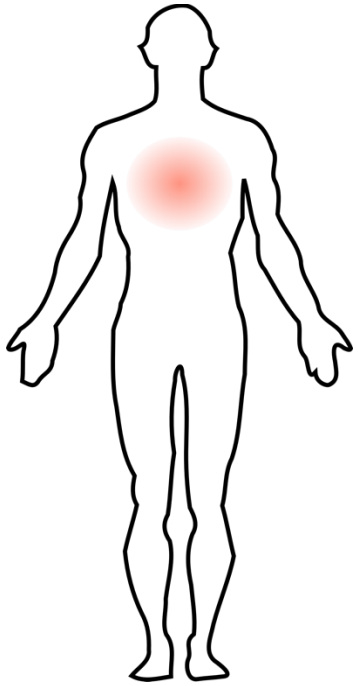
Empowering variant effect prediction with large-scale mutagenesis data

NHGRI Genomic Medicine IX, 2016

Douglas M. Fowler, Ph. D.
Assistant Professor, Genome Sciences
University of Washington

Interpreting coding variation is challenging and important

MEEPQSDPSVEPP
LSQETFSDLWKLL
PENNVLSPLPSQA
MD**E**LMLSPDDIEQ
WFTEDPGPDEAPR
MPEAAPR

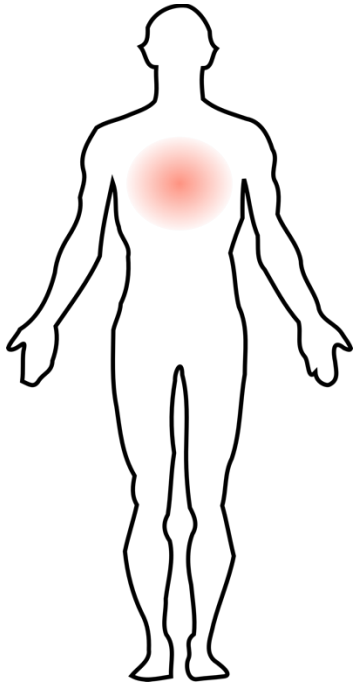


variant

phenotype

Interpreting coding variation is challenging and important

MEEPQSDPSVEPP
LSQETFSDLWKLL
PENNVLSPLPSQA
MD**E**LMLSPDDIEQ
WFTEDPGPDEAPR
MPEAAPR



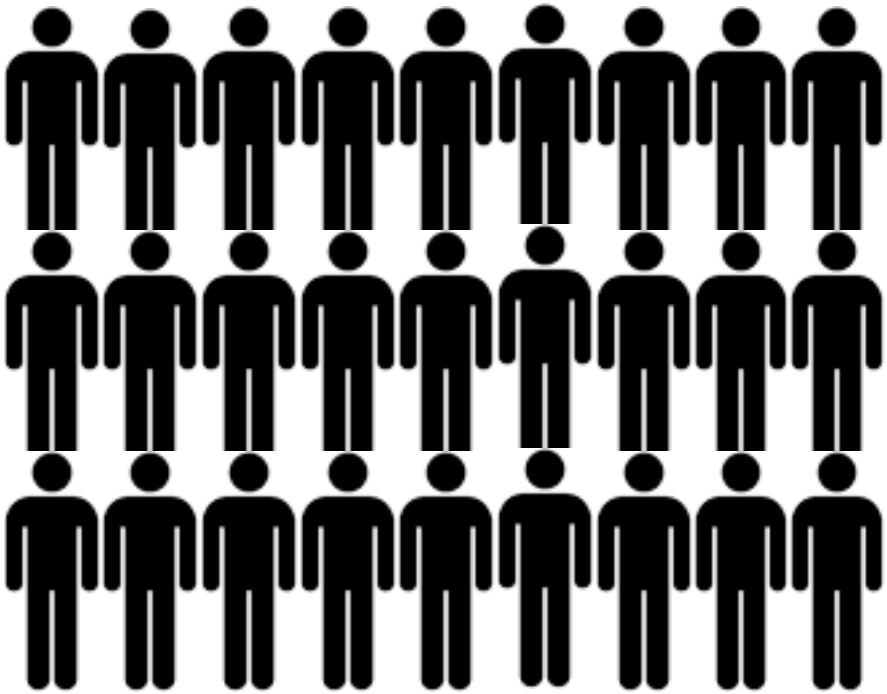
variant

phenotype

interpreting rare variation is especially hard

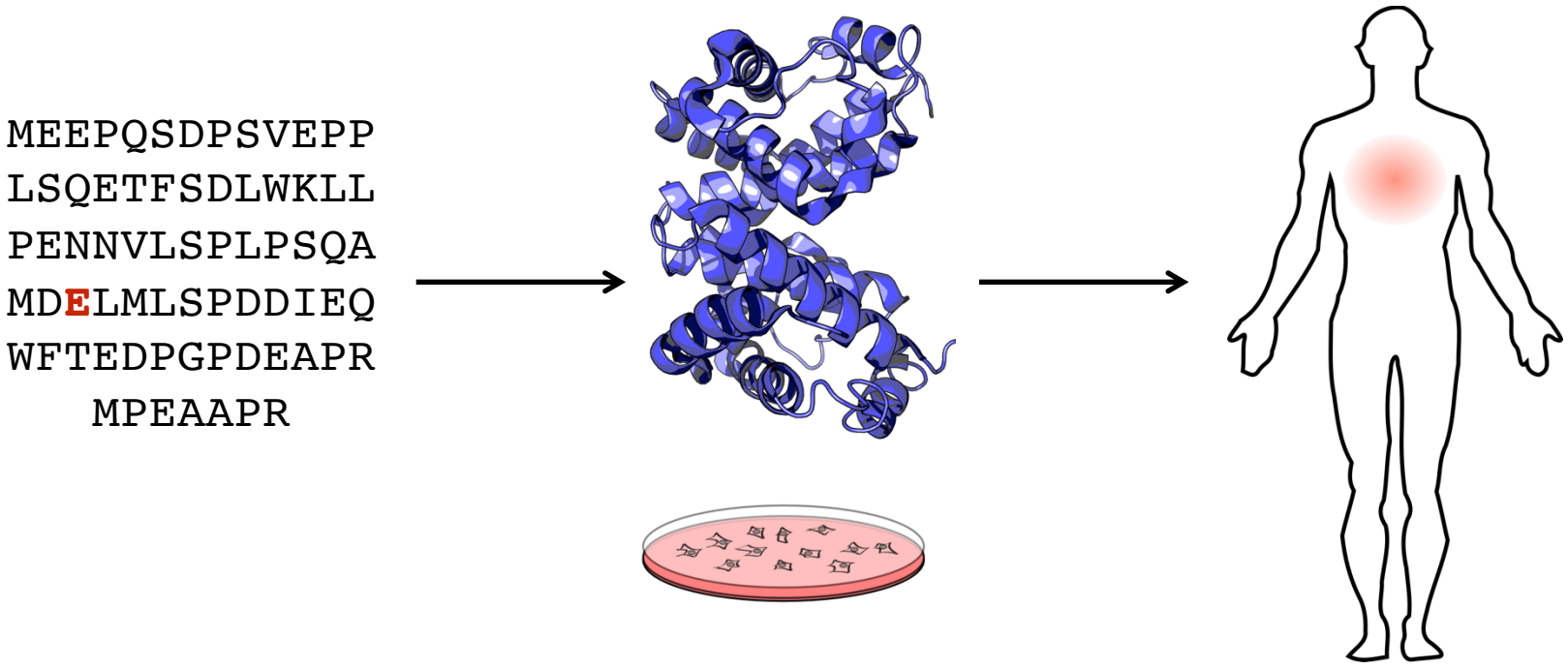
Every possible variant in the human genome exists

Given a mutation rate of ~1e-8, each person harbors **~60 de novo variants**



That means in ~7e9 of us alive now, there are ~4e11 de novo variants or **~44 instances of every possible SNV**

Interpreting coding variation is challenging and important



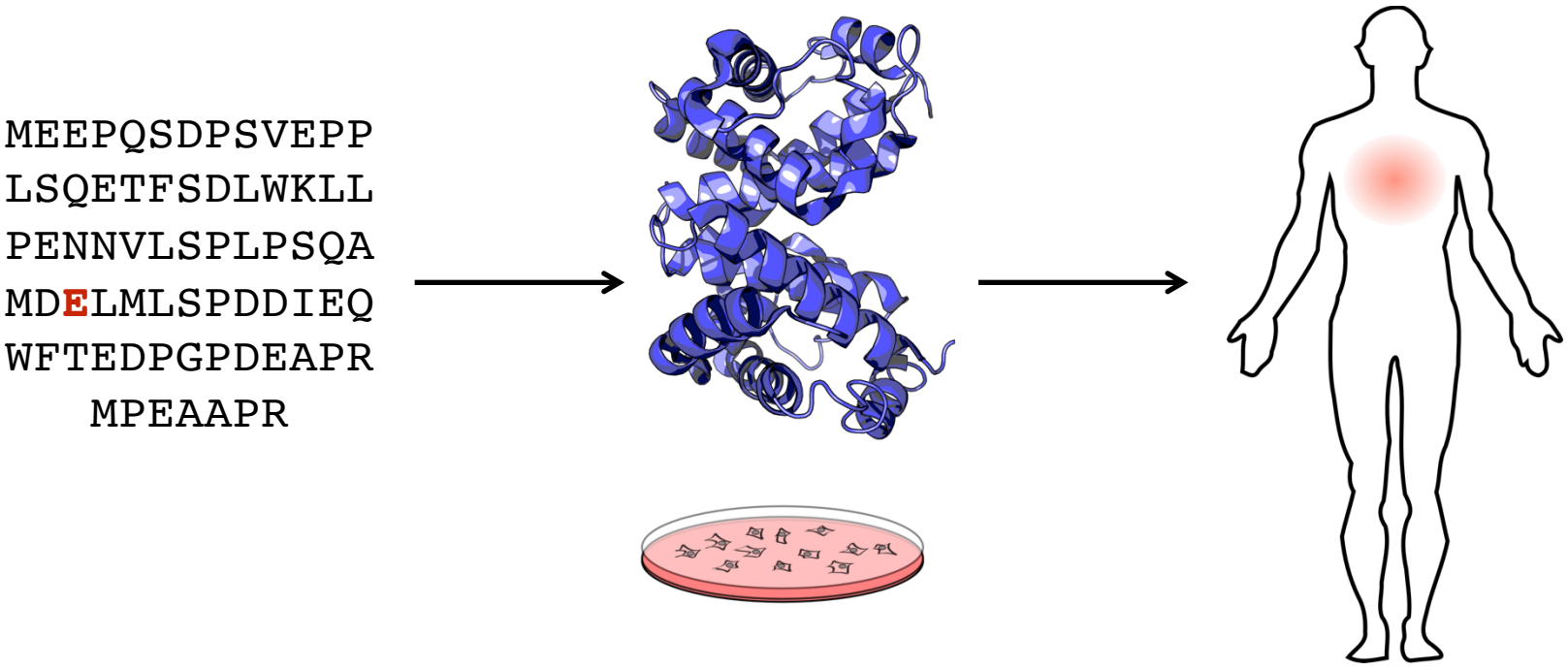
variant

functional effect

phenotype

mutagenesis is a way to interrogate functional effect…

Interpreting coding variation is challenging and important



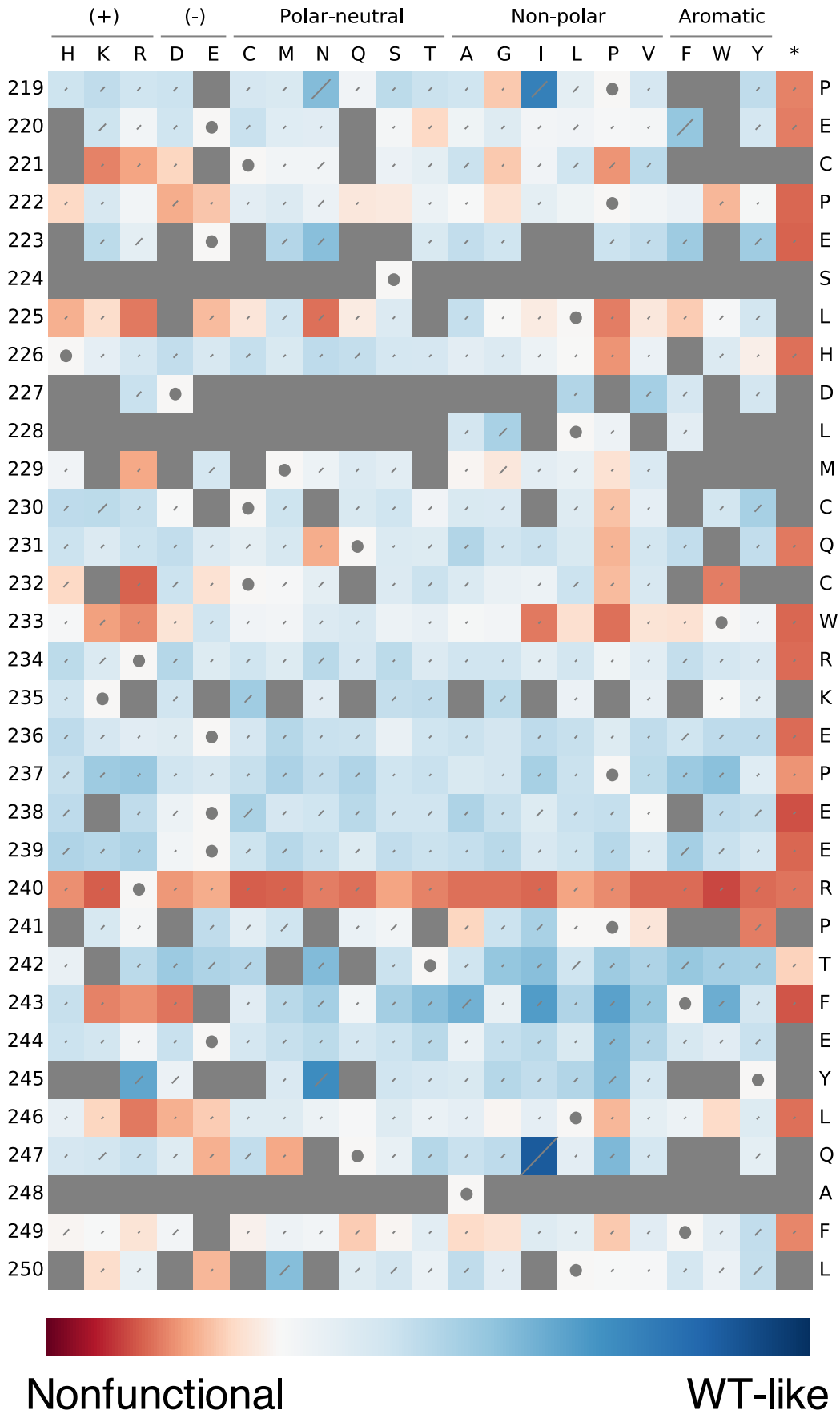
variant

functional effect

phenotype

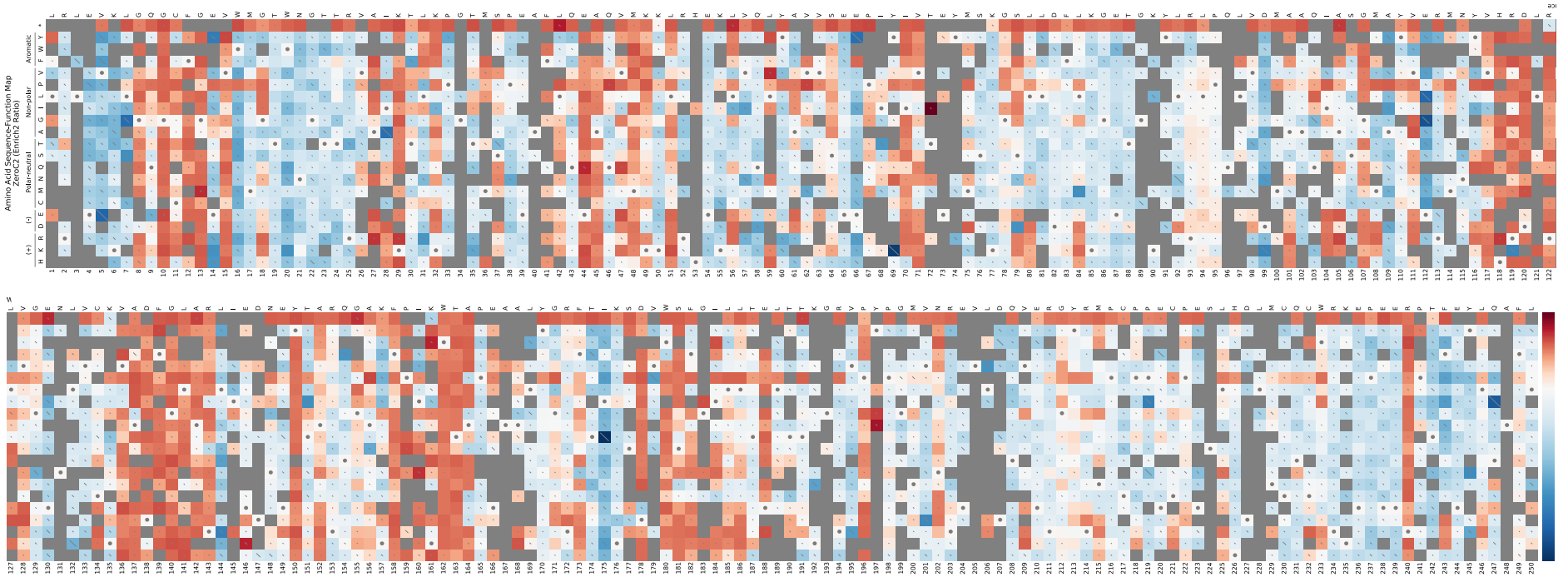
…but sequence space is vast

Deep mutational scanning to measure protein function

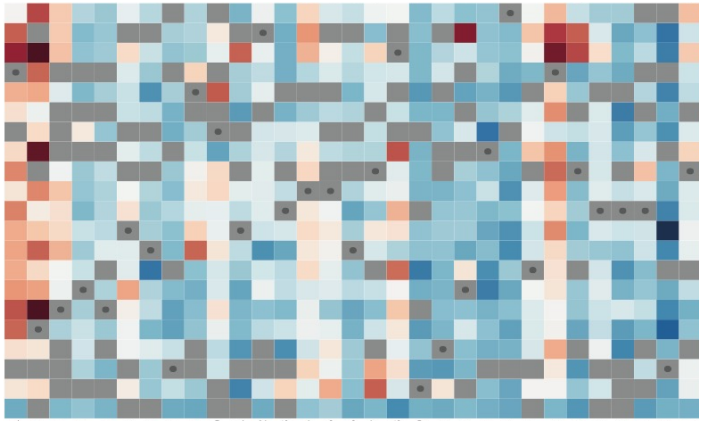


Src kinase sequence-function map

Src kinase sequence-function map

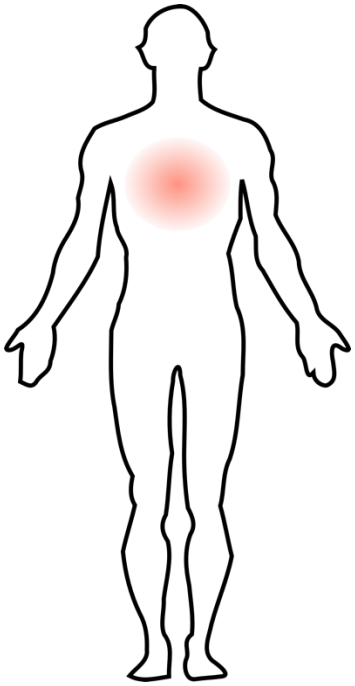


Large-scale data-driven prediction of protein variant effects



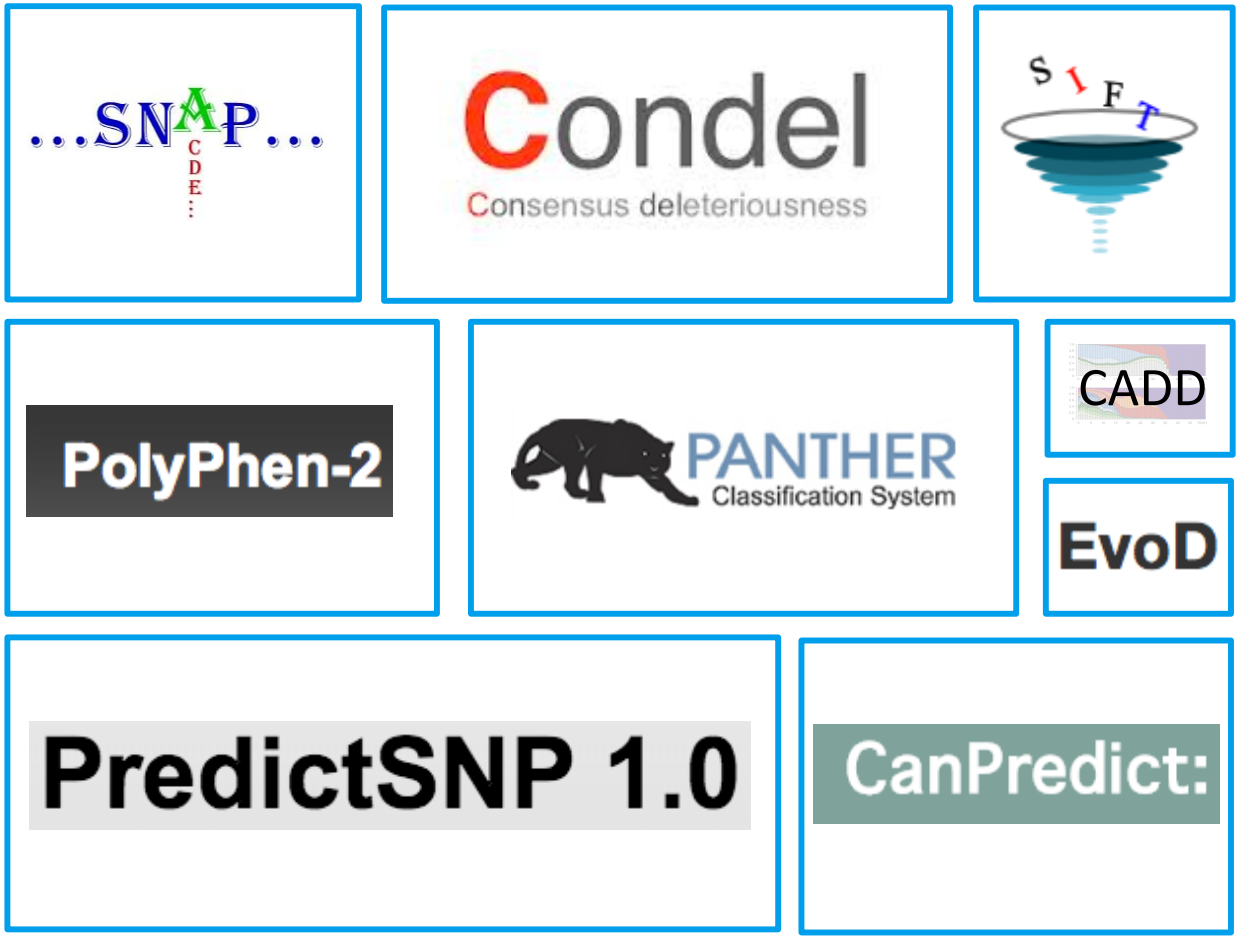
Sequence-function map

Model trained with
variants of known effect



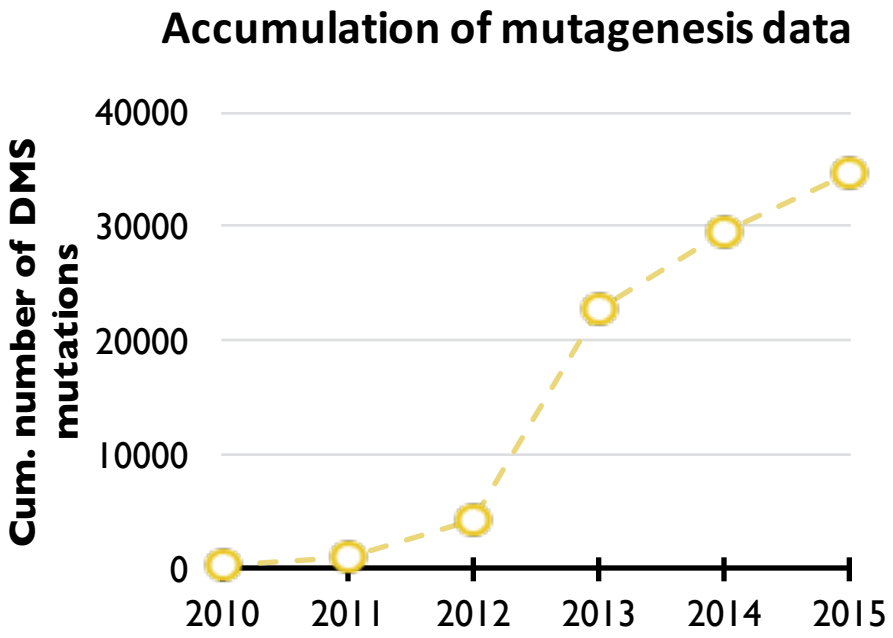
Accurate predictions
for nearly all variants in
protein of interest

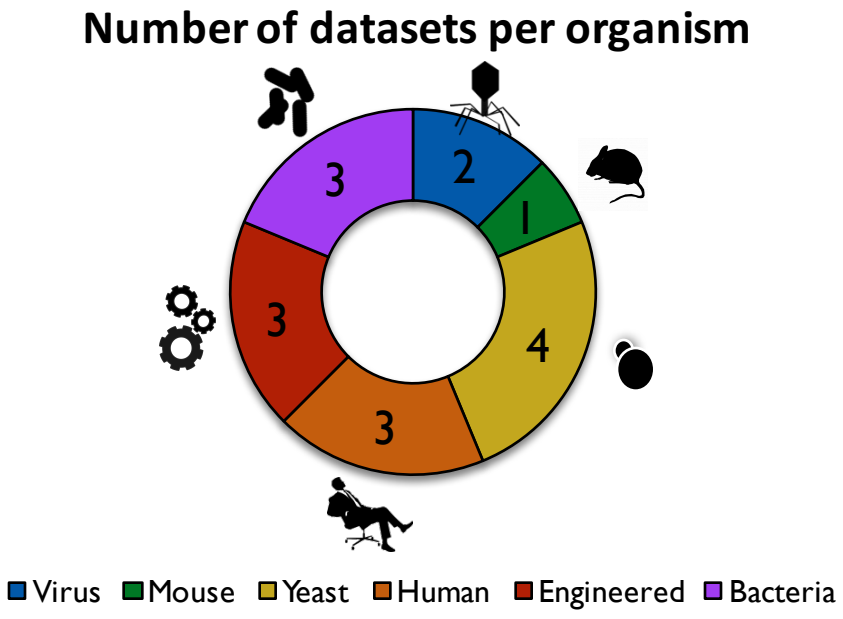
Challenge: there are many disease-associated proteins



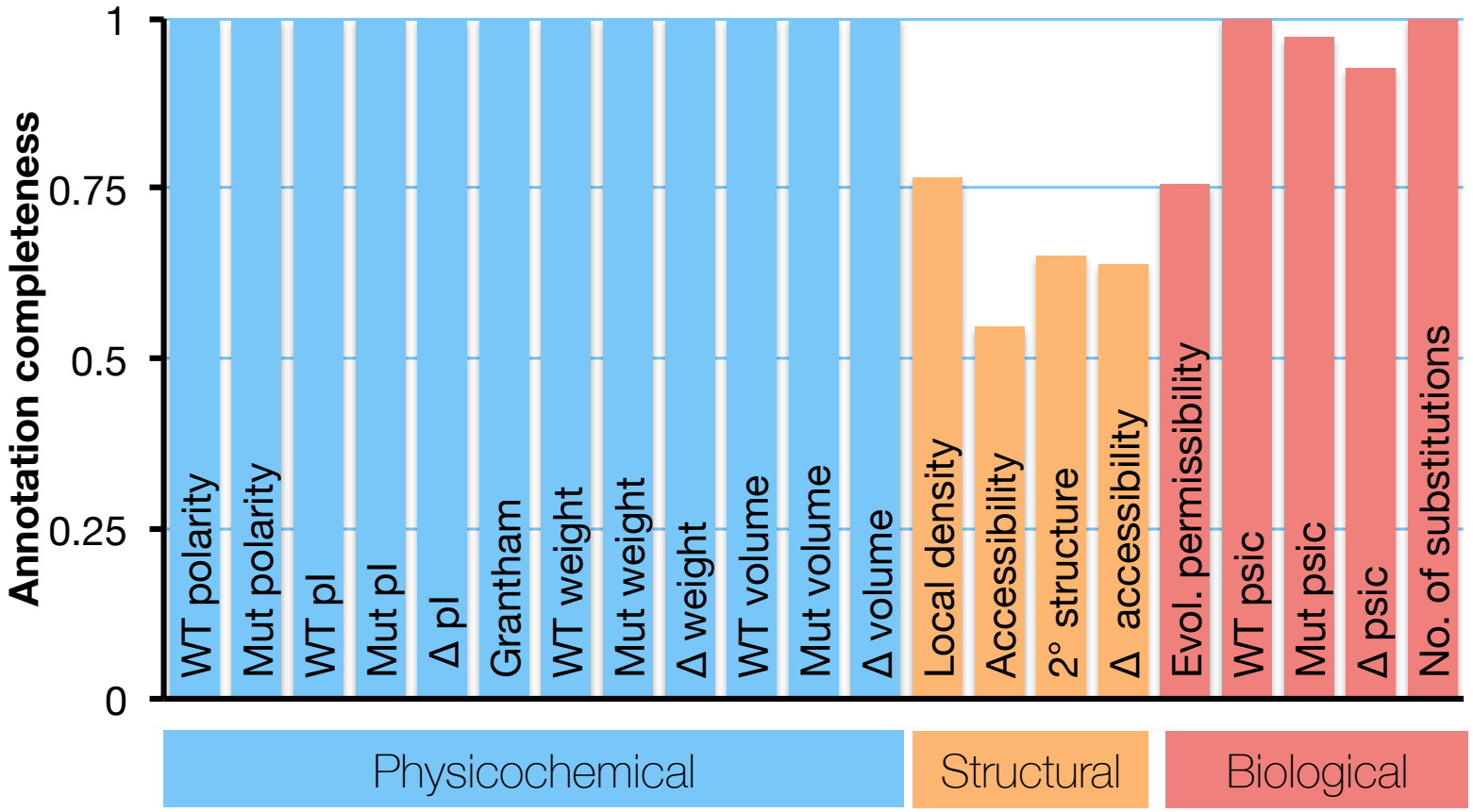
Focus on human disease variants
Unable to capture activity-enhancing variants

Large-scale mutagenesis data are becoming increasingly available

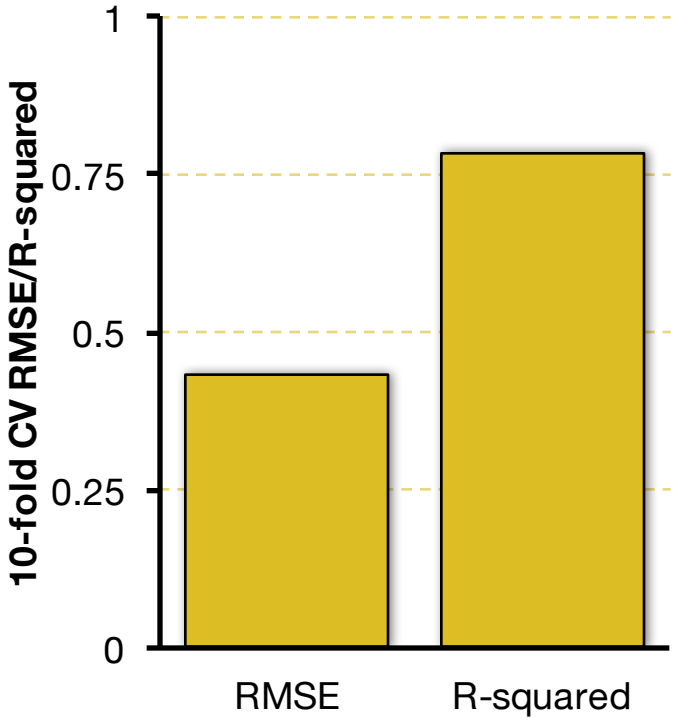


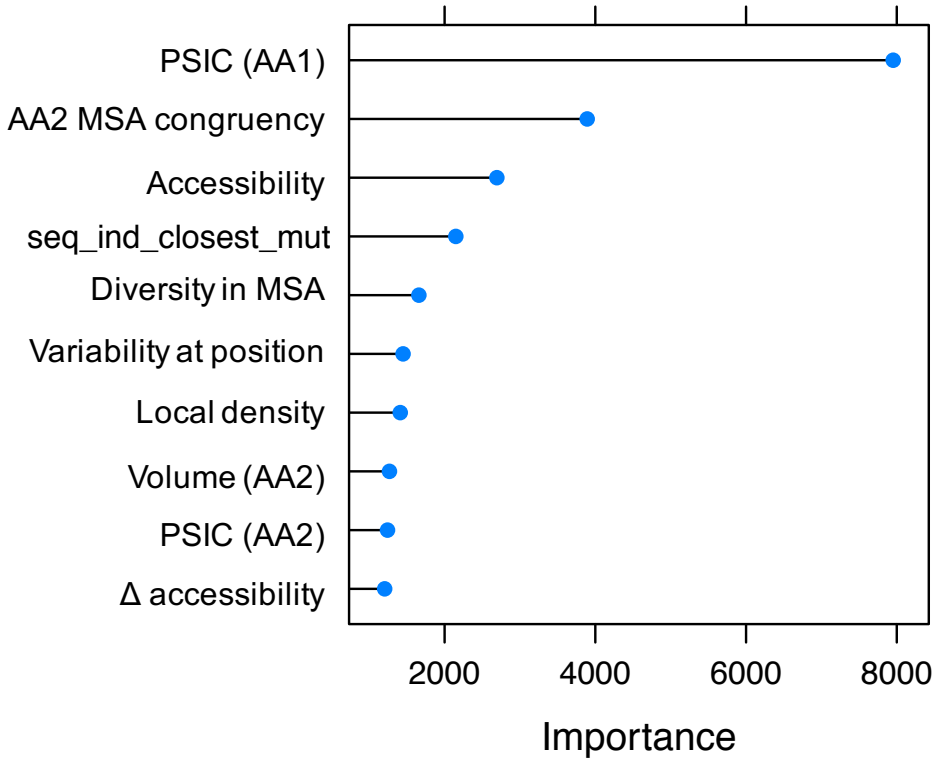


Mutations are annotated with 3 types of descriptive features



Global regression model is accurate





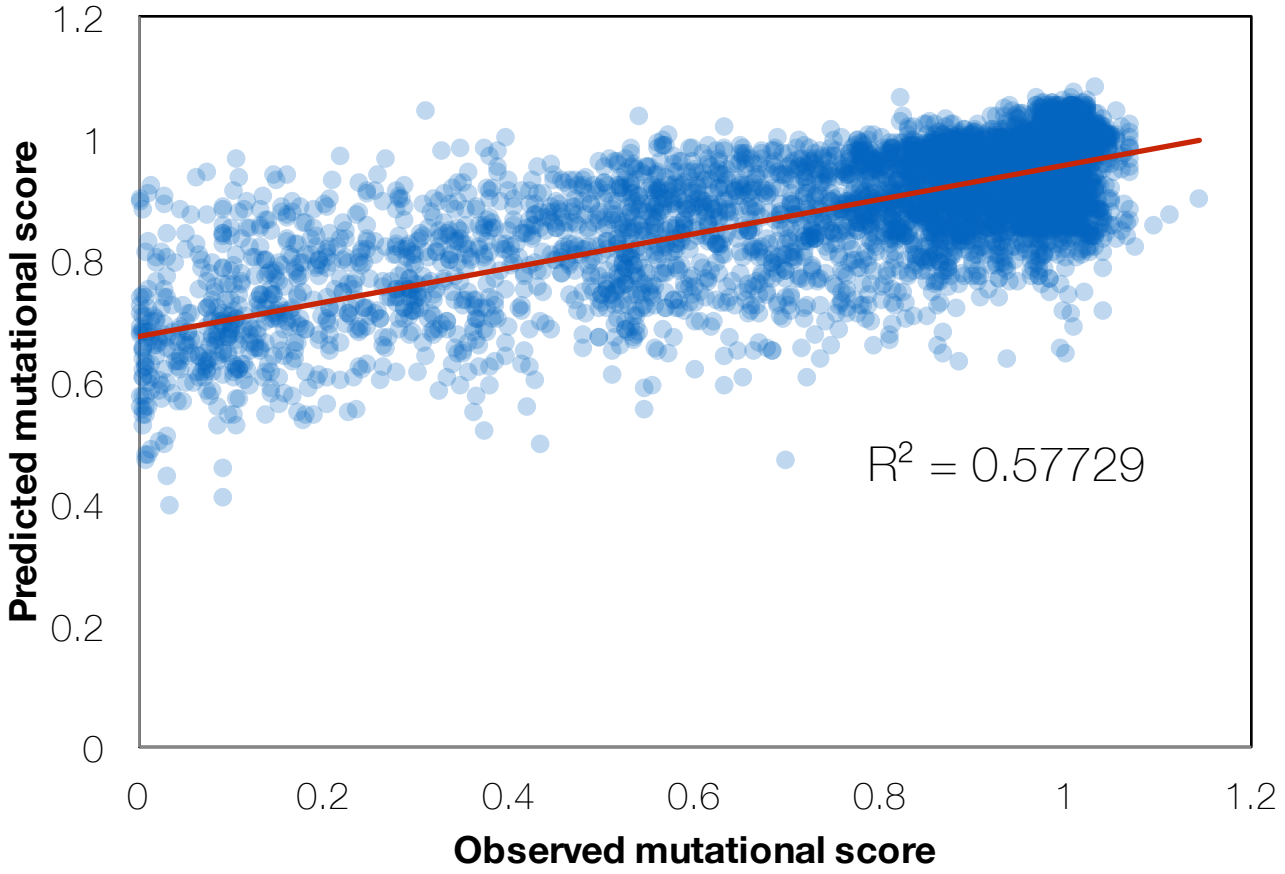
Regression model is modestly generalizable to unseen datasets

Training data

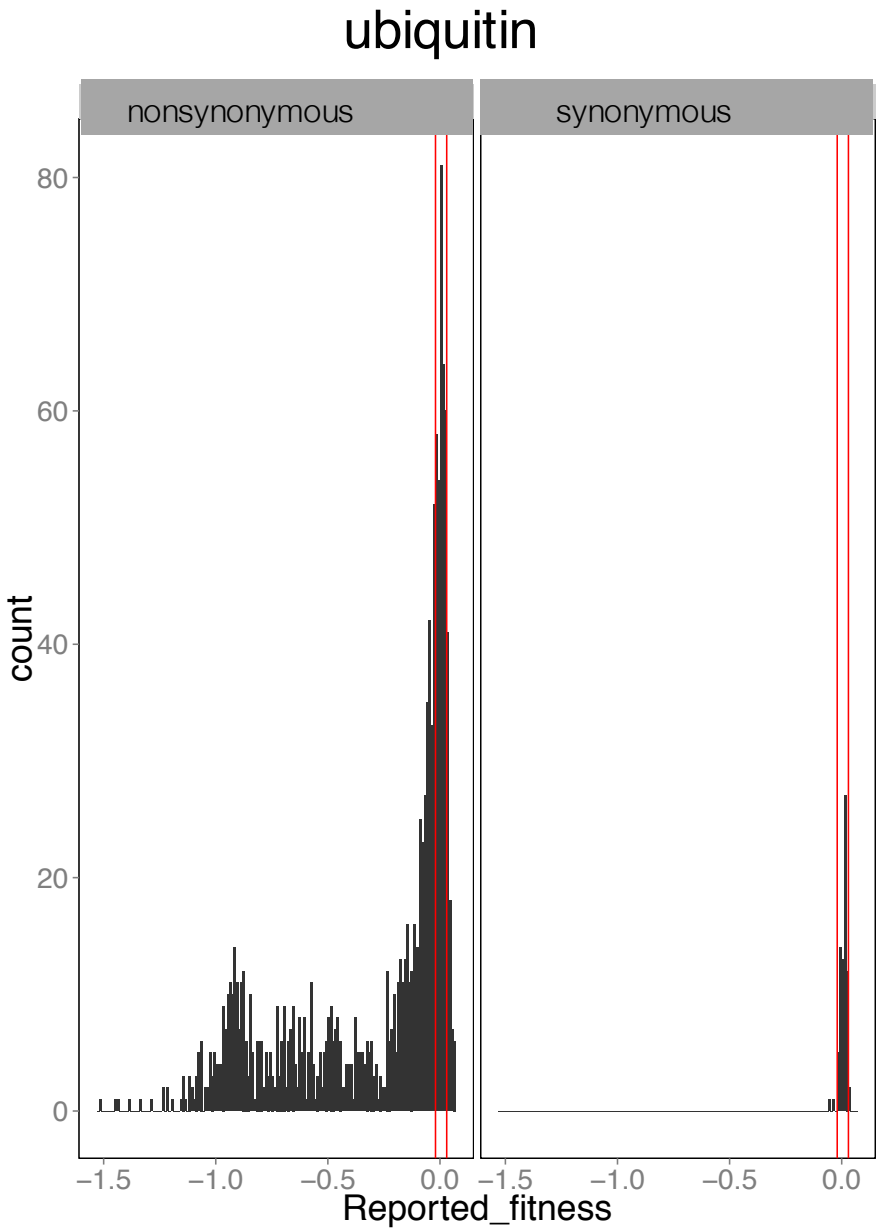
BRCA1
Hsp90
Ubiquitin
WW-domain
E3-ligase
pab1

Testing data

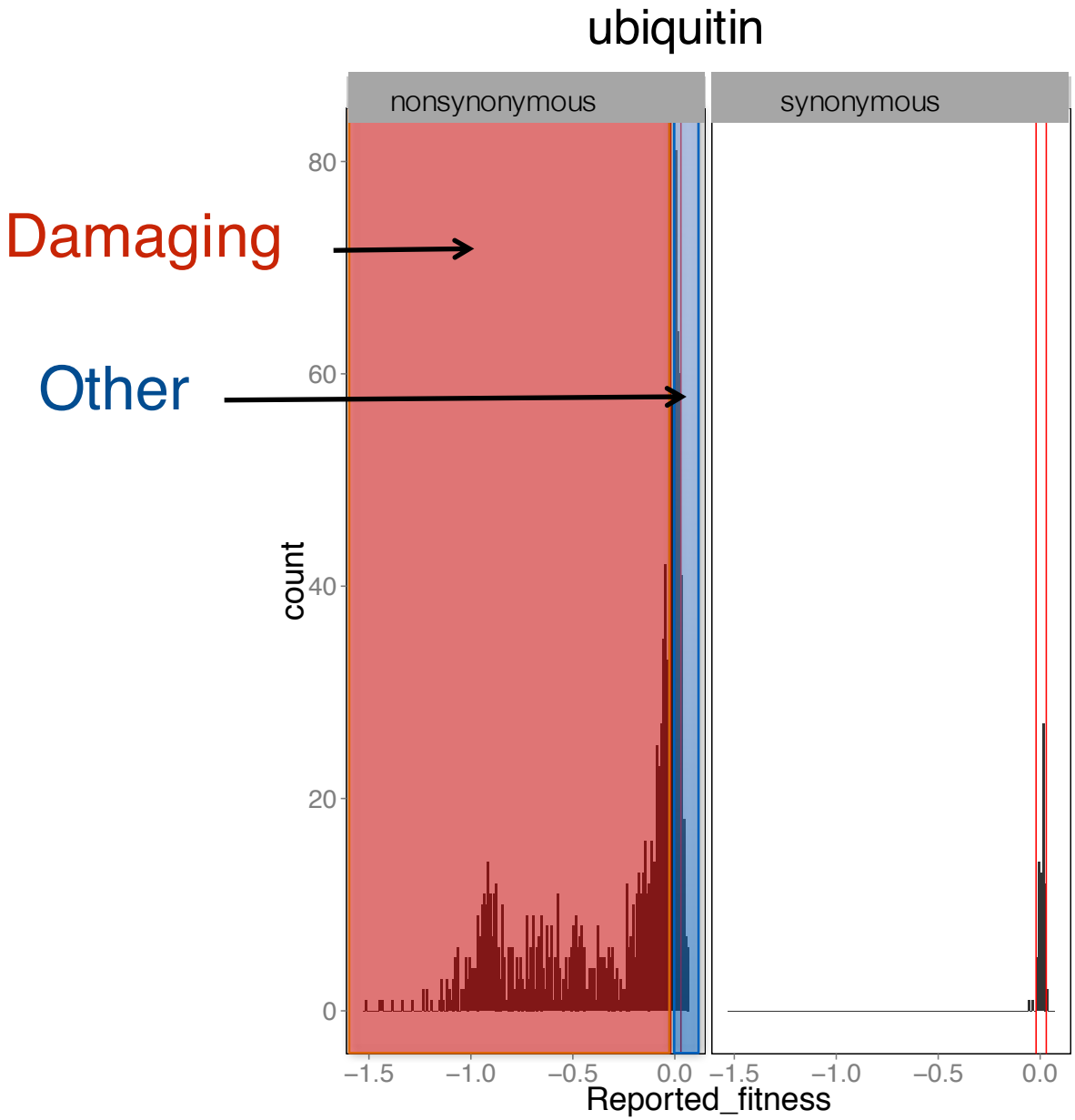
β-lactamase



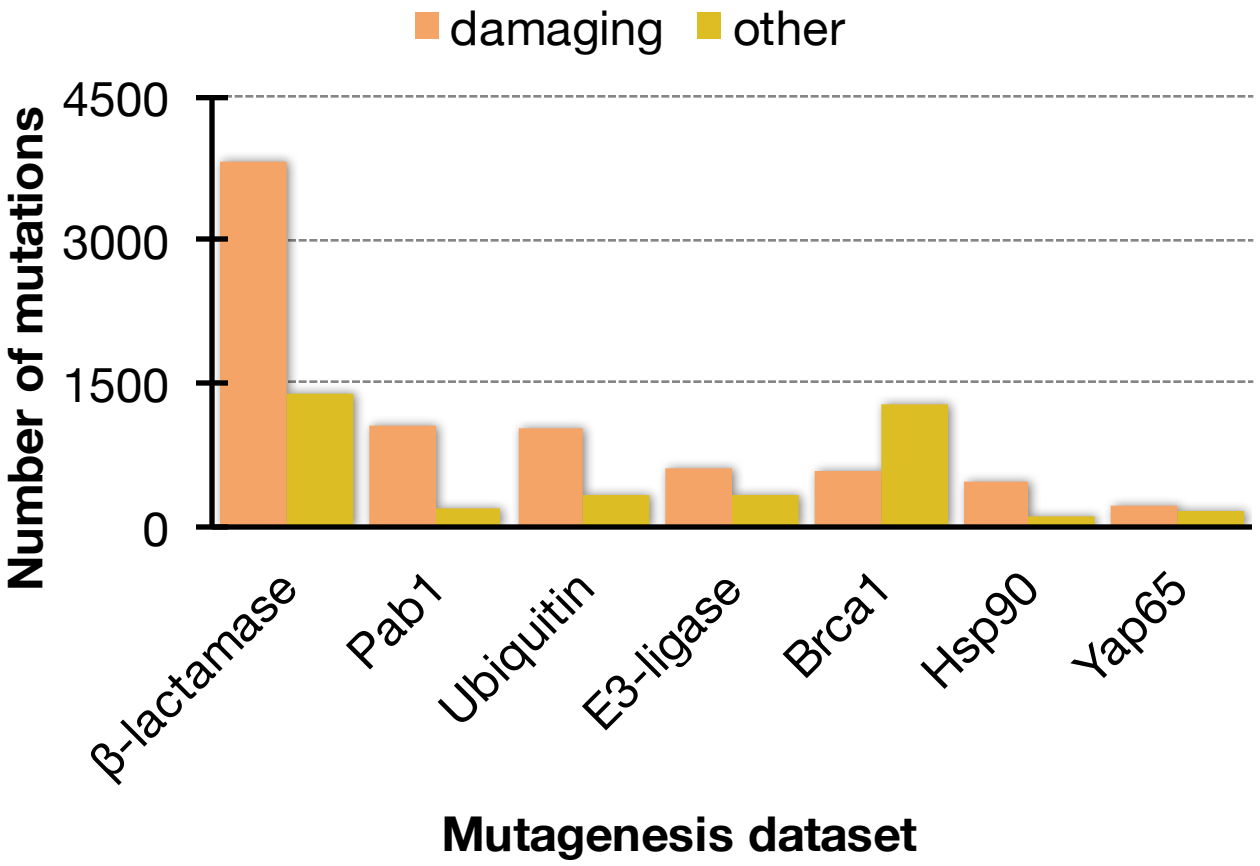
Discretization of functional scores



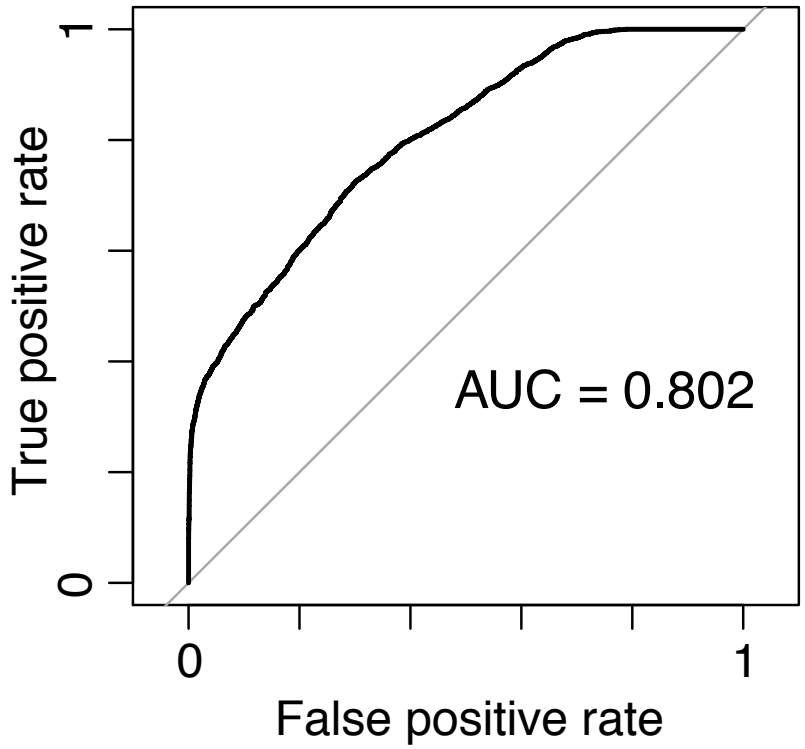
Discretization functional scores

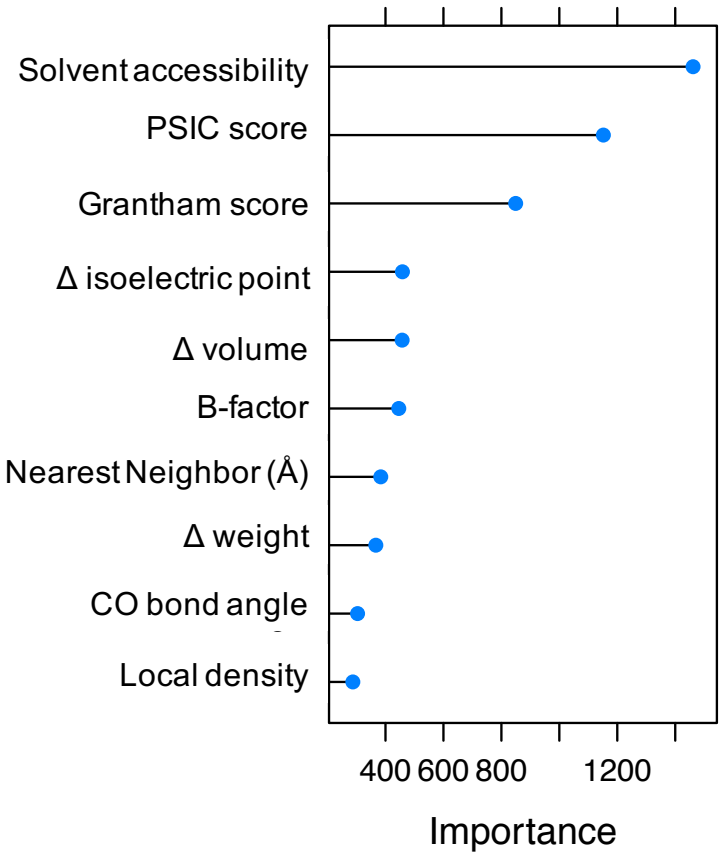


Most mutations are damaging



Classification model accurately predicts damaging mutations



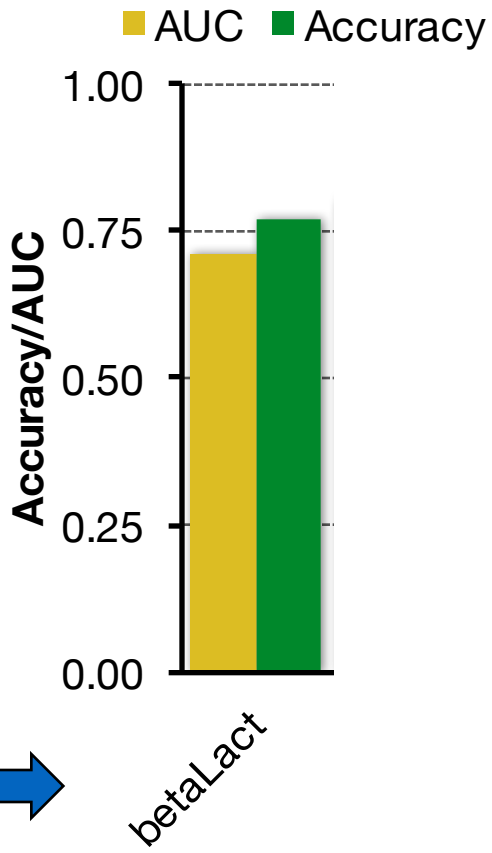


Accurate classification of unseen damaging mutations

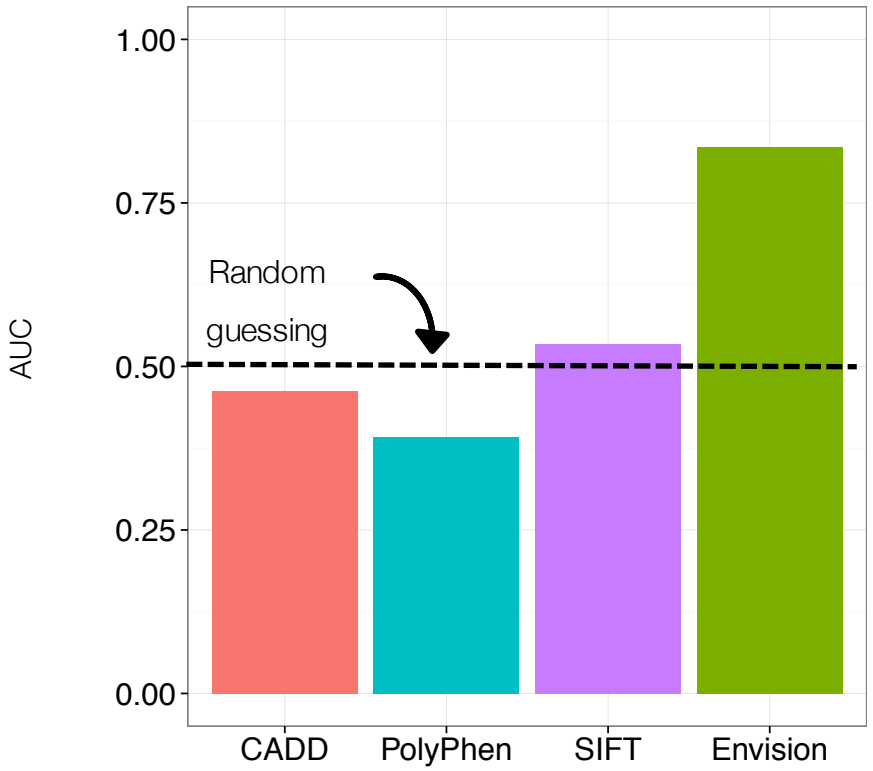
Training data

BRCA1
Hsp90
Ubiquitin
WW-domain
E3-ligase
pab1

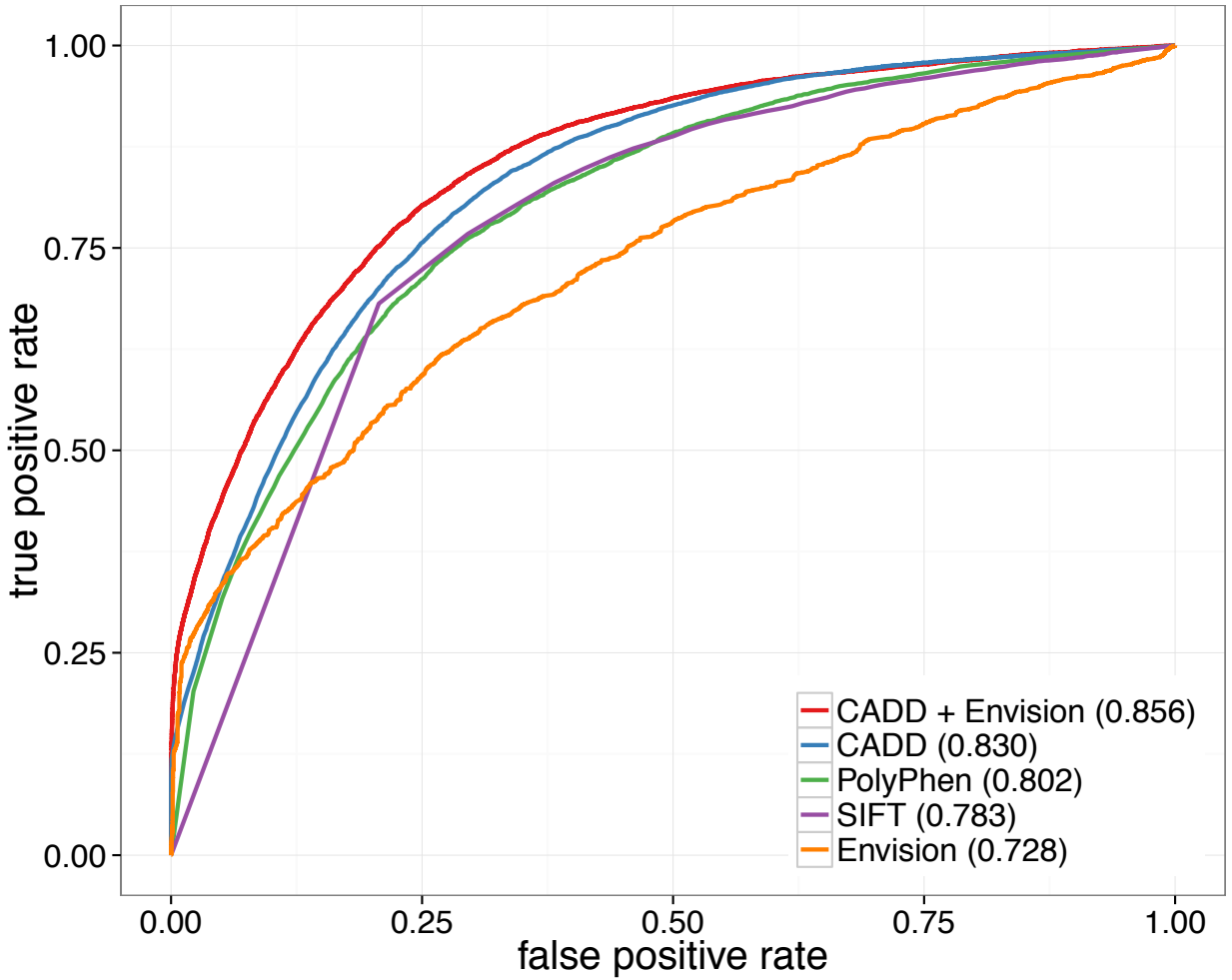
Testing data



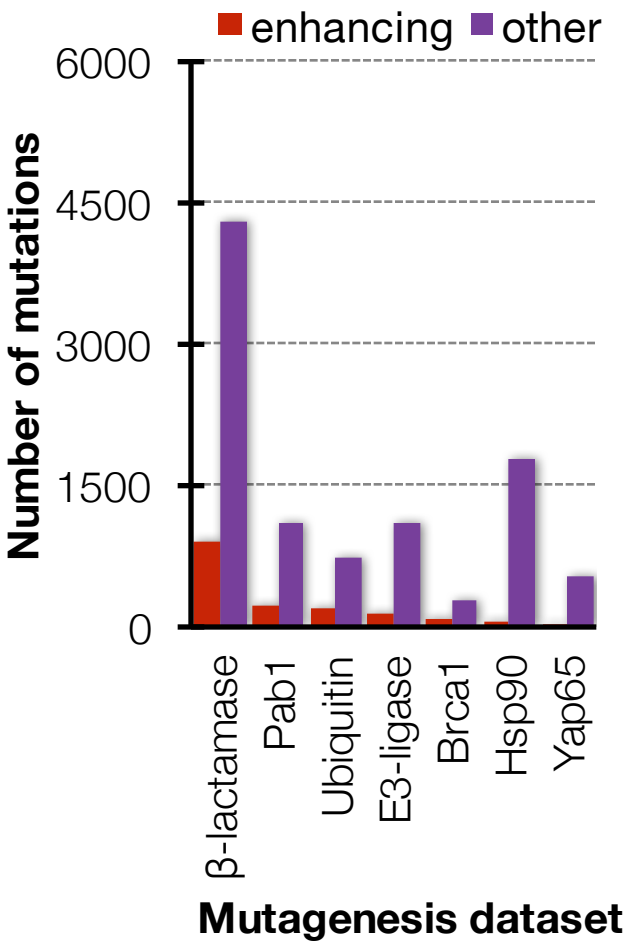
Envision, our best model, compares favorably with other variant effect predictors

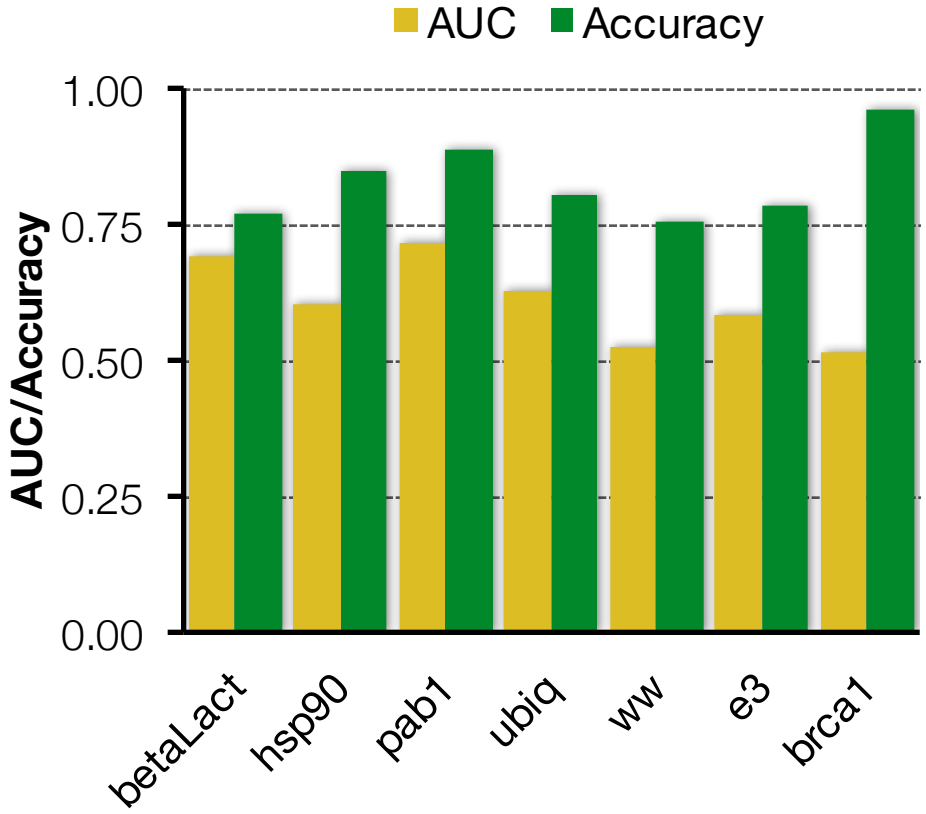


Envision can improve pathogenicity predictions

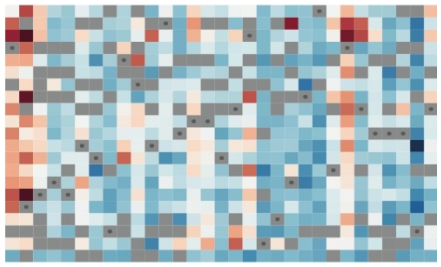


Envision predicts function-enhancing mutations



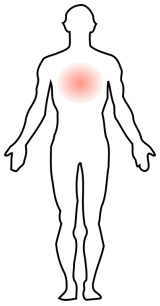


Large-scale data-driven prediction of protein variant effects

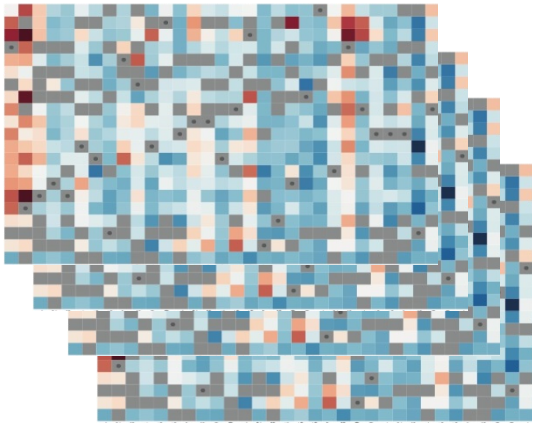


Sequence-function map

Protein-specific model



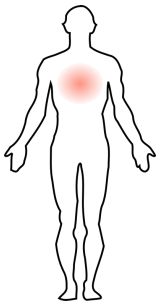
Accurate predictions for nearly all variants in protein of interest



Sequence-function maps

Envision

Global model



Accurate predictions for any variant in any protein

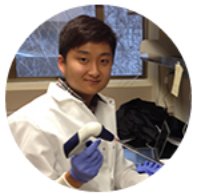
Acknowledgements


Jason Stephany

Kate Sitko



Barbara Taskinen

Ethan Ahler



Vanessa Gray

Melissa Chiasson



Kenny Matreyek

Evandro Ferrada


Miriam Williamson



Hannah Gelman

Funding

NIGMS

alzheimer's association

PGRN

Collaborators

Terry Speed
Alan Rubin

Jay Shendure
Ron Hause
Jens Leubeck

Stan Fields

Lea Starita

Jay Hesselberth
Molly Gasperini