

# sabinaHSBM: An R package for link prediction and network reconstruction using Hierarchical Stochastic Block Models

Supplementary material

Herlander Lima, Jennifer Morales-Barbero, Ruben G. Mateo,  
Ignacio Morales-Castilla, Miguel A. Rodríguez

## Appendix 1: MCMC sampling process in *sabinaHSBM*

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.

In the “full\_reconstruction” method, marginal probabilities are computed by averaging link presence across all network configurations visited during the 10 proposals per MCMC iteration. In contrast, the “binary\_classifier” method computes conditional probabilities from the final configuration of each iteration, after completing the 10 proposals. This reflects how likely the link is under each sampled block structure.

## Appendix 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	0.47	3267
2	0.98	0.0016	178	0.37	2775
3	0.99	0.0022	178	0.35	2542
4	0.99	0.002	178	0.42	3084
5	0.99	0.0016	178	0.45	3119
6	0.99	0.0024	179	0.3	2698
7	0.99	0.0019	178	0.33	2533
8	0.99	0.0019	178	0.46	3185
9	0.99	0.0015	179	0.37	2696
10	0.99	0.001	179	0.42	3214

<b>Average</b>	<b>0.99</b>	<b>0.0018</b>	<b>0.39</b>	<b>2911.3</b>
----------------	-------------	---------------	-------------	---------------

**Table S2.** Literature search for most probable interactions

Host	Parasite	Prob	SD	Evidence	Ref
Neovison vison	Mesocestoides lineatus	0.19	0.27	Confirmed	[1]
Lynx rufus	Carnivore protoparvovirus 1	0.12	0.31	Confirmed	[2]
Procyon lotor	Yersinia pestis	0.08	0.08	Confirmed	[3]
Meles meles	Eucoleus aerophilus	0.06	0.06	Confirmed	[4]
Lynx rufus	Felid alphaherpesvirus 1	0.06	0.04	Plausible. Artificial inoculation resulted in asymptomatic infection.	[5]
Procyon lotor	Ctenocephalides felis	0.04	0.07	Confirmed	[6]
Nyctereutes procyonoides	Capillaria aerophila	0.03	0.03	Confirmed for parasite synonym name Eucoleus aerophilus	[7]
Lutra lutra	Molineus patens	0.03	0.03	Confirmed	[8]
Nyctereutes procyonoides	Toxoplasma gondii	0.03	0.02	Confirmed	[9]
Neovison vison	Macracanthorhynchus catulinus	0.03	0.02	No evidence found. Infection of other members of Mustela genus.	[10]

**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
<b>Level 1 (nr groups: 12)</b>						
	32	102.419	101.345	2.133	0.659	
	19	61.710	101.138	3.327	0.001	*
	22	103.088	101.366	2.823	0.709	
	10	82.213	101.143	6.058	0.011	*
	15	88.617	101.339	4.165	0.013	*
<b>Level 2 (nr groups: 4)</b>						
	69	99.232	101.248	1.094	0.049	*
	27	61.427	101.329	2.440	0.001	*
	39	83.234	101.322	1.838	0.001	*
<b>Level 3 (nr groups: 2)</b>						
	135	100.806	101.262	0.195	0.021	*
	135	100.806	101.268	0.189	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
<b>Level 1 (nr groups: 12)</b>						
	32	26.387	25.326	3.048	0.629	
	19	23.968	31.870	5.420	0.069	
	22	28.264	29.734	4.542	0.377	
	10	22.180	44.941	10.622	0.014	*
	15	25.960	35.512	6.747	0.067	
<b>Level 2 (nr groups: 4)</b>						
	69	19.928	18.837	1.400	0.770	
	27	22.459	27.386	3.717	0.096	
	39	15.087	23.465	2.615	0.001	*
<b>Level 3 (nr groups: 2)</b>						
	135	15.216	15.374	0.251	0.250	
	135	15.216	15.369	0.250	0.256	

**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	3/4	0/1	0/1
4	2/4	1/3	0/1	NA
5	2/4	4/4	1/1	NA
6	3/5	3/3	1/1	1/1
7	5/6	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
<b>Average</b>	<b>0.53</b>	<b>0.78</b>	<b>0.3</b>	<b>0.67</b>

**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
<b>Average</b>	<b>0.1</b>	<b>0.22</b>	<b>0.2</b>	<b>0</b>

## References

- [1] Jana Zschille, Dietrich Heidecke, and Michael Stubbe. “Verbreitung und Ökologie des Minks-*Mustela vison* Schreber, 1777 (Carnivora, Mustelidae)-in Sachsen-Anhalt”. In: *Hercynia-Ökologie und Umwelt in Mitteleuropa* 37.1 (2004), pp. 103–126.
- [2] Andrew B Allison et al. “Frequent cross-species transmission of parvoviruses among diverse carnivore hosts”. In: *Journal of virology* 87.4 (2013), pp. 2342–2347.
- [3] JR Clover et al. “Serologic evidence of *Yersinia pestis* infection in small mammals and bears from a temperate rainforest of north coastal California”. In: *Journal of Wildlife Diseases* 25.1 (1989), pp. 52–60.
- [4] RL Byrne et al. “The helminth parasite community of European badgers (*Meles meles*) in Ireland”. In: *Journal of Helminthology* 94 (2020), e37.
- [5] Richard Eberle et al. “Feline herpesvirus infections in bobcats (*Lynx rufus*): disease in experimentally inoculated animals”. In: *Journal of Zoo and Wildlife Medicine* (1991), pp. 175–183.
- [6] Meysam Sharifdini et al. “The first record of ectoparasites of raccoons (*Procyon lotor*)(Carnivora, Procyonidae) in Iran”. In: *Persian Journal of Acarology* 10.1 (2021), pp. 41–54.

- [7] Leidi Laurimaa et al. “Alien species and their zoonotic parasites in native and introduced ranges: the raccoon dog example”. In: *Veterinary Parasitology* 219 (2016), pp. 24–33.
- [8] N Takeuchi-Storm et al. “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”. In: *Parasitology International* 84 (2021), p. 102418.
- [9] Natalia Osten-Sacken et al. “Prevalence of *Toxoplasma gondii* Antibodies and Risk Factors in Two Sympatric Invasive Carnivores (*Procyon lotor* and *Nyctereutes procyonoides*) from Zgorzelec County, Poland”. In: *Pathogens* 13.3 (2024), p. 210.
- [10] VV Shimalov and VT Shimalov. “Helminth fauna of the European polecat (*Mustela putorius* Linnaeus, 1758) in Belorussian Polesie”. In: *Parasitology research (1987)* 88.3 (2002), pp. 259–260.