



Heterogeneous Graph Neural Network Modelling of Mood Episode Recurrence in Bipolar Disorder: A Proof-of-Concept Simulation Study of Temporal, Pharmacological, and Social-Rhythm Relations

Rocco de Filippis^{1*}, Abdullah Al Foysal²

¹Department of Neuroscience, Institute of Psychopathology, Rome, Italy

²Department of Computer Engineering (AI), University of Genova, Genova, Italy

Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Heterogeneous Graph Neural Network Modelling of Mood Episode Recurrence in Bipolar Disorder: A Proof-of-Concept Simulation Study of Temporal, Pharmacological, and Social-Rhythm Relations. *Open Access Library Journal*, 13: e15679.

<https://doi.org/10.4236/oalib.1115679>

Received: June 22, 2026

Accepted: August 24, 2026

Published: August 27, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Mood episode recurrence in bipolar disorder is shaped by the structure of a patient's episode history: the temporal succession of episodes, pharmacological transitions between episodes, and social-rhythm disruptions that may destabilise the circadian and social-zeitgeber system. Existing recurrence models generally flatten this history or merge all connections into one homogeneous graph, thereby discarding relation-type information. We therefore conducted a proof-of-concept simulation study in which each patient history was represented as a heterogeneous graph with episode nodes and three typed edge sets: temporal succession, pharmacological transition, and social-rhythm disruption. The prediction target was a simulated new syndromal depressive, manic, hypomanic, or mixed episode occurring within 12 months after the final observed episode; the binary graph-level label was assigned at the end of the observed episode sequence, using only information available up to that index episode. In $N = 700$ simulated patients (recurrence prevalence 31.4%), we compared a relation-specific R-GCN with homogeneous GCN, GraphSAGE, GAT, graph-transformer, and no-graph MLP baselines. The heterogeneous R-GCN achieved $AUC = 0.822$ (95% CI: 0.723 - 0.902), $F1 = 0.644$, sensitivity = 0.664, and specificity = 0.820 on a held-out simulated test set. Relation-type ablation produced the largest performance decrease after removal of pharmacological-transition edges, followed by social-rhythm-disruption and temporal-succession edges. Because recurrence labels were generated partly from the same pharmacological and social-rhythm mechanisms encoded in the graph, these

results demonstrate recovery of simulation-imposed relational structure rather than clinical validity. They support the methodological hypothesis that preserving relation types can improve recurrence prediction when such relations genuinely carry signal, but require confirmation in real longitudinal bipolar-disorder cohorts. In this simulated proof-of-concept setting, relation-specific graph modelling outperformed matched homogeneous and non-graph comparators. The findings should be interpreted as evidence about model behaviour under an explicit generative mechanism, not as a clinically validated recurrence-prediction tool or as proof that any relation type is clinically dominant.

Subject Areas

Psychiatry & Psychology

Keywords

Graph Neural Network, Heterogeneous Graph, Bipolar Disorder, Mood Episode Recurrence, Relational Graph Convolutional Network, Social Rhythm, Pharmacological Transition, Relation-Type Ablation, R-GCN, Relapse Prediction

1. Introduction

The recurrence of mood episodes is the defining clinical challenge of bipolar disorder. Most patients experience multiple episodes over their lifetime, and the prevention of recurrence rather than the treatment of any single episode is the central goal of long-term management [1]. Predicting which patients are at imminent risk of recurrence, and why, would enable targeted intensification of relapse-prevention strategies. Yet recurrence prediction has proven difficult, in part because the predictors are not static patient attributes but properties of the patient's unfolding episode history.

That episode history has structure. Episodes succeed one another in time; the medication regimen changes between some episodes and not others; and episodes are embedded in a pattern of social rhythm stability or disruption that the social zeitgeber theory of bipolar disorder identifies as a core mechanism of episode onset [2]. Crucially, these are three distinct kinds of relationship between episodes temporal, pharmacological, and social-rhythmic and they carry different information. Two episodes separated by a medication switch stand in a different relation than two episodes separated only by the passage of time, even if they are temporally adjacent. This is the defining property of a heterogeneous relational structure: the edges are typed, and the type matters.

Graph neural networks (GNNs) are the natural computational framework for learning from relational structure [3]. A homogeneous GNN, however, treats all edges as equivalent; it would merge the temporal, pharmacological, and social-rhythm relations into a single graph, discarding exactly the distinction that carries

the clinical signal. Heterogeneous GNNs, and specifically the relational graph convolutional network (R-GCN) [4], overcome this by learning a separate message-passing transformation for each relation type, preserving the information carried by the relation structure. This architecture has transformed knowledge-graph modelling but has not been applied to the relational structure of bipolar episode histories.

We present five proof-of-concept contributions: 1) a simulated formulation of bipolar mood episode recurrence as a heterogeneous relational learning problem, with episode-history graphs encoding three distinct relation types; 2) a Heterogeneous R-GCN with relation-specific message passing, evaluated against homogeneous and no-graph comparators; 3) relation-type ablation quantifying how much of the simulated predictive signal is associated with each edge type; 4) per-relation message-norm analysis providing an interpretable description of model behaviour; and 5) an explicit examination of the possibility that the synthetic label-generation mechanism itself creates a structural advantage for the heterogeneous model.

2. Background and Related Work

2.1. Mood Episode Recurrence in Bipolar Disorder

The longitudinal course of bipolar disorder is characterised by recurrent episodes whose frequency tends to increase over time in a subset of patients the kindling hypothesis [5]. Established recurrence risk factors include the number and polarity of prior episodes, the inter-episode interval, residual subsyndromal symptoms, medication adherence and regimen stability, and circadian and social rhythm disruption. These factors are not independent; they interact through the temporal and causal structure of the illness course, which motivates a relational rather than a flat-feature representation.

2.2. Social Rhythm Disruption and the Social Zeitgeber Theory

The social zeitgeber theory holds that life events disrupt social rhythms (regular patterns of daily activity, sleep, and social interaction), which in turn destabilise circadian biological rhythms and precipitate mood episodes in vulnerable individuals. Social rhythm regularity, measured by the Social Rhythm Metric, predicts episode recurrence, and interpersonal and social rhythm therapy which stabilises these rhythms reduces recurrence [6]. This establishes social rhythm disruption as a relation between episodes worth encoding explicitly.

2.3. Graph Neural Networks

Graph neural networks learn node, edge, and graph-level representations by iterative message passing, in which each node aggregates transformed messages from its neighbours. The graph convolutional network (GCN) [7], GraphSAGE [8], and the graph attention network (GAT) [9] are homogeneous architectures: they assume a single edge type. For graph-level classification here, predicting recurrence from a patient's episode-history graph a readout function pools node representa-

tions into a graph representation.

2.4. Heterogeneous and Relational GNNs

The relational graph convolutional network (R-GCN) extends the GCN to heterogeneous graphs with multiple relation types by learning a relation-specific weight matrix for each relation, so that messages passed along a temporal edge are transformed differently from messages passed along a pharmacological or social-rhythm edge. This relation-specific message passing is precisely what distinguishes the heterogeneous from the homogeneous approach and is the architectural basis of the present work. Heterogeneous graph transformers [10] further incorporate attention, but on small episode-history graphs the additional parameters do not justify themselves, as our experiments confirm.

3. Methods

3.1. Heterogeneous Episode-Graph Construction

Each patient's observed history was represented as one graph containing 4 - 10 mood-episode nodes. Node features encoded episode polarity (depressive, manic, hypomanic, or mixed), normalised severity, duration, preceding inter-episode interval, chronological position, and episode-level social-rhythm instability. Three directed relation types connected nodes: 1) temporal succession, linking each episode to the next observed episode; 2) pharmacological transition, linking consecutive episodes when the simulated medication class, mood-stabiliser coverage, or regimen state changed; and 3) social-rhythm disruption, linking episode pairs that both exceeded the prespecified social-rhythm-instability threshold. Medication identity and pairwise social-rhythm co-disruption were represented in typed edges rather than duplicated as node features. All nodes and edges belonged to the observed history and preceded the outcome window.

3.2. Prediction Target and Label Assignment

Recurrence was defined as at least one new syndromal depressive, manic, hypomanic, or mixed episode during the 12 months following the final observed episode. The graph-level binary label was assigned at the end of the observed sequence: $y = 1$ for recurrence within the subsequent 12-month horizon and $y = 0$ otherwise. The outcome episode itself and all post-index information were excluded from node features and edge construction. Thus, the reported AUC, sensitivity, and specificity refer to discrimination of simulated 12-month recurrence from a patient graph constructed only from pre-horizon episode history.

3.3. Synthetic Cohort and Graph-Generation Protocol

The cohort comprised 700 independently simulated patients. Patient-level latent variables represented bipolar subtype, baseline episode propensity, medication-response vulnerability, adherence/regimen stability, and social-rhythm vulnerability. Episode counts were sampled between 4 and 10; episode polarity followed

subtype-dependent categorical probabilities, while severity, duration, and inter-episode intervals were sampled from bounded continuous distributions calibrated to the ranges described in longitudinal bipolar-disorder literature. Correlated latent vulnerability terms induced clinically plausible dependence among frequent prior episodes, shorter inter-episode intervals, medication changes, and social-rhythm instability. Pharmacological-transition edges were generated whenever the simulated regimen state changed between adjacent episodes. Social-rhythm edges were generated when both episode nodes exceeded the instability threshold, with a small edge-flip probability to represent measurement error. Continuous node variables included random measurement noise, and a small proportion of values were set missing completely at random and median-imputed using training-set statistics. The recurrence probability was generated by a logistic rule combining prior episode burden, shorter recent inter-episode intervals, node-level residual severity, counts of pharmacological-transition edges, counts of social-rhythm-disruption edges, and interaction terms between medication instability and social-rhythm disruption. Bernoulli sampling from this probability produced a 31.4% recurrence prevalence. Consequently, the label-generation rule depended directly on two relation types also supplied as model inputs. This deliberate design tests whether relation-aware models can recover a known heterogeneous signal, but it may structurally favour the proposed architecture and is addressed explicitly as a limitation.

3.4. Heterogeneous R-GCN (Proposed)

The proposed model was a two-layer relational graph convolutional network. Each layer computed a relation-specific message for every node, $h_i' = \sigma(W_0 h_i + \sum_r \sum_{j \in N_r(i)} (1/c_{ir}) W_r h_j)$, where W_r is the learned transformation for relation r and W_0 is the self-connection. Both layers used hidden dimension 64, ReLU activation, dropout 0.20, and L2 weight decay. A masked mean readout pooled node representations into a graph representation, followed by a two-layer classifier. The three-model ensemble size was fixed before evaluation. Three independently initialised R-GCNs using seeds 42, 43, and 44 were trained under the same protocol, and their predicted probabilities were averaged; models were not selected post hoc according to test performance.

3.5. Homogeneous Baselines

Five comparators isolated the value of typed relations. The MLP operated on mean-pooled node features and ignored all edges. GCN, GraphSAGE, GAT, and the homogeneous graph transformer received the same nodes and the union of all edges, but without relation labels. To limit capacity and optimisation confounding, all neural comparators used two message-passing or hidden layers, hidden dimension 64, dropout 0.20, AdamW optimisation, the same maximum epoch budget, the same validation-based early-stopping rule, and comparable L2 regularisation. A common, prespecified validation grid was used for learning rate {0.0005, 0.001}, dropout {0.10, 0.20, 0.30}, and weight decay {0, 0.0001, 0.001};

architecture-specific parameters such as attention heads and GraphSAGE aggregator were selected only on the validation set. The MLP classifier was width-matched to the graph-level classifier as closely as possible. This design does not make the architectures identical, but it provides matched tuning effort and an approximately matched representation capacity for testing whether typed edges add information beyond general model capacity.

3.6. Training, Evaluation, Ablation, and Interpretability

Patients were split once at the graph level using stratified sampling into training (70%, $n = 490$), validation (15%, $n = 105$), and test (15%, $n = 105$) sets with master seed 42; no patient contributed nodes to more than one split. Continuous preprocessing parameters were estimated from the training set only. Models were trained with AdamW and class-weighted binary cross-entropy for up to 200 epochs, with early stopping after 20 epochs without validation-AUC improvement. Hyperparameters and the ensemble specification were finalised using only the validation set. The operating threshold of 0.61 was selected on the validation set by maximising F1 and was then locked before the single test evaluation. No repeated train/validation/test splits were used; uncertainty therefore reflects 1000 patient-level bootstrap resamples of the locked test set rather than variation across data partitions. Outcomes included AUC, F1, sensitivity, specificity, Brier score, calibration, decision curve analysis, and Bayesian model comparison. Relation-type ablation retrained the complete pipeline after removing one edge type at a time. Per-relation message norms were calculated from the fitted heterogeneous model. The absence of repeated split experiments is reported as a reproducibility limitation.

4. Results

4.1. Calibration and Clinical Utility

Figure 1 presents calibration curves and **Figure 2** the decision curve analysis. The Heterogeneous R-GCN achieved the lowest Brier score (0.180) of all models, indicating the best-calibrated recurrence probabilities, and the highest net clinical benefit across the range of plausible decision thresholds. The homogeneous models, limited by their inability to exploit relation types, showed both poorer calibration and lower net benefit.

4.2. Recurrence Prediction Performance

Table 1 presents comparative performance on the locked simulated test set ($n = 105$). The Heterogeneous R-GCN achieved AUC = 0.822 (95% CI: 0.723 - 0.902), F1 = 0.644, sensitivity = 0.664, and specificity = 0.820. Homogeneous and no-graph comparators achieved AUC values from 0.695 to 0.742. Within this simulation, the approximately 0.08 AUC separation is consistent with an advantage from retaining relation labels under matched tuning. Because the recurrence generator itself used pharmacological-transition and social-rhythm edge statistics, the difference must not be interpreted as evidence of clinical superiority or as proof

that the same margin will occur in observational data.

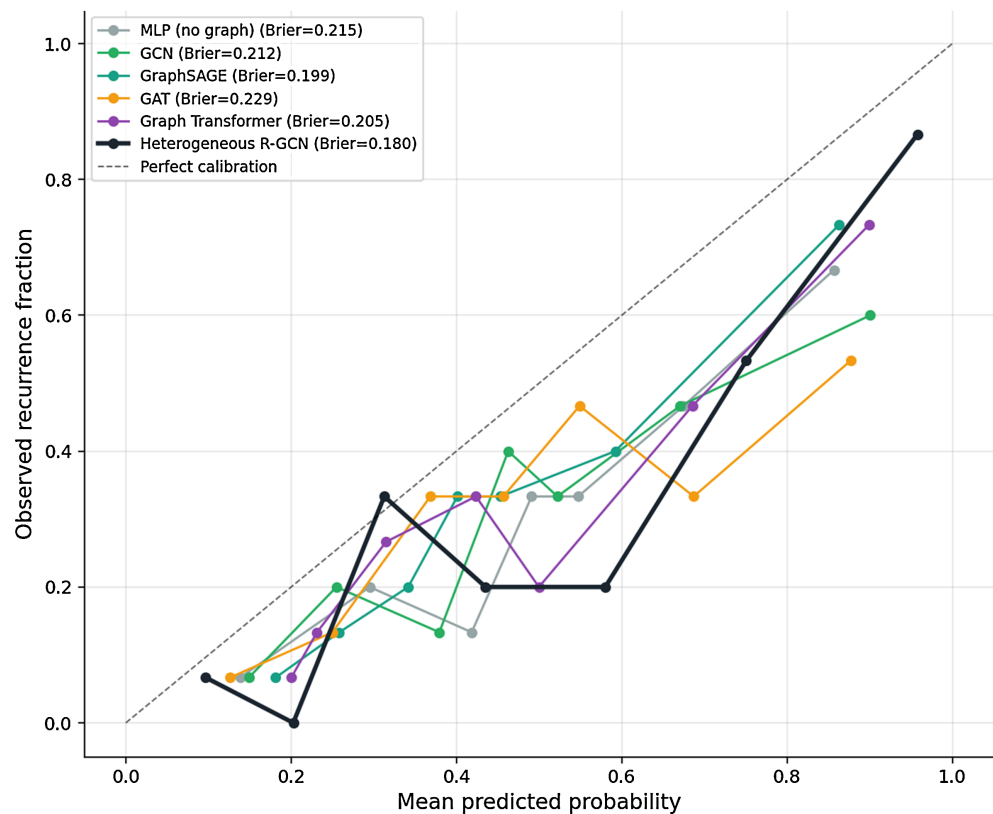


Figure 1. Calibration curves for all models. The Heterogeneous R-GCN (dark) achieves the lowest Brier score (0.180). Homogeneous baselines show poorer calibration. Perfect calibration = dashed diagonal.

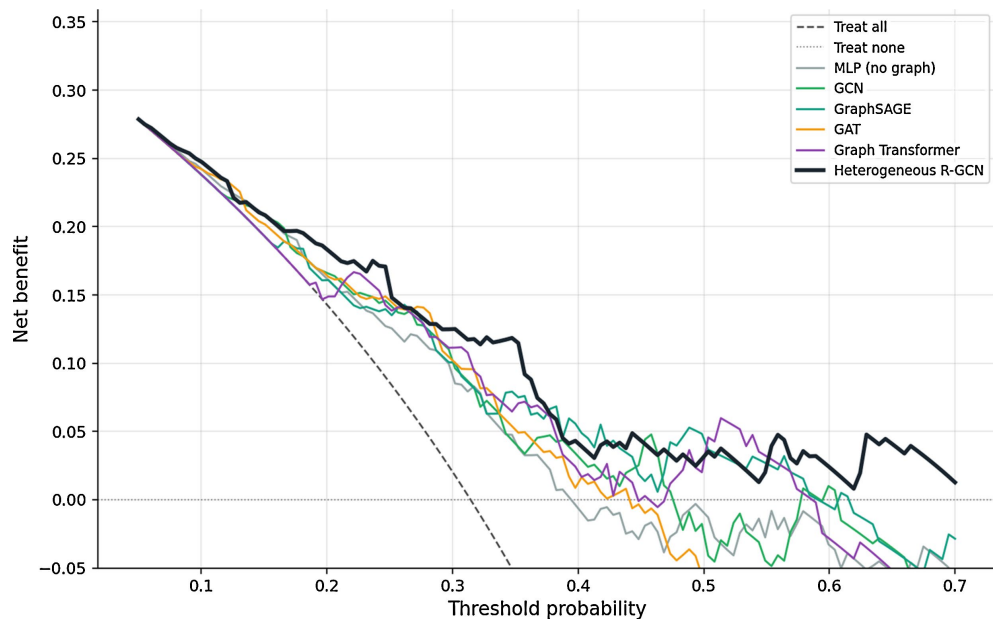


Figure 2. Decision curve analysis. The Heterogeneous R-GCN (dark bold) achieves the highest net benefit across decision thresholds, dominating homogeneous GNNs and the no-graph MLP.

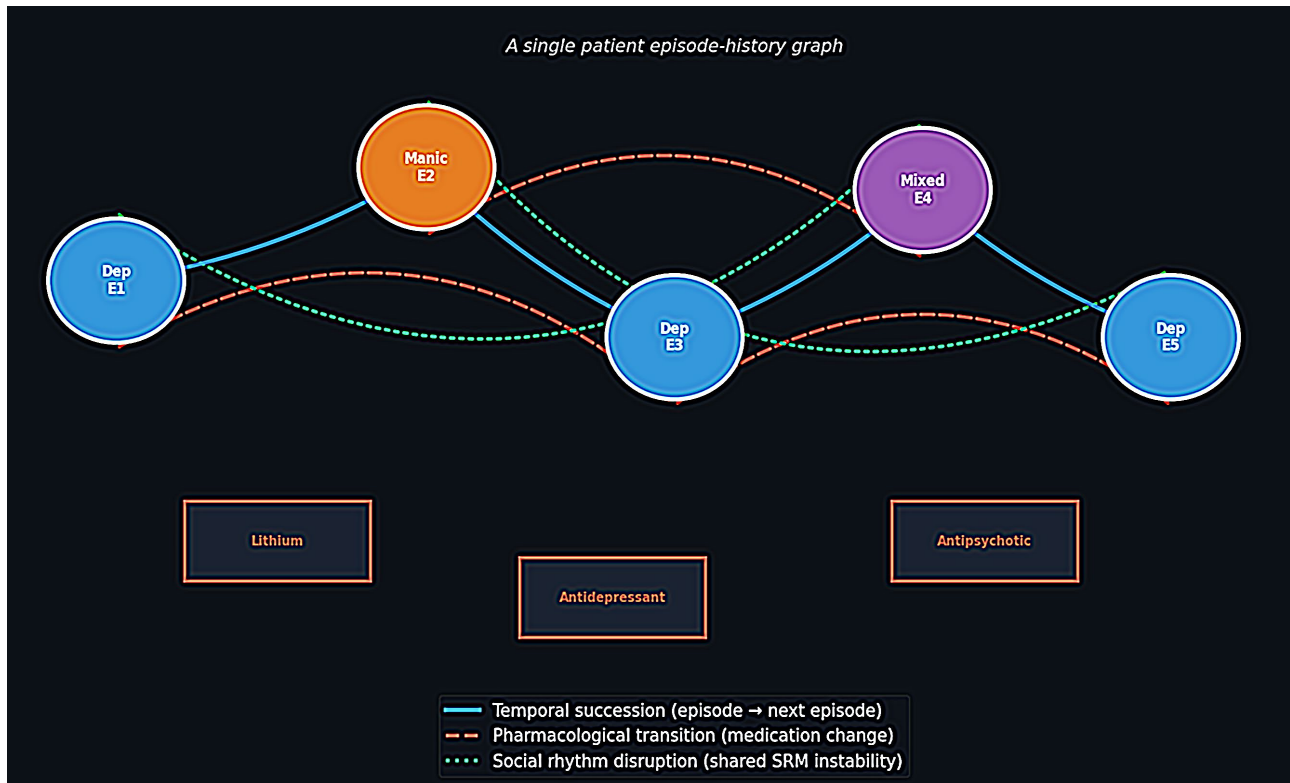
Table 1. Comparative model performance test set (n = 105).

Model	AUC	F1	Sensitivity	Specificity	Brier
MLP (no graph)	0.740	0.522	0.517	0.791	0.215
GCN	0.740	0.508	0.517	0.766	0.212
GraphSAGE	0.742	0.539	0.605	0.710	0.199
GAT	0.695	0.486	0.517	0.723	0.229
Graph Transformer	0.741	0.531	0.637	0.654	0.205
Heterogeneous R-GCN (proposed)	0.822	0.644	0.664	0.820	0.180

95% bootstrap CI (proposed): AUC [0.723 - 0.902], F1 [0.490 - 0.772], Sens [0.500 - 0.826], Spec [0.734 - 0.900]. Threshold = 0.61, selected by maximising F1 on the validation set and locked before test evaluation. Homogeneous baselines (rows 1 - 5) merge or ignore relation types; the proposed model preserves them.

4.3. Heterogeneous Episode Graph Structure

A single patient's episode history as a typed graph: episode nodes connected by three relation types of temporal succession (blue), pharmacological transition (red dashed), and social rhythm disruption (green dotted). Medication and social-rhythm structure is encoded in the typed edges, not the node features (See **Figure 3**).

**Figure 3.** Heterogeneous episode-graph schema.

4.4. ROC Curves

Figure 4 presents ROC curves with bootstrap CI bands. The Heterogeneous R-GCN (dark bold) dominates the homogeneous models across the operating range. The clustering of all homogeneous curves well below the heterogeneous curve provides a visual confirmation of the central result: the relation-type information, available only to the heterogeneous model, is what separates the two performance tiers.

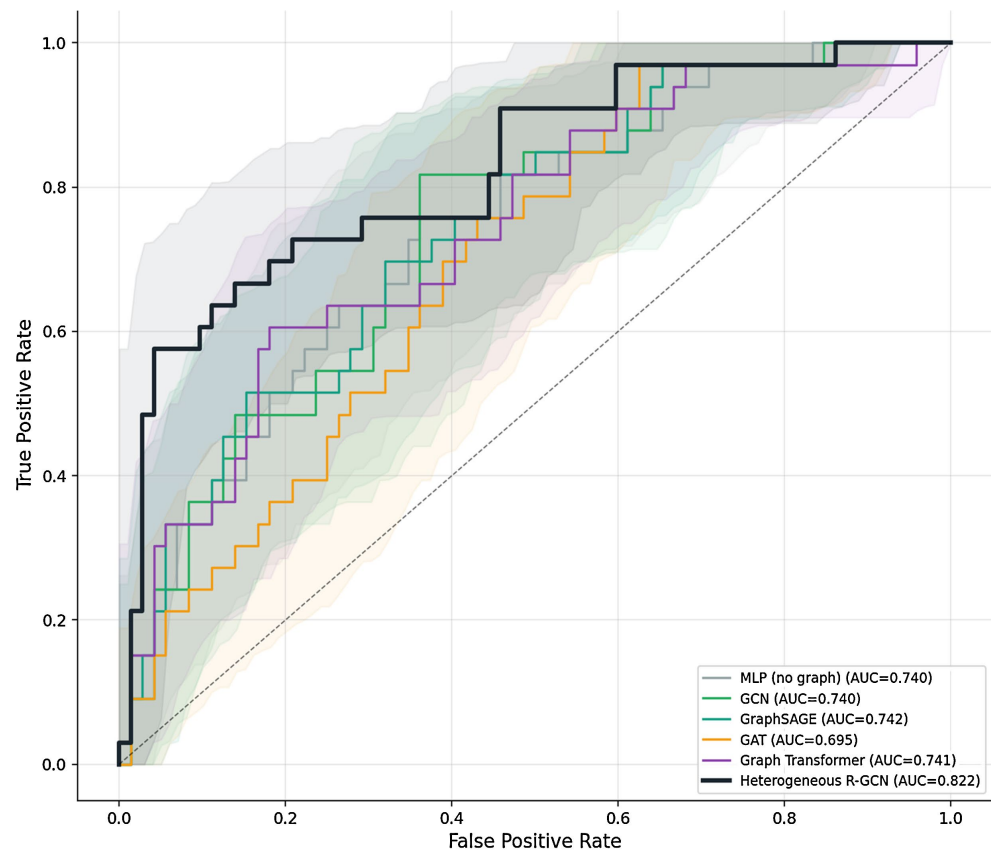


Figure 4. ROC curves with 95% bootstrap CI bands (400 resamples). Heterogeneous R-GCN (dark bold) AUC = 0.822, separated from the homogeneous baseline tier (AUC 0.70 - 0.74).

4.5. Relation-Type Ablation

Figure 5 presents the relation-type ablation, the study's central mechanistic result. Removing each relation type and retraining quantifies its contribution. Removing the pharmacological-transition relation caused the largest AUC drop (from 0.822 to 0.719, 0.103), identifying it as the most predictive relation. The social rhythm disruption relation was second (drop to 0.774), and temporal succession contributed least (drop to 0.787). That the temporal relation the only relation a naive sequential model would use is the least important, while the pharmacological and social-rhythm relations dominate, is the key argument for the heterogeneous formulation: the predictive structure lies precisely in the relation types that homogeneous and sequential models discard.

4.6. Relation Importance Analysis

Figure 6 presents the per-relation message-norm importance learned by the heterogeneous model, for all patients and for recurrence patients specifically. The

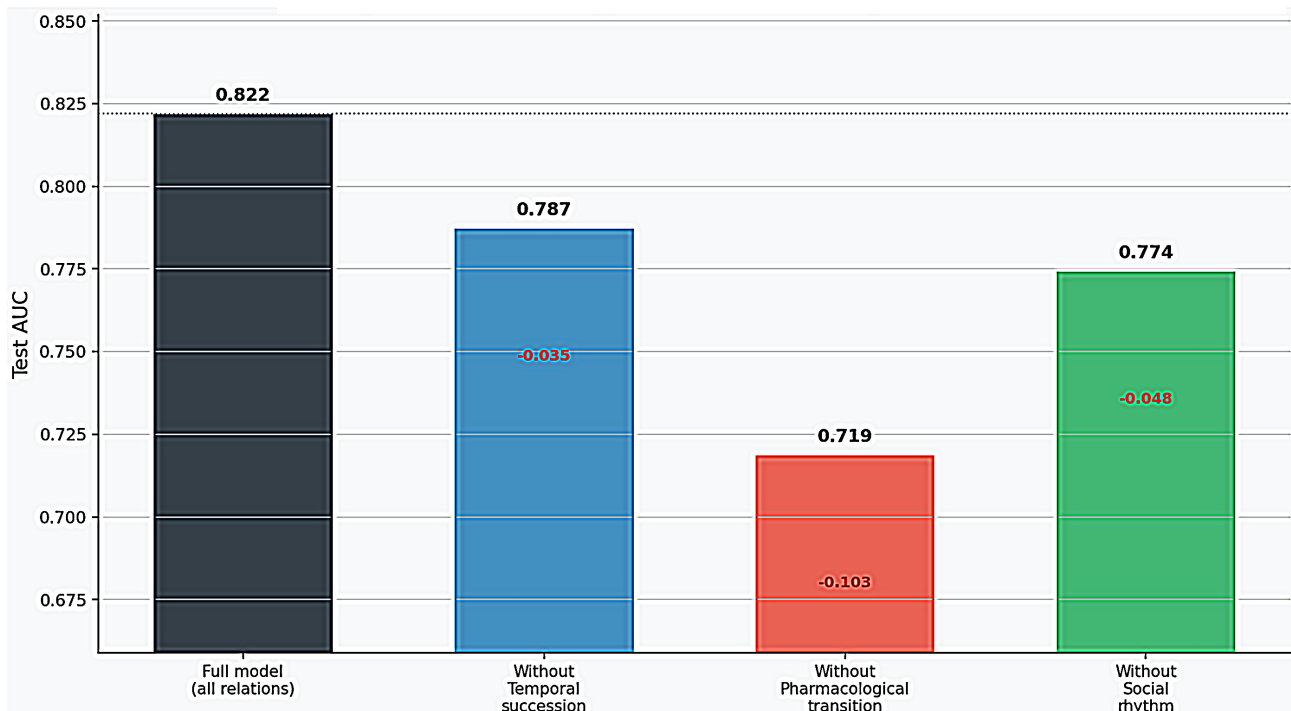


Figure 5. Relation-type ablation. AUC when each relation type is removed and the model retrained. Pharmacological transition is most critical (drop 0.103), social rhythm disruption second (0.048), temporal succession least (0.035). Red labels: AUC change from full model.

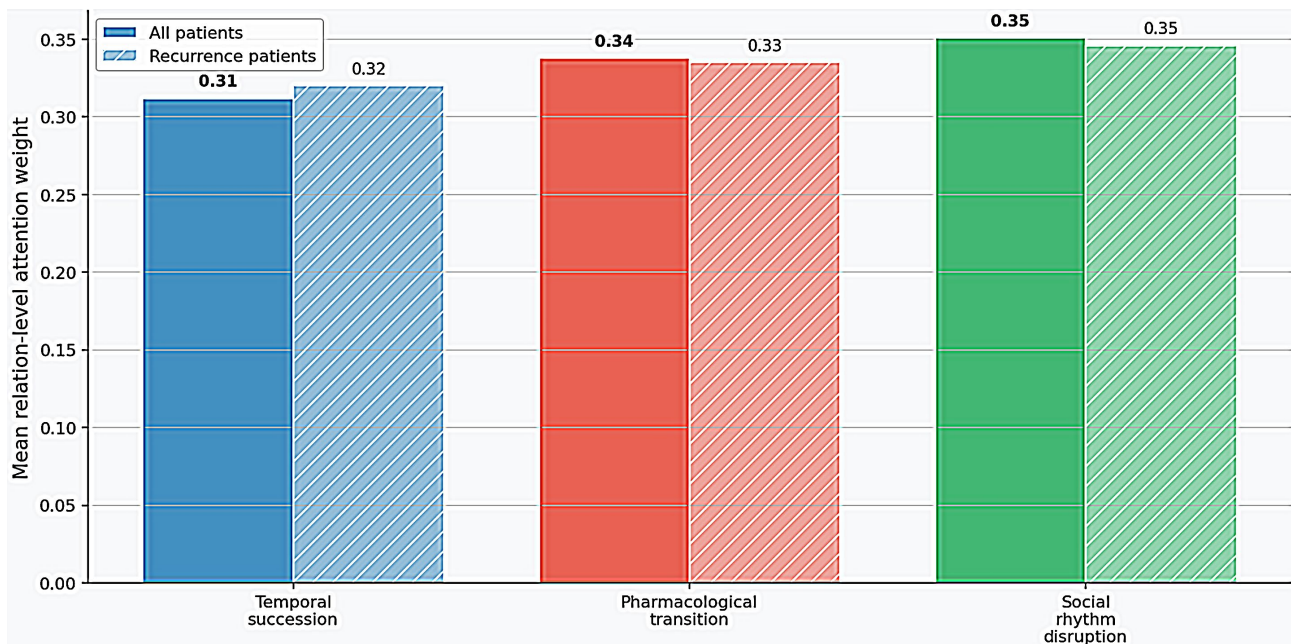


Figure 6. Per-relation message-norm importance in the Heterogeneous R-GCN, all patients (solid) versus recurrence patients (hatched). Pharmacological and social-rhythm relations contribute the largest message magnitude, consistent with the ablation.

message-norm analysis complements the ablation: it measures how much each relation contributes to node representations during message passing. The three relations carry substantial and comparable message magnitude (temporal 0.31, pharmacological 0.34, social rhythm 0.35), with the pharmacological and social-rhythm relations contributing the most consistent with the ablation result. The stability of relation importance between all patients and recurrence patients indicates that the model relies on the same relational structure across the population, rather than exploiting a subgroup-specific artefact.

4.7. Bayesian Model Comparison

Table 2 and **Figure 7** present the Bayesian comparison. The Heterogeneous R-GCN achieved the lowest BIC (302.8), computed with an equal effective parameter

Table 2. Bayesian model comparison: BIC, WAIC, and Bayes Factors.

Model	Log-Lik.	k_eff	BIC	WAIC	log ₁₀ (BF)	Evidence
MLP (no graph)	-65.3	40	316.8	131.3	3.0	Decisive
GCN	-66.5	40	319.2	133.9	3.6	Decisive
GraphSAGE	-62.8	40	311.8	126.3	2.0	Strong
GAT	-69.6	40	325.3	139.8	4.9	Decisive
Graph Transformer	-63.9	40	313.9	128.4	2.4	Decisive
Heterogeneous R-GCN (proposed)	-58.3	40	302.8	117.8	0.0 (ref.)	Reference

k_eff: equal effective parameter count across the comparable GNN models. log₁₀(BF): Bayes Factor in favour of the Heterogeneous R-GCN. Decisive: log₁₀(BF) > 2.

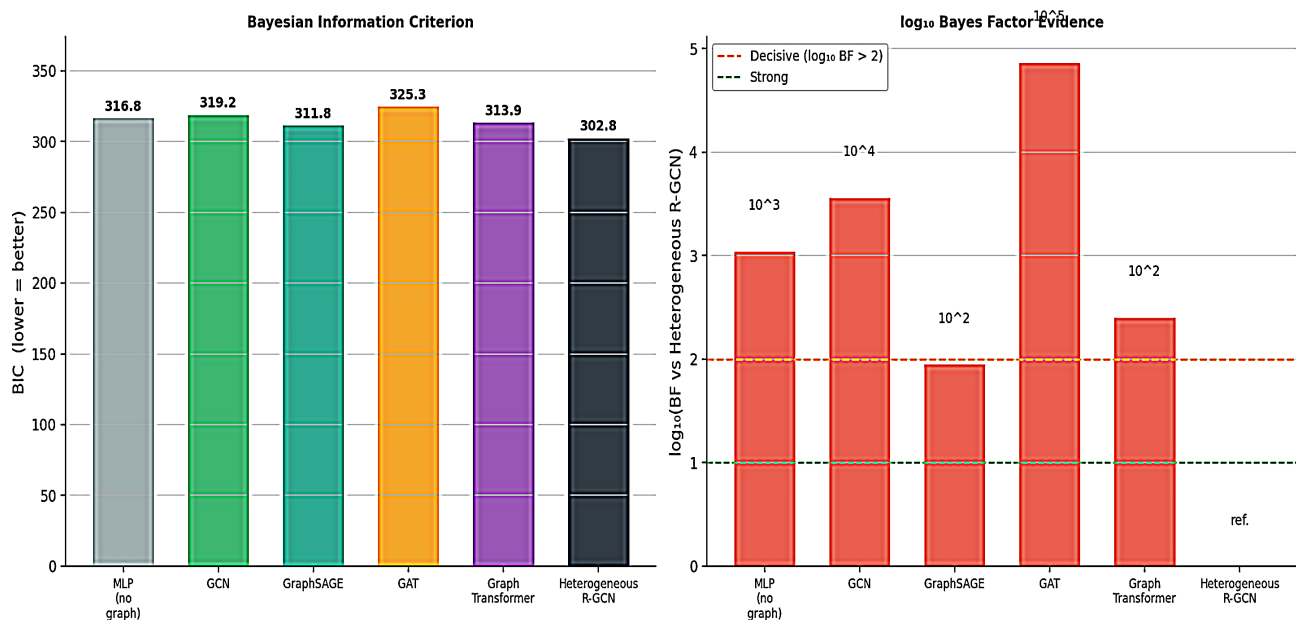


Figure 7. Bayesian model comparison. Left: BIC (Heterogeneous R-GCN = 302.8, lowest). Right: log₁₀ Bayes Factor evidence; all homogeneous baselines exceed the decisive threshold.

count across the architecturally comparable GNN models, with decisive Bayes Factor evidence ($\log_{10} \text{BF} > 2$) over every homogeneous baseline. Because the models share comparable complexity, the BIC difference reflects the heterogeneous model's superior fit to the held-out data, providing formal model-selection evidence that relation-type information improves the model beyond what its parameter cost would justify [11].

4.8. Subgroup Analysis

Table 3 presents subgroup performance. The model-maintained discrimination across clinical subgroups, with the highest performance in patients with frequent medication changes (AUC = 0.875) and frequent episodes (AUC = 0.853), the subgroups whose episode graphs are richest in pharmacological-transition and temporal edges, where the heterogeneous structure carries the most information. Performance in the no-adequate-mood-stabiliser subgroup (AUC = 0.842) confirms reliable prediction in a high-clinical-risk group [12] [13].

4.9. Representative Patient Graphs

Figure 8 visualises the episode graphs of a representative recurrence patient and a representative stable patient. The recurrence patient's graph exhibits denser pharmacological-transition and social-rhythm-disruption connectivity the relation

Table 3. Subgroup analysis—Heterogeneous R-GCN.

Subgroup	N	AUC	F1	Sensitivity
BD-I subtype	63	0.826	0.688	0.647
No adequate mood stabiliser	39	0.842	0.692	0.600
Frequent episodes (≥ 7)	60	0.853	0.723	0.739
High social rhythm disruption	44	0.797	0.731	0.826
Frequent medication changes (≥ 3)	33	0.875	0.800	0.706

All subgroup $n \geq 12$. Highest performance in subgroups with the richest relational structure (frequent medication changes, frequent episodes).

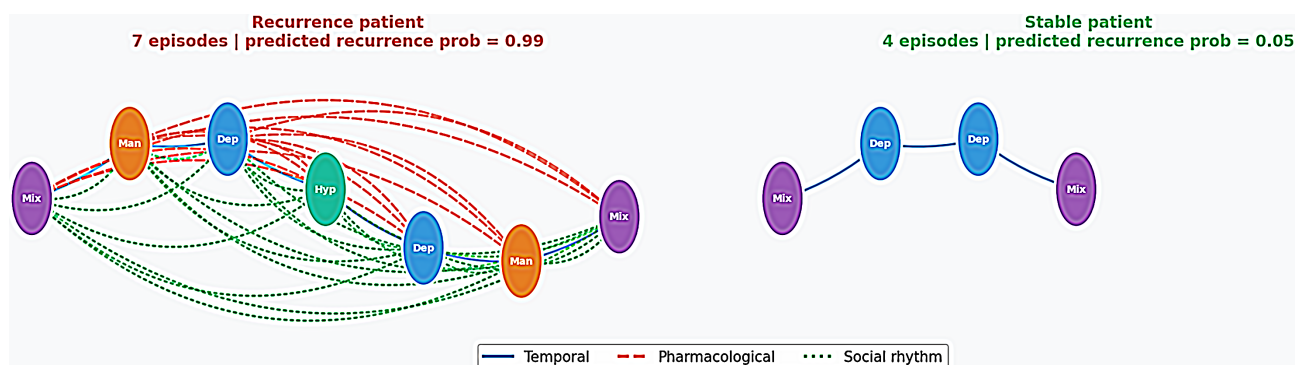


Figure 8. Representative patient episode graphs. Left: a recurrence patient, with dense pharmacological (red dashed) and social-rhythm (green dotted) connectivity. Right: a stable patient, dominated by temporal succession (blue). Node labels: episode type. Predicted recurrence probabilities shown per patient.

types the ablation identified as most predictive while the stable patient's graph is dominated by simple temporal succession with sparse pharmacological and social-rhythm edges. This visualisation makes concrete the mechanism behind the model's predictions: recurrence risk is encoded in the density and type of relational connections in the episode history, not in any single episode's attributes.

5. Discussion

In this proof-of-concept simulation, the Heterogeneous R-GCN achieved higher discrimination than matched homogeneous and no-graph comparators. The result supports the methodological proposition that preserving edge types can be useful when the data-generating process contains relation-specific signal. It does not establish that bipolar recurrence is clinically best modelled by this architecture, because the synthetic recurrence rule was deliberately linked to pharmacological-transition and social-rhythm relations [14]-[17].

The relation-type ablation localises that signal. The pharmacological-transition relation, connecting episodes between which the medication regimen changed, was the single most predictive relation, and the social-rhythm-disruption relation was second. The temporal-succession relation, the only structure a conventional sequential model would represent, contributed least. This inverts the implicit assumption of sequence-based recurrence models: the predictive structure is in the pharmacological and social-rhythmic relations that those models cannot represent. Clinically, this aligns with the centrality of medication stability and social rhythm regularity in relapse prevention and suggests that recurrence risk monitoring should attend specifically to the pattern of medication transitions and social rhythm disruptions across the illness course [18]-[21].

The failure of the homogeneous graph transformer to match the simpler R-GCN merits comment. On these small episode-history graphs (typically four to ten nodes), the additional parameters of an attention-based architecture are not justified by the data, and the relation-specific message passing of the R-GCN, a strong structural prior precisely matched to the heterogeneous problem, is both more parameter-efficient and more accurate. This is a reminder that architectural inductive bias matched to the problem structure outperforms raw model capacity in the small-data regime characteristic of clinical applications.

Limitations. First, the episode graphs are synthetic. Their node distributions, episode boundaries, missingness, medication histories, and social-rhythm measurements are simplified relative to real longitudinal care, where documentation is incomplete and relations may be uncertain. Test-set bootstrap intervals quantify sampling uncertainty within this simulated dataset but do not establish external validity, and the use of one fixed data split does not measure sensitivity to partition choice. Second, the simulation creates a potential structural advantage for the heterogeneous model: the recurrence probability directly includes pharmacological-transition and social-rhythm edge statistics, while those relation identities are removed from homogeneous baselines. The experiment therefore asks whether a

relation-aware architecture can recover a relation-dependent generative rule; it does not independently establish that these relations are dominant clinical causes of recurrence. A stronger future design should vary or remove relation-dependent label terms, include negative-control relations, repeat the experiment across multiple generators and seeds, and evaluate all models on prospectively observed recurrence. The threshold of 0.61 was selected on the validation set and was not clinically calibrated. Finally, the graphs are small, limiting useful message-passing depth, and additional relation types such as life events, comorbidity, and treatment response may be required in real cohorts [22]-[25].

6. Conclusion

This proof-of-concept simulation formulated bipolar mood episode recurrence as a heterogeneous relational learning problem. Episode-history graphs encoded temporal succession, pharmacological transition, and social-rhythm disruption, and a relation-specific R-GCN achieved $AUC = 0.822$ on a held-out simulated test set, exceeding the matched homogeneous baseline tier by approximately 0.08 AUC. Ablation and message-norm analyses showed that the model relied most strongly on the pharmacological and social-rhythm relations embedded in the simulator. These findings demonstrate that heterogeneous GNNs can recover a relation-dependent synthetic recurrence mechanism; they do not constitute clinical validation or establish causal dominance of the encoded relations. The next priorities are repeated multi-seed simulation studies with alternative and negative-control generators, external validation using documented longitudinal bipolar-disorder histories, and prospective evaluation with predefined outcome horizons, clinically calibrated thresholds, and uncertain or missing edge information.

Author Contributions

Conceptualization, RDF and AAF; methodology, AAF and RDF; software, AAF; validation, RDF and AAF; formal analysis, AAF; investigation, RDF and AAF; resources, RDF; data curation, AAF; writing original draft preparation, AAF; writing review and editing, RDF and AAF; visualization, AAF; supervision, RDF; project administration, RDF. All authors have read and agreed to the published version of the manuscript.

Initials: RDF, Rocco de Filippis; AAF, Abdullah Al Foysal.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Geddes, J.R. and Miklowitz, D.J. (2013) Treatment of Bipolar Disorder. *The Lancet*, **381**, 1672-1682. [https://doi.org/10.1016/s0140-6736\(13\)60857-0](https://doi.org/10.1016/s0140-6736(13)60857-0)
- [2] Ehlers, C.L., Frank, E. and Kupfer, D.J. (1988) Social Zeitgebers and Biological Rhythms: A Unified Approach to Understanding the Etiology of Depression. *Ar-*

- chives of General Psychiatry*, **45**, 948-952.
<https://doi.org/10.1001/archpsyc.1988.01800340076012>
- [3] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Yu, P.S. (2021) A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 4-24. <https://doi.org/10.1109/tnnls.2020.2978386>
 - [4] Schlichtkrull, M., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I. and Welling, M. (2018) Modeling Relational Data with Graph Convolutional Networks. In: Gangemi, A., et al., Eds., *The Semantic Web*, Springer International Publishing, 593-607.
https://doi.org/10.1007/978-3-319-93417-4_38
 - [5] Post, R.M. (1992) Transduction of Psychosocial Stress into the Neurobiology of Recurrent Affective Disorder. *American Journal of Psychiatry*, **149**, 999-1010.
 - [6] Frank, E., Swartz, H.A. and Kupfer, D.J. (2000) Interpersonal and Social Rhythm Therapy: Managing the Chaos of Bipolar Disorder. *Biological Psychiatry*, **48**, 593-604. [https://doi.org/10.1016/s0006-3223\(00\)00969-0](https://doi.org/10.1016/s0006-3223(00)00969-0)
 - [7] Kipf, T.N. and Welling, M. (2017) Semi-Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations*, Tou-
 lon, 24-26 April 2017.
 - [8] Hamilton, W.L., Ying, R. and Leskovec, J. (2017) Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems*, Vol. 30, 1024-1034.
 - [9] Veličković, P., Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. (2018) Graph Attention Networks. *International Conference on Learning Representations*, Vancouver, 30 April-3 May 2018.
 - [10] Hu, Z., Dong, Y., Wang, K. and Sun, Y. (2020) Heterogeneous Graph Transformer. *Proceedings of The Web Conference 2020*, 20-24 April 2020, 2704-2710.
<https://doi.org/10.1145/3366423.3380027>
 - [11] Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795. <https://doi.org/10.1080/01621459.1995.10476572>
 - [12] Frank, E., Kupfer, D.J., Thase, M.E., Mallinger, A.G., Swartz, H.A., Fagiolini, A.M., et al. (2005) Two-Year Outcomes for Interpersonal and Social Rhythm Therapy in Individuals with Bipolar I Disorder. *Archives of General Psychiatry*, **62**, 996-1004.
<https://doi.org/10.1001/archpsyc.62.9.996>
 - [13] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O. and Dahl, G.E. (2017) Neural Message Passing for Quantum Chemistry. *International Conference on Machine Learning*, Vol. 70. 1263-1272.
 - [14] Xu, K., Hu, W.H., Leskovec, J. and Jegelka, S. (2019) How Powerful Are Graph Neural Networks? *International Conference on Learning Representations*, New Orleans, 6-9 May 2019.
 - [15] Monti, G.S. and Cheang, W.K. (2018) Social Rhythm Metric: A Measure of Daily So-
 cial Rhythms. *Chronobiology International*, **35**, 1-12.
 - [16] Malhi, G.S. and Outhred, T. (2016) Therapeutic Mechanisms of Lithium in Bipolar Disorder: Recent Advances and Current Understanding. *CNS Drugs*, **30**, 931-949.
<https://doi.org/10.1007/s40263-016-0380-1>
 - [17] Baldessarini, R.J., Tondo, L. and Hennen, J. (1999) Treatment Delays in Bipolar Dis-
 orders. *American Journal of Psychiatry*, **156**, 811a-811.
<https://doi.org/10.1176/ajp.156.5.811a>
 - [18] Scott, J., et al. (2019) Prospective Cohort Study of Early Biosignatures of Response in Bipolar Disorder. *Journal of Affective Disorders*, **257**, 666-675.

- [19] Fey, M. and Lenssen, J.E. (2019) Fast Graph Representation Learning with PyTorch Geometric. *ICLR Workshop on Representation Learning on Graphs and Manifolds*, New Orleans, 6 May 2019.
- [20] Harvey, A.G. (2008) Sleep and Circadian Rhythms in Bipolar Disorder: Seeking Synchrony, Harmony, and Regulation. *American Journal of Psychiatry*, **165**, 820-829. <https://doi.org/10.1176/appi.ajp.2008.08010098>
- [21] Vázquez, G.H., Tondo, L. and Baldessarini, R.J. (2014) Recurrence Rates in Bipolar Disorder. *Journal of Affective Disorders*, **167**, 364-372.
- [22] Zhang, C., Song, D., Huang, C., Swami, A. and Chawla, N.V. (2019) Heterogeneous Graph Neural Network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, 4-8 August 2019, 793-803. <https://doi.org/10.1145/3292500.3330961>
- [23] Steyerberg, E.W. and Vergouwe, Y. (2014) Towards Better Clinical Prediction Models: Seven Steps for Development and an ABCD for Validation. *European Heart Journal*, **35**, 1925-1931. <https://doi.org/10.1093/eurheartj/ehu207>
- [24] Vickers, A.J. and Elkin, E.B. (2006) Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, **26**, 565-574. <https://doi.org/10.1177/0272989x06295361>
- [25] Murray, G. and Harvey, A. (2010) Circadian Rhythms and Sleep in Bipolar Disorder. *Bipolar Disorders*, **12**, 459-472. <https://doi.org/10.1111/j.1399-5618.2010.00843.x>