Appendix: Gaussian Graphical Model exploration and selection in high dimension low sample size setting

Thomas Lartigue, Simona Bottani, Stéphanie Baron, Olivier Colliot, Stanley Durrleman, Stéphanie Allassonnière for the Alzheimer's Disease Neuroimaging Initiative

Appendix A
Additional Oracle metrics

In this section, we present the results, evaluated with new oracle metrics, of the experiment on synthetic data of section 4.3 "Model selection on a deterministic path with the GGMselect criterion and with the Cross Validated Cross Entropy" of the main paper. The new metrics are the widely used l_2 reconstruction of the True Σ, $\left\|\Sigma - \widehat{\Sigma}_{\mathcal{G}}(S_{train})\right\|_F$, and the oracle nodewise regression l_2 recovery $\left\|\Sigma^{\frac{1}{2}}(I_p - \Theta_{\mathcal{G}}(\underline{X}_{train}))\right\|_F$ (the oracle metric of the GGMselect authors [Giraud et al., 2012]). Fig. 1 shows the results in terms of performance and sparsity over 1000 simulations. In the top row, the performances are measured with the KL, in the middle row with the matrix l_2 and in the bottom row, with the local regression l_2. The observations wit the new metrics are the same as with the KL: the solution selected by the Out of Sample criterion of the Composite method (shades of red) is better in average (an in some case significantly better) than GGMselect (in green) according to the metric in question. Moreover, these solution are closer in terms of number of edges to the best (in blue) and real (with 35 edges) graphs. We observe that this particularly true when the fraction of data reserved for the *validation* set is large (35% or 40% of the *training* data, the lighter nuances of red in the figure, closest to yellow). Indeed, the graphs selected with a smaller *validation* set (20%) have a larger variance in terms of their size, and in terms of their performances when the metric is not the KL. They are also further away from the real graph in average and have the worst average performances of the solutions selected by CVCE. They are still better in average than the solutions selected with the GGMSC though.

Appendix B
Cortex visualisation

We display different perspectives of the graphs proposed by GGMselect (281 edges) and the Composite method ($\sim$ 600 edges) as well as two GLASSO solutions on the Alzheimer's Disease data (ADNI) of Section 5.1 "Experiment on Alzheimer's Disease patients" of the main paper. The first GLASSO solution corresponds to a value of the penalty parameter ρ that gives it a number of edges similar to GGMselect (364 here). This allow to visually compare the sparse GLASSO graphs with GGMselect. The second GLASSO solution is the best encountered on the GLASSO path in terms of Out of Sample KL. It is much larger, with ~ 3500 edges.

As explain in section 5.1 of the main paper, each graph is made of 343 nodes, representing both different areas of the brain and different measured modalities. To visually represent such a complex graph, we choose to display different subsets of its many edges. The following Fig. 2 to 7 correspond to three of these subsets of edges.

The sub-graph containing only the inferred conditional correlations in-between MRI measures are represented on Fig. 2 for the GGMselect and Composite solutions, and Fig. 3 for the two GLASSO solutions. The best GLASSO has so many edges that the graph is hard to interpret. The other graphs possess many connections between symmetric areas of the cortex. With the Composite graph having comparatively more intra-hemispheric edges, and the sparse GLASSO comparatively less edges overall, on this part of the graph.

The sub-graph containing only the inferred conditional correlations in-between PET measures are represented on Fig. 4 for the GGMselect and Composite solutions, and Fig. 5 for the two GLASSO solutions. The observations are mostly the same, but the sparse GLASSO has many more edges than between the MRI measure. It has even more edges than GGMselect and Composite on this part of the graph.

Finally on Fig. 6 and 7, we represent the inter-modality edges between MRI (red) and PET (yellow) nodes. We observe that neither GGMselect nor the sparse GLASSO put any edges in this part of the graph, both methods hence excluding inter-modality conditional correlation from the estimated model. On the other hand, the best GLASSO solution has as many edges as the Composite method on this sub-part of the graph, despite having a considerably larger amount of edges everywhere else.

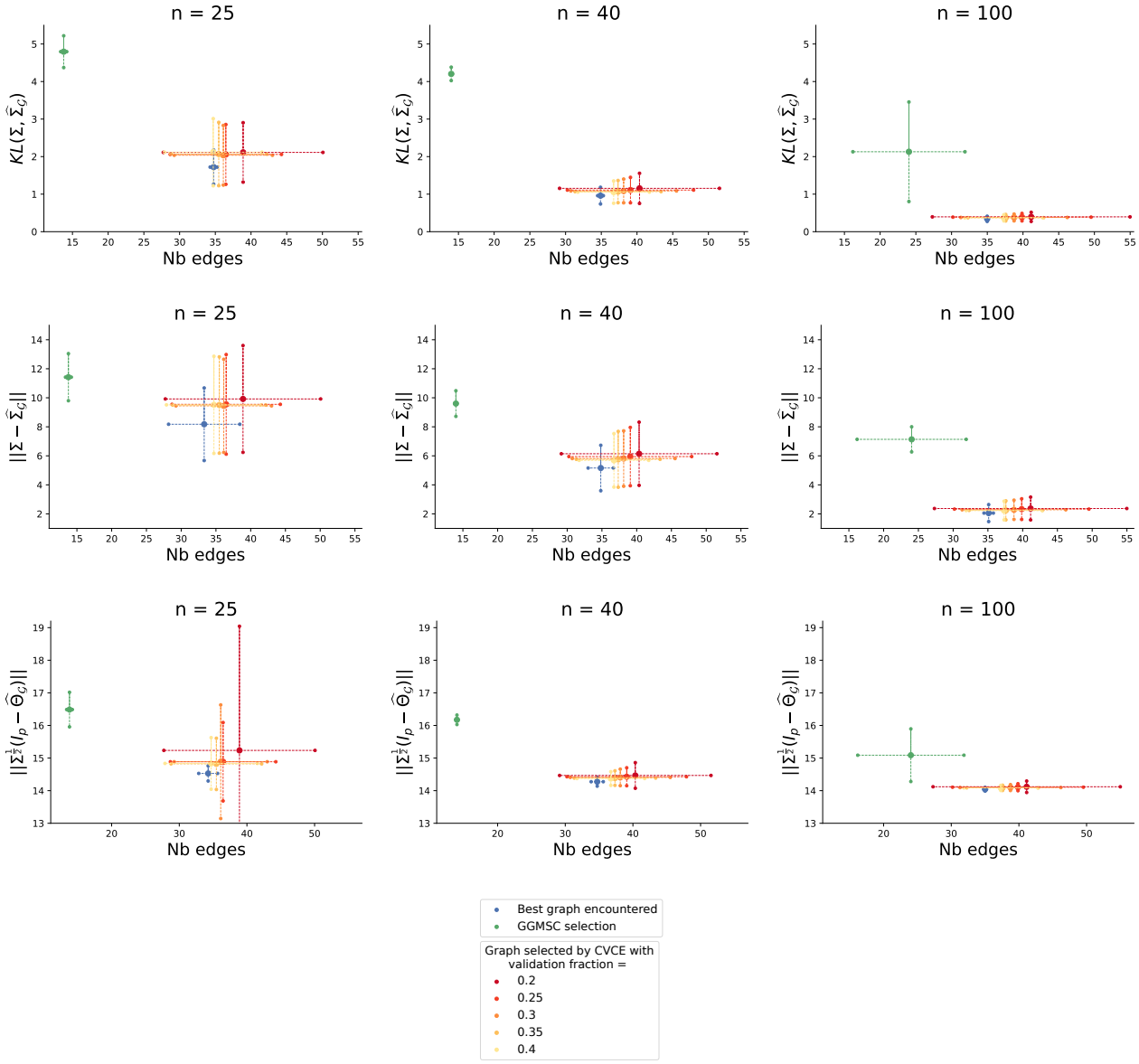

Fig. 1. Results of the experiment of section 4.3 of the main paper evaluated with the KL (top), as well as two alternative metrics: the l_2 recovery of Σ (middle) and the Oracle metric of GGMselect, the nodewise l_2 recovery (bottom). The behaviour and conclusion are the same as with the KL. We also observe that when the validation set is too small (only 20% of the training set), there is a lot of variance on the selection by CVCE, and the performances suffer.

APPENDIX C
GLASSO SOLUTIONS ON THE NEPHROLOGY PATIENTS

In this section, we display, see Fig. 8, the path of GLASSO solutions applied to the neprhology patients of Section 5.2 "Experiments on neprhology patients" of the main paper. The bottom left matrix, $\rho = 0.8$, is the one compared to the GGMselect and Composite solution in the main paper. It was chosen as a representative because the other GLASSO solution have many more edges than GGMselect and Composite. This decision allowed us to compare the first edges selected by GLASSO on its path of solutions as ρ decreases with the edges highlighted by GGMselect. The medical analysis in the main paper concluded that the GGMselect edges were much more consistent with the domain knowledge. The other GLASSO solutions displayed here feature some of the important edges missed by the first, sparse, solution, but these edges appear later in the path, alongside many other a priory irrelevant edges.

REFERENCES

[Giraud et al., 2012] Giraud, C., Huet, S., and Verzelen, N. (2012). Graph selection with ggmselect. *Statistical applications in genetics and molecular biology*, 11(3).

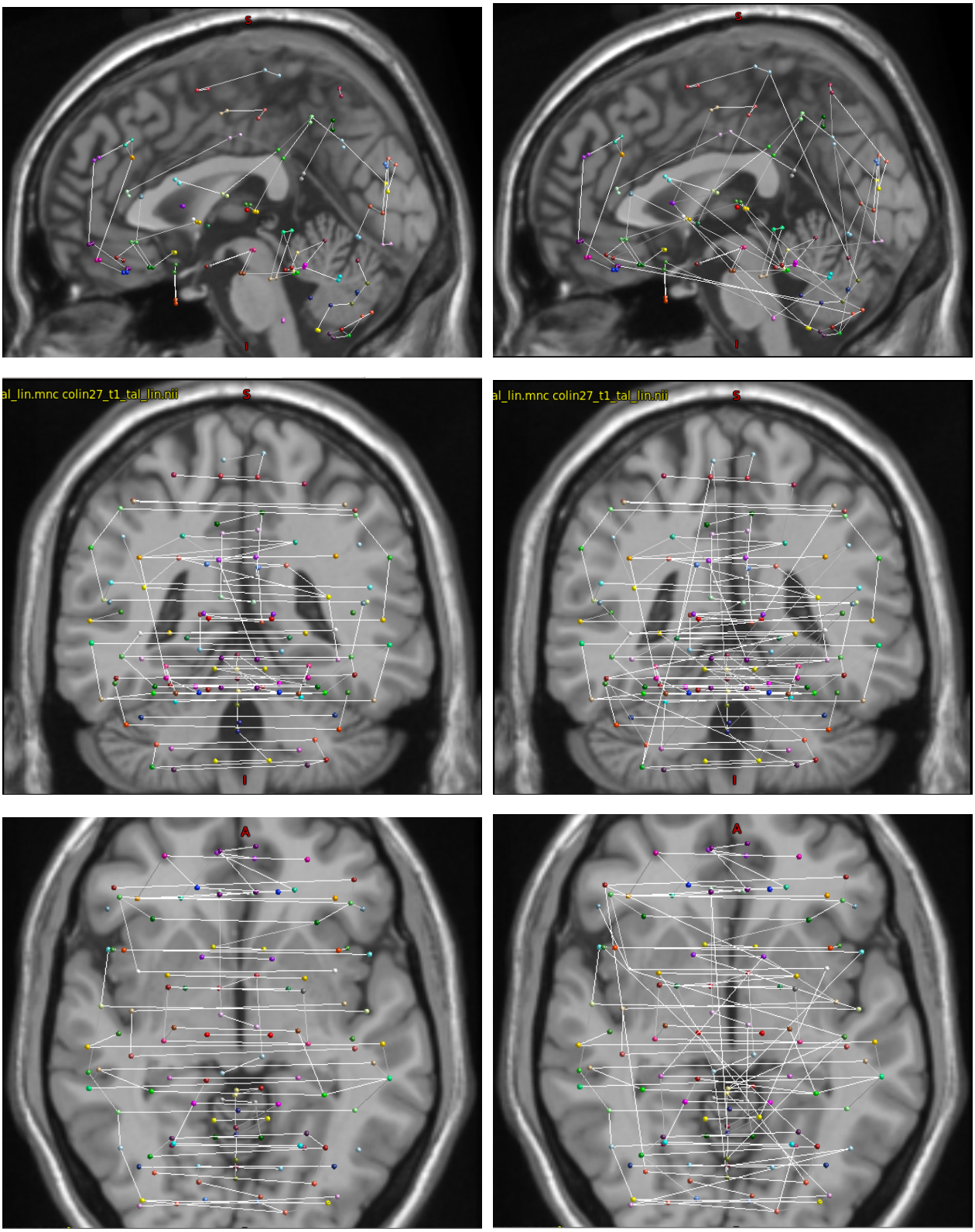

Fig. 2. **I.1 Sub-graph in-between MRI measures from the GGMselect (left) and Composite (right) solutions.** The GGMselect full graph has 281 edges in total, and the Composite around 600. The sagittal, frontal and transverse views of the Cortex are displayed. With both methods, many of the connections are inter-hemispheric, between symmetrical areas. Although the Composite solution proposes a number of new, intra-hemispheric, edges.

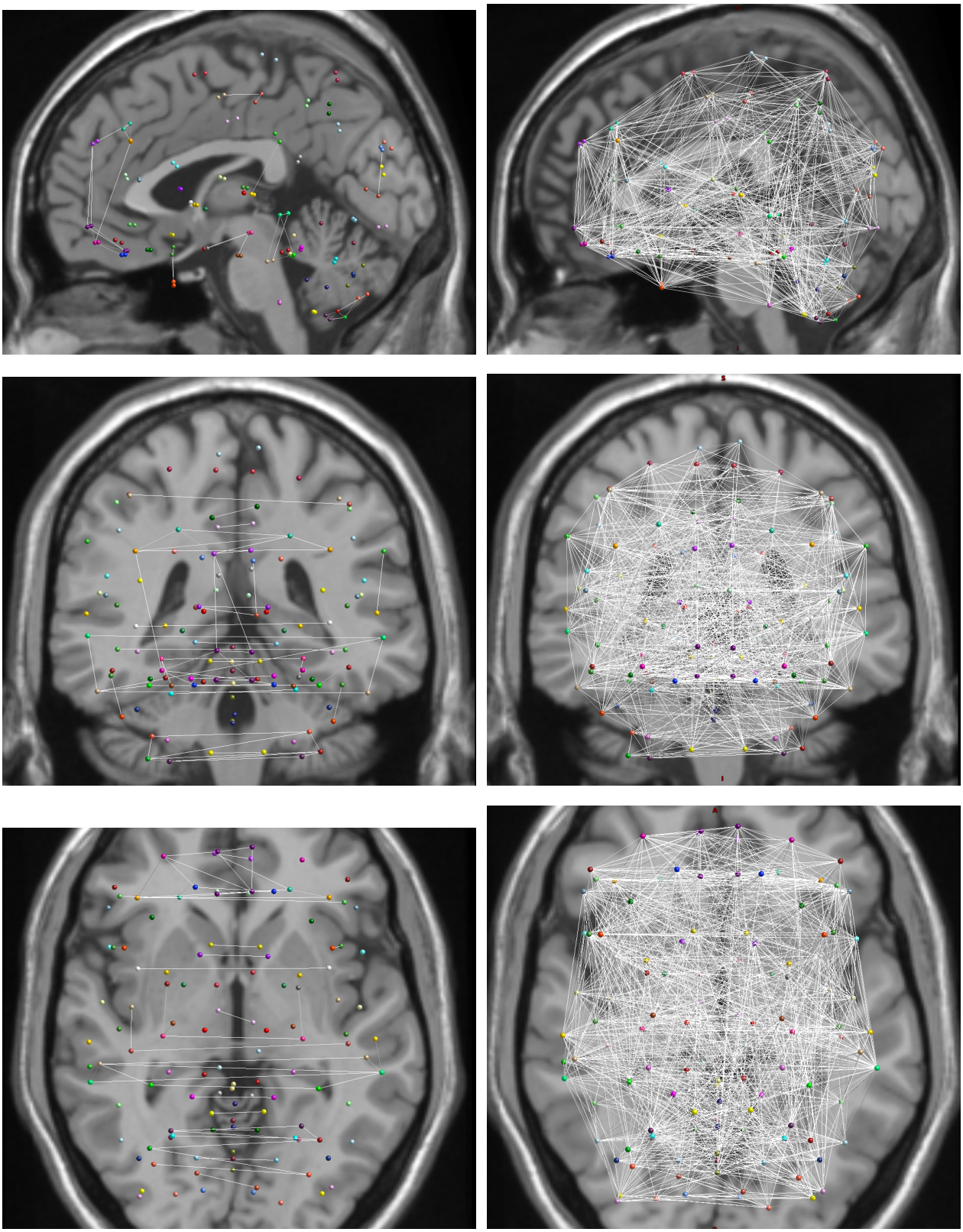

Fig. 3. **I.2 Sub-graph in-between MRI measures from a sparse GLASSO solution (left) and the best Out of Sample GLASSO solution in KL (right).** The full graph of the sparse GLASSO solution has 364 edges in total. A number chosen to be close to the GGMselect solution. The best OSL GLASSO solution features 3546 edges in total. The sparse solution features mostly inter-hemispheric connections between symmetrical areas. The larger solution, with better performances, is mostly unreadable.

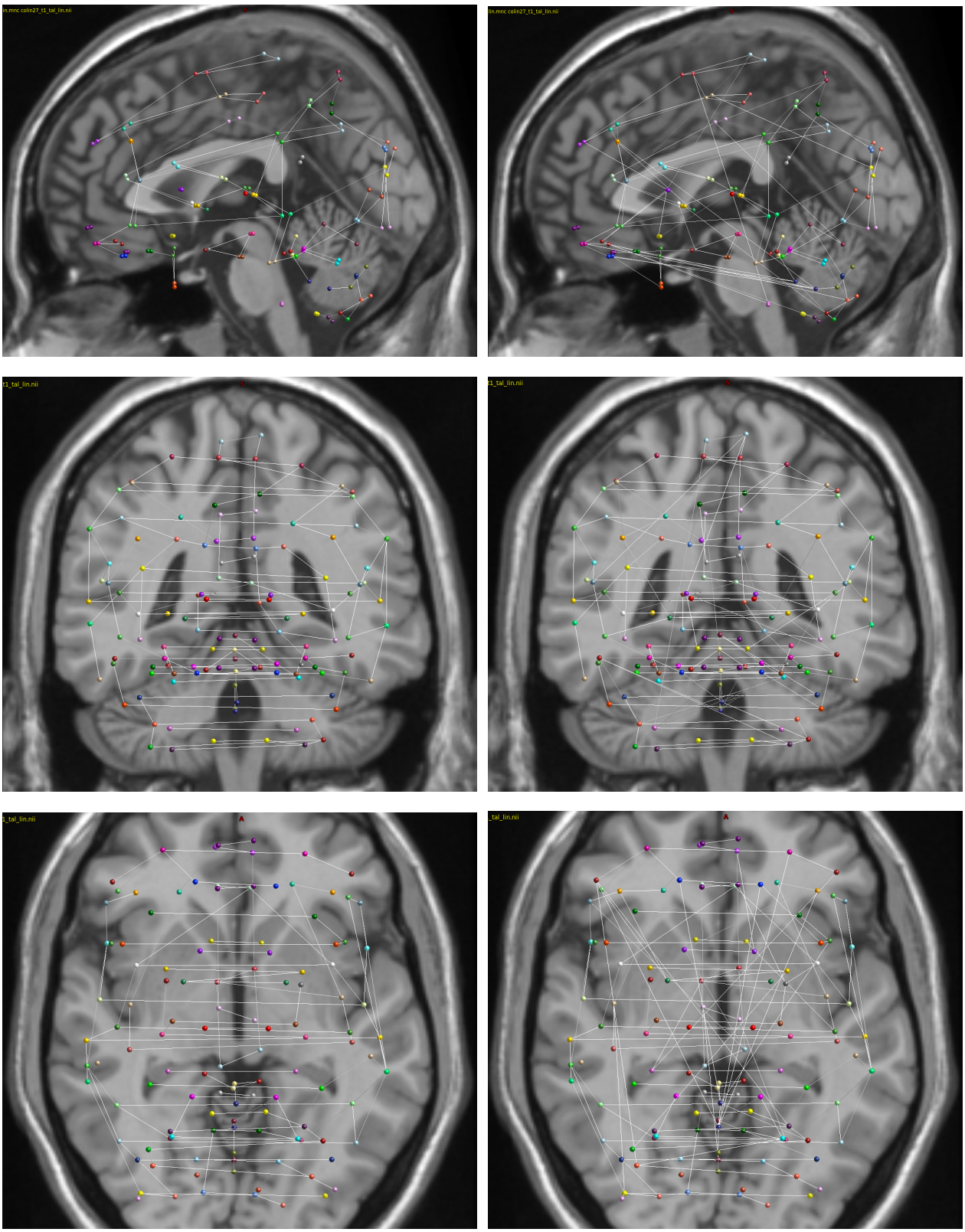

Fig. 4. **II.1 Sub-graph in-between PET measures from the GGMselect (left) and Composite (right) solutions.** The observations are similar to the MRI sub-graph: many inter-hemispheric connections between symmetric regions, with new intra-hemispheric connections in the Composite solution.

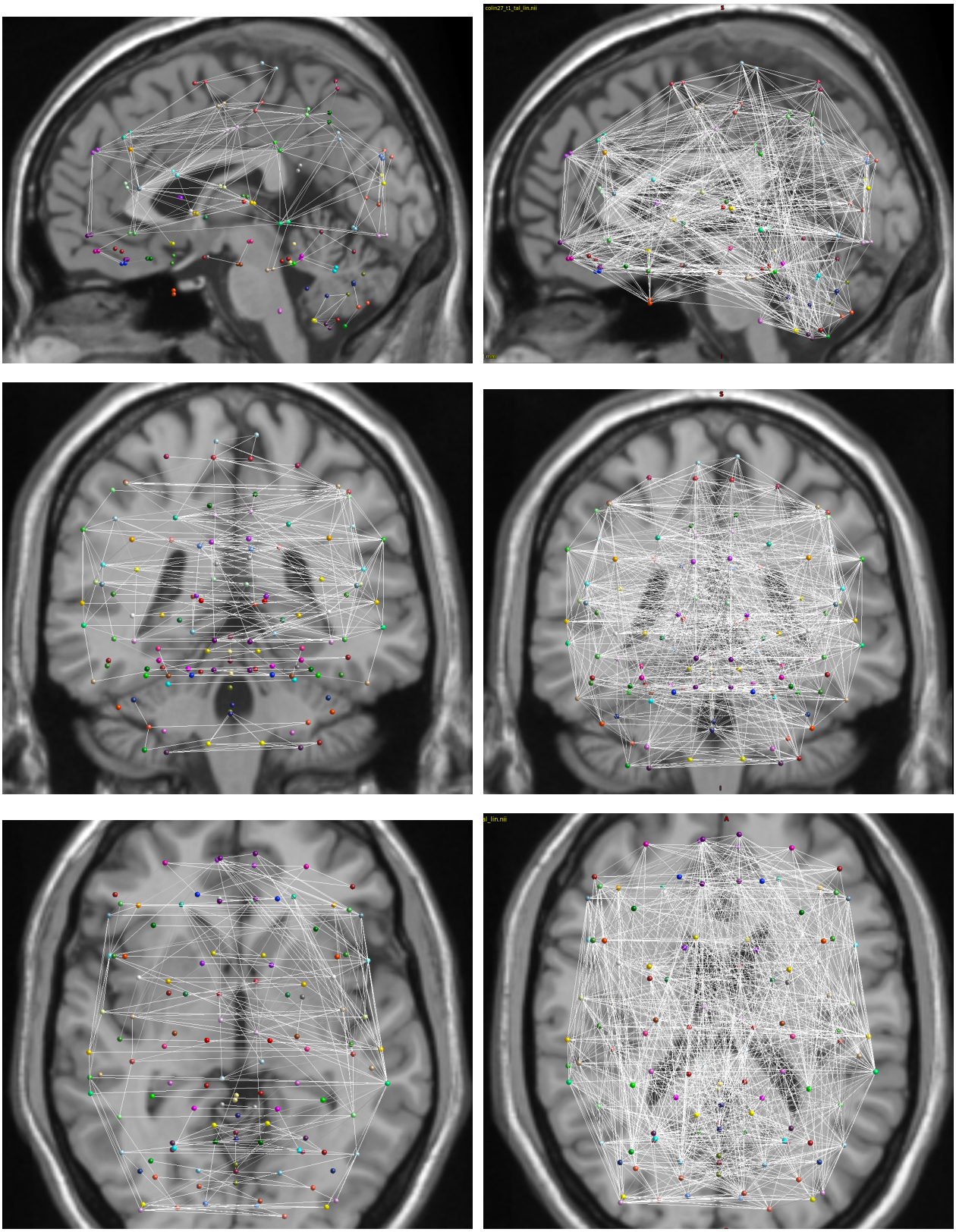

Fig. 5. **II.2 Sub-graph in-between PET measures from a sparse GLASSO solution (left) and the best Out of Sample GLASSO solution in KL (right).** This figure reveals that the sparse GLASSO solution possesses more edges in-between PET measures than in-between MRI measures. The larger GLASSO is still mostly unreadable.

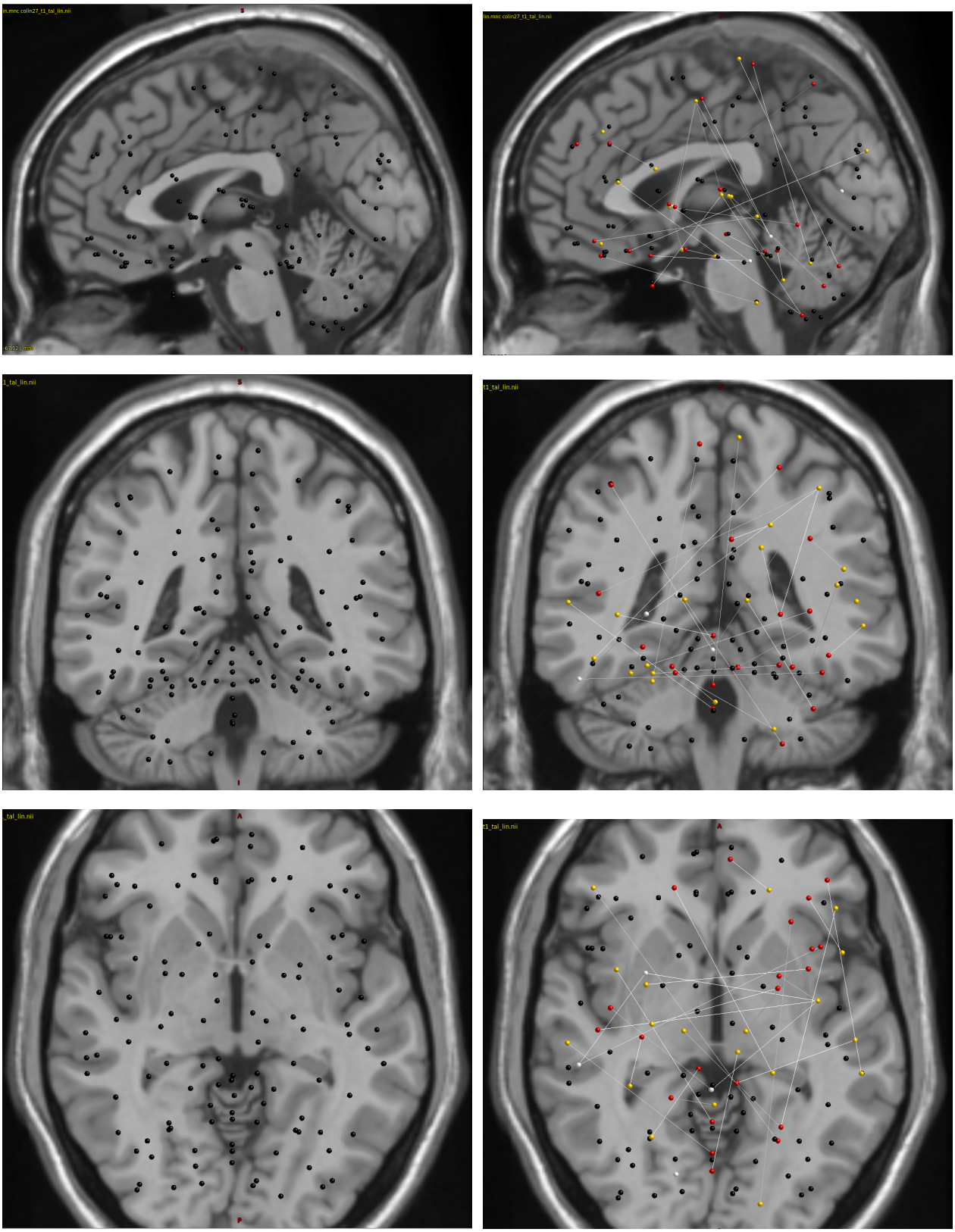

Fig. 6. **III.1 Sub-graph of the edges between MRI (red) and PET (yellow) measures from the GGMselect (left) and Composite (right) solutions.** Unlike the Composite method, the GGMselect graph proposes no edge between any MRI and PET measures.

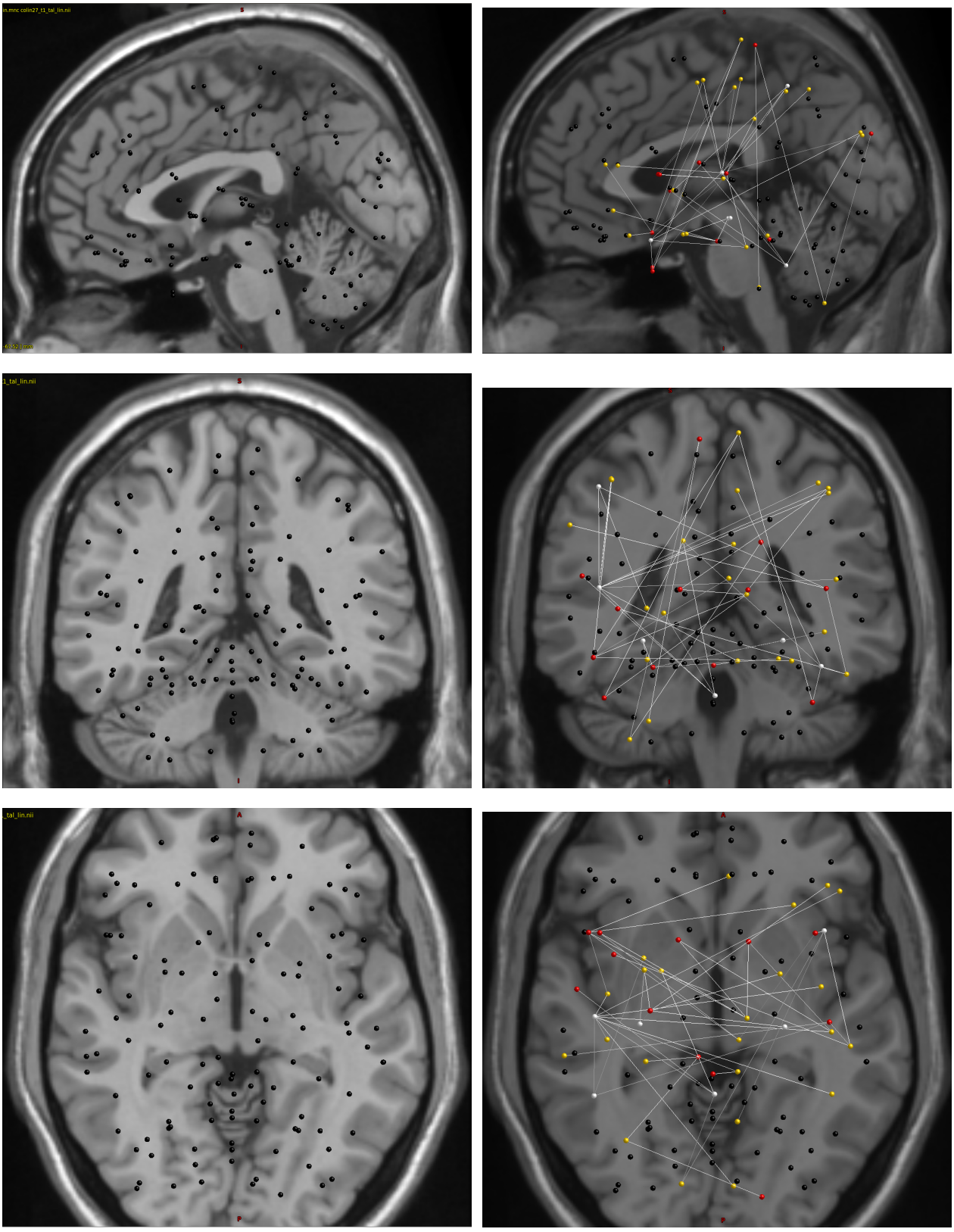

Fig. 7. **III.2 Sub-graph of the edges between MRI (red) and PET (yellow) measures from a sparse GLASSO solution (left) and the best Out of Sample GLASSO solution in KL (right).** Like with GGMselect, there is no edge in this part of the sparse GLASSO graph. The larger GLASSO solution has edges in this sub-part of the graph. Unlike in the other regions however, the large GLASSO features a number of edges similar to the corresponding Composite method sub-graph.

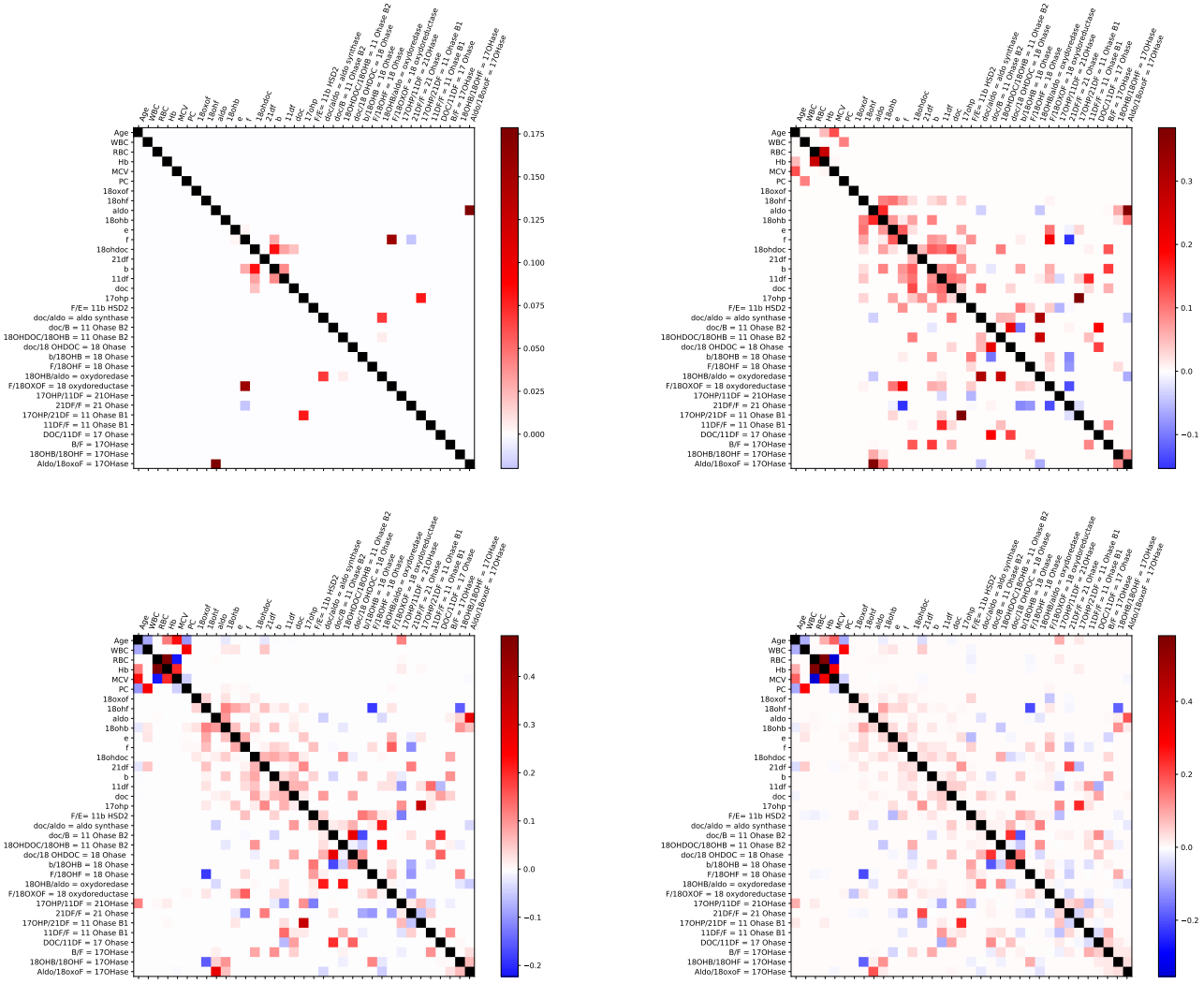

Fig. 8. Matrices of the pairwise conditional correlations corresponding to the GLASSO solutions applied to the nephrolgy patients. The value considered for the parameter ρ are 0.8 (top left), 0.4 (top right), 0.2 (bottom left) and 0.1 (bottm right)