# Gaussian Process Density Estimation to Obtain Ramachandran Probability Distribution

Agustina Arroyuelo[1*], Jorge A Vila[1,2] and Osvaldo A Martin[1]

1 IMASL-CONICET, Italia 1556, 5700-San Luis, Argentina.
2 Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, USA.

* **aarroyuelo@unsl.edu.ar**

## Introduction

Protein dihedral degrees of freedom play a central role in simulation and structural analysis of these biomolecules with $\phi$ and $\psi$ angles as the main torsionals influencing protein 3D structure. The Ramachandran map displays the sterically allowed regions and forbidden ones for the torsional angle combination ($\phi$, $\psi$) [1]. Therefore featuring enlightenment about the conformation that every possible aminoacid in a protein structure can present. Sampling of protein conformational space using a bayesian machinery requires the inclussion of a prior for the torsional angles ($\phi$, $\psi$). Taking these torsionals from a uniform distribution is a very poor prior for this task. In the present work we explore the implementation of a non parametric model, such as Gaussian Process to *learn* a Ramachandran Probability Distribution for its use in Bayesian Inference. Our implementation uses Python's probabilistic programming module, PyMC3 [2].

## Methods

Gaussian processes are often used as priors over functions, generally applied in regression and classification [3]. In this work we make use of a Logistic Gaussian process as prior for probability density functions as an alternative to other nonparametric methods such as Dirichlet process and Kernel Density Estimation [4]. A dataset was elaborated with a large number of torsional angle values (> 600000) obtained from structures from the Protein Data Bank. The structures selected for this dataset are high quality structures. The torsional angles were read using PyMOL [5]. A grid or 2D histogram, of gridsize 18, was constructed from this data. The gridsize was selected arbitrarily, but under the premise that smaller gridsize will produce better density estimation. The gaussian process was defined as $f(x) \sim \mathcal{GP}(0, \kappa(x, x'))$ with exponential quadratic covariance function. For the estimation of the Ramachandran Probability Distribution $p(x)$, where $p(x) = softmax(f(x))$ [6].

## Results

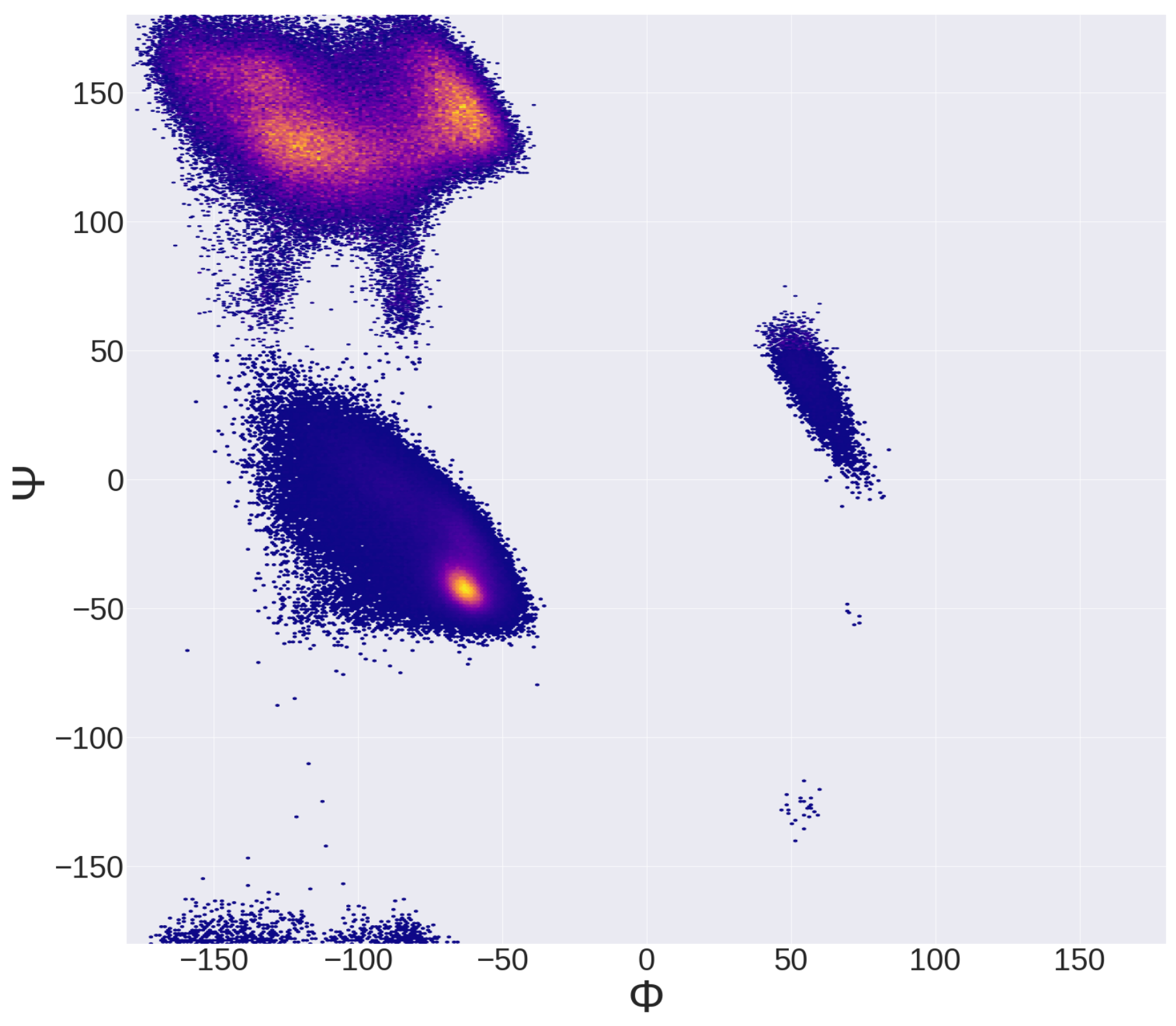

Figure 1: Ramachandran map showing torsional angle pairs in the dataset. Although glycine residues were included in the analisys, these are not shown.

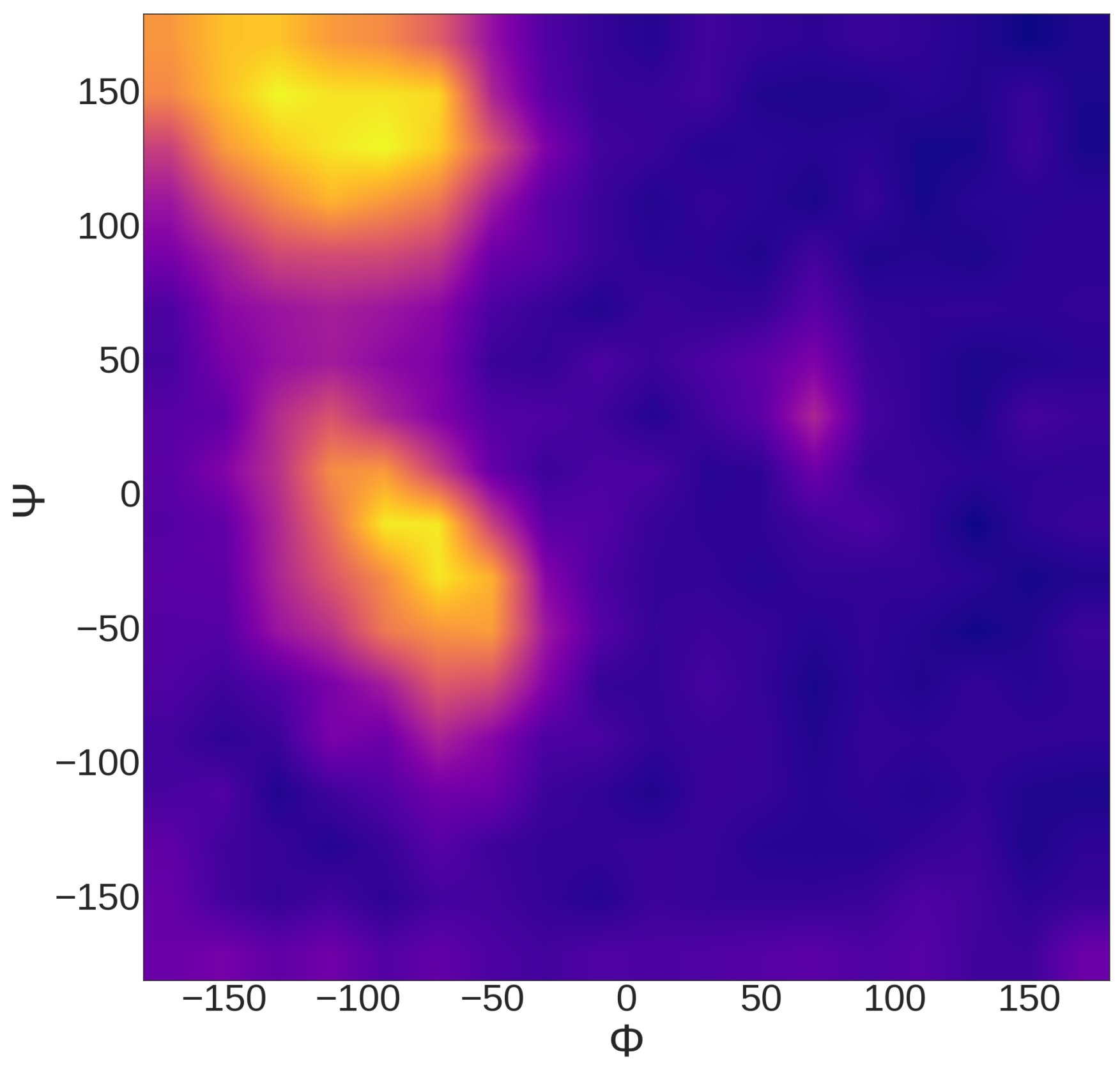

Figure 3: Bidimensional density estimation of the $\phi$ and $\psi$ torsional angles on a 18x18 grid.

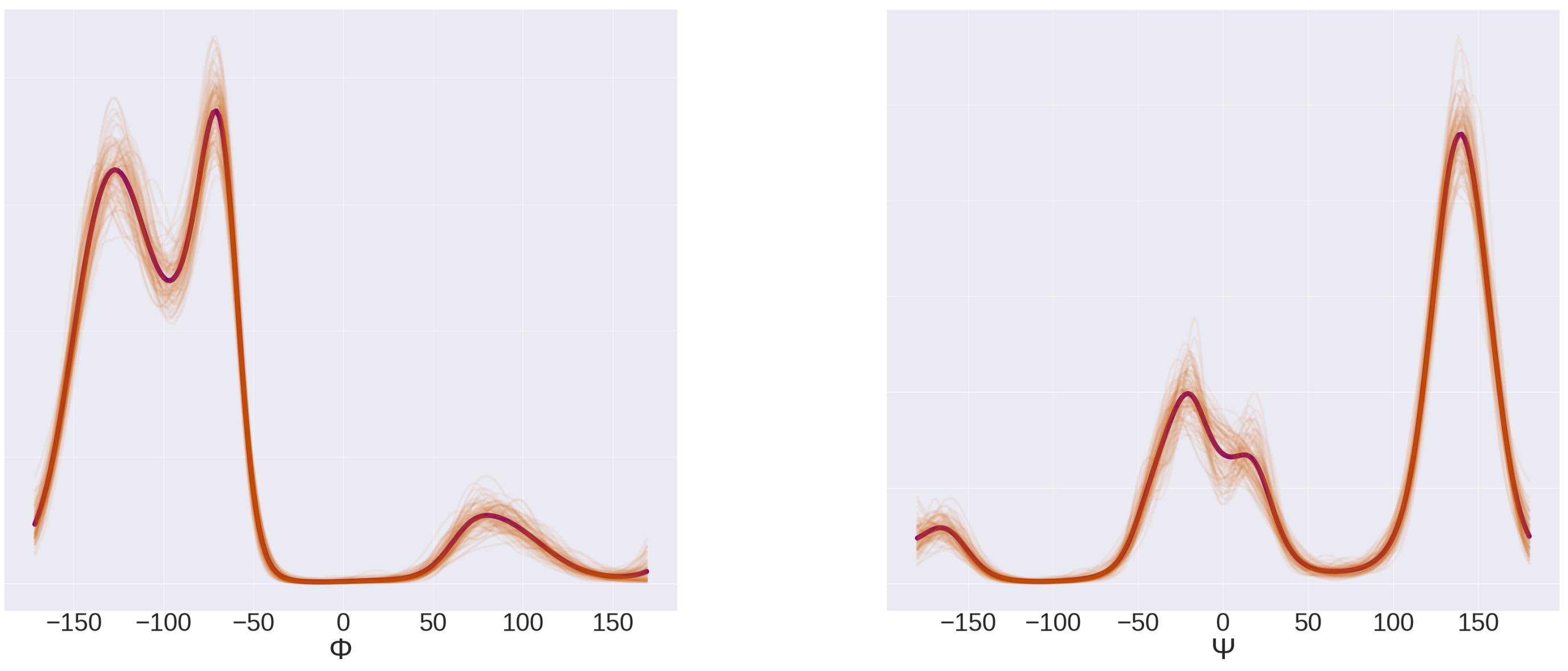

Figure 2: One-dimensional density estimation of the $\phi$ and $\psi$ torsional angles.

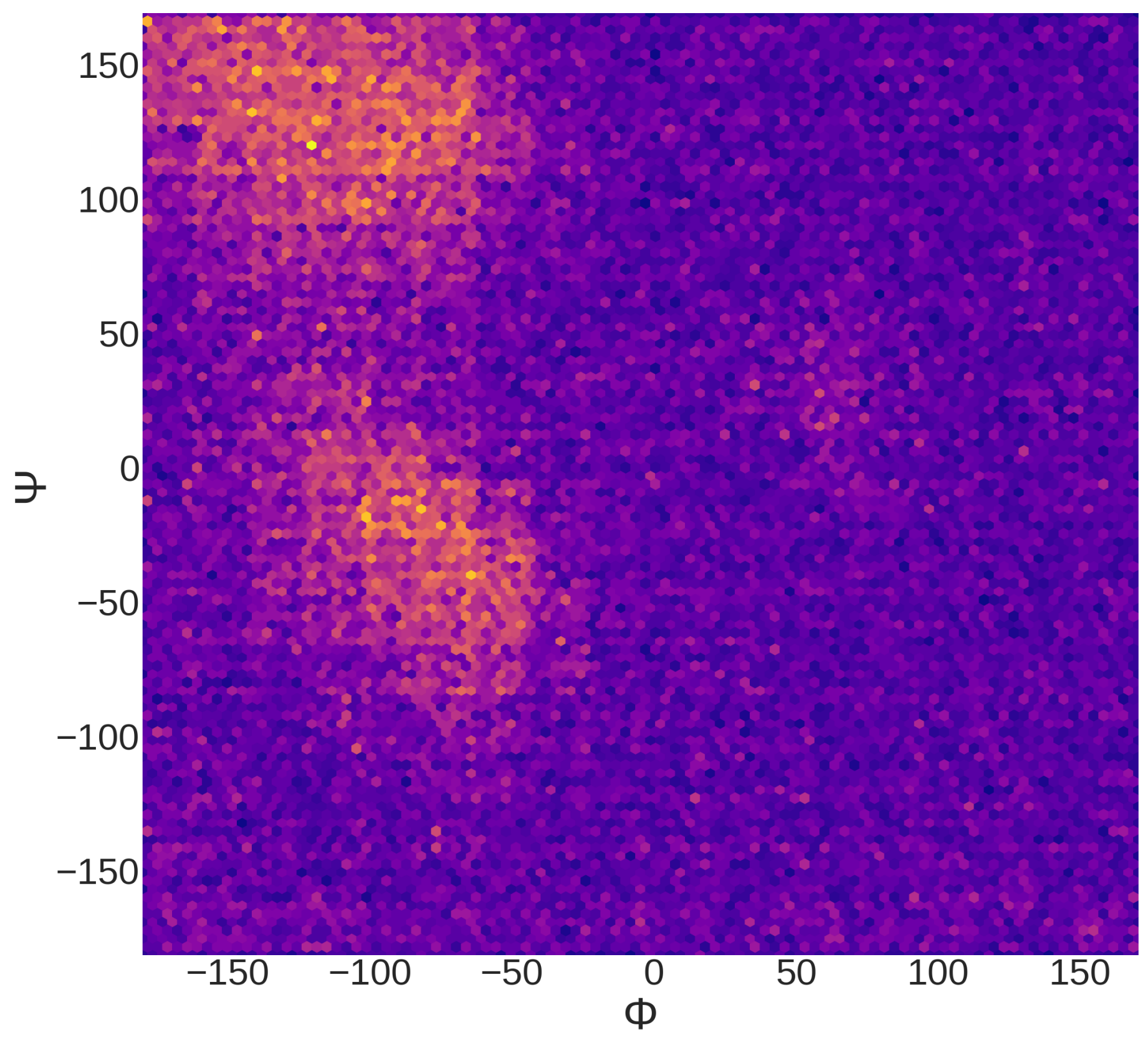

Figure 4: Ramachandran Posterior Predictive Distribution with added jitter.

## Future Work

The density estimation method we present, should be applied to each of the 20 aminoacids separately, in orden to obtain a posterior predictive distribution for every residue. We will incorporate the estimated densities as priors in molecular simulations. Results will be compared to those using a univariate Von Mises Distribution as prior. Future approaches will include the implementation of Laplace's Approximation for Logistic gaussian processes. Given that Dirichlet Processes are also used for density estimation, we could note the similarity of our method and the DP density estimation [8]. Moreover, this method should be adapted to capture the Ramachandran's map circular space, an issue particularly noticeable for the $\Phi$ torsional angle.

## References

[1] C. Ramakrishnan and G.N. Ramachandran, Stereochemical Criteria for Polypeptide and Protein Chain Conformations: II. Allowed Conformations for a Pair of Peptide Units, 1965.
[2] Salvatier J, Wiecki TV, Fonnesbeck C., "Probabilistic programming in Python using PyMC3". PeerJ Computer Science, 2016.
[3] Rasmussen, Carl Edward and Williams, Christopher K. I., "Gaussian Processes for Machine Learning". Adaptive Computation and Machine Learning, 2005.
[4] By Ryan P. Adams, Iain Murray and David J.C. MacKay, "Nonparametric Bayesian Density Modeling With Gaussian Processes", 2009.
[5] Schrödinger, LLC, "PyMOL, The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC.", 2015.
[6] Jaakko Riihimäki, Aki Vehtari, "Laplace approximation for logistic Gaussian process density estimation and regression", 2012. *arXiv:1211.0174*
[7] Tom Leonard, "Density Estimation, Stochastic Processes and Prior Information". Journal of the Royal Statistical Society. Series B (Methodological),1978.
[8] Steven N. Maceachern and Peter Müller "Estimating Mixture of Dirichlet Process Models", 1998.

## Acknowledgements

This work was supported by grant PICT-0218 PICT-Joven Plan Argentina Innovadora 2020: Probabilistic programing for Structural Bioinformatics