A. Appendix

Below we detail the results on several datasets using different combinations of transformations and autoregressive conditional models. Each additive coupling transformation uses a fully connected network with two hidden layers of 256 units. RNN transformations use a hidden state with 16 units. SingleInd conditional models modeled each dimension's conditional as a standard Gaussian. MultiInd modeled each dimension's conditional as independent mixtures with 40 components (each with mean, scale, and weight parameter). RAM, LAM, and Tied conditional models each had a hidden state with 120 units that was fed through two fully connected layers each with 120 units to produce the parameters of the mixtures with 40 components. The RAM hidden state was produced by a GRU with 256 units. LAM and Tied hidden states came through a linear map as discussed above.

Table 2. Held out test log-likelihoods for the Markovian dataset. The superscripts denote rankings of log-likelihoods on the validation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	**LAM**	**RAM**	**TIED**	**MultiInd**	**SingleInd**
None	14.319	−29.950	−0.612	−41.472	− − −
L None	**15.486**$^{(9)}$	14.538	10.906	5.252	−9.426
RNN	14.777	−37.716	11.075	−30.491	−37.038
L RNN	**15.658**$^{(5)}$	10.354	10.910	5.370	3.310
2xRNN	14.683	13.698	11.493	−18.448	−34.268
L 2xRNN	**15.474**$^{(8)}$	**15.752**$^{(3)}$	12.316	5.385	3.739
4xAdd+Re	15.269	12.257	12.912	12.446	11.625
L 4xAdd+Re	**15.683**$^{(6)}$	12.594	13.845	12.768	12.069
4xSRNN+Re	14.829	14.381	11.798	11.738	12.932
L 4xSRNN+Re	15.289	**16.202**$^{(1)}$	12.748	**15.415**$^{(10)}$	13.908
RNN+4xAdd+Re	15.171	12.991	14.455	11.467	10.382
L RNN+4xAdd+Re	15.078	12.655	14.415	12.886	12.315
RNN+4xSRNN+Re	14.968	**16.216**$^{(2)}$	12.590	**15.589**$^{(4)}$	14.231
L RNN+4xSRNN+Re	15.429	**15.566**$^{(7)}$	14.179	14.528	13.961

Table 3. Held out test log-likelihoods for star 32d dataset. The superscript denotes ranking of log-likelihood on cross validation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	**LAM**	**RAM**	**TIED**	**MultiInd**	**SingleInd**
None	−2.041	2.554	−10.454	−29.485	− − −
L None	5.454	8.247	−7.858	−26.988	−38.952
RNN	−1.276	2.762	−6.292	−25.946	−41.275
L RNN	7.775	6.335	−1.157	−25.986	−34.408
2xRNN	3.705	8.032	−0.565	−25.100	−38.490
L 2xRNN	**14.878**$^{(3)}$	9.946	0.901	−23.772	−33.075
4xAdd+Re	**13.278**$^{(6)}$	**11.561**$^{(9)}$	7.146	−16.740	−21.332
L 4xAdd+Re	**15.728**$^{(2)}$	**12.444**$^{(7)}$	9.031	−6.091	−11.225
4xSRNN+Re	3.496	8.429	−1.380	−15.590	−23.712
L 4xSRNN+Re	**16.042**$^{(1)}$	**9.939**$^{(10)}$	5.598	−12.530	−16.889
RNN+4xAdd+Re	**14.071**$^{(5)}$	**14.123**$^{(4)}$	6.868	−14.773	−20.483
L RNN+4xAdd+Re	**11.819**$^{(8)}$	9.253	2.638	−7.662	−14.530
RNN+4xSRNN+Re	−0.679	3.320	−6.172	−12.879	−19.204
L RNN+4xSRNN+Re	7.433	7.324	3.554	−10.427	−15.243

Table 4. Held out test log-likelihood for Star 128d dataset.The superscript denotes ranking of log-likelihood on crossvalidation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd
None	15.671	15.895	−83.115	−128.238	− − −
L None	57.881	−82.100	−28.206	−123.939	−159.391
RNN	18.766	48.295	−22.485	−113.181	−178.641
L RNN	**66.070**$^{(9)}$	−49.084	31.136	−107.083	−155.324
2xRNN	27.295	45.834	−11.930	−113.210	−178.331
L 2xRNN	**85.681**$^{(3)}$	−84.524	30.974	−105.368	−162.635
4xAdd+Re	**77.195**$^{(6)}$	**61.947**$^{(10)}$	16.062	−75.206	−111.542
L 4xAdd+Re	**88.837**$^{(1)}$	−21.882	20.234	−65.694	−96.071
4xSRNN+Re	33.577	−98.796	3.256	−88.912	−98.936
L 4xSRNN+Re	**86.375**$^{(2)}$	**76.968**$^{(5)}$	33.481	−85.590	−93.086
RNN+4xAdd+Re	**66.540**$^{(8)}$	−57.861	−16.277	−75.491	−114.729
L RNN+4xAdd+Re	**80.063**$^{(4)}$	32.104	21.944	−71.933	−100.384
RNN+4xSRNN+Re	21.719	−87.335	−6.517	−76.459	−85.422
L RNN+4xSRNN+Re	**72.463**$^{(7)}$	56.201	26.269	−71.843	−91.695

Table 5. Average performance percentage score for each model across all datasets. Note that this measure is not over a logarithmic space.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd	MAX
None	0.218	0.118	0.006	0.000	0.000	0.218
L None	0.154	0.179	0.026	0.051	0.001	0.179
RNN	0.086	0.158	0.014	0.001	0.000	0.158
L RNN	0.173	**0.540**	0.014	0.040	0.013	0.540
2xRNN	0.151	0.101	0.045	0.001	0.000	0.151
L 2xRNN	0.118	0.330	0.015	0.045	0.025	0.330
4xAdd+Re	0.036	0.047	0.015	0.010	0.006	0.047
L 4xAdd+Re	0.153	0.096	0.025	0.014	0.009	0.153
4xSRNN+Re	0.086	0.051	0.031	0.010	0.008	0.086
L 4xSRNN+Re	0.109	0.143	0.023	0.021	0.018	0.143
RNN+4xAdd+Re	0.121	0.096	0.023	0.011	0.011	0.121
L RNN+4xAdd+Re	0.336	0.165	0.024	0.016	0.013	0.336
RNN+4xSRNN+Re	0.102	0.151	0.017	0.012	0.014	0.151
L RNN+4xSRNN+Re	0.211	0.288	0.024	0.018	0.016	0.288
MAX	0.336	0.540	0.045	0.051	0.025	

Table 6. Held out test log-likelihood for `forest` dataset.The superscript denotes ranking of log-likelihood on crossvalidation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd
None	0.751	−1.383	−0.653	−12.824	− − −
L None	1.910	1.834	−0.243	−7.665	−11.062
RNN	1.395	0.053	0.221	−5.130	−15.983
L RNN	**2.189**$^{(8)}$	1.747	−0.087	−4.001	−5.807
2xRNN	1.832	1.830	0.448	−6.162	−9.095
L 2xRNN	**2.240**$^{(6)}$	**2.432**$^{(3)}$	0.264	−3.956	−5.125
4xAdd+Re	1.106	1.430	0.420	−0.021	−0.492
L 4xAdd+Re	2.043	1.979	0.909	0.365	−0.088
4xSRNN+Re	1.178	1.428	0.187	−0.029	−0.212
L 4xSRNN+Re	**2.089**$^{(9)}$	**2.061**$^{(10)}$	0.611	0.754	0.593
RNN+4xAdd+Re	1.962	**2.226**$^{(7)}$	0.857	0.081	0.086
L RNN+4xAdd+Re	**2.389**$^{(4)}$	**2.672**$^{(1)}$	0.852	0.450	0.251
RNN+4xSRNN+Re	1.599	1.545	0.510	0.182	0.369
L RNN+4xSRNN+Re	**2.297**$^{(5)}$	**2.443**$^{(2)}$	0.804	0.600	0.480

Table 7. Held out test log-likelihood for `pendigits` dataset. The superscript denotes ranking of log-likelihood on crossvalidation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd
None	**6.923**$^{(1)}$	**3.911**$^{(8)}$	1.437	−14.138	− − −
L None	**4.104**$^{(9)}$	2.911	−2.872	−9.997	−15.617
RNN	**5.464**$^{(3)}$	3.273	−1.676	−10.144	−19.719
L RNN	**4.072**$^{(6)}$	1.398	−2.299	−10.840	−13.103
2xRNN	**6.376**$^{(5)}$	**3.896**$^{(7)}$	−4.002	−12.132	−16.576
L 2xRNN	2.987	0.871	−3.977	−10.890	−12.711
4xAdd+Re	−1.924	−3.087	−3.172	−5.010	−6.498
L 4xAdd+Re	−1.796	−1.438	−2.288	−4.951	−7.834
4xSRNN+Re	**5.854**$^{(2)}$	2.146	−2.827	−5.970	−7.084
L 4xSRNN+Re	3.758	−1.020	−3.370	−5.885	−12.978
RNN+4xAdd+Re	−2.357	−2.869	−2.187	−5.454	−8.053
L RNN+4xAdd+Re	−2.687	−2.103	−2.185	−4.742	−6.941
RNN+4xSRNN+Re	**5.207**$^{(4)}$	2.425	−2.126	−5.147	−8.859
L RNN+4xSRNN+Re	**3.466**$^{(10)}$	0.496	−2.761	−7.205	−13.897

Table 8. Held out test log-likelihood for `susy` dataset.The superscript denotes ranking of log-likelihood on crossvalidation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd
None	9.736	−14.821	−5.721	−21.369	− − −
L None	15.731	**16.930**$^{(8)}$	6.410	−8.846	−17.130
RNN	12.784	3.347	6.114	−18.575	−44.273
L RNN	16.381	**18.389**$^{(2)}$	6.772	−5.744	−11.489
2xRNN	11.052	14.362	3.595	−16.478	−33.126
L 2xRNN	14.523	**17.373**$^{(7)}$	10.687	−6.884	−10.420
4xAdd+Re	9.835	8.033	7.238	6.031	4.245
L 4xAdd+Re	**17.673**$^{(3)}$	**16.500**$^{(10)}$	11.613	10.941	9.034
4xSRNN+Re	8.798	13.235	1.234	6.936	3.378
L 4xSRNN+Re	14.242	**17.870**$^{(5)}$	15.397	12.161	13.413
RNN+4xAdd+Re	15.408	12.480	9.409	7.619	5.446
L RNN+4xAdd+Re	**17.474**$^{(6)}$	16.376	13.765	10.951	8.269
RNN+4xSRNN+Re	14.066	**17.691**$^{(4)}$	9.136	10.088	7.656
L RNN+4xSRNN+Re	**16.627**$^{(9)}$	**18.941**$^{(1)}$	13.469	12.105	12.349

Table 9. Held out test log-likelihood for `higgs` dataset.The superscript denotes ranking of log-likelihood on crossvalidation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd
None	−6.220	−5.848	−13.883	−25.793	− − −
L None	**−3.798**$^{(8)}$	−10.651	−9.084	−16.025	−36.051
RNN	−5.800	**−2.600**$^{(3)}$	−10.797	−25.760	−66.223
L RNN	**−3.975**$^{(9)}$	**−0.340**$^{(1)}$	−8.574	−18.607	−32.753
2xRNN	−6.456	−4.833	−9.192	−25.398	−60.040
L 2xRNN	−5.866	**−3.222**$^{(5)}$	−8.216	−16.083	−30.730
4xAdd+Re	−6.502	−10.491	−9.356	−13.678	−15.138
L 4xAdd+Re	−5.377	−5.611	−8.006	−12.106	−14.129
4xSRNN+Re	−7.422	−6.863	−11.033	−11.878	−12.182
L 4xSRNN+Re	−5.999	−9.329	−8.474	−8.223	−8.926
RNN+4xAdd+Re	**−4.242**$^{(10)}$	−4.804	−9.187	−12.321	−15.261
L RNN+4xAdd+Re	**−3.396**$^{(6)}$	**−3.049**$^{(4)}$	−8.052	−12.246	−13.765
RNN+4xSRNN+Re	−5.262	**−2.116**$^{(2)}$	−10.105	−12.307	−9.388
L RNN+4xSRNN+Re	**−3.756**$^{(7)}$	−4.773	−8.097	−9.378	−7.721

Table 10. Held out test log-likelihood for `hepmass` dataset.The superscript denotes ranking of log-likelihood on crossvalidation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd
None	2.328	**3.710**$^{(6)}$	−4.948	−19.771	− − −
L None	**3.570**$^{(7)}$	2.517	−4.052	−9.266	−35.042
RNN	2.088	**4.935**$^{(1)}$	−1.639	−19.851	−47.686
L RNN	**2.869**$^{(10)}$	**5.047**$^{(2)}$	−2.920	−16.032	−30.210
2xRNN	1.774	0.902	−1.909	−15.440	−36.754
L 2xRNN	2.053	**3.680**$^{(5)}$	−2.150	−15.457	−24.079
4xAdd+Re	1.678	1.873	−4.046	−9.117	−11.387
L 4xAdd+Re	1.961	2.543	−2.259	−6.907	−9.275
4xSRNN+Re	1.443	2.156	−2.904	−6.091	−7.186
L 4xSRNN+Re	2.072	2.730	−3.014	−5.747	−6.245
RNN+4xAdd+Re	2.817	0.912	−2.514	−6.003	−9.284
L RNN+4xAdd+Re	**3.906**$^{(3)}$	−1.869	−3.847	−6.339	−9.103
RNN+4xSRNN+Re	2.663	**3.586**$^{(8)}$	−0.863	−7.146	−3.939
L RNN+4xSRNN+Re	**3.759**$^{(4)}$	**3.487**$^{(9)}$	−0.239	−7.522	−6.102

Table 11. Held out test log-likelihood for `satimage2` dataset.The superscript denotes ranking of log-likelihood on crossvalidation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd
None	**-1.716**$^{(9)}$	**-1.257**$^{(3)}$	−9.296	−50.507	− − −
L None	−20.164	**-1.079**$^{(4)}$	−2.635	**-1.570**$^{(5)}$	−5.972
RNN	−7.728	−4.949	−5.466	−6.047	−16.521
L RNN	−31.296	**-0.773**$^{(2)}$	−3.944	**-1.824**$^{(8)}$	−2.977
2xRNN	−12.283	**-2.193**$^{(7)}$	−2.137	−5.447	−8.075
L 2xRNN	−20.968	**-0.550**$^{(1)}$	−5.140	**-1.699**$^{(6)}$	**-2.276**$^{(10)}$
4xAdd+Re	−19.931	−7.539	−11.826	−18.901	−17.977
L 4xAdd+Re	−21.128	−9.944	−12.336	−21.677	−24.070
4xSRNN+Re	−7.519	−11.368	−2.549	−7.730	−7.232
L 4xSRNN+Re	−18.170	−7.709	−5.533	−17.085	−15.347
RNN+4xAdd+Re	−19.278	−11.789	−12.837	−21.249	−22.786
L RNN+4xAdd+Re	−20.899	−12.949	−12.867	−26.164	−28.302
RNN+4xSRNN+Re	−13.476	−3.951	−6.284	−15.025	−16.443
L RNN+4xSRNN+Re	−20.179	−12.128	−7.258	−18.065	−18.125

Table 12. Held out test log-likelihood for `music` dataset.The superscript denotes ranking of log-likelihood on crossvalidation dataset. Note that NADE is TIED conditional with None Transform and NICE is Add+Re Transformation with SingleInd Conditional. In parenthesis is the top-10 picks using valiation set.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd
None	−57.873	−97.925	−98.047	−113.099	− − −
L None	**−52.954**$^{(4)}$	−74.220	−72.441	−82.866	−104.287
RNN	**−54.933**$^{(10)}$	−80.436	−74.361	−106.219	−144.735
L RNN	**−52.710**$^{(3)}$	−59.815	−66.536	−82.731	−98.813
2xRNN	−56.958	−85.359	−77.456	−104.440	−133.898
L 2xRNN	**−53.956**$^{(8)}$	−57.611	−65.016	−82.678	−96.542
4xAdd+Re	−56.349	−69.302	−67.064	−73.886	−83.524
L 4xAdd+Re	**−53.169**$^{(5)}$	−59.282	−59.093	−69.887	−79.330
4xSRNN+Re	−57.670	−68.116	−74.006	−78.032	−121.197
L 4xSRNN+Re	**−53.879**$^{(7)}$	−55.665	−63.894	−77.564	−81.188
RNN+4xAdd+Re	**−53.177**$^{(6)}$	−67.377	−63.372	−73.882	−84.032
L RNN+4xAdd+Re	**−51.572**$^{(1)}$	−56.190	−58.885	−69.484	−79.555
RNN+4xSRNN+Re	**−54.065**$^{(9)}$	−61.204	−76.437	−71.814	−81.087
L RNN+4xSRNN+Re	**−52.617**$^{(2)}$	−68.756	−65.061	−83.292	−78.997

Table 13. Held out test log-likelihood for `wordvecs` dataset.The superscript denotes ranking of log-likelihood on validation dataset. Due to time constraints only models with linear transformations were trained.

Transformation	LAM	RAM	TIED	MultiInd	SingleInd
L None	**−252.659**$^{(6)}$	−279.788	−278.789	−332.474	−387.341
L RNN	**−252.894**$^{(7)}$	−278.795	−278.663	−332.689	−386.700
L 2xRNN	**−250.285**$^{(4)}$	−275.508	−277.848	−333.234	−386.649
L 4xAdd+Re	**−247.440**$^{(1)}$	**−272.371**$^{(8)}$	−274.205	−331.148	−374.563
L 4xSRNN+Re	**−248.393**$^{(2)}$	−300.666	**−273.372**$^{(9)}$	−308.735	0.000
L RNN+4xAdd+Re	**−249.980**$^{(3)}$	−280.938	**−273.976**$^{(10)}$	−331.316	−380.031
L RNN+4xSRNN+Re	**−251.468**$^{(5)}$	−280.325	−274.082	−308.148	−395.084