldsymbol{x}')$$

such that

$$\hat{f}(\boldsymbol{x}) \neq \hat{f}(\boldsymbol{x}') \text{ or } \hat{f}(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}') > \delta$$

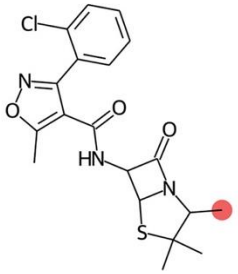
permeate blood-brain barrier?

Base
f(x) = 0.000



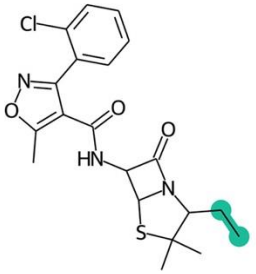
no

Similarity = 0.80
Counterfactual 1
f(x) = 1.000



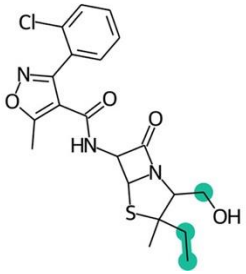
yes

Similarity = 0.77
Counterfactual 2
f(x) = 1.000



yes

Similarity = 0.71
Counterfactual 3
f(x) = 1.000



yes

Counterfactual Analysis

Counterfactual analysis aims to provide an explanation by comparison. Given a particular example/behavior, the idea is to find something that is close in all respects except the behavior to be explained.

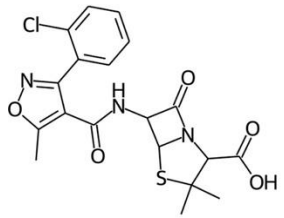
$$\min d(\boldsymbol{x}, \boldsymbol{x}')$$

such that

$$\hat{f}(\boldsymbol{x}) \neq \hat{f}(\boldsymbol{x}') \text{ or } \hat{f}(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}') > \delta$$

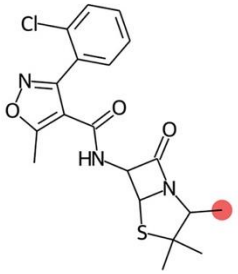
permeate blood-brain barrier?

Base
f(x) = 0.000



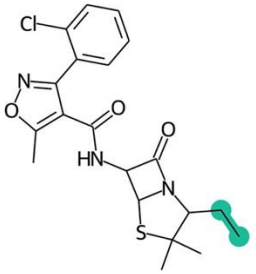
no

Similarity = 0.80
Counterfactual 1
f(x) = 1.000



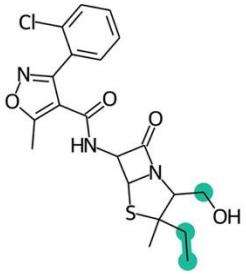
yes

Similarity = 0.77
Counterfactual 2
f(x) = 1.000



yes

Similarity = 0.71
Counterfactual 3
f(x) = 1.000



yes

Counterfactual Analysis

Counterfactual analysis aims to provide an explanation by comparison. Given a particular example/behavior, the idea is to find something that is close in all respects except the behavior to be explained.

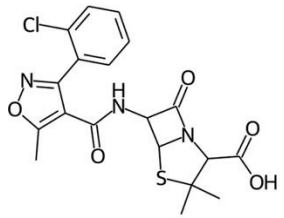
$$\min d(\boldsymbol{x}, \boldsymbol{x}')$$

such that

$$\hat{f}(\boldsymbol{x}) \neq \hat{f}(\boldsymbol{x}') \text{ or } \hat{f}(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}') > \delta$$

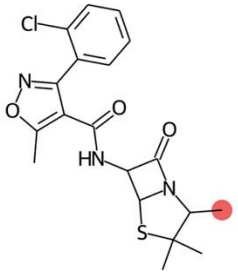
permeate blood-brain barrier?

Base
f(x) = 0.000



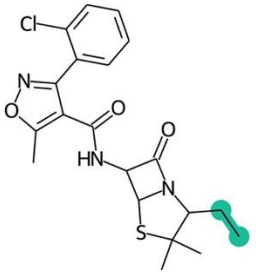
no

Similarity = 0.80
Counterfactual 1
f(x) = 1.000



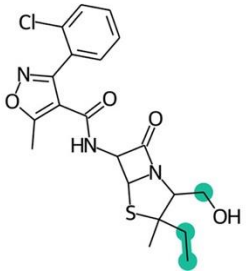
yes

Similarity = 0.77
Counterfactual 2
f(x) = 1.000



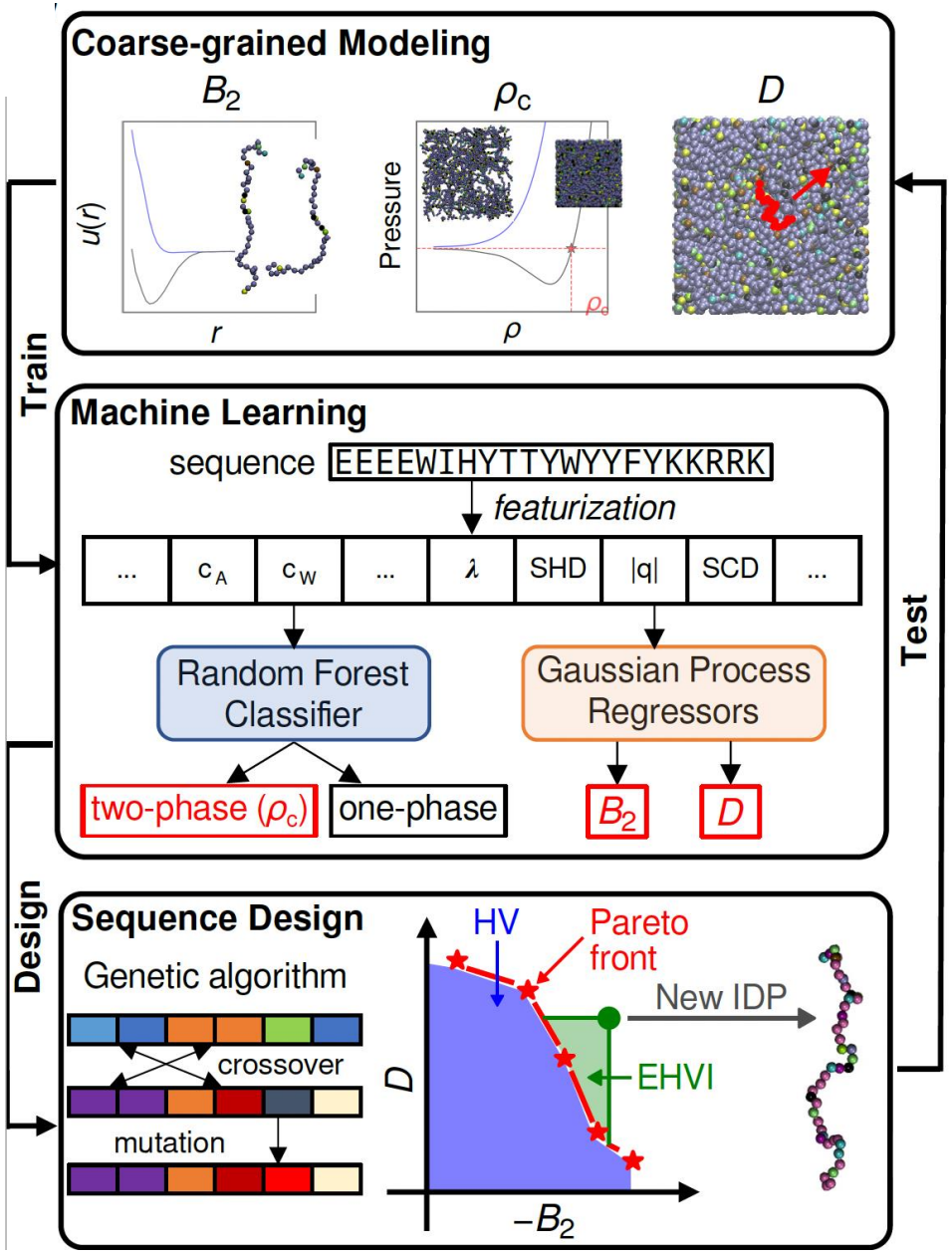
yes

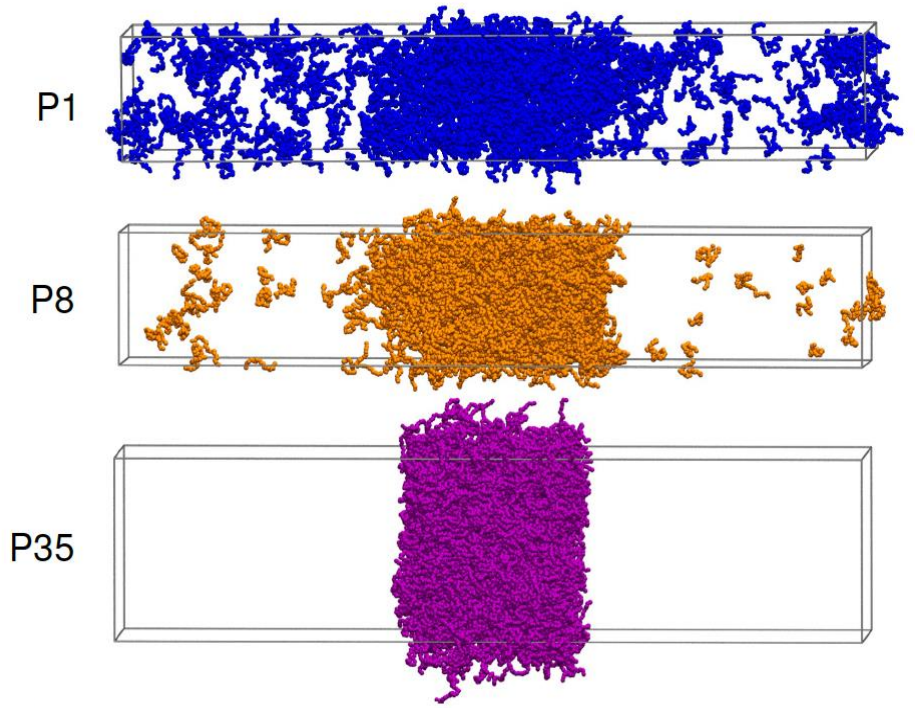
Similarity = 0.71
Counterfactual 3
f(x) = 1.000

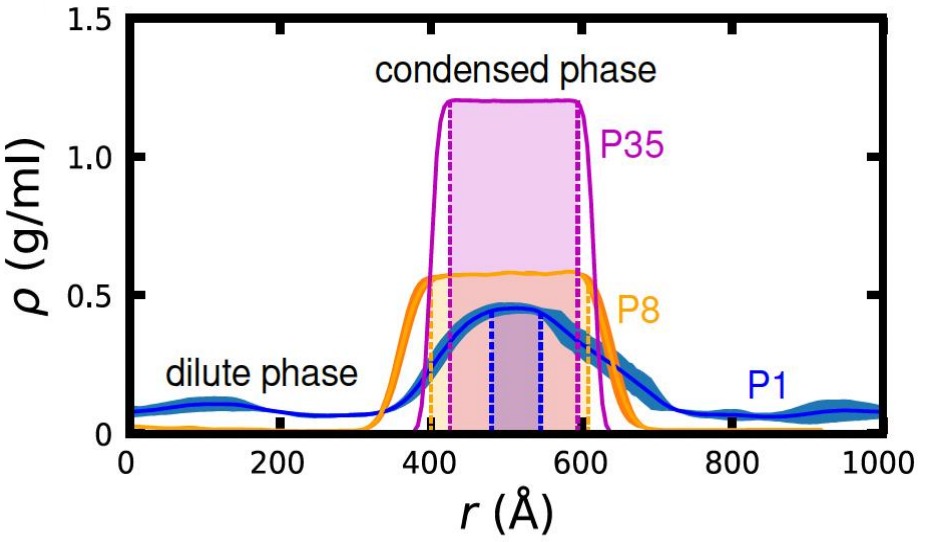


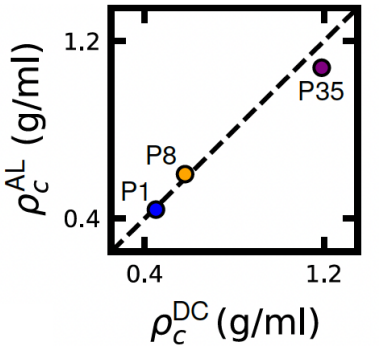
yes

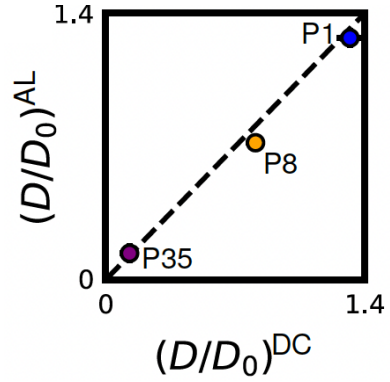
Counterfactual Analysis











brute-force

3 replicas
1000 chains
~10 μs simulation

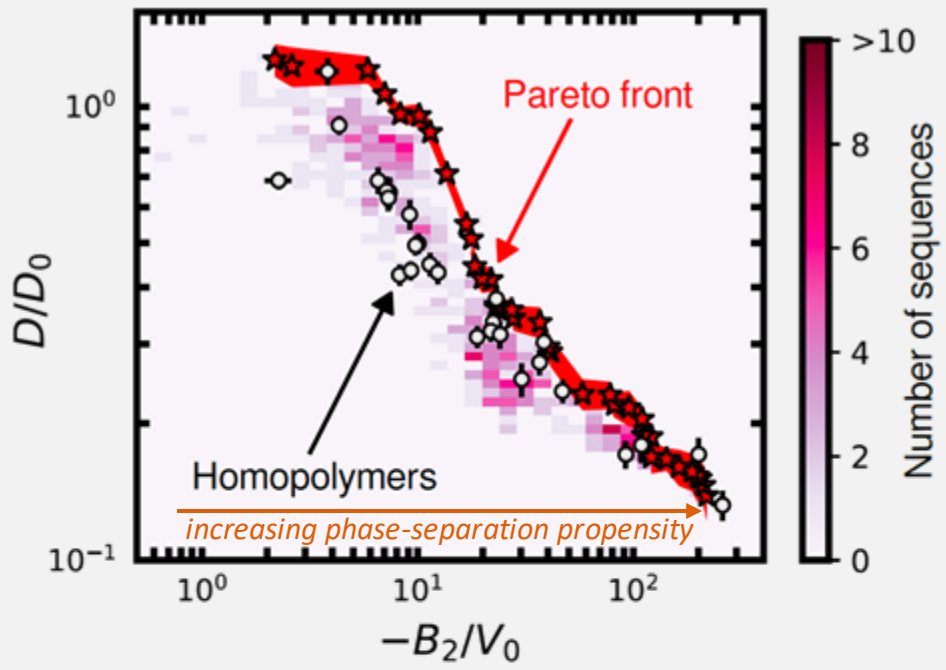
vs.

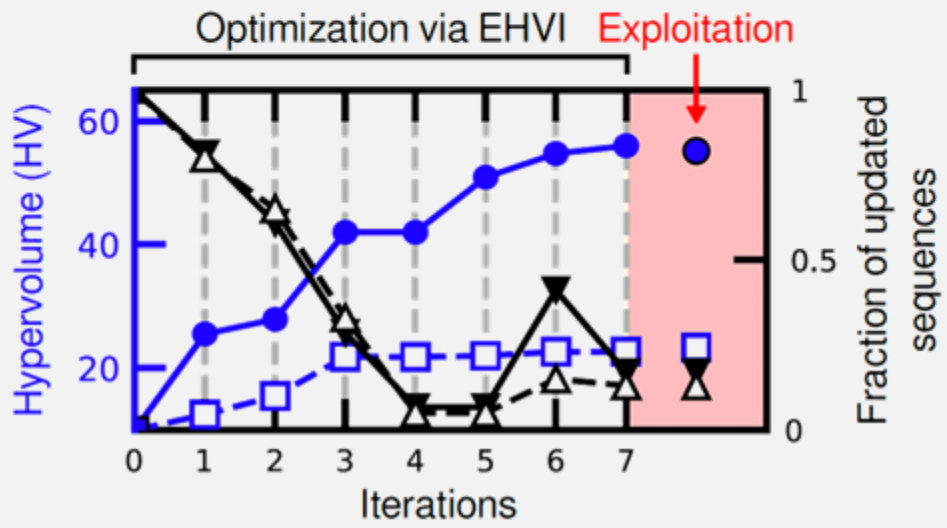
ML-assisted

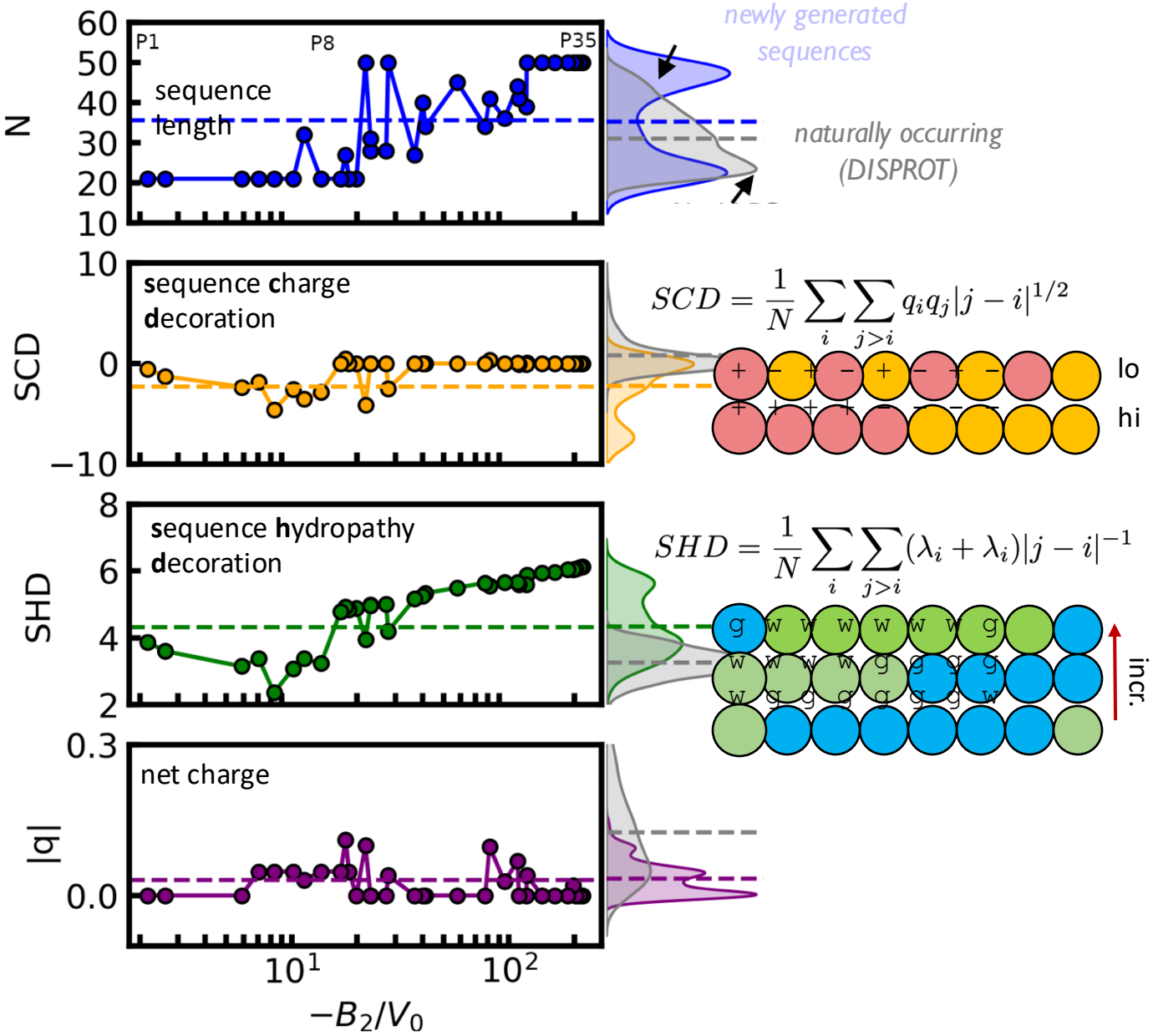
30 replicas
100 chains
100 ns simulation

150x faster

Counterfactual Analysis

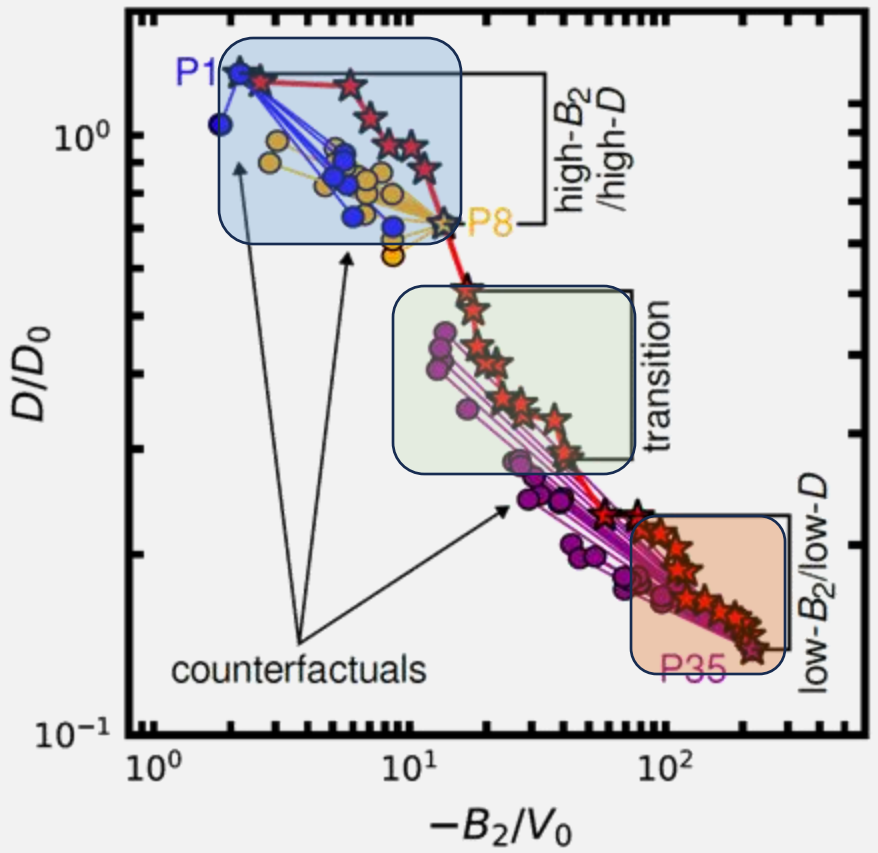






Counterfactual Analysis

*The notion of a **counterfactual** can provide some insight as to why any particular sequence is "on the Front" compared to a "close" sequence that is not*



- **high-high**: polyampholytes with subtle sequence variations
- **transition**: highly degenerate region with many options
- **low-low**: neutral sequences with strong hydrophobic patterning

$$\langle \Delta x_k \rangle \equiv \frac{1}{n_{\mathcal{C}_{ij}}} \sum_{j=1}^{n_{\mathcal{C}_{ij}}} (x_k^{(\mathcal{P}_i)} - x_k^{(\mathcal{C}_{ij})})$$