Supplementary material

1 Technical details of the *hsbm.predict()* function

1.1 Bayesian Model formulation

For the network reconstruction task, the model assumes that the observed network D represents the best available information about the true, unobserved network A. Each entry $D_{ij} = 1$ indicates that a link between node pairs (i, j) was observed, whereas $D_{ij} = 0$ denotes links that were unobserved or removed temporarily as held out during cross-validation. The network D is treated as a noisy measurement of the true network A, potentially containing errors such as false positives (erroneously recorded links) and/or false negatives (missing true links). Following Peixoto (2018), a measurement model is employed where each node pair (i, j) in D is associated with n_{ij} distinct measurements, resulting in x_{ij} observed links, such that $D_{ij} = (n_{ij}, x_{ij})$. Although this model accommodates arbitrary non-negative integers $x_{ij} \leq n_{ij}$, *sabinaHSBM* focuses on the common case of binary adjacency matrices, where only a single measurement per node pair is available or the details of the measurement process are unknown. In this setting, $n_{ij} = 1$ is assumed for all pairs (i, j) and $x_{ij} \in \{0, 1\}$ corresponds directly to the observed values in the adjacency matrix D. The goal is to infer the true network A and the latent community structure b from the observed data $D = (n, x)$, by targeting the joint posterior distribution:

$$P(A, b|D) \propto P(D|A)P(A, b) \tag{1}$$

Here, $P(D|A)$ models the measurement process, and $P(A, b)$ represents the prior joint probability of the network and its structure. Since the model uses the stochastic block model (SBM) as the generative prior, this can be decomposed as $P(A, b) = P(A|b)P(b)$, where $P(A|b)$ is the probability of network A given the block partition b, and $P(b)$ is the prior on the block partition itself. This Bayesian formulation, detailed in Peixoto (2018), integrates uncertainty in the measurement process with the structural correlations captured by the HSBM. To explore the posterior, Markov Chain Monte Carlo (MCMC) sampling is used to generate k configurations $(A^{(k)}, b^{(k)})$ from the joint posterior distribution.

1.2 MCMC Sampling Process

The HSBM inference procedure in *sabinaHSBM* relies on the implementation in the *graph-tool* python module. It uses Markov Chain Monte Carlo (MCMC) sampling to explore the joint posterior distribution of networks A and the latent block structure b. The process unfolds in two phases: an equilibration phase, and a posterior sampling phase.

During the equilibration phase, the model iteratively searches for convergence. In each iteration, the model performs 10 proposals to move nodes between blocks, accepting or rejecting these moves based on improvements in the posterior probability conditioned on the observed network D. After every iteration, the description length (DL) of the current configuration is recorded. Convergence is assessed using the *wait* argument (default: 1,000), which defines the number of successive iterations during which the minimum and maximum values of the DL must remain range stable. This period of DL stability must occur twice to ensure that convergence is not reached by chance or local fluctuations.

Once equilibration is achieved, the sampling phase begins. The model performs a number of additional iterations—defined by the *iter* argument (default: 10,000). As before, each iteration performs 10 node reassignment proposals, leading to a total of 100,000 proposals across the sampling phase. However, only the final state/configuration of each iteration is retained. These retained samples represent samples from the posterior distribution and are used to estimate link probabilities.

1.3 Link Prediction Methods

Two methods for link prediction are available via the *method* argument:

The “conditional_missing” method (conditional probability averaging) is a classical gap-filling approach that focuses exclusively on scoring unobserved links (i.e., those with $D_{ij} = 0$). Specifically, for each unobserved node pair (i, j), it computes the average conditional probability that a link exists, given the inferred block structure. After the initial MCMC equilibration or burn-in (controlled with *wait* argument), a set of N posterior block partitions $b^{(k)}$ is obtained (controlled with *iter* argument). For each partition, the package calls the *MeasuredBlockState.get_edges_prob()* function in *graph-tool* (http://graph-tool.skewed.de/) to compute the conditional probability $P(A_{ij} = 1|b^{(k)})$ and then averages these values across all N samples to yield:

$$p_{ij} = \frac{1}{N}\sum_{k=1}^{N} P(A_{ij} = 1|b^{(k)}) \tag{2}$$

By relying exclusively on the inferred block structure, the “conditional_missing” method implements the classical approach for link prediction assigning each unobserved link a score based on its likelihood under a generative model. It retains the familiar/classical SBM link-prediction paradigm (Ghasemian et al., 2020), allowing direct comparison with other existing algorithms. However, this method does not yield true marginal probabilities, and it is more computationally intensive than the “marginal_all” method. For these reasons, sabinaHSBM restricts this method to estimating probabilities for unobserved links ($D_{ij} = 0$) only.

The “marginal_all” method (marginal posterior probability estimation) is a full network reconstruction approach that computes the marginal posterior probability $P(A_{ij} = 1|D)$ for observed ($D_{ij} = 1$) and unobserved ($D_{ij} = 0$) links. After the MCMC burn-in (controlled by the *wait* argument), we draw N samples (set via *iter* argument). For each sample we use the *MeasuredBlockState.collect_marginal()* function from *graph-tool* to generate a latent/sampled network $A^{(k)}$ and record every link presence $A_{ij}^{(k)} = 1$. These link presences are averaged to obtain:

$$p_{ij} = \frac{1}{N}\sum_{k=1}^{N} A_{ij}^{(k)} \tag{3}$$

where N is the number of posterior samples and $A_{ij}^{(k)} = 1$ if the link (i, j) exists in the k-th latent network, and 0 otherwise. The resulting p_{ij} directly reflects the accumulated evidence for the existence of the link (i, j) across the posterior. The “marginal_all” method delivers a marginal probability matrix that simultaneously uncovers missing links (false negatives) and spurious links (false positives), making it the method of choice for network diagnostics and quality assessment.

Because it samples full networks rather than focusing on blocks, some links may never appear in any sample $A^{(k)}$, yielding $p_{ij} \approx 0$. *sabinaHSBM* lets you handle these cases flexibly via the *na_treatment* argument on the *hsbm.reconstructed()* function, ensuring no uncertainty is lost in downstream analyses.

1.4 Key Function Parameters

The *hsbm.predict()* process is controlled by several arguments:

- *wait*: The number of successive, stable MCMC iterations required to assess convergence during the equilibration phase (default: 1,000).
- *iter*: The number of posterior samples generated after the equilibration phase (default: 10,000).
- *n_cores*: Enables parallel computation across folds, to speed up execution on large datasets or multiple folds (default: 1).
- *rnd_seed*: An optional argument to fix the random seed for the HSBM inference to ensure reproducibility.
- *elist_i*: Selects a specific cross-validation fold for computation. By default, predictions are computed across all folds.

1.5 Additional Function Outputs

Optionally, the function allows saving the results and *graph-tool* objects as Python pickle files with `save_pickle = TRUE`. These files are stored in the working directory under names like hsbm__res__foldi.pkl, where i denotes the fold index. Also, a simple hierarchical edge bundling figure for each fold can be saved in the working directory, using the argument `save_plots = TRUE`.

2 Case study results

Table S1. Results of reconstruction for each fold on the Carnivora dataset

Fold	AUC	Threshold	Nr. Held-out	Pred. held-out ones	Total pred. ones
1	0.99	0.0015	178	0.49	3378
2	0.98	0.0016	178	0.38	2849
3	0.99	0.0022	178	0.36	2583
4	0.99	0.002	178	0.46	3160
5	0.99	0.0016	178	0.47	3224
6	0.99	0.0024	179	0.31	2741
7	0.99	0.0019	178	0.34	2589
8	0.99	0.0019	178	0.48	3262
9	0.99	0.0015	179	0.39	2784
10	0.99	0.001	179	0.45	3388
Average	**0.99**	**0.0018**		**0.41**	**2995.8**

Table S2. Literature search for most likely interactions

Host	Parasite	Prob	SD	Evidence	Ref
Neovison vison	Mesocestoides lineatus	0.19	0.27	Confirmed	Zschille et al. (2004)
Lynx rufus	Carnivore protoparvovirus 1	0.12	0.31	Confirmed	Allison et al. (2013)
Procyon lotor	Yersinia pestis	0.08	0.08	Confirmed	Clover et al. (1989)
Meles meles	Eucoleus aerophilus	0.06	0.06	Confirmed	Byrne et al. (2020)
Lynx rufus	Felid alphaherpesvirus 1	0.06	0.04	Plausible. Artificial innoculation resulted in asymptomatic infection.	Eberle et al. (1991)
Procyon lotor	Ctenocephalides felis	0.04	0.07	Confirmed	Sharifdini et al. (2021)
Nyctereutes procyonoides	Capillaria aerophila	0.03	0.03	Confirmed for parasite synonym name *Eucoleus aerophilus*	Laurimaa et al. (2016)
Lutra lutra	Molineus patens	0.03	0.03	Confirmed	Takeuchi-Storm et al. (2021)
Nyctereutes procyonoides	Toxoplasma gondii	0.03	0.02	Confirmed	Osten-Sacken et al. (2024)
Neovison vison	Macracanthorhynchus catulinus	0.03	0.02	No evidence found. Infection of other members of Mustela genus.	Shimalov and Shimalov (2002)

Table S3. Mean phylogenetic distance for grouping levels found for hosts during Hierarchical Stochastic Block Model inference for fold 6 on the Carnivora dataset. Results shown for groups with at least 10 hosts.

Level	Nr hosts	Obs.	Null mean	Null sd	p	p<0.05
Level 1 (nr groups: 12)						
	32	102.419	101.197	2.166	0.669	
	19	61.710	101.346	3.175	0.001	*
	22	103.088	101.360	2.818	0.704	
	10	82.213	101.422	5.928	0.014	*
	15	88.617	101.362	3.988	0.011	*
Level 2 (nr groups: 4)						
	69	99.232	101.318	1.058	0.040	*
	27	61.427	101.371	2.351	0.001	*
	39	83.234	101.240	1.824	0.001	*
Level 3 (nr groups: 2)						
	135	100.806	101.258	0.194	0.019	*
	135	100.806	101.273	0.188	0.012	*

Table S4. Mean nearest taxon distance for grouping levels found for hosts during Hierarchical Stochastic Block Model inference for fold 6 on the Carnivora dataset. Results shown for groups with at least 10 hosts.

Level	Nr hosts	Obs	Null mean	Null sd	p	p<0.05
Level 1 (nr groups: 12)						
	32	26.387	25.356	3.119	0.623	
	19	23.968	31.671	5.382	0.076	
	22	28.264	29.647	4.274	0.372	
	10	22.180	44.657	10.077	0.014	*
	15	25.960	35.553	6.664	0.066	
Level 2 (nr groups: 4)						
	69	19.928	18.884	1.347	0.787	
	27	22.459	27.296	3.615	0.094	
	39	15.087	23.578	2.579	0.002	*
Level 3 (nr groups: 2)						
	135	15.216	15.358	0.262	0.288	
	135	15.216	15.354	0.262	0.279	

Table S5. Mean phylogenetic distance of inferred groupings compared to null model. G1, G2, G3, G4 are the grouping levels. In each level the numbers refer to how many inferred grouping were found to be significant (p < 0.05) compared to a null model, over all groupings with more than 10 taxa. So 2/4 means that two inferred groupings were found to be significantly clustered on a total of 4 inferred groupings with more than 10 taxa.

Fold	**G1**	**G2**	**G3**	**G4**
1	3/6	2/3	1/1	1/1
2	2/5	2/3	0/1	NA
3	2/4	2/4	0/1	0/1
4	2/4	1/3	0/1	NA
5	2/4	4/4	1/1	NA
6	3/5	2/3	1/1	1/1
7	4/6	2/3	0/1	NA
8	2/5	3/3	0/1	NA
9	2/4	2/3	0/1	NA
10	3/5	4/4	0/1	NA
Average	**0.52**	**0.72**	**0.3**	**0.67**

Table S6. Mean nearest taxon distance of inferred groupings compared to null model. G1, G2, G3, G4 are the grouping levels. In each level the numbers refer to how many inferred grouping were found to be significant (p < 0.05) compared to a null model, over all groupings with more than 10 taxa. So 2/4 means that two inferred groupings were found to be significantly clustered on a total of 4 inferred groupings with more than 10 taxa.

Fold	**G1**	**G2**	**G3**	**G4**
1	0/6	1/3	0/1	0/1
2	0/5	0/3	0/1	NA
3	0/4	1/4	0/1	0/1
4	0/4	0/3	1/1	NA
5	1/4	0/4	0/1	NA
6	1/5	1/3	0/1	0/1
7	2/6	1/3	0/1	NA
8	1/5	2/3	1/1	NA
9	0/4	0/3	0/1	NA
10	0/5	1/4	0/1	NA
Average	**0.1**	**0.22**	**0.2**	**0**

References

Allison, A. B., Kohler, D. J., Fox, K. A., Brown, J. D., Gerhold, R. W., Shearn-Bochsler, V. I., Dubovi, E. J., Parrish, C. R., and Holmes, E. C. (2013). Frequent cross-species transmission of parvoviruses among diverse carnivore hosts. *Journal of virology*, 87(4):2342–2347.

Byrne, R., Fogarty, U., Mooney, A., Harris, E., Good, M., Marples, N., and Holland, C. (2020). The helminth parasite community of european badgers (meles meles) in ireland. *Journal of Helminthology*, 94:e37.

Clover, J., Hofstra, T., Kuluris, B., Schroeder, M., Nelson, B., Barnes, A., and Botzler, R. (1989). Serologic evidence of yersinia pestis infection in small mammals and bears from a temperate rainforest of north coastal california. *Journal of Wildlife Diseases*, 25(1):52–60.

Eberle, R., Baldwin, C. J., Black, D., Kocan, A. A., and Fulton, R. W. (1991). Feline herpesvirus infections in bobcats (lynx rufus): disease in experimentally inoculated animals. *Journal of Zoo and Wildlife Medicine*, pages 175–183.

Ghasemian, A., Hosseinmardi, H., and Clauset, A. (2020). Evaluating overfit and underfit in models of network community structure. *IEEE Transactions on Knowledge and Data Engineering*, 32(9):1722–1735.

Laurimaa, L., Süld, K., Davison, J., Moks, E., Valdmann, H., and Saarma, U. (2016). Alien species and their zoonotic parasites in native and introduced ranges: the raccoon dog example. *Veterinary Parasitology*, 219:24–33.

Osten-Sacken, N., Pikalo, J., Steinbach, P., and Heddergott, M. (2024). Prevalence of toxoplasma gondii antibodies and risk factors in two sympatric invasive carnivores (procyon lotor and nyctereutes procyonoides) from zgorzelec county, poland. *Pathogens*, 13(3):210.

Peixoto, T. P. (2018). Reconstructing networks with unknown and heterogeneous errors. *Phys. Rev. X*, 8:041011.

Sharifdini, M., Norouzi, B., Azari-Hamidian, S., and Karamzadeh, N. (2021). The first record of ectoparasites of raccoons (procyon lotor)(carnivora, procyonidae) in iran. *Persian Journal of Acarology*, 10(1):41–54.

Shimalov, V. and Shimalov, V. (2002). Helminth fauna of the european polecat (mustela putorius linnaeus, 1758) in belorussian polesie. *Parasitology research (1987)*, 88(3):259–260.

Takeuchi-Storm, N., Al-Sabi, M., Chriel, M., and Enemark, H. (2021). Systematic examination of the cardiopulmonary, urogenital, muscular and gastrointestinal parasites of the eurasian otters (lutra lutra) in denmark, a protected species recovering from a dramatic decline. *Parasitology International*, 84:102418.

Zschille, J., Heidecke, D., and Stubbe, M. (2004). Verbreitung und ökologie des minks-mustela vison schreber, 1777 (carnivora, mustelidae)-in sachsen-anhalt. *Hercynia-Ökologie und Umwelt in Mitteleuropa*, 37(1):103–126.