

CAGI 7 slide pack guidelines

I agree to my slides being used according to the following terms:
Gianluca Lombardi

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- *At CAGI 7, use of this slide pack is only required for CAGI dataset providers, assessors, and predictors, but all are encouraged to use the annotations for materials they do not want shared publicly.*
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 7. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted, tooted, posted, X'd unless flagged with "No tweet" icon
- CAGI participants using these slides must explain the "No tweet" icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Combining PRESCOTT and deep learning models for improved zero-shot predictions of missense mutations

Gianluca Lombardi^{1,2}, Mustafa Tekpinar³ and Alessandra Carbone^{1,2}

¹CQSB, Sorbonne Université and CNRS, Paris, France; ²SCAI, Sorbonne Université, Paris, France,

³Department of Physics, Van Yüzüncü Yıl University, Van, Türkiye



Mutations and SNVs

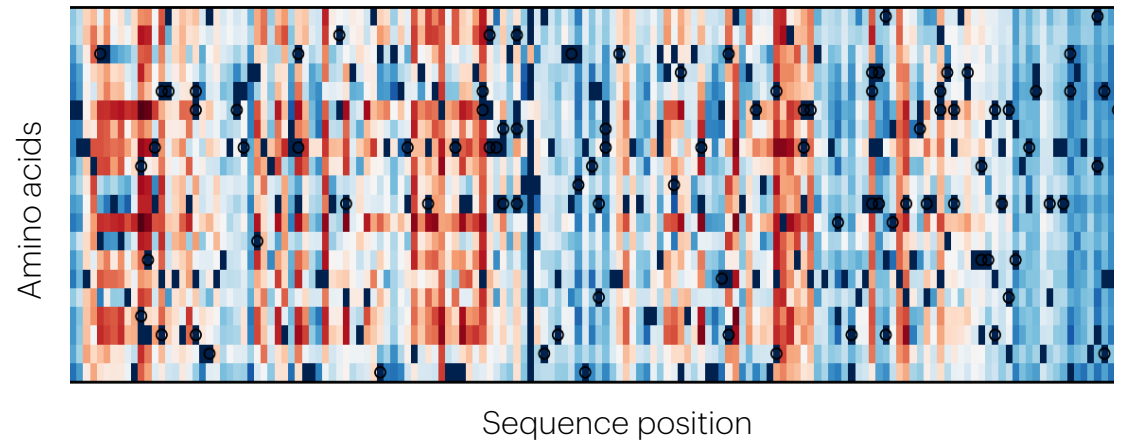
- **71 million** possible missense variants in the human proteome, with more than **4 million observed** missense variants in human populations
- **2%** have been **clinically classified** as **pathogenic** or **benign**, while the vast majority are **VUS**

Mutations and SNVs

- **71 million** possible missense variants in the human proteome, with more than **4 million observed** missense variants in human populations
- **2%** have been **clinically classified** as **pathogenic** or **benign**, while the vast majority are **VUS**

Development of computational methods, based on bioinformatics algorithms and machine learning models, to predict mutational effects **at large scale** and with increasing accuracy

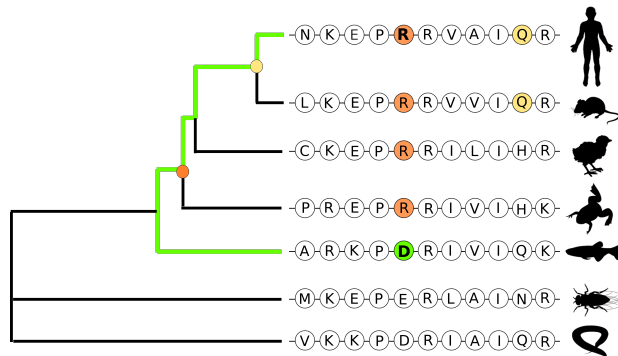
DMS reconstruction



Modelling priors in missense mutations prediction

1. **Evolutionary conservation:** mutations appearing very rarely in nature are likely deleterious
2. **Epistasis:** Effect of a mutations depends on surrounding sequence context

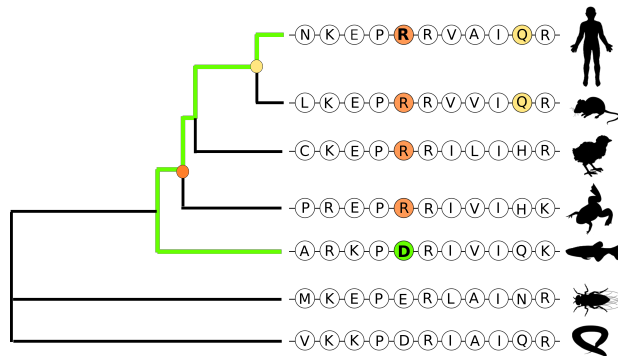
Evolution



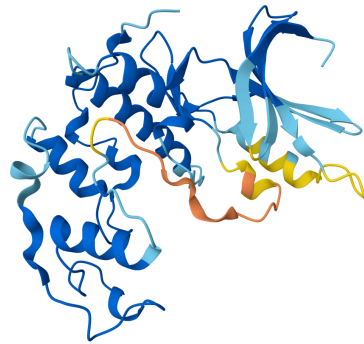
Modelling priors in missense mutations prediction

1. **Evolutionary conservation:** mutations appearing very rarely in nature are likely deleterious
2. **Epistasis:** Effect of a mutations depends on surrounding sequence context
3. **Structure:** mutations in the **core** of a protein and in **interaction surfaces** have higher impact

Evolution



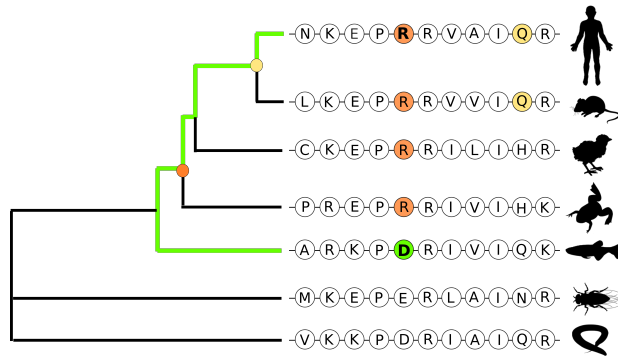
Structure



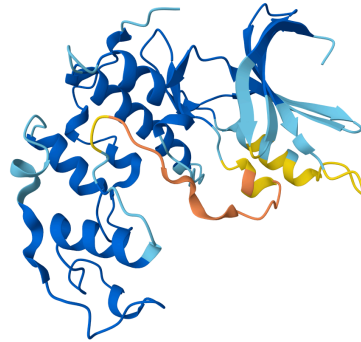
Modelling priors in missense mutations prediction

1. **Evolutionary conservation:** mutations appearing very rarely in nature are likely deleterious
2. **Epistasis:** Effect of a mutations depends on surrounding sequence context
3. **Structure:** mutations in the **core** of a protein and in **interaction surfaces** have higher impact
4. **Spread** of a mutation in human populations is informative on its pathogenicity

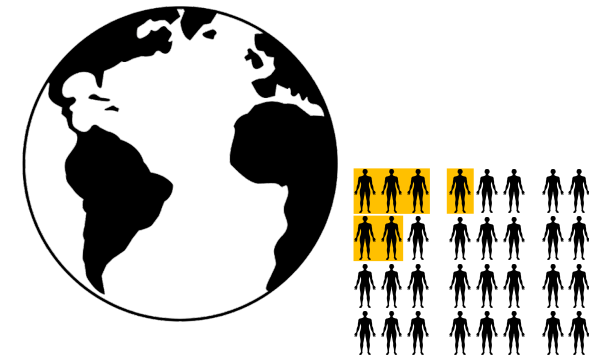
Evolution



Structure



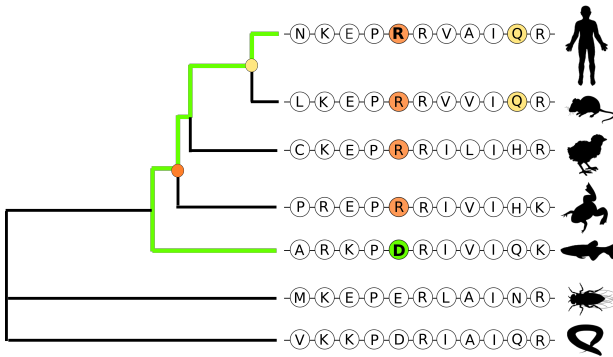
Allele Frequency



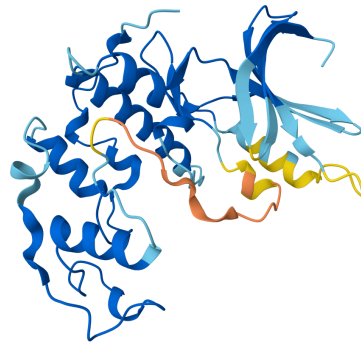
Modelling priors in missense mutations prediction

1. **Evolutionary conservation:** mutations appearing very rarely in nature are likely deleterious
2. **Epistasis:** Effect of a mutations depends on surrounding sequence context
3. **Structure:** mutations in the **core** of a protein and in **interaction surfaces** have higher impact
4. **Spread** of a mutation in human populations is informative on its pathogenicity

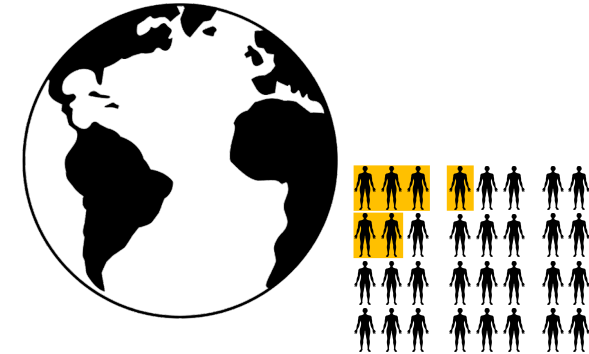
Evolution



Structure

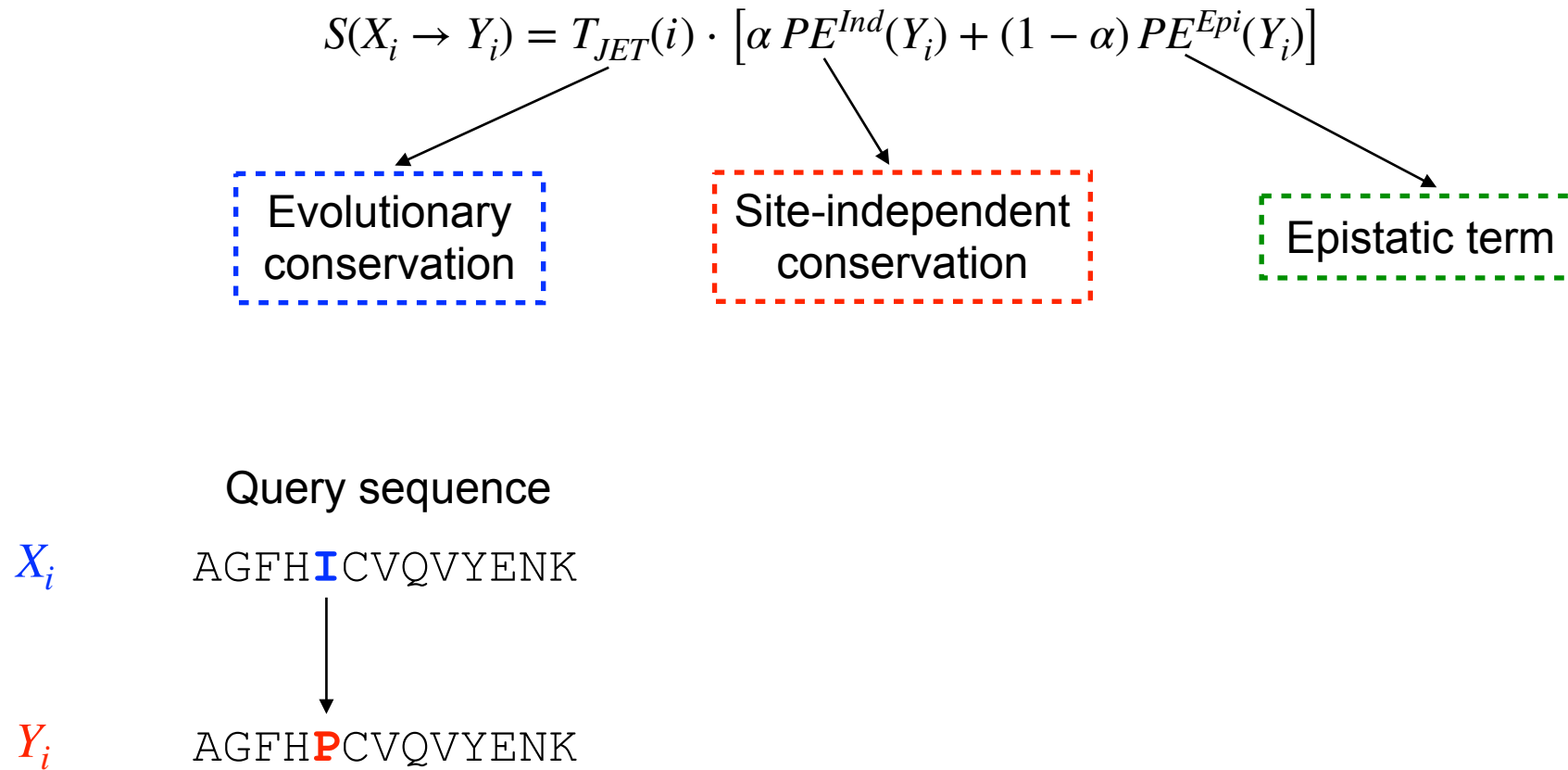


Allele Frequency

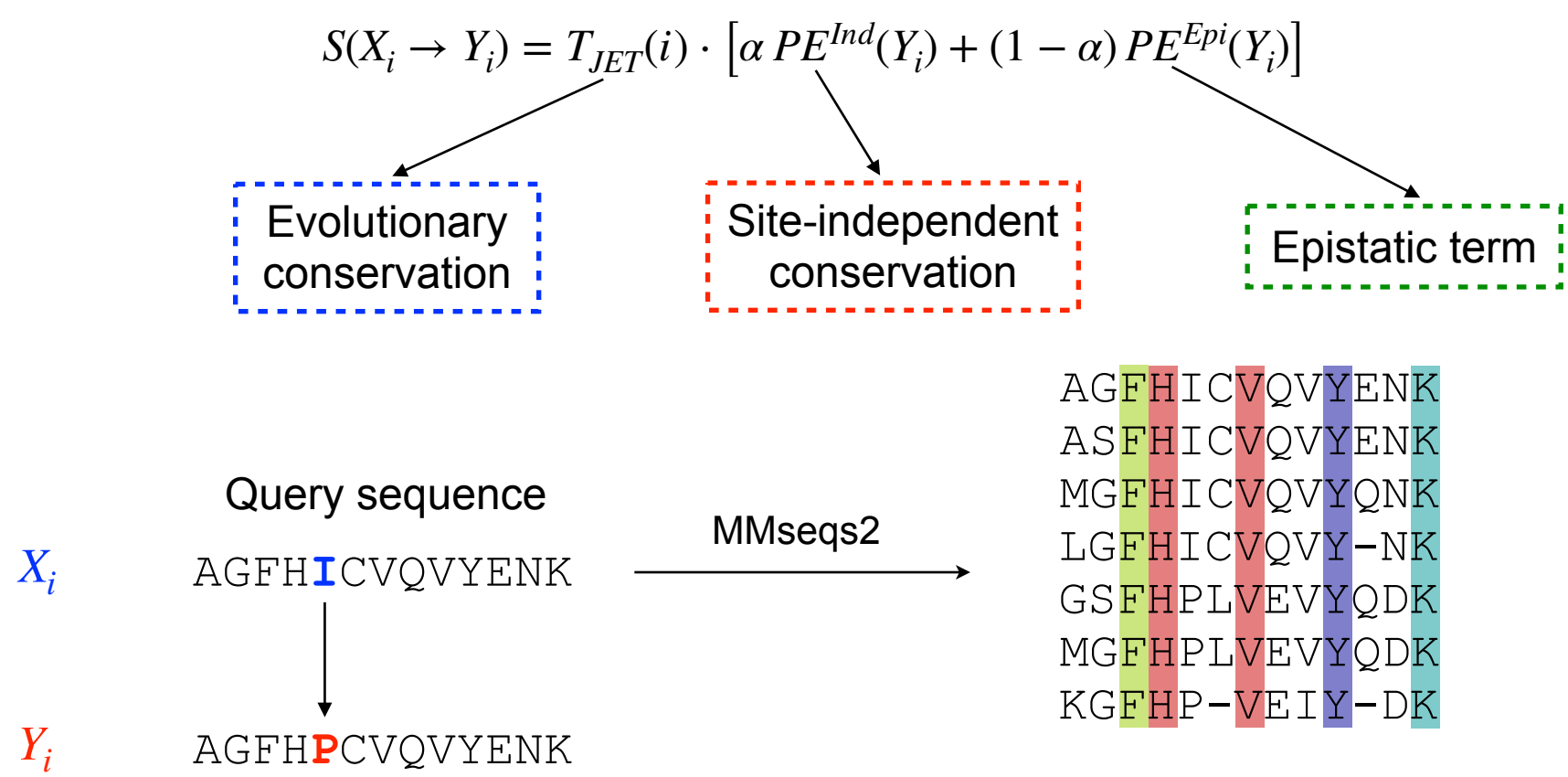


5. **Generalisation** even when alignment is shallow or structure has low confidence

GEMME: Evolution

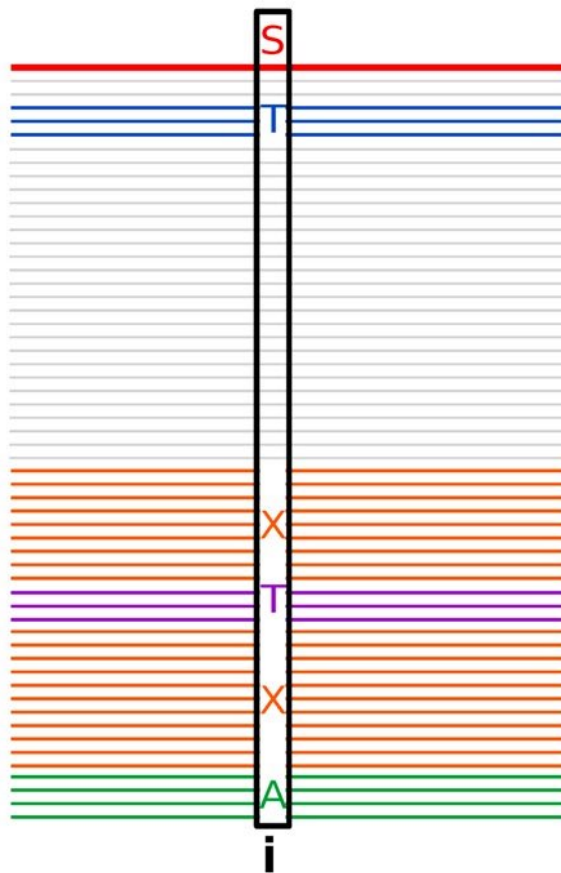


GEMME: Evolution



GEMME: Evolution

$$S(X_i \rightarrow Y_i) = T_{JET}(i) \cdot [\alpha PE^{Ind}(Y_i) + (1 - \alpha) PE^{Epi}(Y_i)]$$

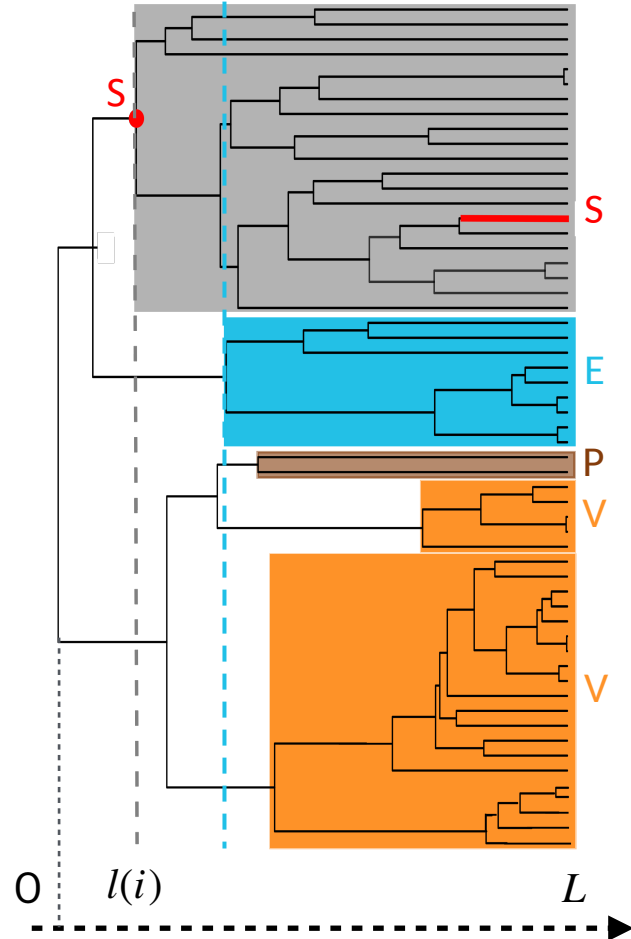


Conservation

$$PE^{Ind}(Y_i) = -\log \frac{\max(1, |S_{Y_i}|)}{|S_{X_i}|}$$

GEMME: Evolution

$$S(X_i \rightarrow Y_i) = \boxed{T_{JET}(i)} \cdot [\alpha PE^{Ind}(Y_i) + (1 - \alpha) PE^{Epi}(Y_i)]$$

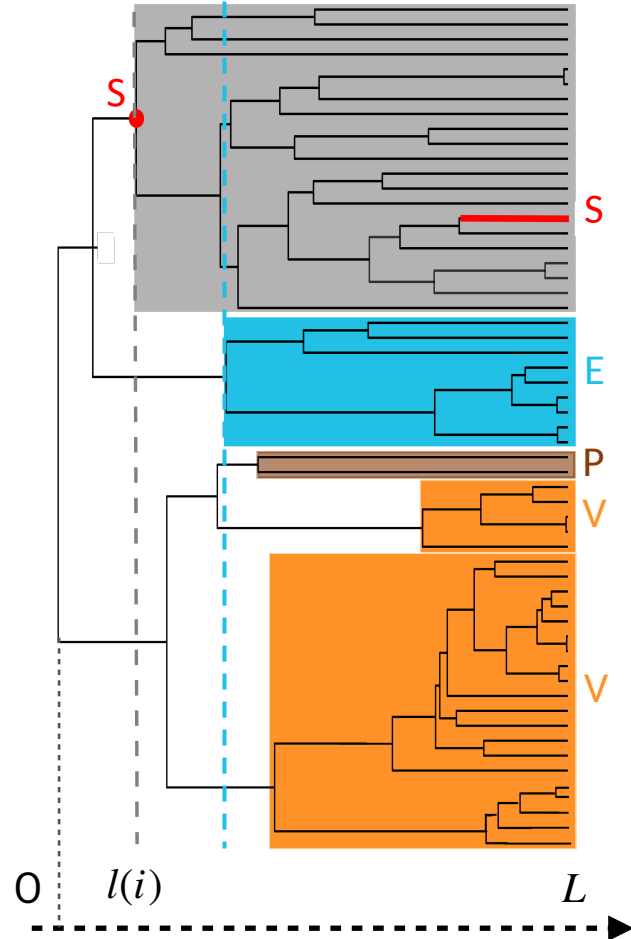


Evolutionary conservation

$$\frac{L - l(i)}{L}$$

GEMME: Evolution

$$S(X_i \rightarrow Y_i) = \boxed{T_{JET}(i)} \cdot [\alpha PE^{Ind}(Y_i) + (1 - \alpha) PE^{Epi}(Y_i)]$$



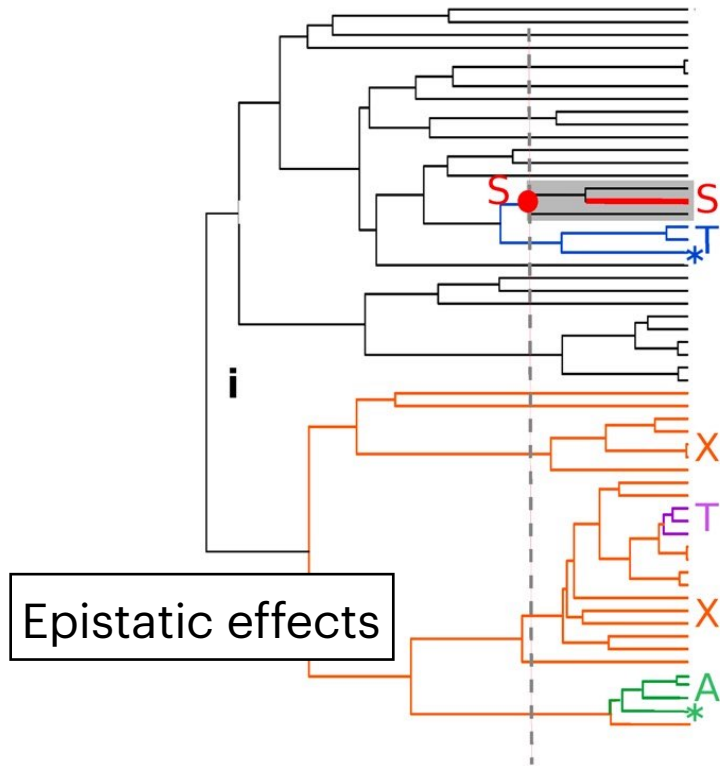
Evolutionary conservation

Construction of $\sqrt{N}$ phylogenetic trees with $\sqrt{N}$ effective sequences from input alignment, to compute residues **trace levels**

$$T_{JET}(i) = \frac{1}{M_i} \sum_{t=1}^{M_i} \frac{L_t - l_t(i)}{L_t}$$

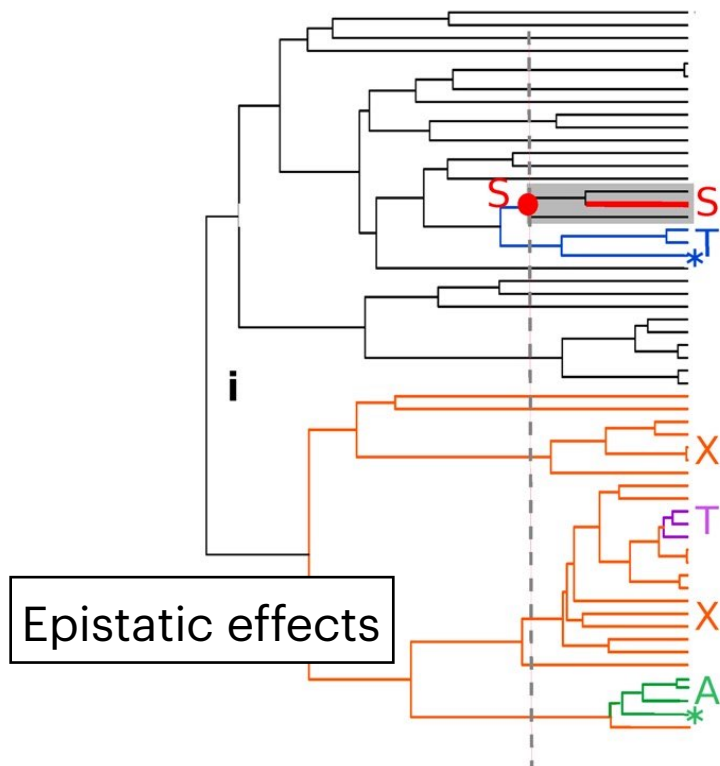
Evolution (GEMME)

$$S(X_i \rightarrow Y_i) = T_{JET}(i) \cdot [\alpha PE^{Ind}(Y_i) + (1 - \alpha) PE^{Epi}(Y_i)]$$



Evolution (GEMME)

$$S(X_i \rightarrow Y_i) = T_{JET}(i) \cdot [\alpha PE^{Ind}(Y_i) + (1 - \alpha) PE^{Epi}(Y_i)]$$

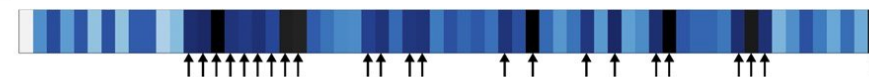


$$D_{evol}(q, s) = \sum_{i=1}^n T_{JET}^2(i) \cdot \mathbb{I}_{X_i^q \neq X_i^s}(i)$$

$$PE^{Epi}(Y_i) = \min_{S_{Y_i}} \left(D_{evol}(q, s_{Y_i}) \right)$$

Conservation levels

q EPRR I V I H R G S T G L G F N I V G G E D G E G I F I S F I L A G G P A D L S G E L R K G D Q I L S V N G V D L R N A S H



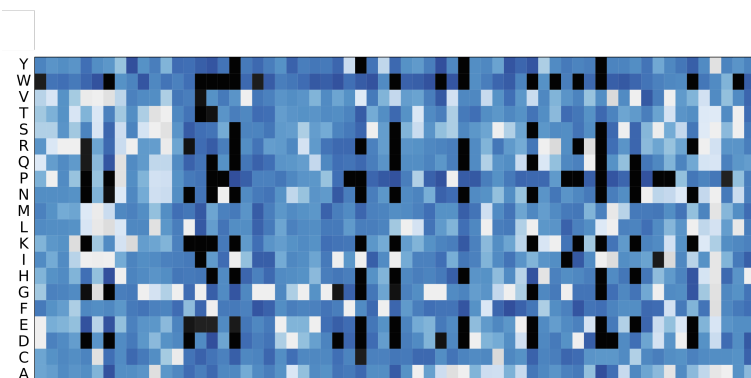
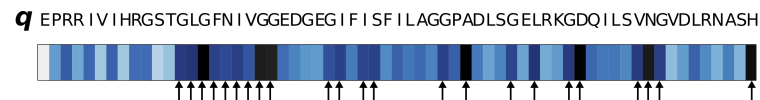
Evolutionary distance

S Q V E Y I D I E R P A G G L G F S V V A V R S H T D I F V K E V Q P G S I A D R D Q R L K E N D Q I L A I N H T P L D R V S H



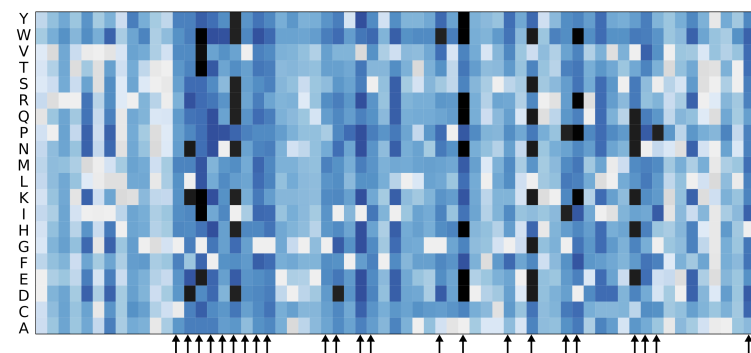
Evolution (GEMME)

$$S(X_i \rightarrow Y_i) = T_{JET}(i) \cdot [\alpha PE^{Ind}(Y_i) + (1 - \alpha) PE^{Epi}(Y_i)]$$



The more conserved a position, the more sensitive to mutations.

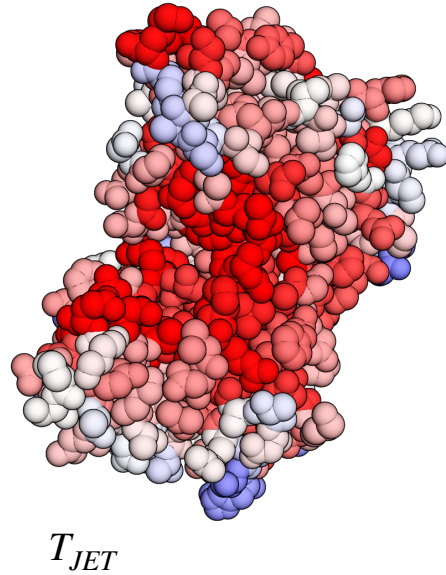
Scores are used as a filter applied on the matrix of predicted effects.



High T_{JET} score positions become highly intolerant to mutations while any substitution at a poorly T_{JET} score position will have a small effect.

ESCOTT: Evolution + Structure

Mutations appearing in the **interfaces** and in the **core** of a protein structure are likely deleterious

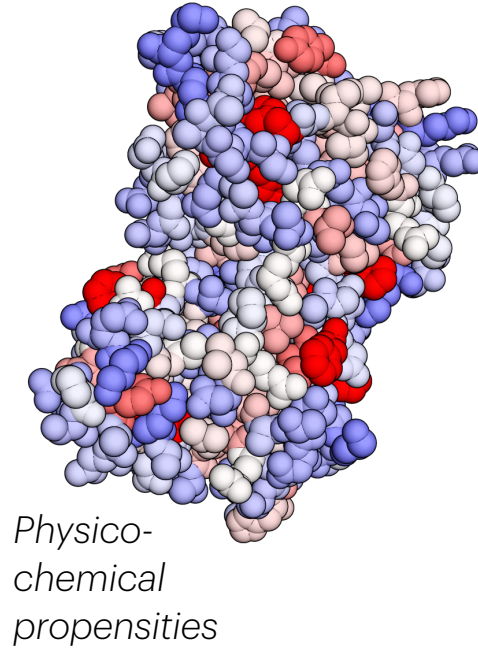
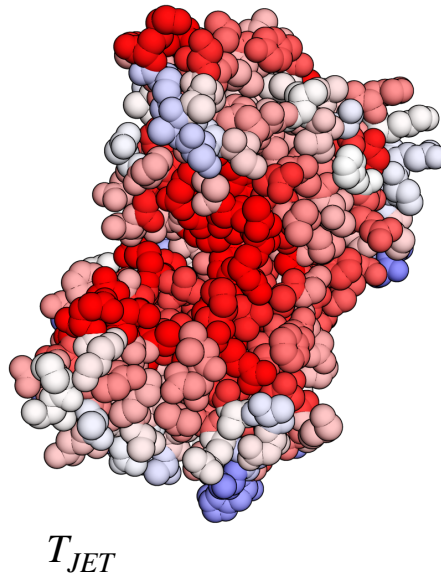


[Engelen et al, 2009]

ESCOTT: Evolution + Structure

Mutations appearing in the **interfaces** and in the **core** of a protein structure are likely deleterious

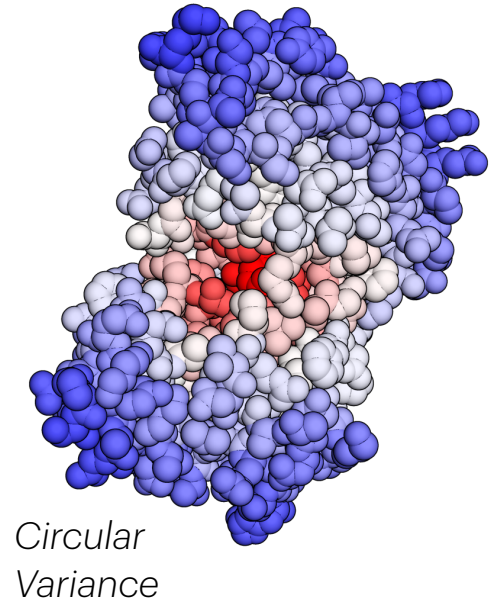
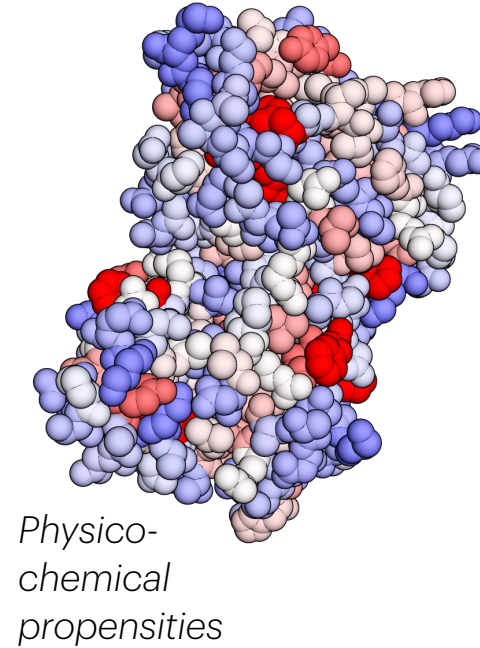
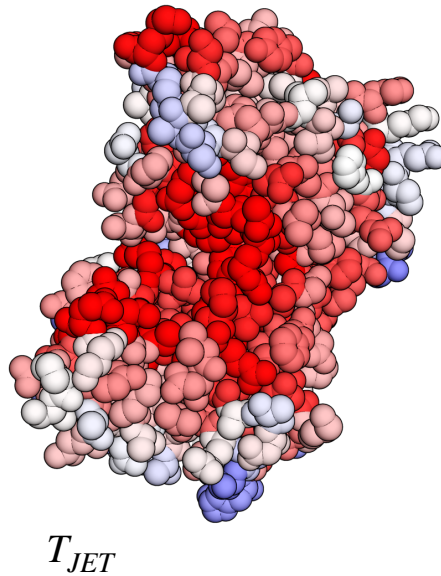
- stronger the **physico-chemical propensity** stronger the deleterious effect



ESCOTT: Evolution + Structure

Mutations appearing in the **interfaces** and in the **core** of a protein structure are likely deleterious

- stronger the **physico-chemical propensity** stronger the deleterious effect
- higher the **packing density** stronger the deleterious effect



$$CV_i = 1 - \frac{1}{n_i} \left| \sum_{j \neq i, r_{ij} < r_c} \frac{\vec{r}_{ij}}{r_{ij}} \right|$$

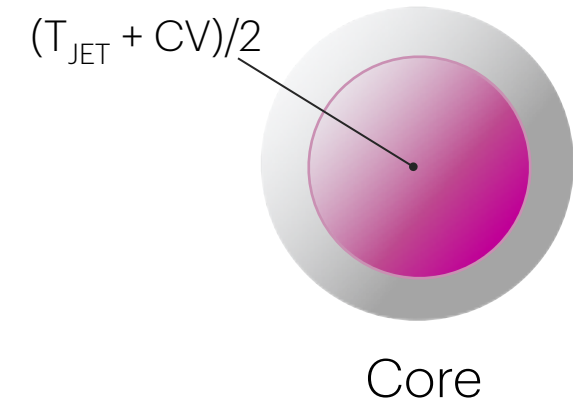
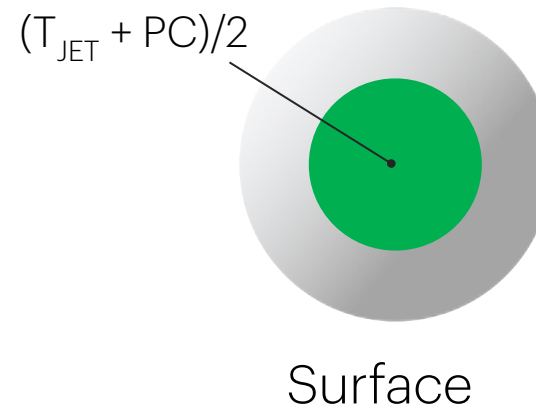
[Ceres et al, 2012]

[Laine and Carbone, 2016]

ESCOTT: Evolution + Structure

Mutations appearing in the **interfaces** and in the **core** of a protein structure are likely deleterious

- stronger the **physico-chemical propensity** stronger the deleterious effect
- higher the **packing density** stronger the deleterious effect



$$S(X_i \rightarrow Y_i) = T_{JET}(i) \cdot [\alpha PE^{Ind}(Y_i) + (1 - \alpha) PE^{Epi}(Y_i)]$$

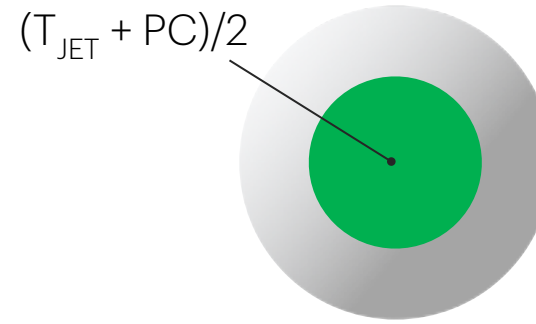
$$T_{JET} \leftarrow \max \left(\frac{T_{JET} + CV}{2}, \frac{T_{JET} + PC}{2} \right)$$

(for non-coiled regions)

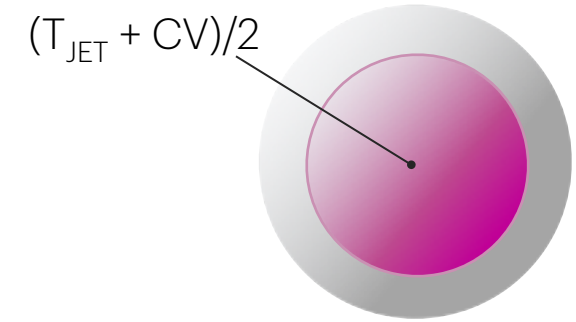
ESCOTT: Evolution + Structure

Mutations appearing in the **interfaces** and in the **core** of a protein structure are likely deleterious

- stronger the **physico-chemical propensity** stronger the deleterious effect
- higher the **packing density** stronger the deleterious effect



Surface



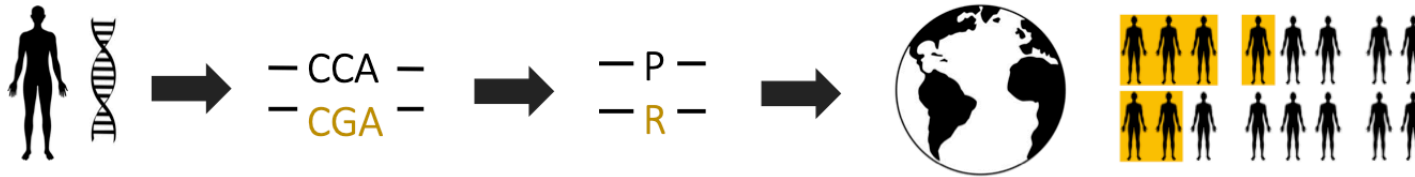
Core

$$S(X_i \rightarrow Y_i) = T_{JET}(i) \cdot [\alpha PE^{Ind}(Y_i) + (1 - \alpha) \boxed{PE^{Epi}(Y_i)}]$$

$$T_{JET} \leftarrow \max \left(\frac{T_{JET} + CV}{2}, \frac{T_{JET} + PC}{2} \right)$$

(for non-coiled regions)

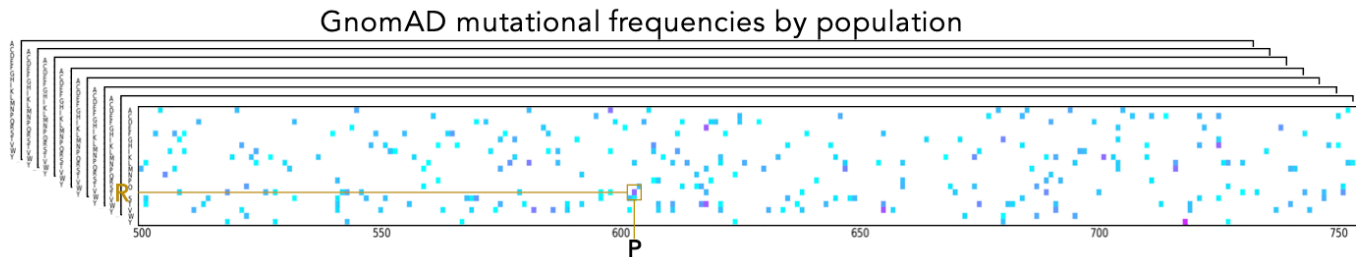
PRESCOTT: Allele frequency from human population data



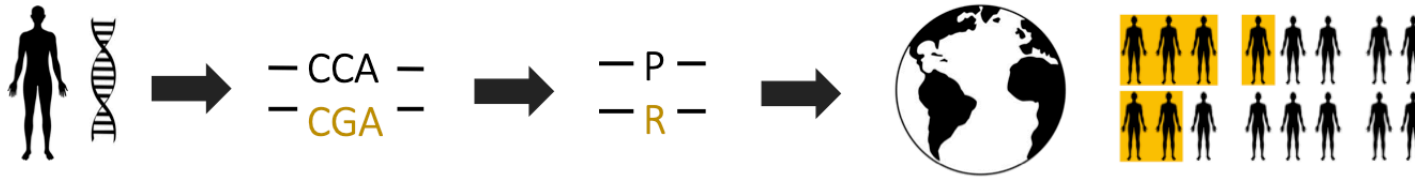
gnomAD v4.0.0

730,947 individuals from 8 populations (>20,000 ind):

- African/African American
- Admixed American
- Ashkenazi Jewish (14804 ind)
- East Asian
- European (Finnish)
- European (non Finnish)
- South Asian
- Middle Eastern (3031 ind)



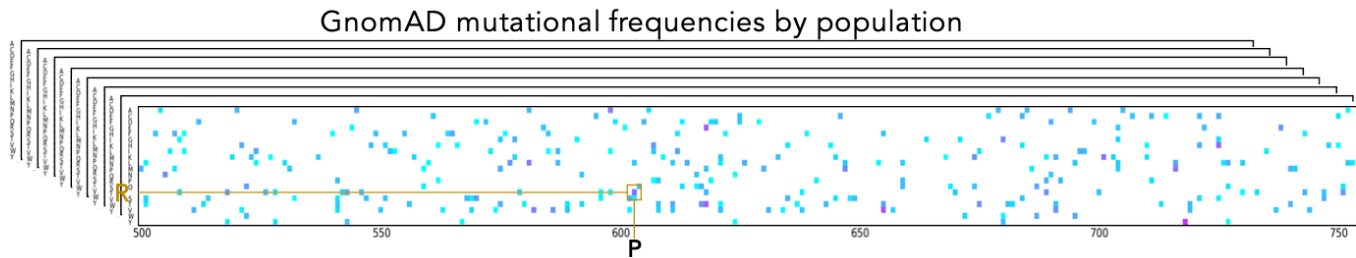
PRESCOTT: Allele frequency from human population data



gnomAD v4.0.0

730,947 individuals from 8 populations (>20,000 ind):

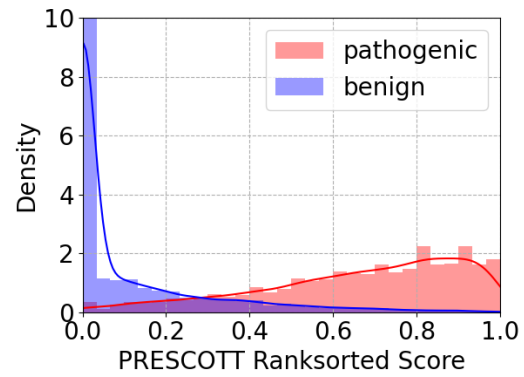
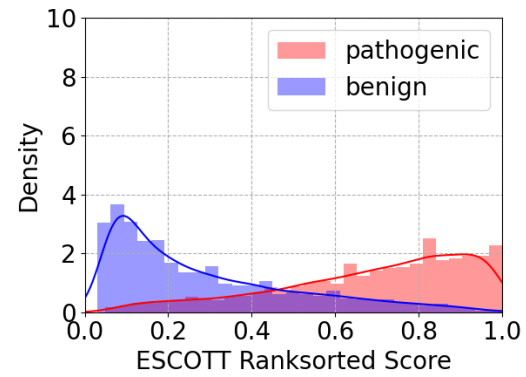
- African/African American
- Admixed American
- Ashkenazi Jewish (14804 ind)
- East Asian
- European (Finnish)
- European (non Finnish)
- South Asian
- Middle Eastern (3031 ind)



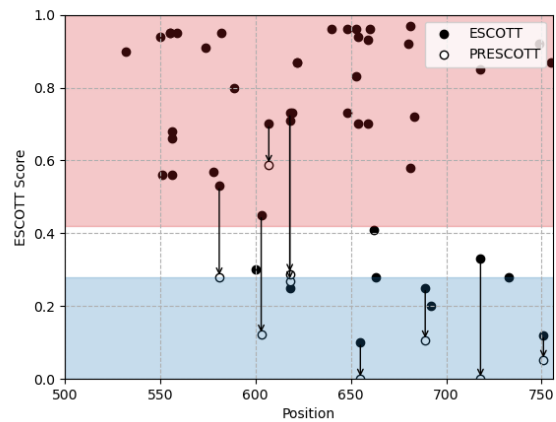
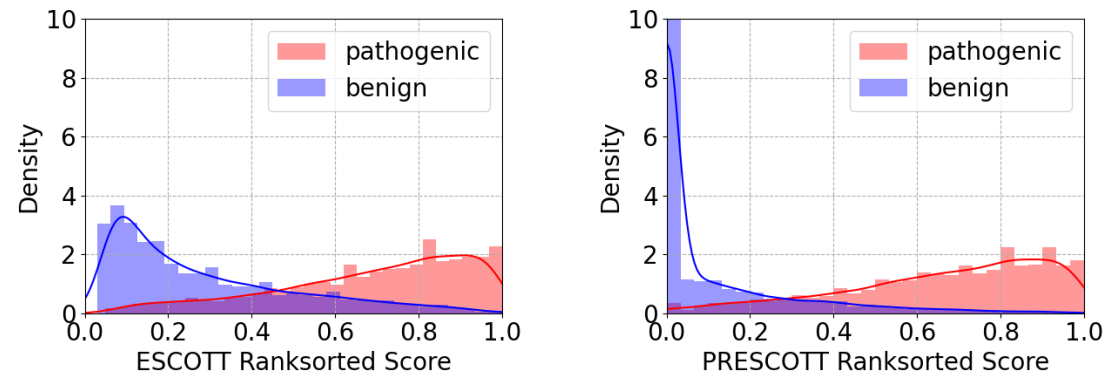
$$S(Y_i) \leftarrow \max \left[0, S'(Y_i) - \theta(f(Y_i) - F_c) \cdot \left(1 - \frac{\log_{10} f(Y_i)}{\log_{10} F_c} \right) \right]$$

Effect: decrease pathogenicity scores for common missense mutations

ESCOTT and PRESCOTT performances

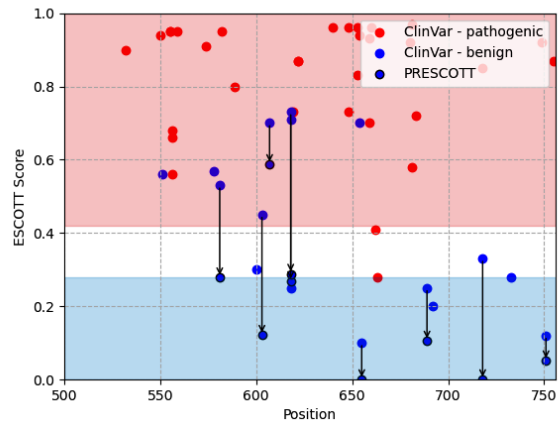
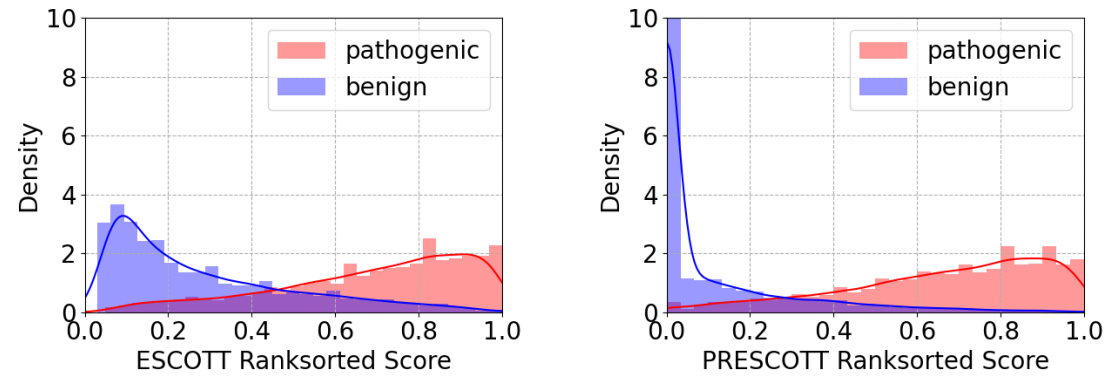


ESCOTT and PRESCOTT performances



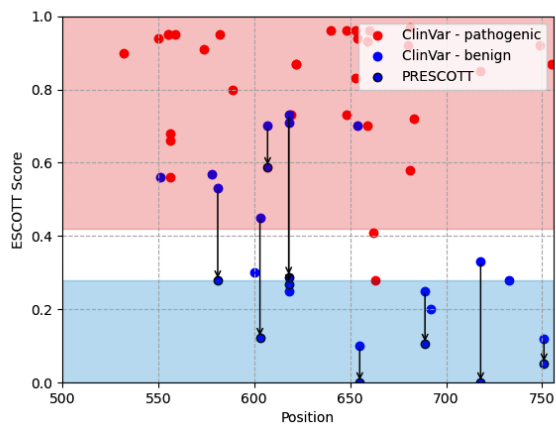
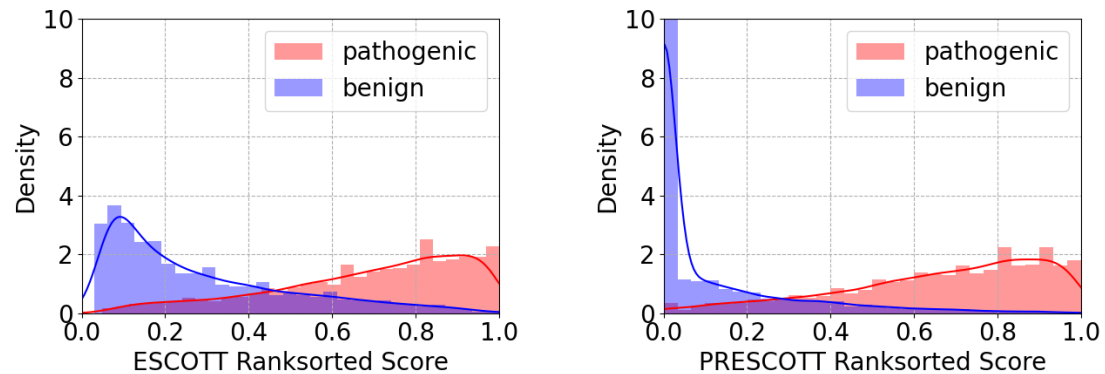
MLH1

ESCOTT and PRESCOTT performances



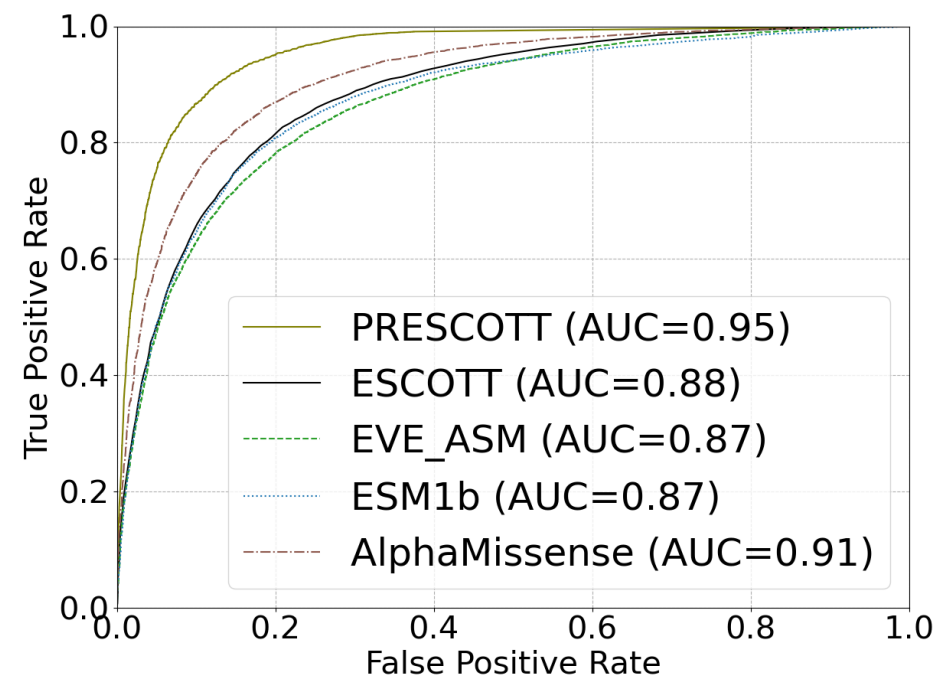
MLH1

ESCOTT and PRESCOTT performances



MLH1

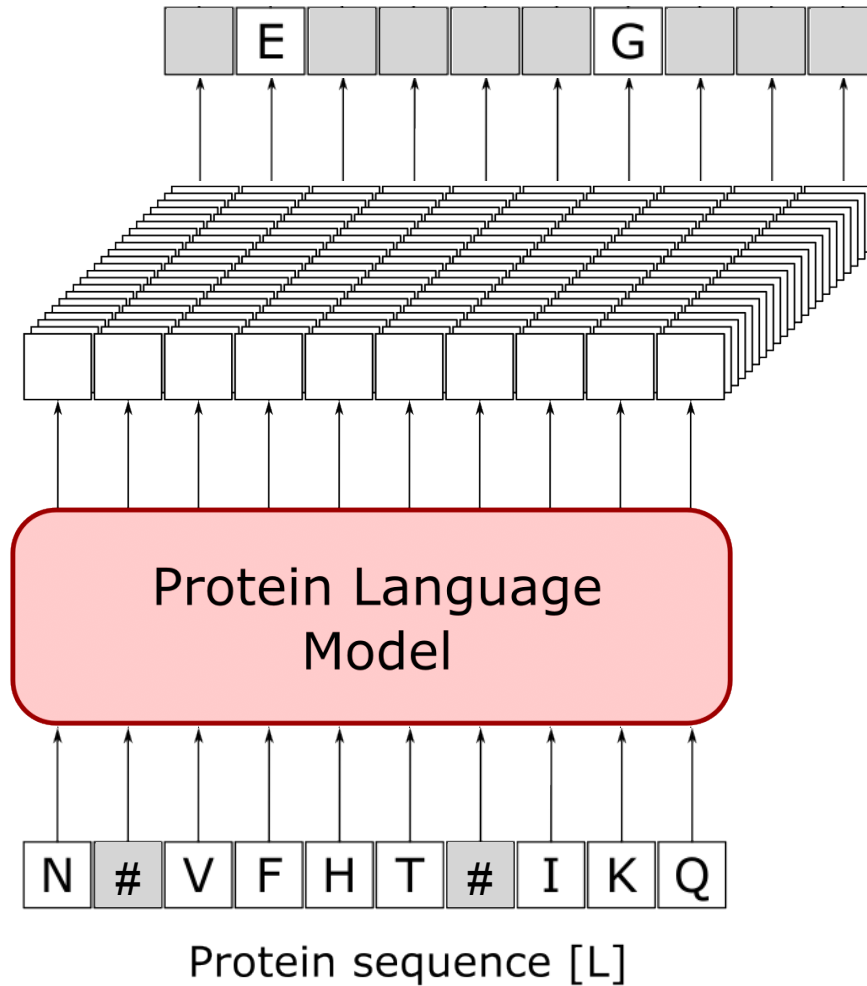
Results on 1883 human proteins from ClinVar, present in gnomAD



Can we further improve generalisation?



Can we further improve generalisation?



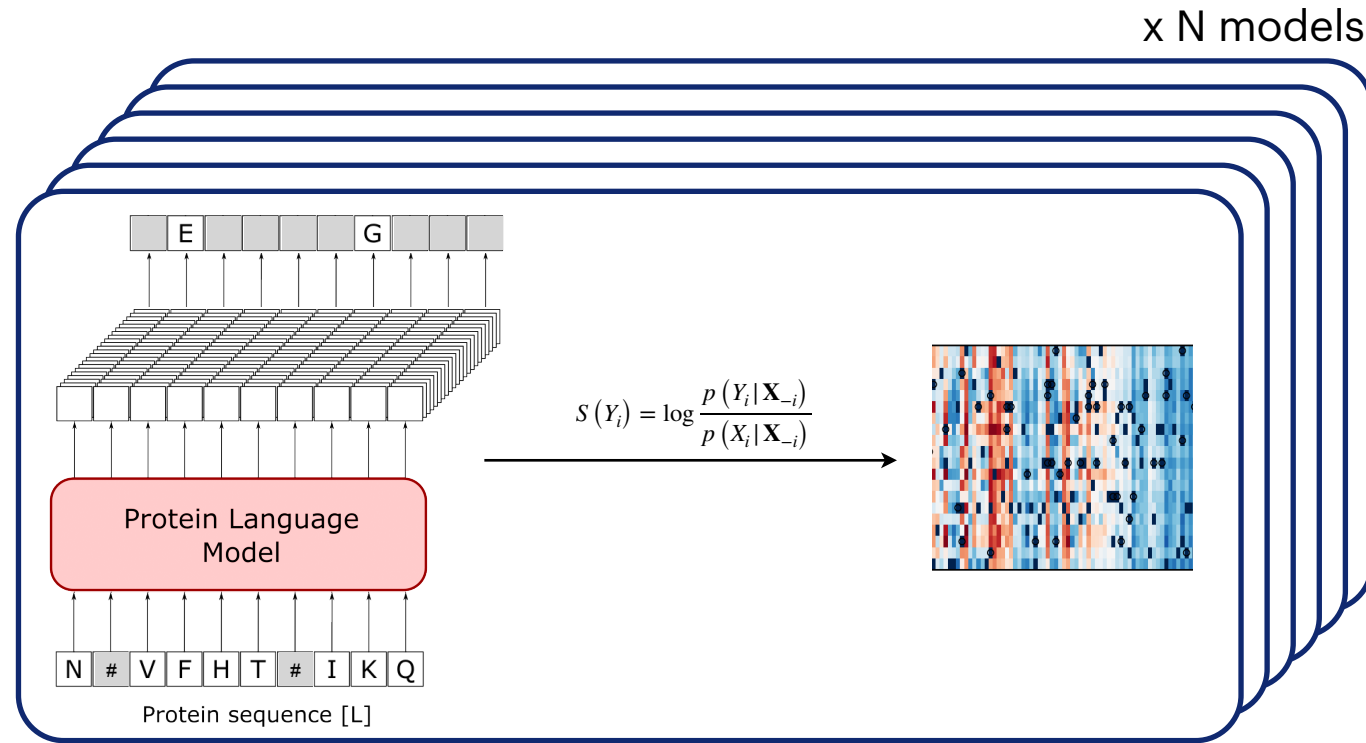
PLMs: large transformer architectures trained on protein sequences/structures

Masked LM objective:

- Mask a small fraction of tokens
- Predict masked tokens based on context

Context-aware embeddings

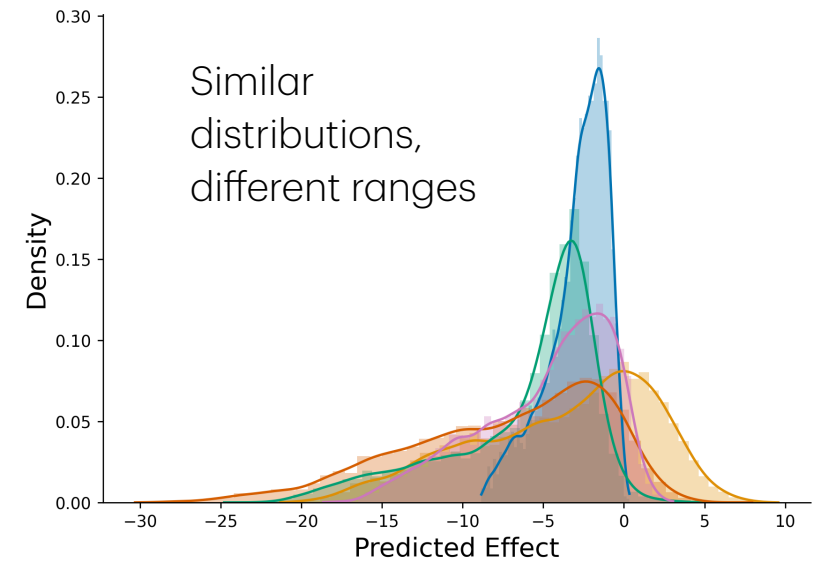
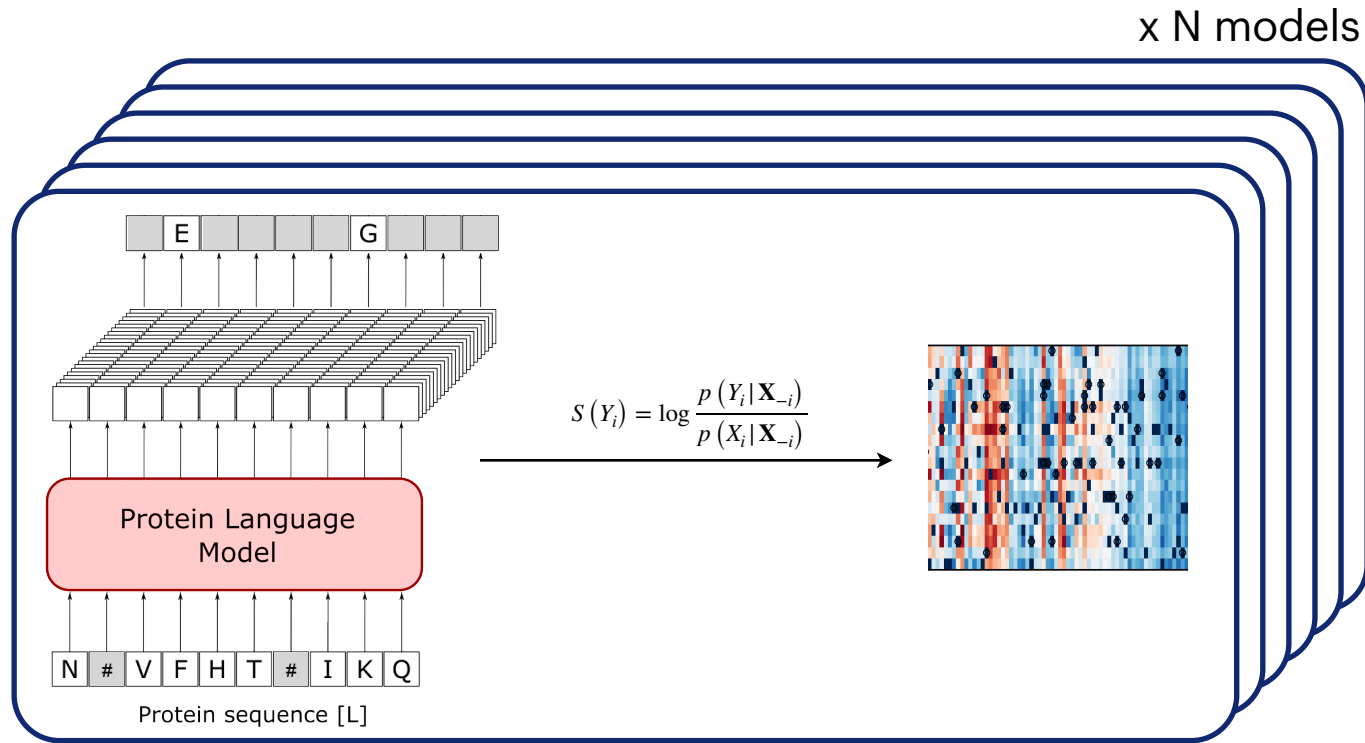
Can we further improve generalisation? Ensemble multiple modalities!



Different modalities:
sequence, structure,
alignments

- AIDO RAG 16B
- ProSST-2048
- ESM2-650M
- ESM-IF1
- AlphaMissense
- ...

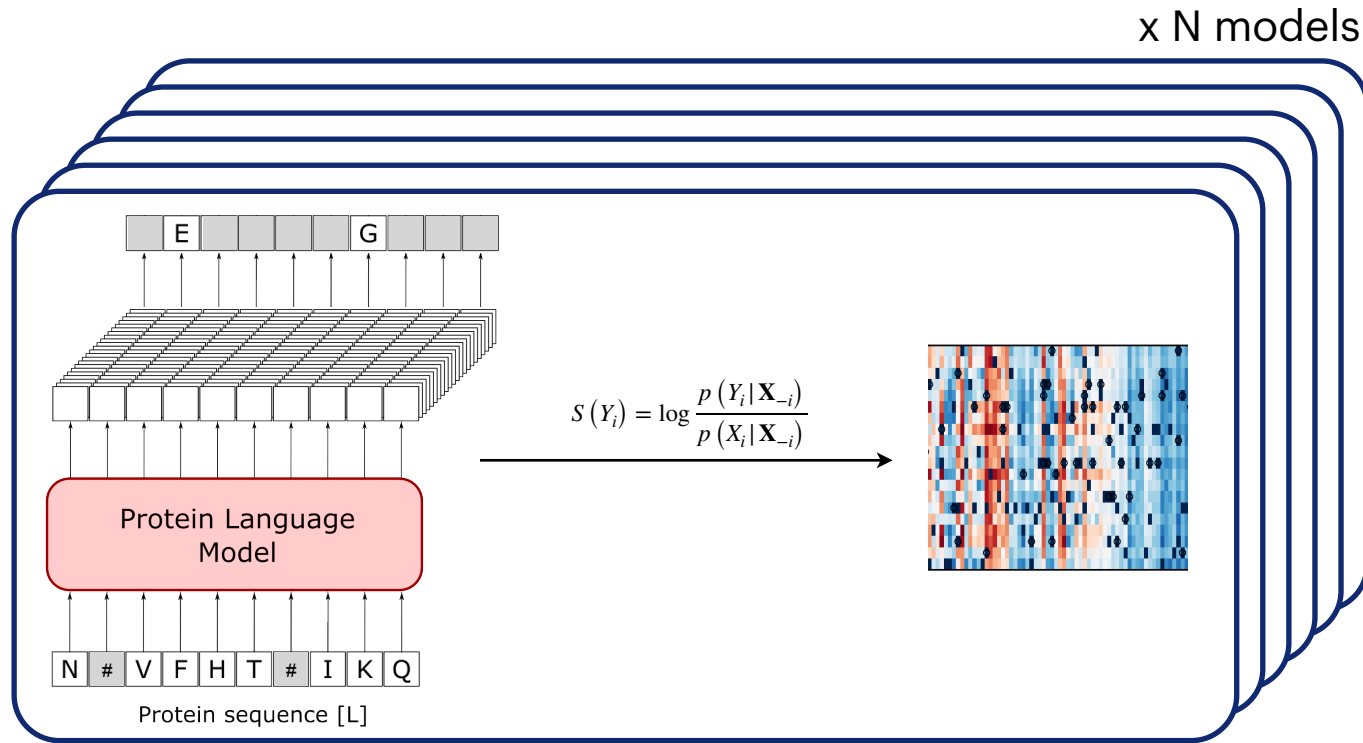
Can we further improve generalisation? Ensemble multiple modalities!



Rescale to ESCOTT range

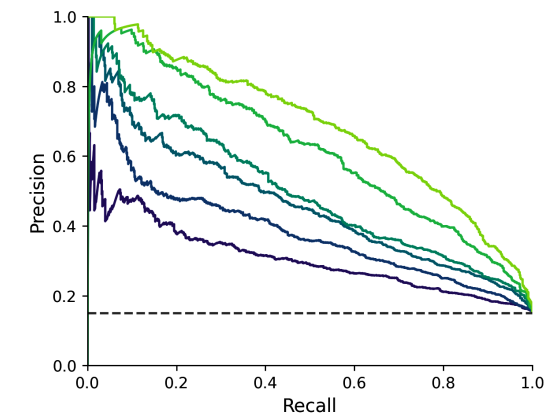
$$S'_M = \frac{S_E^{5\%}}{S_M^{5\%}} S_M$$

Can we further improve generalisation? Ensemble multiple modalities!



Bagging over different combinations of models and selection

- ClinVar
- ProteinGym
- Available validation data



Example: ARSA challenge

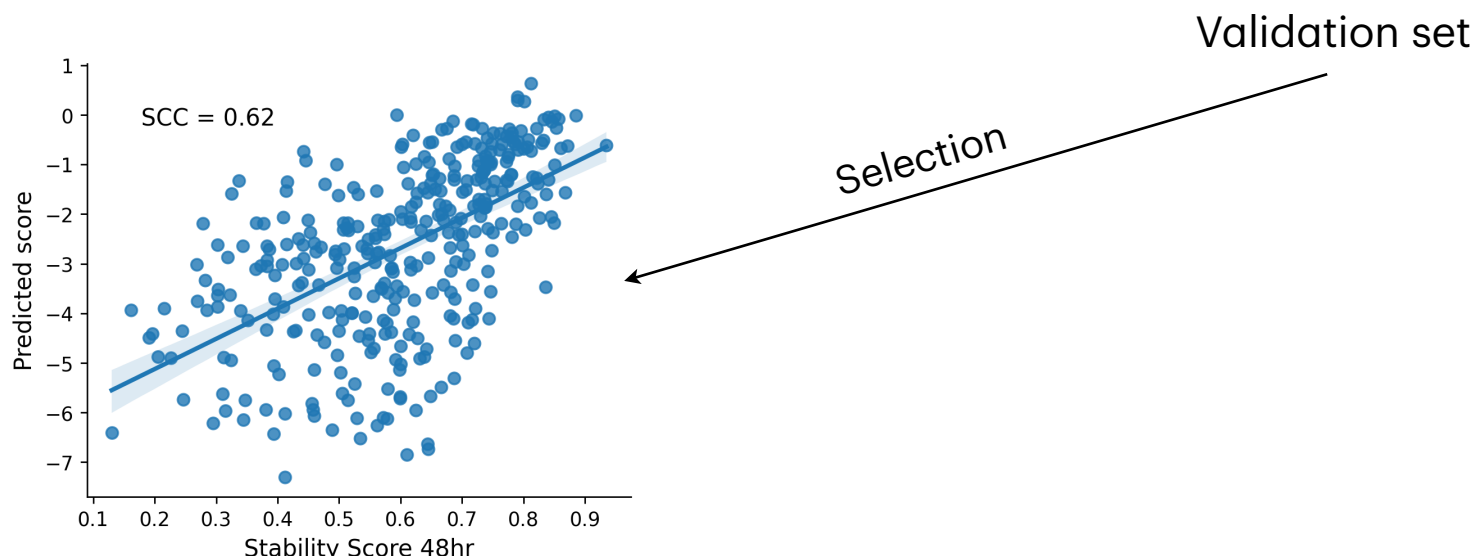
Goal: prediction of stability scores at 48h post expression for (almost) all missense variants

Available data: Subset of experimental scores for 4% mutations

Example: ARSA challenge

Goal: prediction of stability scores at 48h post expression for (almost) all missense variants

Available data: Subset of experimental scores for 4% mutations



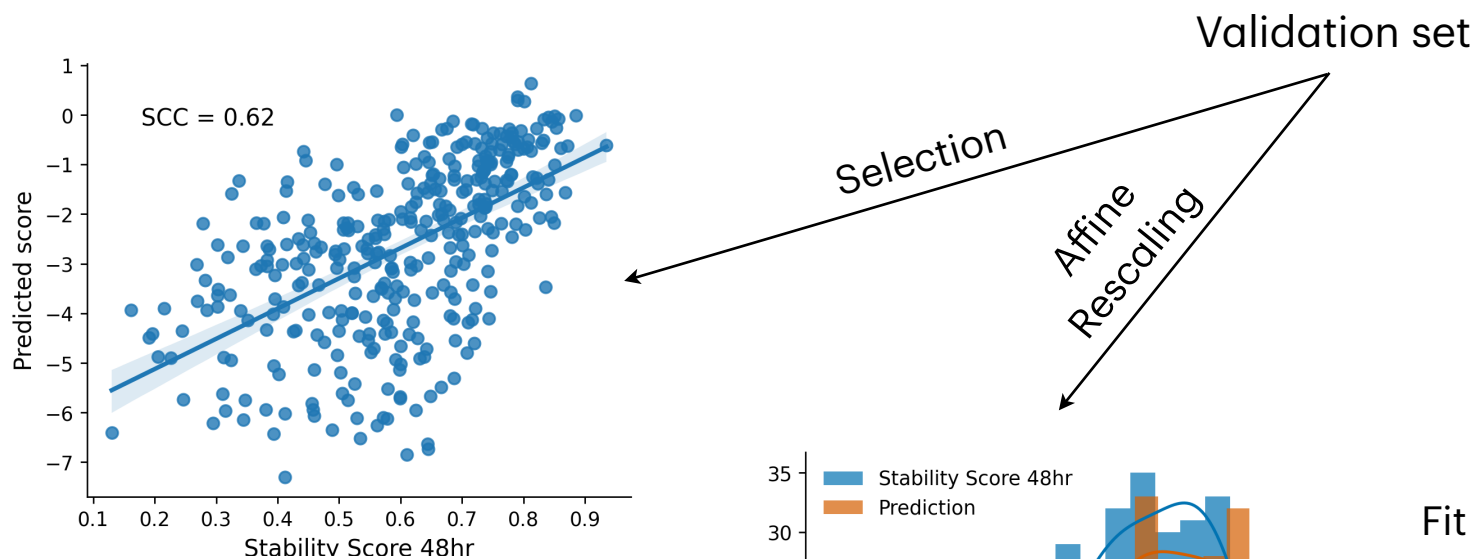
Correlations on
validation data

ESCOTT + AIDO RAG 16B

Example: ARSA challenge

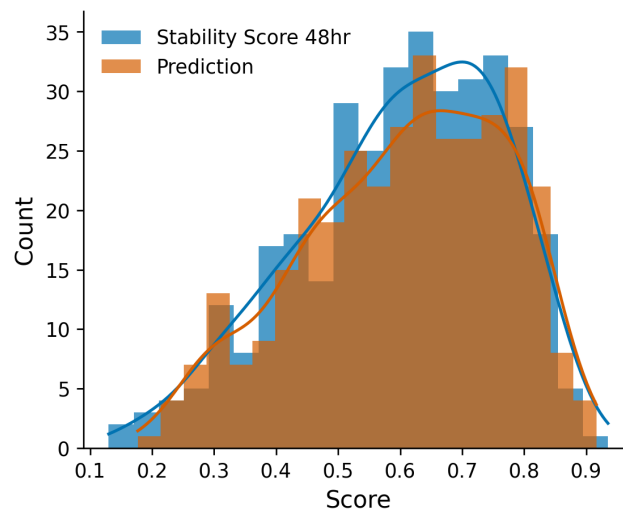
Goal: prediction of stability scores at 48h post expression for (almost) all missense variants

Available data: Subset of experimental scores for 4.4% mutations



Correlations on
validation data

ESCOTT + AIDO RAG 16B



Fit params on
experimental data
+
Ensure low KL
divergence

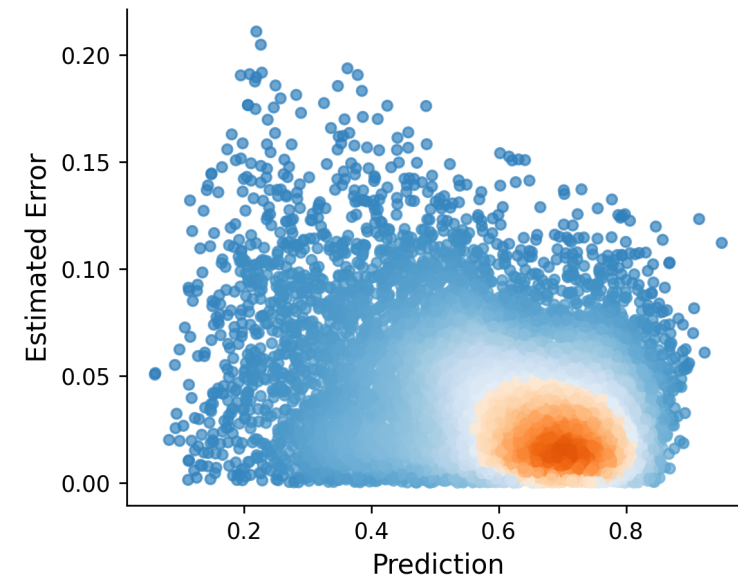
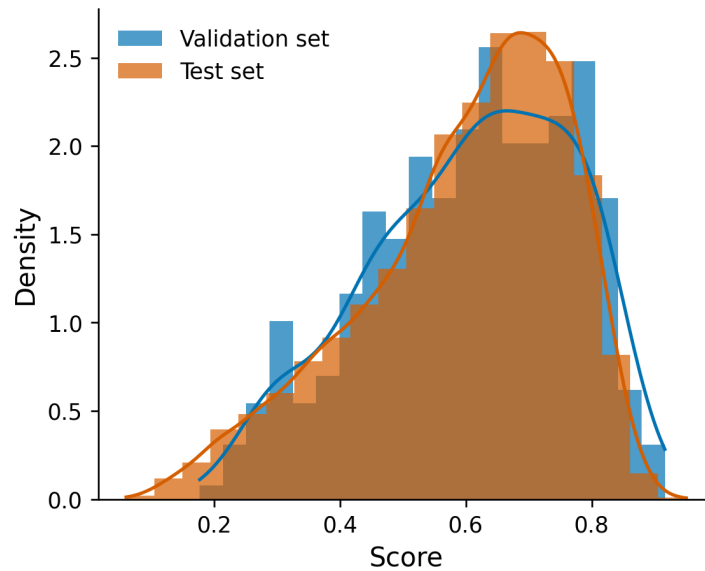
[Sun et al, 2024]

Example: ARSA challenge

Goal: prediction of stability scores at 48h post expression for (almost) all missense variants

Available data: Subset of experimental scores for 4.4% mutations

Apply transformation to test data and errors estimates



Acknowledgments



Alessandra Carbone



Mustafa Tekpinar



Acknowledgments



Alessandra Carbone



Mustafa Tekpinar



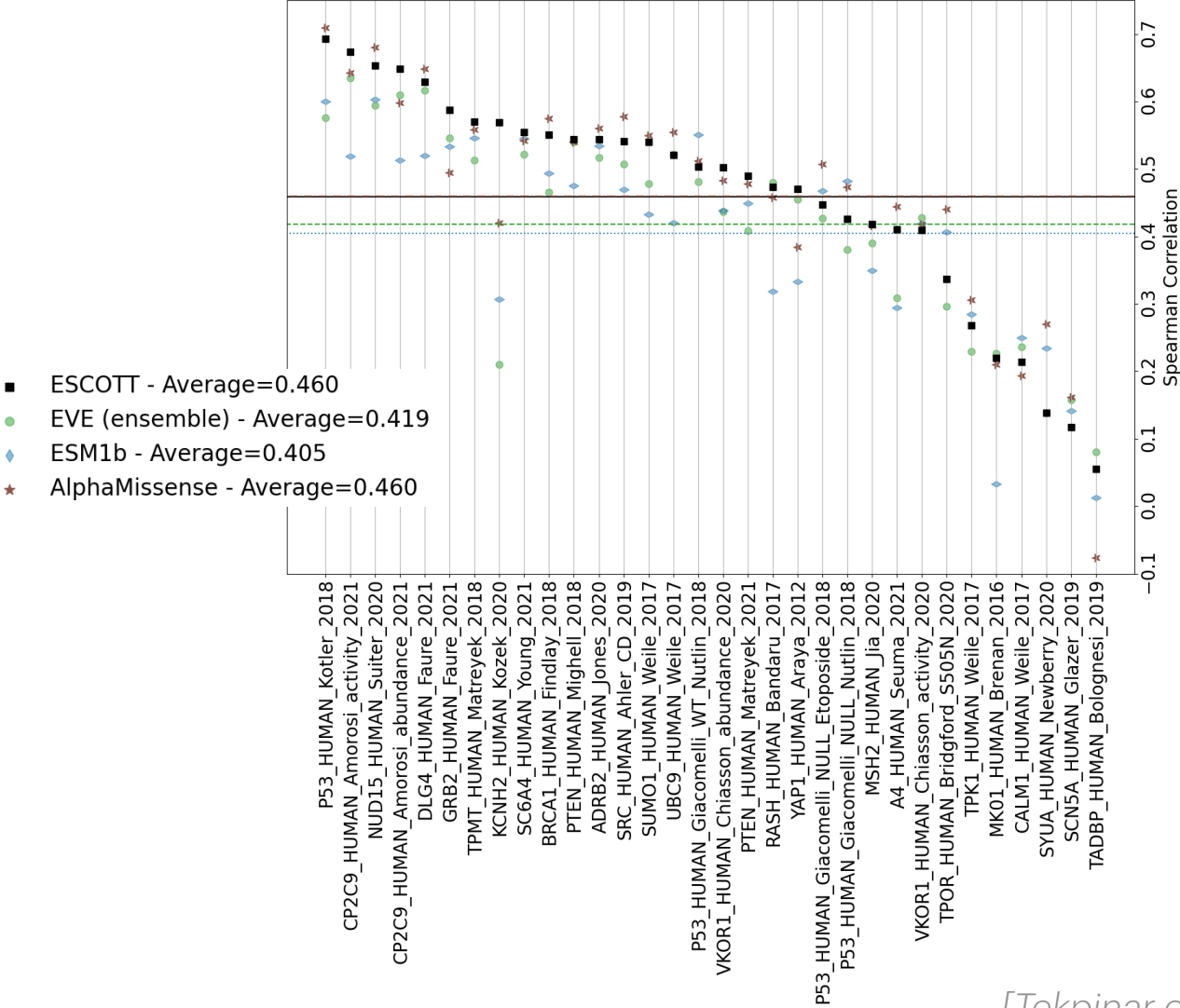
PRESCOTT article



PRESCOTT webserver

ESCOTT and PRESCOTT performances

Deep mutational scanning experiments on human proteins



[Tekpinar et al, 2025]