

# Counterfactual Functional Connectomes for Neurological Classifier Selection

Nicolas Vercheval<sup>\*†</sup>, Marin Benčević<sup>\*‡</sup>, Dario Mužević<sup>§</sup>, Irena Galić<sup>‡</sup> and Aleksandra Pižurica<sup>\*</sup>

<sup>\*</sup>Department of Telecommunications and Information Processing, Ghent University, Belgium

Email: nicolas.vercheval@ugent.be

<sup>†</sup>Department of Electronics and Information Systems, Ghent University, Belgium

<sup>‡</sup>FERIT, J. J. Strossmayer University of Osijek, Croatia

<sup>§</sup>Osijek University Hospital Center, J. J. Strossmayer University of Osijek, Croatia

**Abstract**—Functional connectivity expresses the correlation of brain activity between regions and helps in understanding and diagnosing neurological conditions and disorders. It also provides discriminative features for machine learning classifiers. We propose a model-agnostic method that produces realistic counterfactual functional connectomes by altering the posterior distribution of a hierarchical variational auto-encoder and denoising the result. We evaluate our method on three autism spectrum disorder classifiers for resting state fMRI. The generated counterfactuals include plausible changes in line with medical literature and the brain’s functional anatomy. Our approach strives for explainability and collaboration with medical experts, starting from the model selection.

**Index Terms**—functional connectome, functional connectivity, autism spectrum disorder, counterfactual, XAI

## I. INTRODUCTION

Functional MRI (fMRI) measures brain activity over time and can expose statistical relationships between the brain activity of spatially distant regions. This relationship is called functional connectivity (FC). FC is a valuable tool to analyze and diagnose autism spectrum disorders (ASDs), Alzheimer’s disease, bipolar disorder, Parkinson’s disease, and various other neurological disorders [1], [2]. It is also commonly used to train machine learning (ML) classifiers for these diseases [3].

FC is complex and only partially understood. Machine learning classifiers have the potential to encode these complex relationships [4], but the inner working of complex models is beyond human understanding. Explainable AI (xAI) is an umbrella term for a shared effort by the machine learning community to facilitate the communication between the algorithm and the user. Post-hoc local explanations [5] used to audit the evaluations of specific samples have been particularly appealing when supporting critical decision-making. Between them, counterfactual explanations are rapidly gaining in popularity [6]. A counterfactual explanation consists of the making of a sample that does not exist but is achievable or realistic. The counterfactual sample is very similar to the auditing sample but is classified differently by the same classifier. By looking at the counterfactual sample, experts have a concrete

This research has been partially supported by the Flanders AI Research Programme grant no. 174B09119. This work has been supported in part by the Croatian Science Foundation under the Project UIP-2017-05-4968.

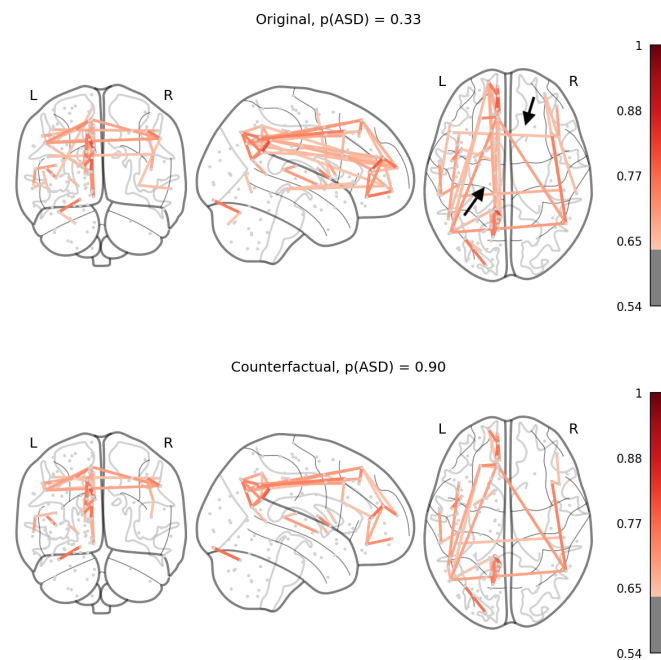


Fig. 1. A real functional connectome compared to its generated counterfactual of the opposite predicted class. Our method allows for functional connectivity matrix generation regardless of the underlying classification method. The intensity represents (positive) correlation strengths. We remove connections with low correlations to improve visual clarity. Arrows show notable connectivity changes.

way of telling if the functioning of a classifier matches their knowledge and if, therefore, the classifier can be trusted.

Despite the advances in counterfactual production, explaining a proposed model may not be enough, as the expert would only be allowed to accept or reject the evaluation without letting them share their knowledge. To unlock the potential of machine learning in medical diagnosis, the development of an algorithm should create a feedback mechanism that keeps medical experts in the loop in all stages. In particular, the expert’s opinion can be most valuable during model selection, which can range between entirely different classifiers.

In this paper, we present counterfactual explanations of different classifiers as functional connectomes, represented as

temporal correlation matrices between functional brain regions. We apply our method to ASD classification from resting state fMRI images. We describe how the FC of a subject would need to change to look typically developed (TD) or affected by ASD according to each classifier. To accommodate all possible types of machine learning, we propose a model-agnostic approach based on hierarchical variational autoencoders [7] (HVAE) that does not access the classifier directly but only its evaluation and is easy to streamline. We extend and adapt previous work on 2D-images [8] to a setting with high-dimensional feature data. We also introduce a simple and effective way of removing the noise from the counterfactual generation caused by posterior collapse, a common drawback of using (H)VAEs. The noise removal, inspired by active noise control, allows a smooth interpolation of the sample that does not lose its sharpness.

We consider three models commonly used for ASD detection: a support vector classifier (SVC), a multi-layer perceptron (MLP), and an MLP fine-tuned from an auto-encoder model. The resulting counterfactuals consistently fool all the classifiers and look similar to the original FC. We illustrate the subtleties that each classifier detects by looking at the changes in the topology of the functional connectomes. This provides an interface to engage the medical expert in the selection of the most reliable classifier for ASD. We share the code on [GitHub](#).

The paper is organized as follows. The rest of this section gives some background on functional connectivity on ASD classification and contextualizes our approach along the literature of counterfactual techniques. Section II explain the method used for counterfactual production in details. Experimental settings, followed by the illustrations of the counterfactual connectomes, is in Section III. Remarks on the findings and future research end the paper in IV.

### A. Functional Connectivity and ASD Classification

Functional connectivity is a temporal correlation between the functional activation of spatially distant brain regions. Mining the scans of several subjects revealed recurring FC patterns [9] and led to the definition of functional atlases, enabling parcellation of the brain into regions of interest (RoIs). A way of quantifying FC is to compute the pairwise correlation matrix between BOLD signal activations in different RoIs of a given atlas. ASD may present both under- and over-connectivity depending on which regions are considered. ASD has also been linked to decreased lateralization, where neural functions are less specialized to one hemisphere than in typically developed subjects [10].

Earlier methods for classifying ASD from FC inputs often relied on traditional statistical learning approaches [11]. Recently, the trend moved to deep learning models with advanced pretraining and feature selection [12], [13]. However, the classifiers are still unreliable in datasets from multiple medical centres and their understanding of FC still needs a dedicated analysis.

### B. Variational Auto-encoders

Variational auto-encoders (VAE) learn a probabilistic mapping  $p(x|z)$  from a densely distributed latent space to the dataset distribution embedded in a higher dimension. The mapping allows the generation of realistic samples from randomly sampled latent variables and dataset extrapolation by exploring the latent space. Introducing a second conditioning variable (cVAE)  $c$ , deterministic or inferred, allows the mapping  $p(x|z; c)$  to disentangle  $z$  from the features that  $c$  represents. Hierarchical VAE (HVAE) presents a more complex probabilistic graph topology. They divide the latent space into several sections, where the distribution of a section depends on the previous one. In particular,  $p(x|z) = p(x|z_L; z_{L-1}; \dots; z_1)$  where the conditional priors  $p_g(z_l) = p(z_l|z_{l-1})$  with  $l = 1 \dots L$  are independent to each other.

NVAE [7] improved image generation using bidirectional inference. They added a deterministic path  $h_d$  to support the prior:  $p_g(z_l) = p(z_l|z_{l-1}; h_d(z_{l-1}))$  with  $h_d(z_{l-1}) = h(z_l; z_{l-1}; \dots; z_1)$ . The inference path  $p_i(z_l) = p(z_l|x; h_d(z_{l-1}))$  updates  $p(z_l)$  with posterior information in the opposite order. The final inference  $p(z_l) = p(z_l|p_g(z_l); p_i(z_l))$  is bidirectional because it combines the prior, which tries to anticipate the inference path, and the inference path, whose information they compress with the Kullback-Leibler divergence.

### C. Counterfactual Explanations

The counterfactual objective and definition also vary according to the overall setting [6]. One approach suitable for complex data employs a generative model to produce realistic samples [14] while a counterfactual loss attacks the classifier. In VAEX [8], the authors show how to produce counterfactuals for images using posterior conditioning. While this model presents many valuable advantages for our use case, the counterfactual shows posterior collapse as the posterior influence attenuates. We propose a different interpolation that does not show this problem.

## II. METHODS

This section explains how we model the counterfactuals, our solutions to problems that arise when using a naive approach and ends with details on the method implementation.

### A. Counterfactual Modelling and Active Noise Cancellation

The proposed approach produces counterfactuals by teaching a cVAE to disentangle its latent space according to the prediction of a given classifier. By doing so, our model learns a correlation between the features the classifier is sensitive to and a variable on which we can intervene, referred to as the condition from here onwards. We perform the intervention as follows: first, we infer the latent space of sample  $x$  and condition  $c$  with  $p(z|x; c)$ ; then, instead of reconstructing the sample  $\hat{x} = p(x|z; c)$ , we switch  $c$  with  $c^d$  corresponding to an opposite evaluation:  $\hat{x}_{c^d} = p(x|z; c^d)$ .

The FC correlation matrices are high-dimensional data, and because of that, they require a complex generative model.

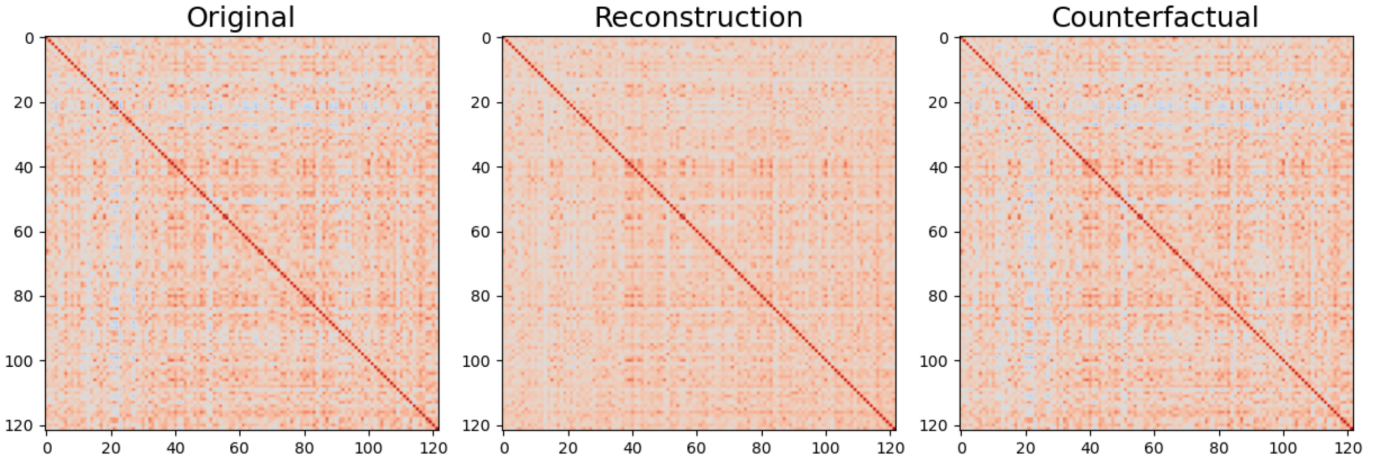


Fig. 2. An FC matrix (red/blue indicates positive/negative correlations) of a typically developed brain correctly classified by the SVM model, its noisy reconstruction and its counterfactual. Note that the noise removal cancellation successfully restores its original sharpness.

Previous work [8] successfully experimented with a more powerful hierarchical cVAE model but suffered from posterior collapse resulting in blurred outputs. As a consequence, the counterfactuals are less realistic and detailed. We propose a noise removal method for counterfactuals inspired by active noise cancellation to prevent that. We assume that noise is unlikely to be correlated to discriminative features by the classifier. Therefore, the noise in the reconstruction should be the same as the noise in the counterfactual. We compute the noise by removing the reconstruction from the original input, then we remove that noise from the counterfactual.

In formulas, for an input  $x$  with condition  $c$  and reconstruction  $\hat{x}$  and a counterfactual  $x_{c'}$ :

$$x_{c'} = \hat{x}_{c'} - (x - \hat{x})$$

We visualize an example in Fig. 2.

### B. Rescaling the Probabilities

Our method allows a consistent comparison of very different models so the distributions of the condition for different classifiers should look as similar as possible. Furthermore, previous work [8] noted that when the evaluated probabilities always take extreme values, as typical in neural networks, the explanatory model extracts less information.

This paper evaluates the method on an SVM model as well as two different neural networks. For the SVM, we obtain estimated probabilities through the standard cross-validation strategy adopted by NumPy [15] and use them as conditions. For the neural network models, the probabilities  $p$  of the neural network models are remapped using the following function, which pushes them away from the extremes:

$$f(p) = \arccos(2p - 1)$$

We rescale the probabilities by applying  $f$  repeatedly until their distribution resembles the one from the SVM model.

### C. Model Architecture and Connectome Visualization

The modelling of our HVAE is the same as in [7], but the implementation is very different as we encode feature data. The inference path consists of linear layers with ReLU activations in between and hidden layers of 512 and 256 neurons. The deterministic path is specular, and the conditional prior is also linear. When conditioning on multiple inputs, they are concatenated together. In particular, we concatenate the condition to the input and all the latent variables. The latent variables  $z_i, i = 1, \dots, 3$  have dimension 32. One dimension of  $z_1$  also has a Gaussian prior whose mean changes with  $c$ . We use the Nilearn library [16] for plotting the connectome.

## III. EXPERIMENTAL RESULTS

We assess our approach by comparing the counterfactual predictions generated by three classifiers: a kernel SVM, a multi-layer perceptron (MLP), and an MLP pre-trained using auto-encoders (AE). We train all models to classify autism spectrum disorders (ASDs) based on functional connectivity matrices extracted from the ABIDE I dataset of 871 resting-state fMRI images preprocessed by the Preprocessed Connectomes Project [17].

To obtain the functional connectivity matrices, we parcellated each preprocessed scan using the BASC122 functional atlas [9], [18]. We computed the full temporal correlation between each ROI pair. The lower triangle of the resulting correlation matrix is flattened into input features for training both the classifiers and the counterfactual model. We kept 20% of the dataset as a held-out test set to evaluate the models.

The kernel SVM is a C-support vector classifier with a radial basis function kernel, using  $C = 100$  and  $\gamma = 1/N$ , where  $N = 7381$  is the number of input features. The MLP model has three fully-connected hidden layers with 512, 256, and 32 neurons, respectively, between ReLU activations. We train it using binary cross-entropy loss, setting the learning rate to  $10^{-4}$  when pre-trained and  $10^{-3}$  otherwise for 150 epochs. We used the same architecture and hyperparameters as in the

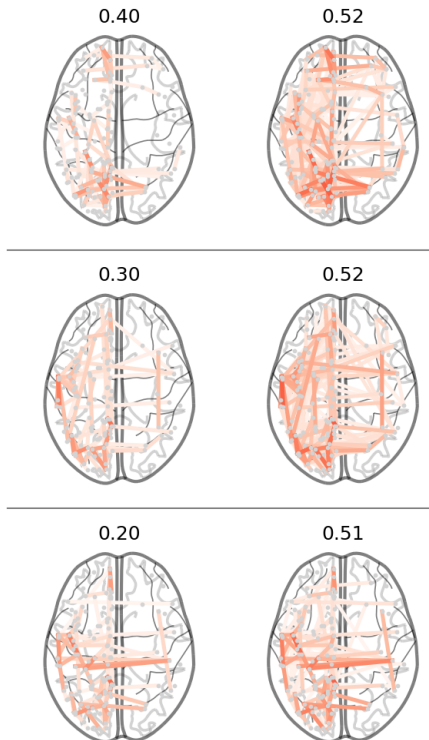


Fig. 3. Counterfactual functional connectome for different subjects. Left: original scans. Right: generated counterfactuals. The ASD probability, estimated by the SVM classifier, is on top of the connectome. We select the target condition until approaching the decision boundary. We only show correlations above the 98th percentile of correlations in the original connectome. The colour intensity indicates correlation strength.

HVAE inference and deterministic path for the AE model. We train the AE model with a learning rate of  $10^{-4}$  for 350 epochs with mean square error loss. Finally, we trained the HVAE with a learning rate of  $10^{-4}$  for 350 epochs with mean square error loss and Kullback-Leibler divergence multiplied by  $5 \cdot 10^5$ . We optimize all models with the AdamW [19] optimizer and weight decay with a  $1 \cdot 10^{-5}$  parameter in batches of 64.

We show the classification results in Table I. We note that all three models have comparable metrics and are consistent with the literature [4], [11].

Fig. 3 shows that the generated counterfactuals exhibit a wide range of changes across subjects, indicating the heterogeneous nature of ASD, which can manifest differently in various patients [20].

TABLE I  
ASD CLASSIFICATION METRICS OF THE EVALUATED CLASSIFIERS  
CALCULATED ON A HELD OUT TEST SET.

Model	Accuracy	Specificity	Sensitivity
SVC	0.66	0.60	0.71
MLP	0.67	0.63	0.71
AE	0.65	0.60	0.69

TABLE II  
QUANTITATIVE RESULTS FOR THE PROPOSED HVAE MODEL.  $D_{KL}$  AND MSE REFER TO THE TEST DATASET’S KULLBACK-LEIBLER DIVERGENCE AND MEAN SQUARE RECONSTRUCTION ERROR.

Model	$D_{KL}$	MSE	CD ( $\times 10^{-2}$ )	DBS ( $\times 10^{-2}$ )
SVC	35.8	1.28	45.9	63.6
MLP	34.9	1.27	54.9	47.1
AE	35.3	1.27	54.9	46.4

#### A. Counterfactual Assessment

When the target condition is the opposite class of the predicted one, the produced counterfactuals successfully confuse all three classifiers for all the test samples. To quantify the influence of a counterfactual on the estimated probabilities, we introduce two metrics: Counterfactual Deviation (CD), the mean square root difference with the original probabilities and Decision Boundary Success (DBS), which is the percentage of success when using a condition equal to 0.5. The first metric shows how much the probabilities change and is best when it is high, but it also depends on the distribution of the classifier (the more extreme, the larger the metric). The second metric depends less directly on the classifier and has 0.5 as the optimal value. As shown in table II, the prediction of each classifier is reliably reverted and is successful more or less half the time when given a neutral condition.

Qualitatively, our method leads to generated connectomes that are visually similar and consistent with original connectomes, as evidenced in Fig. 1. Moreover, the counterfactuals exhibit connection patterns similar to the ones in the dataset. For example, in some ASD-positive counterfactuals, the connection between the middle frontal and right angular gyrus is occasionally removed. In other instances, the TD counterfactual adds connections between the right middle frontal gyrus and the left insular cortex. These connections are consistent with the brain’s functional anatomy and previous research on individuals with ASD [20].

Generating counterfactuals based on different models’ probabilities gives us a better understanding of the models’ internal workings, as demonstrated in Fig. 4. The counterfactuals generated by three different classifiers exhibit remarkable similarity, suggesting that the models are sensitive to similar FC features. Although subtle, there are variations in connection strength between the counterfactuals of different models. Comparing these variations is a valuable tool for model selection, as it allows us to evaluate how well each model captures the underlying relationships in the data. In the reported experiment, the similarity of the counterfactuals is evidence of the robustness of all three models. At the same time, specific variations may be relevant to the medical expert.

#### IV. CONCLUSIONS

This paper presents a method for producing sharp and realistic counterfactuals for arbitrary classifiers that has a hierarchical variational auto-encoder at its core. The counterfactuals are varied, reliably successful and realistic thanks

