

# FRANCISCO JAVIER GUZMÁN-VEGA

Bioinformatician ~ Biologist ~ Developer

 [guzmanfj.com](https://guzmanfj.com)

 [francisco.guzmanvega@kaust.edu.sa](mailto:francisco.guzmanvega@kaust.edu.sa)

 KAUST, Thuwal

 [www.linkedin.com/in/guzmanfj](https://www.linkedin.com/in/guzmanfj)

## SUMMARY

Trained as a Biotechnology Engineer, passionate in Bioinformatics and Computer Science. I am committed to the application of Bioinformatics and AI to solve problems in research and healthcare. I am an independent learner who's obsessed with achieving a deep understanding of the problems around me and how to solve them.

## SKILLS

**Languages:** Python, R.

**Technologies:** Linux, HPC cluster, Docker, AlphaFold.

**Interests:** Structural Biology, Statistics.

## EDUCATION

- 6/2019 - 5/2024 **Ph.D. in Bioscience** KAUST  
Finish date May 2024. Thesis titled "Large Scale Approaches for Protein Research with Machine Learning-Enabled Bioinformatic Tools". As part of the Structural Biology and Engineering (StruBE) lab, led by Professor Stefan Arold.
- 8/2017 - 5/2019 **M.Sc. in Bioscience** KAUST  
Thesis titled "Applications in computational structural biology: the generation of a protein modelling pipeline and the structural analysis of patient-derived mutations". As part of the Structural Biology and Engineering (StruBE) lab, led by Professor Stefan Arold.
- 8/2012 - 5/2017 **B.Sc. in Biotechnology Engineering** Tecnológico de Monterrey, Campus Monterrey  
GPA of 96.8 (/100). Scholarship to the Academic Talent.

## SOFTWARE PROJECTS

### AlphaCRV

<https://github.com/strubelab/AlphaCRV>

Python package that helps to identify correct interactors in a one-against-many AlphaFold-Multimer screen by clustering, ranking, and visualizing conserved binding topologies. It was able to identify the correct binding topologies in 3/3 tested proteome-wide interaction screens (40,000+ models each).

### AlphaFold-Ibex

<https://github.com/strubelab/alphafold>

AlphaFold wrapper that allows for easy parallelization in massive modeling tasks in the Slurm scheduler for HPC clusters. Separates CPU and GPU-bound steps to optimize resources, and provides quality scores and plots to visualize results. Used by 20+ people in KAUST from 4+ research groups

### Multiprot

<https://github.com/strubelab/multiprot>

Python-based pipeline to model proteins by connecting two or more structured domains with disordered linkers. Supports single or multiple-chain models and symmetric structures. Used by several members of our lab.

### Variant annotator

Work in progress

Pipeline to annotate protein variants with structural features from AlphaFold models, as well as sequence-based features such as conservation and pathogenicity scores. Aimed to assist in the interpretation of protein variants related to disease. Annotated several million variants from multiple databases (manuscript in preparation, see "Current Research Projects" below).

### CoolerCodonOpt

<https://github.com/strubelab/CoolerCodonOpt>

DNA codon optimization tool using the Python `dnachise1` library. It provides an interface to run codon optimization on one or more protein or DNA sequences. Used by several members of our lab.

## CURRENT RESEARCH PROJECTS

### Automating the annotation of protein variants and their relation to disease.

Manuscript in preparation

- Created pipeline to annotate protein variants with structure and sequence-based features.
- Studied the structural landscape of pathogenic vs benign variants in different datasets, and identified features characteristic to both classes.
- Project in collaboration with two Master's students. Manuscript in preparation.
- In the future I would really like to turn this project into a tool or service that assists clinicians in the interpretation of protein variants related to disease, with the integration of LLMs.

Python / Linux-based tools

## Folding and function of the rice "de novo-ome"

Manuscript in preparation

- Studied the structural and physicochemical characteristics of *de novo* proteins in rice with bioinformatic methods, and found differences and similarities with random and canonical rice protein groups.
- Performed proteome-wide screens of selected *de novo* proteins vs the rice proteome, and found potential binding partners.
- Obtained insights into the evolution and function of *de novo* proteins in rice.  
Python / Linux-based tools / AlphaFold

## SELECT PUBLICATIONS

---

- 2023 **Leveraging AI Advances and Online Tools for Structure-Based Variant Analysis**  
**Guzmán-Vega, F. J.**, González-Álvarez, A. C., Peña-Guerra, K. A., Cardona-Londoño, K. J., & Arold, S. T. *Current Protocols*, 3, e857. doi: 10.1002/cpz1.857.  
*I led the design and drafting of a protocol to help clinicians and researchers interpret protein variants from a structural perspective.*
- 2023 **An open-source, automated, and cost-effective platform for COVID-19 diagnosis and rapid portable genomic surveillance using nanopore sequencing.**  
Ramos-Mandujano, G., Grünberg, R., Zhang, Y., Bi, C., **Guzmán-Vega, F. J.** et al. *Sci Rep* 13, 20349. <https://doi.org/10.1038/s41598-023-47190-w>.  
*I helped implement an automated RNA extraction protocol on a Tecan EVO-200 liquid handling robot.*
- 2020 **Rapid Evolution of Plastic-degrading Enzymes Prevalent in the Global Ocean.**  
Alam, I., Aalismail, N., Martin, C., Kamau, A., **Guzmán-Vega, F. J.** et al. *bioRxiv*. doi: <https://doi.org/10.1101/2020.09.07.285692>.  
*I performed 3D modeling of oceanic PETase-like sequences and designed a scoring system to assess their putative performance with respect to the PETase from *I. sakaiensis*.*
- 2024 **Novel Homozygous Variants of SLC13A5 Expand the Functional Heterogeneity of a Homogeneous Syndrome of Early Infantile Epileptic Encephalopathy.**  
Alsemari, A., **Guzmán-Vega, F. J.**, Meyer, B. F., Arold, S. T. *Pediatric Neurology*, 151, 68-72. doi: 10.1016/j.pediatrneurol.2023.10.005.  
*I modeled the 3D structure of the SLC13A5 dimer and evaluated disease-associated variants that strongly suggested an impairment of the sodium/citrate transport of SLC13A5.*
- 2023 **Expanding the genotype-phenotype landscape of PDE10A-associated movement disorders.**  
Bohlega S, Abusrair AH, Al-Qahtani Z, **Guzmán-Vega FJ**, ..., Arold ST. *Parkinsonism Relat Disord*. 2023 Mar;108:105323. doi: 10.1016/j.parkreldis.2023.105323. Epub 2023 Feb 13. PMID: 36805523.  
*I modeled the 3D structure of the PDE10A dimer and evaluated disease-associated variants that suggested differences between the dominant and recessive disease mechanisms.*

## SELECT CONFERENCES

---

- 2023 **ISMB-ECCB2023**  
Presented poster for the rice "de novo-ome" project (above).
- 2022 **ECCB2022**  
Presented poster for the rice "de novo-ome" project, and contributed to poster of the protein variant annotation project (above).

## TEACHING

---

- Have mentored and supervised 4+ Masters' students and visiting students during my time at KAUST.
- Gave a lesson for the Biomolecular Structure and Function course at KAUST.
- Have given workshops to lab members about the use of Ibbex and AlphaFold, with my own teaching material.

## AWARDS

---

- 2022 BESE Dean's Award

## EXTRACURRICULAR

---

- 2019-2023 Member of the KAUST Volunteer Fire Department

## LANGUAGES

---

**English** - C1+, **German** - A2, **Spanish** - native