APPENDIX

A ANALYSIS OF THE uGLAD ARCHITECTURE

A.1 uGLAD PARAMETER DETAILS

For the sake of completeness of understanding the GLAD architecture, we graciously borrow the algorithm (see Alg.2) and neural network design details from Shrivastava et al. (2020b) here. GLAD parameter settings details are: ρ_{nn} was a 4 layer neural network and Λ_{nn} was a 2 layer neural network. Both used 3 hidden units in each layer. The non-linearity used for hidden layers was $\tanh$, while the final layer had sigmoid (σ) as the non-linearity for both, ρ_{nn} and Λ_{nn} (refer to Figure 2). The learnable offset parameter of initial Θ_0 was set to $t = 1$. It was unrolled for $L = 30$ iterations.The optimizer used was adam with the learning rates were chosen between $[0.001, 0.005]$.

B CASE STUDY: ANAEROBIC DIGESTION

Our algorithm development was inspired by a practical problem of domain exploration in anaerobic digestion. Anaerobic digestion is a growing field addressing waste management with both environmental benefits (reduced odor and pathogens, improved soil health, reduction in methane emissions) and economic value from use of captured methane gas. Despite numerous studies, the dynamics of organisms' growth in digesters, their dependence on conditions (temperature, pH, feedstock mix, nitrogen to carbon ratio, etc.) and their impact on methane yield are not well understood.

Algorithm 2: GLAD

Function GLADcell$(\widehat{\Sigma}, \Theta, Z, \lambda)$:
- $\lambda \leftarrow \Lambda_{nn}(\|Z - \Theta\|_F^2, \lambda)$
- $Y \leftarrow \lambda^{-1}\widehat{\Sigma} - Z$
- $\Theta \leftarrow \frac{1}{2}\big(-Y + \sqrt{Y^\top Y + \frac{4}{\lambda}I}\big)$
- **For** *all* i, j **do**
  - $\rho_{ij} = \rho_{nn}(\Theta_{ij}, \widehat{\Sigma}_{ij}, Z_{ij})$
  - $Z_{ij} \leftarrow \eta_{\rho_{ij}}(\Theta_{ij})$
- **return** Θ, Z, λ

Function GLAD$(\widehat{\Sigma})$:
- $\Theta_0 \leftarrow (\widehat{\Sigma} + tI)^{-1}$, $\lambda_0 \leftarrow 1$
- **For** $k = 0$ *to* $K - 1$ **do**
  - $\Theta_{k+1}, Z_{k+1}, \lambda_{k+1}$
  - $\leftarrow$GLADcell$(\widehat{\Sigma}, \Theta_k, Z_k, \lambda_k)$
- **return** Θ_K, Z_K

Data comes from a 6-year study of 46 digesters located at 22 Danish treatment plants. We have three types of digesters, operating at different temperatures (mesophilic, mesophilic with thermal hydrolysis pretreatment, thermophilic). Digesters operate continuously with sludge retention rate of 15.8 to 35.6 days. Samples were taken at 3-month and 6-month intervals, so they can be treated as i.i.d. We have a total of 1,010 sludge samples, 418 used to sequence archaea and 592 bacteria, performed using 16S rRNA gene amplicon sequencing. Analysis resulted in identification of 33,047 bacterial and 878 archaeal unique amplicon sequence variants (ASVs). 70% of genera and 93% of the species were determined to be novel or previously unclassified. This presents problems for all approaches attempting to utilize existing databases to determine organisms' function for the purpose of grouping and feature selection. In fact, one of the best ways to determine an organism's function is based on checking properties of organisms whose abundance numbers in the digester best correlate with the given organism's numbers.

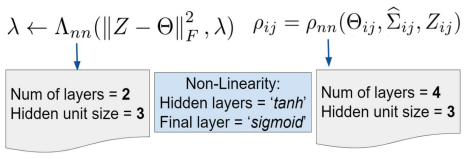

Figure 2: **Minimalist** neural network architectures designed for GLAD

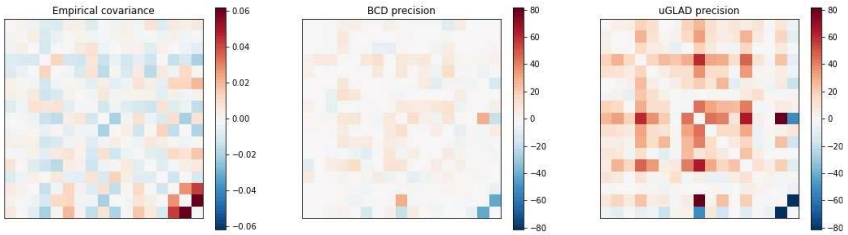

Figure 3: uGLAD recovered precision matrix compared to empirical covariance and precision matrix recovered by the BCD algorithm for archaea at **family** level

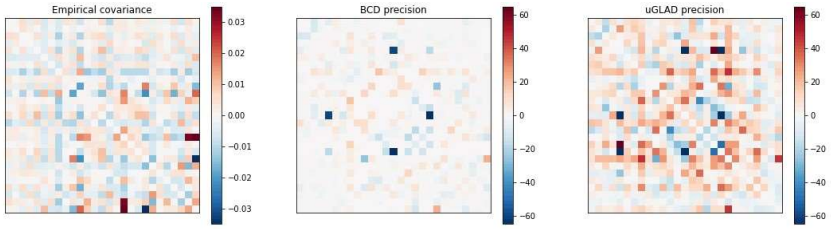

Figure 4: uGLAD recovered precision matrix compared to empirical covariance and precision matrix recovered by the BCD algorithm for archaea at **species** level

Our algorithm works with any input, including: ASVs filtered by frequency, ASVs rolled up to higher taxonomy levels (species, genus, family), ASVs abundance normalized in various ways Badri et al. (2020). We calculate the partial correlation matrix from the precision matrix. Each entry of the partial correlation matrix P_{ij} shows the correlation of the feature x_i, x_j given the values of the other features are observed. This helps us obtain the direct dependence of the features.

$$\mathrm{P}_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} \tag{9}$$

We use networkx package to visualize the graphs, presenting positive correlations in green and negative in red, with edge weights corresponding to the strength of the correlation.

Figures 4 and 3 present precision matrices recovered by our algorithm and BCD with empirical covariance shown for comparison. Figures 5 and 6 show corresponding graphs for archaea at family and species level.

Figures 8 and 7 show a result of multitask learning based on digester type: mesophilic (operating at temperature 38°C) and thermophilic (operated at temperature 53.6°C). The two graphs' edges are filtered to show only edges common to both graphs, which is a small fraction of all edges. Note that in some cases, the sign of the correlation (and the color of the edge) changes depending on digester's type.

Our model is being used by domain experts to gain insight into the domain of anaerobic digestion. One use case is to understand properties of newly discovered bacteria and archaea by analyzing which known organisms their abundance in digester samples correlates with (positively or negatively). That can lead to focusing attention on a smaller organism set. Another use case centers around understanding the role of digester conditions and feedstock mix on organisms' growth and methane yield. The results presented in recovered graphs lead to new hypotheses and new experiments being designed to test them.

C Case study: Infant Mortality

We used uGLAD to recover the graph for the domain of infant mortality. The dataset is based on CDC Birth Cohort Linked Birth – Infant Death Data Files of Health et al.. It describes pregnancy and birth variables for all live births in the U.S. together with an indication of an infant's death before the first birthday. We used the data for 2015 (latest available), which includes information about 3,988,733 live births in the US during 2015 calendar year.

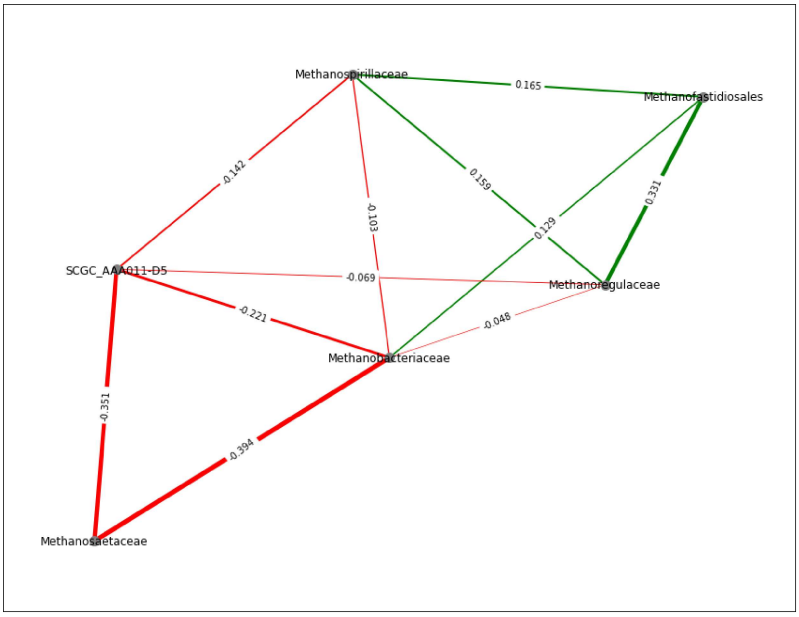

Figure 5: uGLAD graph for archaea at family level. Edge color indicates the sign of the correlation: green - positive, red - negative, edge weight corresponds to correlation's strength.

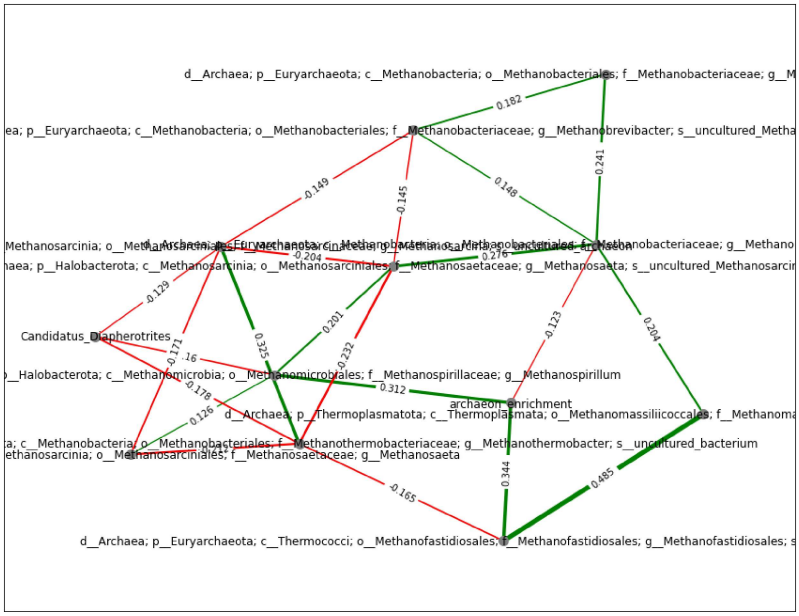

Figure 6: uGLAD graph for archaea at species level. Edge color indicates the sign of the correlation: green - positive, red - negative, edge weight corresponds to correlation's strength.

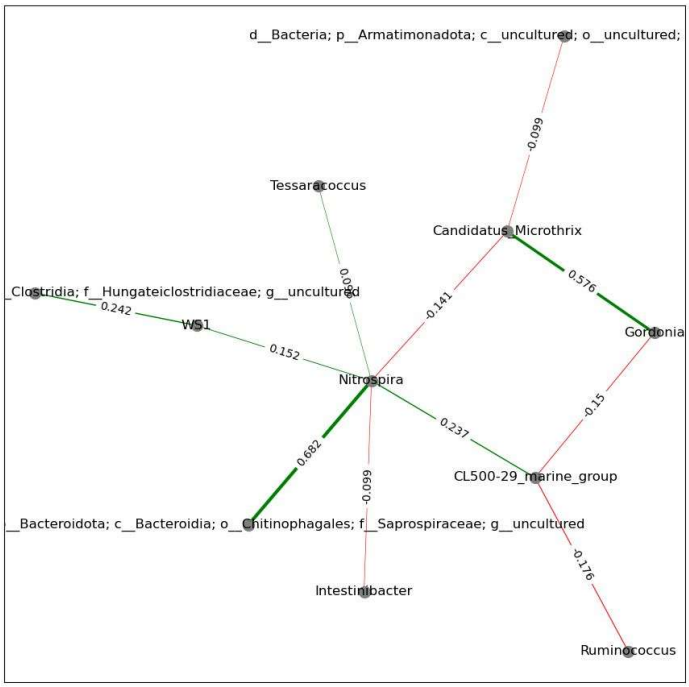

Figure 7: Example of **multitask** learning applied to different digester types. uGLAD graph for bacteria at genus level for thermophilic digesters showing only edges common to all digester types. Note that edge colors (correlation signs) are different from the corresponding graph for mesophilic digesters. Edge color indicates the sign of the correlation: green - positive, red - negative, edge weight corresponds to correlation's strength.

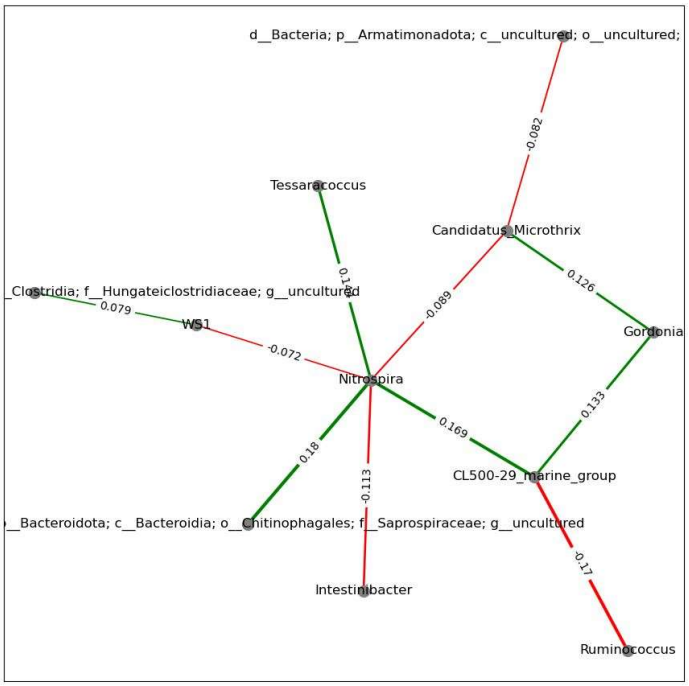

Figure 8: Example of **multitask** learning applied to different digester types. uGLAD graph for bacteria at genus level for mesophilic digesters showing only edges common to all digester types. Edge color indicates the sign of the correlation: green - positive, red - negative, edge weight corresponds to correlation's strength.

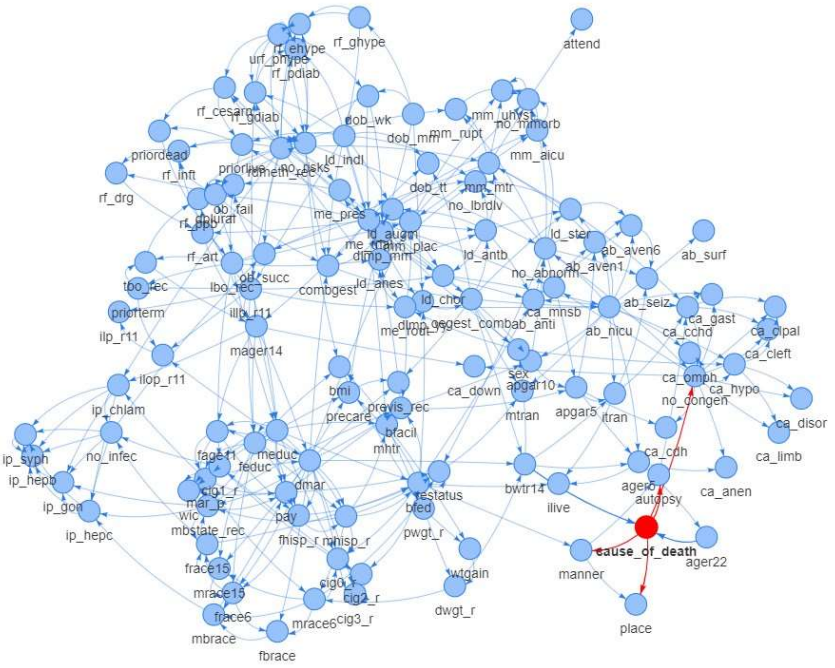

Figure 9: The Bayesian network graph learned using score-based method for the Infant Mortality 2015 data.

We recovered the graph strucure of the dataset using `uGLAD` and using the Bayesian network package `bnlearn` Scutari (2010) with Tabu search and AIC score. The graphs are shown in Fig. 10 and 9 respectively. Since `bnlearn` does not support networks containing both continuous and discrete variables, all variables were converted to categorical for `bnlearn` structure learning and inference. In contrast, `uGLAD` is equipped to work with mixed types of variables and was trained on the dataset prior to conversion.

Both graphs show similar sets of clusters with high connectivity within each cluster:

- describing both parents' race and ethnicity (`mrace` and `frace` variables),
- related to mother's bmi, height (`mhtr`) and weight, both pre-pregnancy (`pwgt_r`) and at delivery (`dwgt_r`),
- consisting of maternal morbidity variables marked with `mm` prefix (e.g., unplanned hysterectomy),
- showing pregnancy related complications such as hypertension and diabetes (variables prefixed with `rf` and `urf`),
- consisting of variables related to parents' STD infections (`ip` prefix),
- related to delivery complications and interventions (variables prefixed with `ld`),
- showing interventions after delivery (`ab` prefix) such as ventilation or neonatal ICU,
- describing congenital anomalies diagnosed in the infant at the time of birth (variables prefixed with `ca`),
- related to infant's death: age at death, place, autopsy, manner, etc.

Apart from these clusters, there are a few highly connected variables in both graphs: gestational age (`combgest` and `oegest`), delivery route (`rdmeth_rec`), Apgar score, type of insurance (`pay`), parents' ages (`fage` and `mage` variables), birth order (`tbo` and `lbo`), and prenatal care.

With all these similarities, however, the total number of edges varies greatly between the two graphs and the number of edges unique to each graph outnumbers the number of edges the two graphs have in common (see Figure 11). Obviously, the two graph recovery algorithms are very different in both algorithmic approach and objective function. One additional reason for the differences lies in the continuous-to-categorical conversion performed prior to Bayesian network structure discovery and training.

New edges discovered by `uGLAD` and missed by `bnlearn` merit futher exploration.

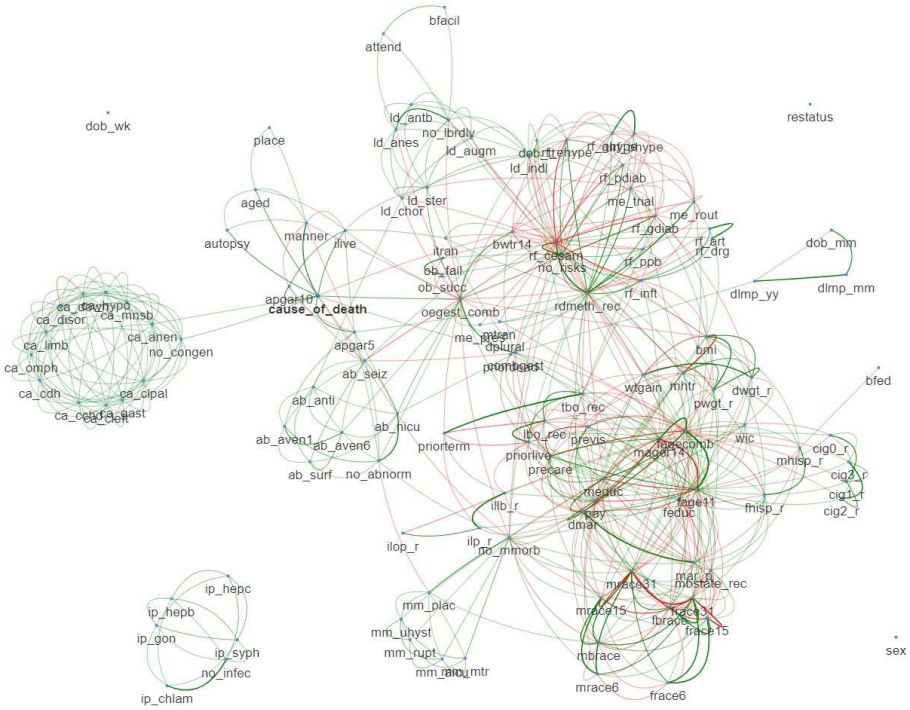

Figure 10: The CI graph recovered by uGLAD for the Infant Mortality 2015 data.

D Potential other applications of the uGLAD model

Listing some more applications for the uGLAD model for which we feel that it can help improve the current state-of-the-art performance.

- Protein Structure recovery: PSICOV Jones et al. (2012) uses graphical lasso based approach to predict the contact matrix, which then eventually gives the 3D protein structure. uGLAD can be substituted for predicting the contact matrix from the input correlation matrix between the amino acid sequences.
- Finance & Healthcare: Finding correlations between stocks to see how companies compare Hallac et al. (2017). Similarly systems for finding connection between important body vitals of ICU patients Shrivastava et al. (2021); Bhattacharya et al. (2019).
- Class imbalance handling: We can potentially use uGLAD to find correlation between the features. This correlation graph can be helpful in sampling down useful feature clusters. This will help identify key features and in-turn improve performance in cases where there is less data or imbalanced data (more data points for one class over another). Some of the methods for class imbalance handling on which uGLAD model can act as a preprocessing steps are Rahman & Davis (2013); Shrivastava et al. (2015); Bhattacharya et al. (2017).
- Gaussian processes & time series problems: uGLAD can be extended to this interesting work by Chatrabgoun et al. (2021) on combining graphical lasso with Gaussian processes for learning gene regulatory networks. Similarly, in a recent work on including negative data points for the Gaussian processes Shrivastava et al. (2020a), uGLAD can be used for narrowing down the relevant features for doing the GP regression. Our model can be used for time-series modeling, refer Jung et al. (2015) for an example.
- Video sequence predictions: uGLAD can be integrated into the pipeline for latest models used for generating unseen future video frames Denton & Fergus (2018); Shrivastava & Shrivastava (2021). Specifically, the parameters of the uGLAD can be learned to narrow down the potential future viable frames from the generated ones.

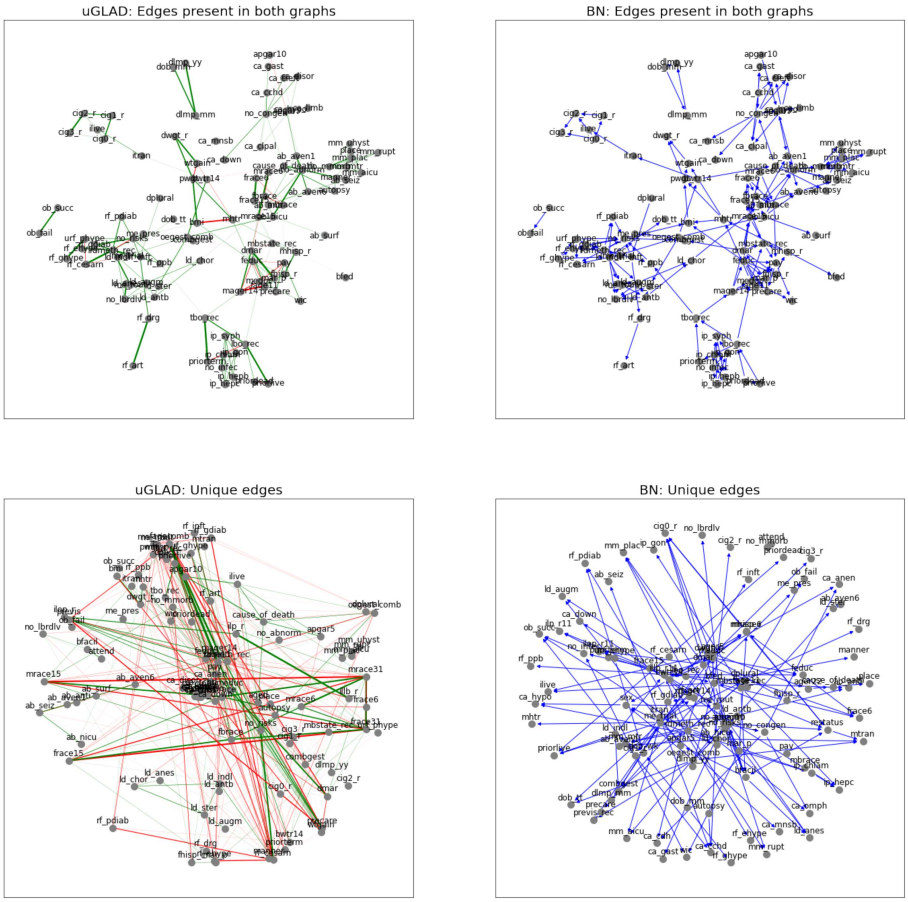

Figure 11: Comparing the graphs recovered by uGLAD and Bayesian Network recovery package (Scutari, 2010).