Table 1: Evaluating the predictive performance across 50 synthetic and 17 real-world datasets for scenarios 1-4 in terms of Root Mean Squared Error (RMSE). The best result among all fully Bayesian methods is marked in **bold**. For the fully Bayesian approaches, we use the posterior mean as a point estimate for the response. MAP denotes the predictive performance of the model with the maximum a posteriori estimate for the latents. TabPFN gives the best overall performance, which is expected since it is not limited to the GLM structure. Besides that, the MAP approach obtains consistently the best results, while our ICL method performs on par (scenarios 1,2,3) or better than (scenario 4) compared to the fully Bayesian methods on the real-world data. On the synthetic data there is no significant difference to the other fully Bayesian methods, except for the real-world data in scenario where HMC is clearly the best method.

Scenario	Model	RMSE Real-World (↓)	RMSE Synthetic (↓)
Scenario 1	HMC	**0.591** (± 0.023)	**0.510** (± 0.040)
	Laplace Approximation	**0.594** (± 0.023)	**0.510** (± 0.040)
	VI: DiagonalNormal	**0.591** (± 0.023)	**0.509** (± 0.040)
	VI: MultivariateNormal	**0.591** (± 0.023)	**0.510** (± 0.040)
	VI: Structured Normal	**0.629** (± 0.017)	0.555 (± 0.039)
	VI: IAF	**0.593** (± 0.023)	**0.510** (± 0.040)
	ICL (ours)	**0.593** (± 0.020)	**0.524** (± 0.038)
	MAP	0.555 (± 0.024)	0.491 (± 0.038)
	TabPFN	0.483 (± 0.036)	0.453 (± 0.036)
Scenario 2	HMC	**0.559** (± 0.023)	**0.556** (± 0.049)
	Laplace Approximation	**0.561** (± 0.022)	**0.557** (± 0.049)
	VI: DiagonalNormal	**0.560** (± 0.023)	**0.557** (± 0.049)
	VI: MultivariateNormal	**0.559** (± 0.023)	**0.556** (± 0.049)
	VI: Structured Normal	**0.604** (± 0.016)	0.685 (± 0.054)
	VI: IAF	**0.563** (± 0.023)	**0.557** (± 0.049)
	ICL (ours)	**0.561** (± 0.019)	**0.653** (± 0.049)
	MAP	0.513 (± 0.023)	0.522 (± 0.048)
	TabPFN	0.449 (± 0.034)	0.498 (± 0.047)
Scenario 3	HMC	**0.684** (± 0.027)	**0.512** (± 0.040)
	Laplace Approximation	**0.688** (± 0.026)	0.516 (± 0.040)
	VI: DiagonalNormal	**0.686** (± 0.027)	**0.513** (± 0.040)
	VI: MultivariateNormal	**0.685** (± 0.027)	**0.512** (± 0.040)
	VI: Structured Normal	**0.733** (± 0.016)	0.607 (± 0.043)
	VI: IAF	**0.686** (± 0.027)	**0.512** (± 0.040)
	ICL (ours)	**0.690** (± 0.023)	**0.588** (± 0.045)
	MAP	0.646 (± 0.028)	0.495 (± 0.039)
	TabPFN	0.556 (± 0.041)	0.462 (± 0.037)
Scenario 4	HMC	**0.642** (± 0.027)	**0.559** (± 0.051)
	Laplace Approximation	0.737 (± 0.048)	2.457 (± 0.493)
	VI: DiagonalNormal	0.751 (± 0.038)	2.046 (± 0.399)
	VI: MultivariateNormal	**0.690** (± 0.037)	2.155 (± 0.454)
	VI: Structured Normal	**0.686** (± 0.015)	3.019 (± 0.545)
	VI: IAF	**0.643** (± 0.027)	1.751 (± 0.422)
	ICL (ours)	**0.649** (± 0.023)	1.464 (± 0.151)
	MAP	0.626 (± 0.038)	2.377 (± 0.529)
	TabPFN	0.522 (± 0.037)	0.496 (± 0.047)

Table 2: Evaluating the predictive performance across 50 synthetic and 17 real-world datasets for scenarios 5 and 7 in terms of Root Mean Squared Error (RMSE). The best result among all fully Bayesian methods is marked in **bold**. For the fully Bayesian approaches, we use the posterior mean as a point estimate for the response. MAP denotes the predictive performance of the model with the maximum a posteriori estimate for the latents. TabPFN gives the best overall performance, which is expected since it is not limited to the GLM structure, follow by the MAP approach. On scenario 5 (gamma prior on the regression coefficients), the in-context learner performs significantly worse than all other methods, while this difference is less pronounced for scenario 7.

Scenario	**Model**	**RMSE Real-World** ($\downarrow$)	**RMSE Synthetic** ($\downarrow$)
Scenario 5	HMC	**0.699** ($\pm$ 0.022)	**0.490** ($\pm$ 0.036)
	Laplace Approximation	**0.699** ($\pm$ 0.022)	**0.491** ($\pm$ 0.036)
	VI: DiagonalNormal	**0.702** ($\pm$ 0.022)	**0.491** ($\pm$ 0.036)
	VI: MultivariateNormal	**0.698** ($\pm$ 0.021)	**0.491** ($\pm$ 0.036)
	VI: Structured Normal	1.507 ($\pm$ 0.089)	0.741 ($\pm$ 0.053)
	VI: IAF	**0.699** ($\pm$ 0.022)	**0.490** ($\pm$ 0.036)
	ICL (ours)	0.769 ($\pm$ 0.020)	0.701 ($\pm$ 0.049)
	MAP	0.658 ($\pm$ 0.022)	0.471 ($\pm$ 0.035)
	TabPFN	0.534 ($\pm$ 0.040)	0.442 ($\pm$ 0.035)
Scenario 7	HMC	**0.953** ($\pm$ 0.015)	**0.719** ($\pm$ 0.041)
	Laplace Approximation	**0.950** ($\pm$ 0.016)	**0.719** ($\pm$ 0.041)
	VI: DiagonalNormal	**0.954** ($\pm$ 0.015)	**0.718** ($\pm$ 0.041)
	VI: MultivariateNormal	**0.953** ($\pm$ 0.015)	**0.718** ($\pm$ 0.041)
	VI: Structured Normal	1.082 ($\pm$ 0.026)	1.028 ($\pm$ 0.118)
	VI: IAF	**0.954** ($\pm$ 0.014)	**0.720** ($\pm$ 0.041)
	ICL (ours)	1.019 ($\pm$ 0.017)	**0.765** ($\pm$ 0.041)
	MAP	0.945 ($\pm$ 0.017)	0.686 ($\pm$ 0.048)
	TabPFN	0.817 ($\pm$ 0.040)	0.654 ($\pm$ 0.039)

Table 3: Evaluating the predictive performance across 50 synthetic and 17 real-world datasets for scenarios 5 and 7 in terms of accuracy (Acc.). The best result among all fully Bayesian methods is marked in **bold**. For the fully Bayesian approaches, we use the posterior mean as a point estimate for the response. MAP denotes the predictive performance of the model with the maximum a posteriori estimate for the latents. Here, TabPFN has a substantially better performance than all other methods. The MAP approach performs on average better than all fully Bayesian methods, which themselves do not differ significantly.

Scenario	**Model**	**Acc. Real-World** ($\uparrow$)	**Acc. Synthetic** ($\uparrow$)
Scenario 6	HMC	**0.694** ($\pm$ 0.028)	**0.546** ($\pm$ 0.015)
	Laplace Approximation	**0.692** ($\pm$ 0.027)	**0.547** ($\pm$ 0.015)
	VI: DiagonalNormal	**0.700** ($\pm$ 0.028)	**0.546** ($\pm$ 0.015)
	VI: MultivariateNormal	**0.691** ($\pm$ 0.029)	**0.546** ($\pm$ 0.015)
	VI: Structured Normal	**0.686** ($\pm$ 0.028)	**0.546** ($\pm$ 0.015)
	VI: IAF	**0.689** ($\pm$ 0.029)	**0.545** ($\pm$ 0.015)
	ICL (ours)	**0.688** ($\pm$ 0.027)	**0.545** ($\pm$ 0.015)
	MAP	0.723 ($\pm$ 0.025)	0.610 ($\pm$ 0.016)
	TabPFN	0.862 ($\pm$ 0.021)	0.673 ($\pm$ 0.011)