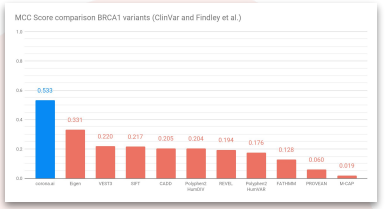
Missense variant effect prediction

with

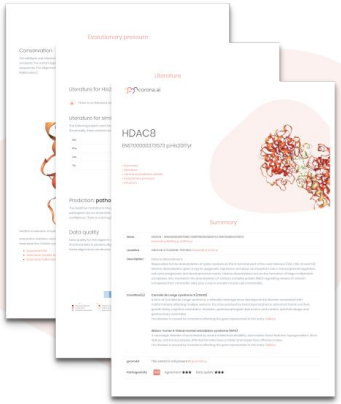
Helix

Overview

1. **Pathogenicity predictor for missense variants**
    - Filtering, ranking, disease association, …
    - Complete proteome coverage
2. **In-depth variant interpretation**
    - Understanding mechanisms, binding effects, PPI, …



improved predictions compared to REVEL, PolyPhen, CADD, etc

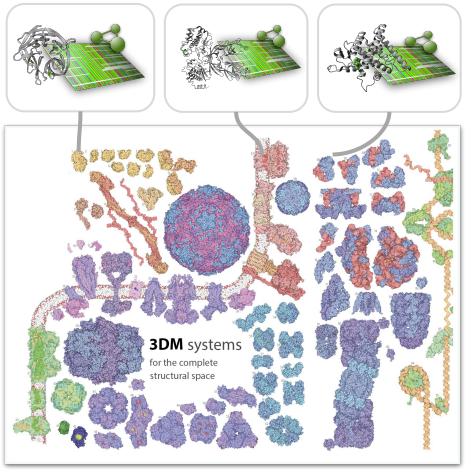


insights into individual variants based on protein structures, evolution, literature, etc

Applications

- Clinical diagnostics
- Disease association studies
- Patient stratification
- Strain engineering
- Pathway investigation

Training Helix

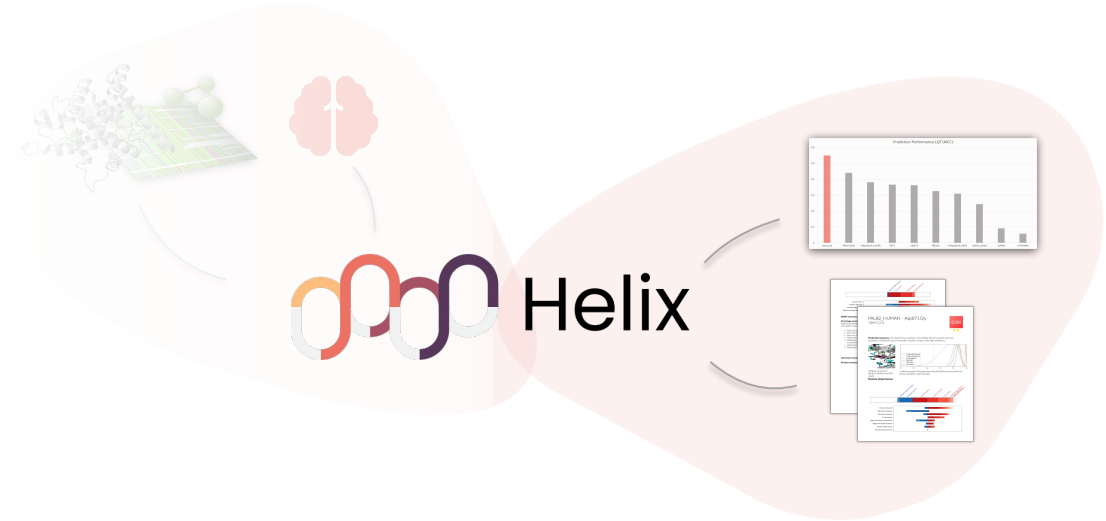


1. **Reference variants and their classes**
   Helix was built using reference variants obtained from gnomAD, ClinVar and the VKGL dataset. The full dataset consists of **~400k variants**, covering roughly **8000 genes related to genetic disorders.**

2. **Features describing variants**
   Variant annotations are mainly built using 3DM.
   - very **accurate and deep alignments**
   - allow for **more and better metrics** describing evolutionary constraints
   - extensive structure-based features

   Helix does **not** use any existing predictor scores as features.
   Helix does **not** use any MAF data as features.

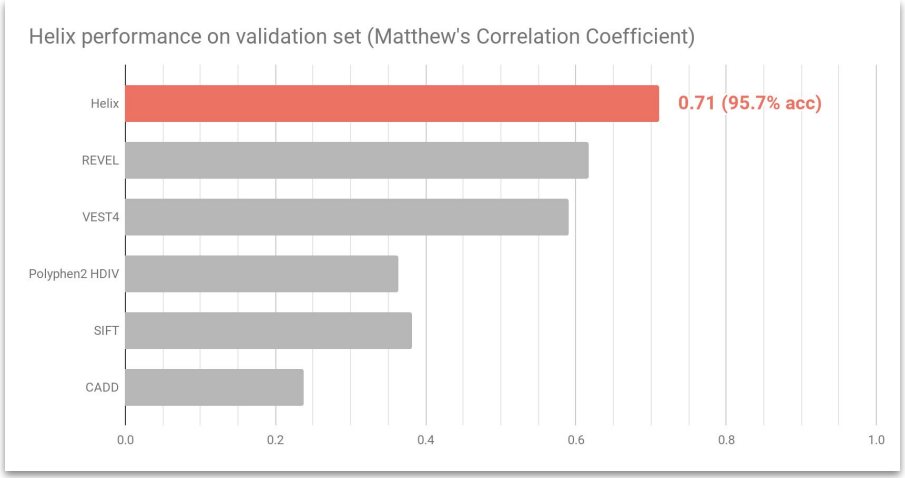
3. **Machine learning**
   Helix is built around a number of state of the art machine learning technologies, applying a ensemble approach based on neural nets and gradient boosting algorithms.



Helix offers world-class performance

Increased predictive performance over the whole human exome

Validation set - 10 fold cross-validation



Helix predictions were obtained by using 10-fold cross validation.

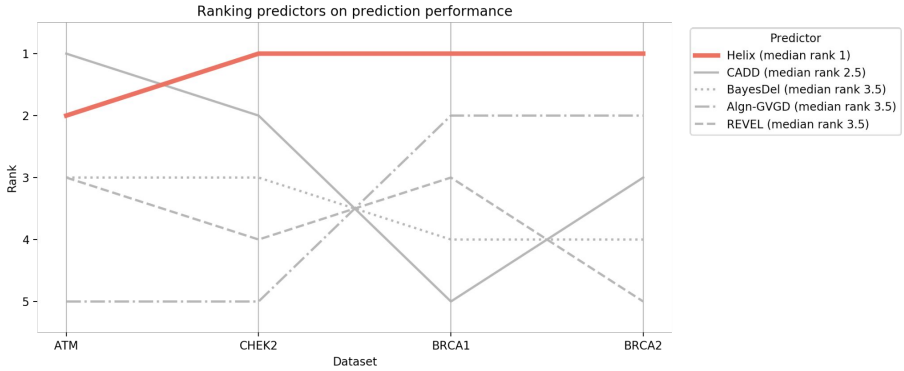
In case of Helix prediction scores were obtained by testing on **unseen** genes (not in training set). Even though this is often **not** the case for the other predictors Helix outperforms all competitors.

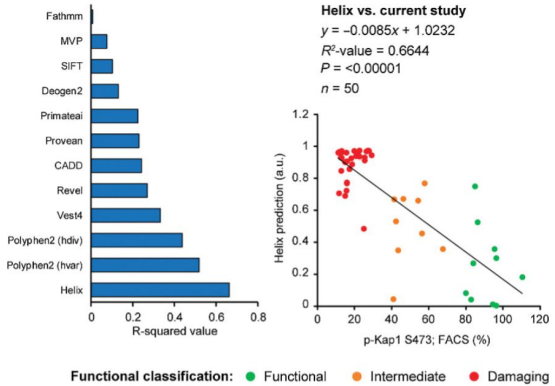
Scores displayed are the Matthew's Correlation Coefficient. This metric is a score that balances precision and recall. Performance close to random scores 0 in this metric, while a perfect score is equal to 1. In addition, accuracy (acc) is indicated for the Helix predictions.

Helix

Verified on novel experimental datasets

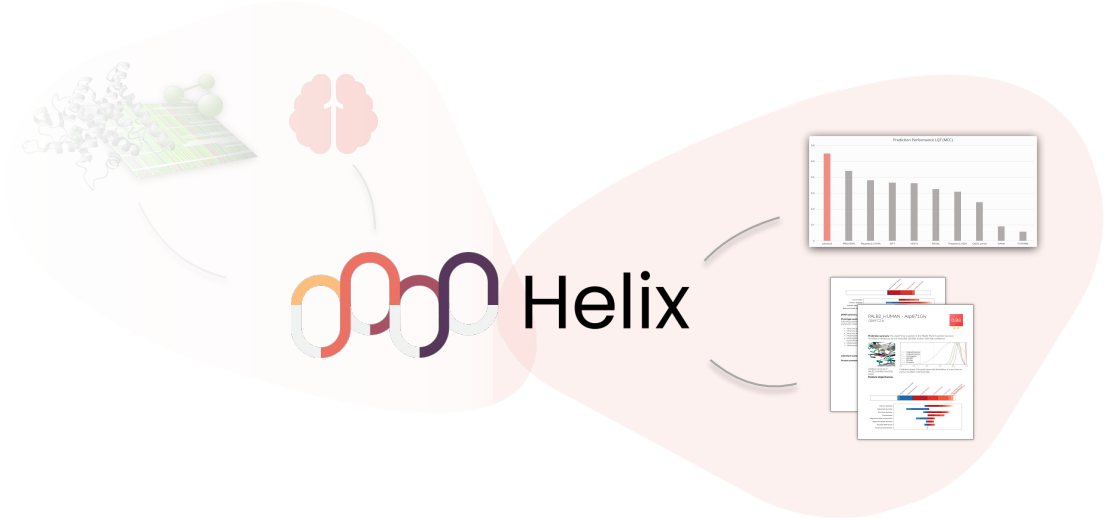
Helix was confirmed to be the best predictor of cancer causing missense variants in two separate studies.
Helix consistently ranks at the top of predictors. Scores shows high correlation with measured deleteriousness.





Boonen et al. Functional Analysis Identifies Damaging CHEK2 Missense Variants Associated with Increased Cancer Risk, MOLECULAR CELL BIOLOGY | FEBRUARY 15 2022
Dorling et al. Breast cancer risks associated with missense variants in breast cancer susceptibility genes, GENOME MEDICINE, 2022

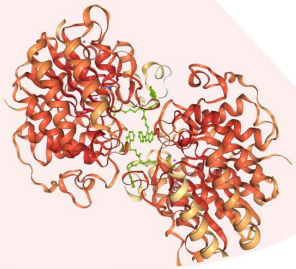
Helix



Detailed reports for all possible missense variants

HDAC8

ENST00000373573 p.His201Tyr



Summary

Gene	HDAC8 - ENSG00000147099 | ENSP00000362674 | ENST00000373573 Ensembl RefSeq UniProt
Location	GRCh38 X:72329516-72573103 Ensembl UCSC
Description	histone deacetylase 8 Responsible for the deacetylation of lysine residues on the N-terminal part of the core histones (H2A, H2B, H3 and H4). Histone deacetylation gives a tag for epigenetic repression and plays an important role in transcriptional regulation, cell cycle progression and developmental events. Histone deacetylases act via the formation of large multiprotein complexes. Also involved in the deacetylation of cohesin complex protein SMC3 regulating release of cohesin complexes from chromatin. May play a role in smooth muscle cell contractility.
Condition(s)	**Cornelia de Lange syndrome 5 (CDLS5)** A form of Cornelia de Lange syndrome, a clinically heterogeneous developmental disorder associated with malformations affecting multiple systems. It is characterized by facial dysmorphisms, abnormal hands and feet, growth delay, cognitive retardation, hirsutism, gastroesophageal dysfunction and cardiac, ophthalmologic and genitourinary anomalies. The disease is caused by mutations affecting the gene represented in this entry: OMIM **Wilson-Turner X-linked mental retardation syndrome (WTS)** A neurologic disorder characterized by severe intellectual disability, dysmorphic facial features, hypogonadism, short stature, and truncal obesity. Affected females have a milder phenotype than affected males. The disease is caused by mutations affecting the gene represented in this entry: OMIM
gnomAD	This variant is not present in gnomAD.
Pathogenicity	97% Agreement: ★★★ Data quality: ★★★

Literature

Literature for His201Tyr in HDAC8

There is no literature available for this specific variant

Literature for similar variants in homologous proteins

The following papers were found describing similar variants in proteins homologous to HDAC8.
Structurally, these variants are located at an equivalent position compared to His201Tyr.

Ala

2007-11-03	Q841U6_THECA	H183A	**Functional analysis of a histone deacetylase-like protein of Thermus caldophilus GK24 in mammalian cell.** Kim YS, Song YM, Kwon HJ *(Biochem. Biophys. Res. Commun., 2007-11-03)*	17767915
2001-06-29	HDAC1_HUMAN	H199A	**SMRTE inhibits MEF2C transcriptional activation by targeting HDAC4 and 5 to nuclear domains.** Wu X, Li H, Park EJ, Chen JD *(J. Biol. Chem., 2001-06-29)*	11304536
1998-03-31	HDAC1_HUMAN	H199A	**A role for histone deacetylase activity in HDAC1-mediated transcriptional repression.** Hassig CA, Tong JK, Fleischer TC, Owa T, Grable PG, Ayer DE, Schreiber SL *(Proc. Natl. Acad. Sci. U.S.A., 1998-03-31)*	9520398

Phe

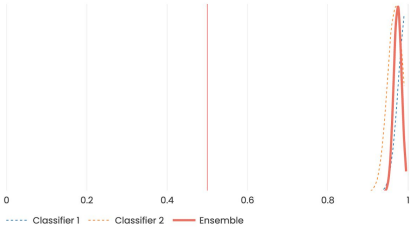
Leu

Tyr

Helix prediction details

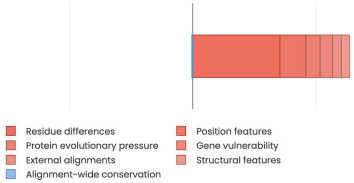
Prediction: **pathogenic** 97%

The His201Tyr mutation in the protein has been classified as pathogenic by our ensemble classifier system, with very high confidence. There is a 98% agreement between all subclassifiers.



Data quality

Data quality for this region is considered **good**. This means that enhanced, deep alignments are present and there is a variety of data for the algorithm to predict from.



Prediction factors

External models have estimated which sets of features contributed primarily to the classification. These sets of features are listed here.

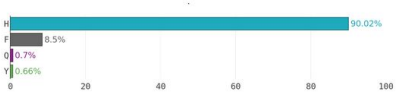
Primary contributing factors

- Residue differences point towards pathogenic
- Protein evolutionary pressure points towards pathogenic
- Position features point towards pathogenic
- Gene vulnerability points towards pathogenic

Evolutionary pressure

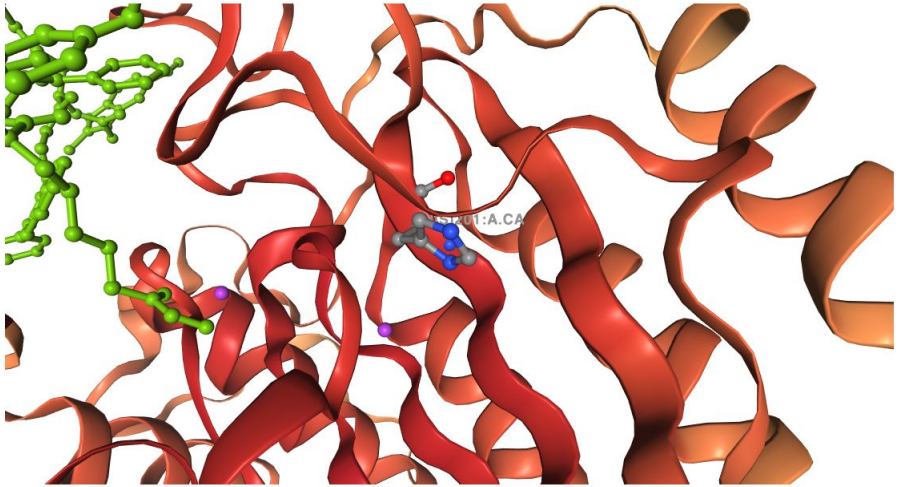
Conservation

The wildtype was observed in 90% of the 15999 sequences analyzed. The variant type was observed in < 1% of observed sequences. The alignment type is structural (based on 3DM PDBClusters).



Structure

□ Toggle interactions

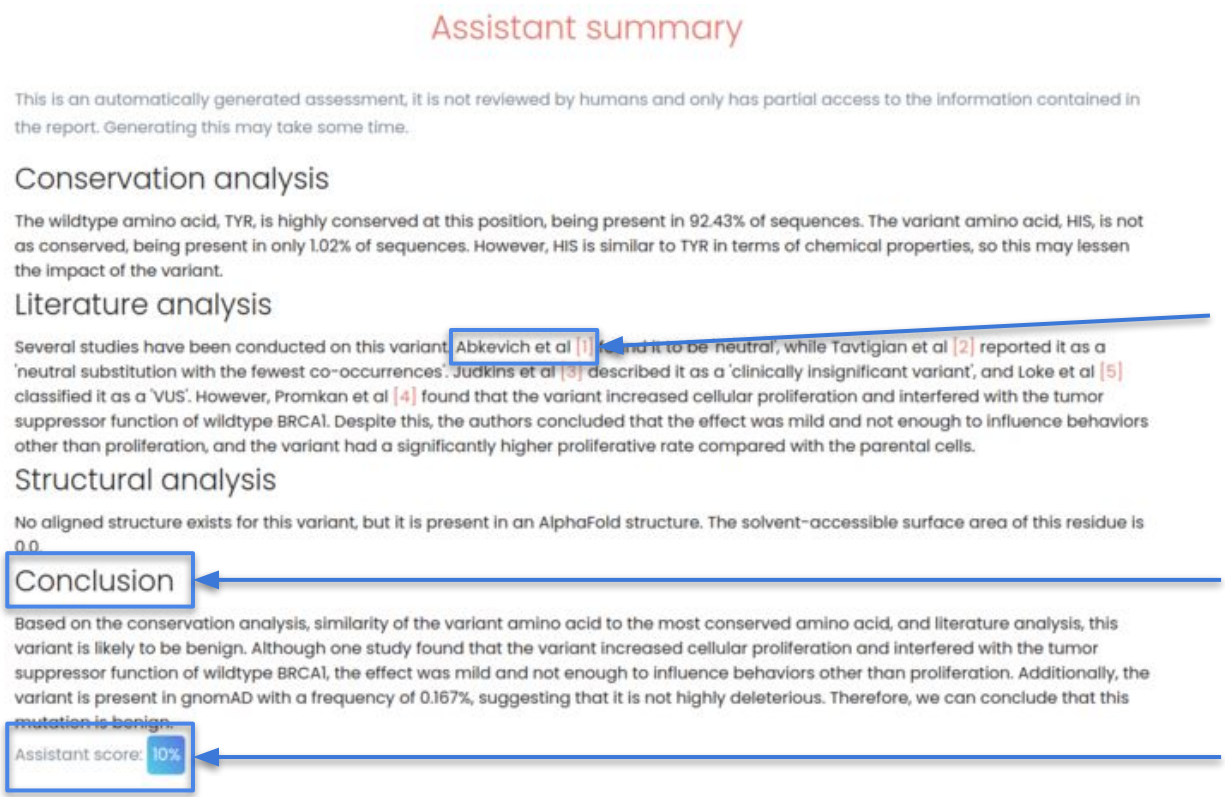


His201 is involved in 4 hydrogen bonds, 1 PiPi interaction and 2 hydrophobic interactions with neighbouring residues.

Interaction statistics were calculated using advanced molecular optimization techniques and may not be visible in the plain PDB file. Please download the YASARA scene to explore the interactions in more detail.

- Download PDB
- Download YASARA Scene
- Download PyMol Scene

- Integrates literature
- Uses state of the art NLP methods to explain mutation effects



Cites authoritative literature

Detailed explanation based on combination of data

Accurate summary scores

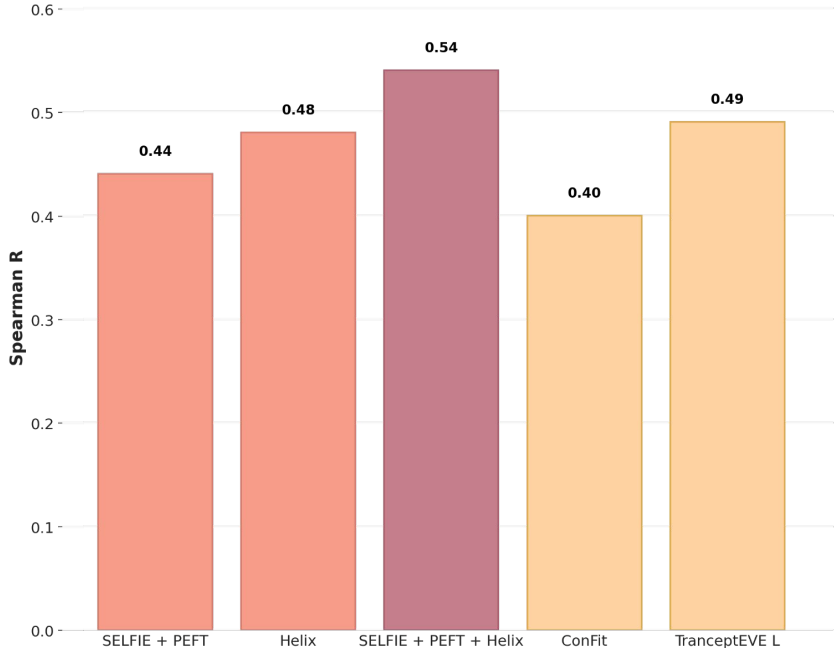
Helix APIs for integration with in-house pipelines

Helix Prediction API: provides variant predictions and related statistics. Suitable for integration with in-house variant annotation pipelines for ranking, filtering, etc.

Helix & Protein Engineering

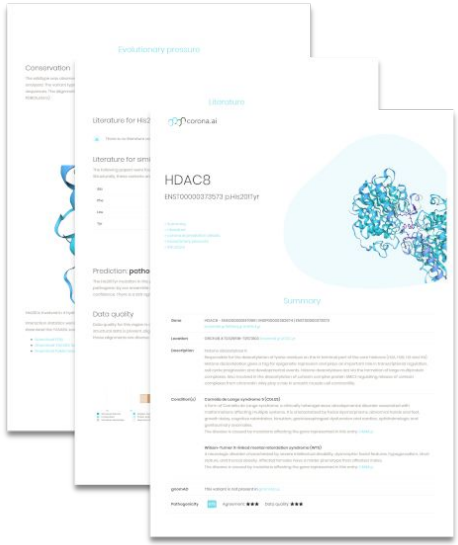
Helix integrates with our Protein Engineering efforts. Helix Predictions can be made for any organism, which makes it an excellent addition to our protein Engineering toolkit.

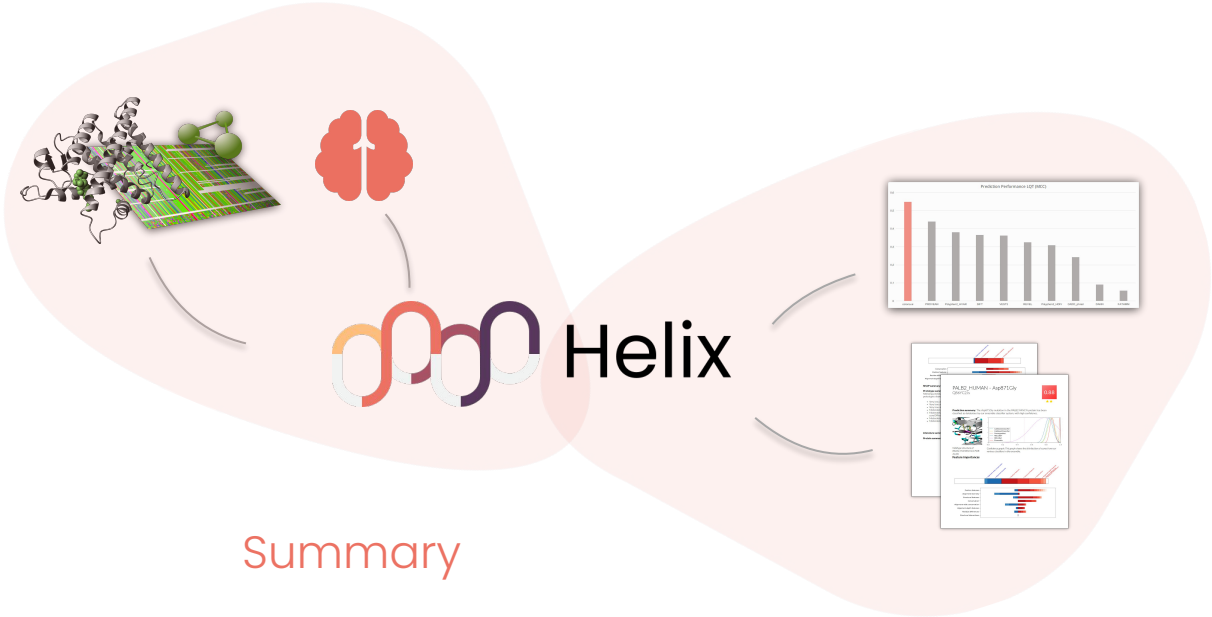
Our Pathogenicity predictions perform excellently on their own & in conjunction with targeted Protein Engineering solutions (SELFIE).



Average spearman correlation score on the ProteinGym Protein Engineering benchmark dataset. Helix can be combined with our Protein Engineering solutions to provide state of the art fitness prediction performance.

- We can build Helix systems for any organism, even custom genomes
- Built on 3DM alignments
- Fully integrated AlphaFold 2 structures where needed
- Access to literature
- Useful for strain engineering
    - Existing customers have achieved **significant improvements in organism fitness** using Helix LUCA information
    - Users have noted that they are **extremely impressed by the interface** and the integration of structure with the predictions
- Helix LUCA systems are built specifically for you on project basis





Summary

- State of the art prediction performance
- Interpretable predictions
- Extensive reporting on variants (**https://helixlabs.ai/sample**)
- Literature integration
- Well-documented APIs for predictions
- State of the art integration for Protein Engineering