

---

# Generative Models for Screening Genetic Mutations in Marfan Syndrome

Antonin Della Noce<sup>\*1,2</sup>, Vesna Lukic<sup>3</sup>, Hakim Benkirane<sup>3,4</sup>, Pauline Arnaud<sup>5,6,7</sup>, Agathe Le Galiot, Nadine Hanna<sup>5,6,7</sup>, Olivier Milleron<sup>7</sup>, Carine Le Goff<sup>5</sup>, Catherine Boileau<sup>5,6</sup>, Paul-Henry Cournède<sup>3</sup>, and Guillaume Jondeau<sup>5,7</sup>

<sup>1</sup>MATHERIALS, Paris – INRIA – France

<sup>2</sup>CERMICS, École des Ponts, Marne-la-Vallée – IP Paris – France

<sup>3</sup>MICS, CentraleSupélec, Gif-sur-Yvette – Université Paris-Saclay – France

<sup>4</sup>Inserm U1018 Oncostat, Villejuif – Inserm – France

<sup>5</sup>LVTS, U1148 – Inserm, Université Paris-Cité – France

<sup>6</sup>Département de Génétique – Hopital Bichat, APHP – France

<sup>7</sup>Centre de référence pour le syndrome de Marfan et apparentés – Hopital Bichat, APHP – France

## Résumé

**Background:** Genetic disorders within the connective tissue spectrum, such as Marfan, Loeys-Dietz, and Ehlers-Danlos syndromes, are associated with a heightened risk of early-onset cardiovascular complications, including thoracic aortic aneurysm and dissection. Early detection of mutations, particularly in genes such as *FBN1*, *TGFBR1*, and *TGFBR2*, is crucial for timely prophylactic interventions to improve patient outcomes. However, reference centers specializing in these rare disorders face budgetary and capacity constraints that prevent them from testing every patient referred due to suspected connective tissue disorders. Current screening strategies, like those based on the Ghent nosology, struggle to reliably differentiate between individuals carrying pathogenic mutations and those who do not. Our goal is to enhance this screening strategy by modeling the distribution of mutation types based on an arbitrary set of clinical characteristics employed by the Ghent nosology.

**Methods:** We analyzed a cohort of 3,982 patients referred to the Reference Center for Marfan Syndrome and Related Disorders at Bichat-Claude Bernard Hospital, Paris, between 1988 and 2018. Genetic sequencing was performed for 36 connective tissue-associated genes, with identified mutations classified into three phenotype categories: mutations on *FBN1*, mutations on *TGFBR1* or *TGFBR2*, and mutations on the other genes of the panel. Clinical covariates such as age, sex, height, aortic dimensions, and skeletal features were collected. 954 wild-type individuals, sequenced as part of a family investigation, served as controls. Given that the proportion of control patients in the cohort does not reflect the general population, a traditional classifier trained on this data would be limited to screening patients within the reference center. To extend screening to a general population, where the proportion of control individuals is higher, we applied Bayes' theorem, converting the classification problem into the estimation of the multivariate phenotype distribution for each mutation type. We developed an arbitrary conditioning generative model using a series of residual neural networks to output the conditional distributions for both continuous and categorical variables.

---

\*Intervenant

**Results:** The generative model achieved a global reconstruction coefficient of determination ( $R^2$ ) of 0.9 on the validation set. This joint distribution allows for the estimation of mutation probabilities based on clinical profiles, enabling a more precise and effective screening strategy.