Supplemental 4

Feature Similarity

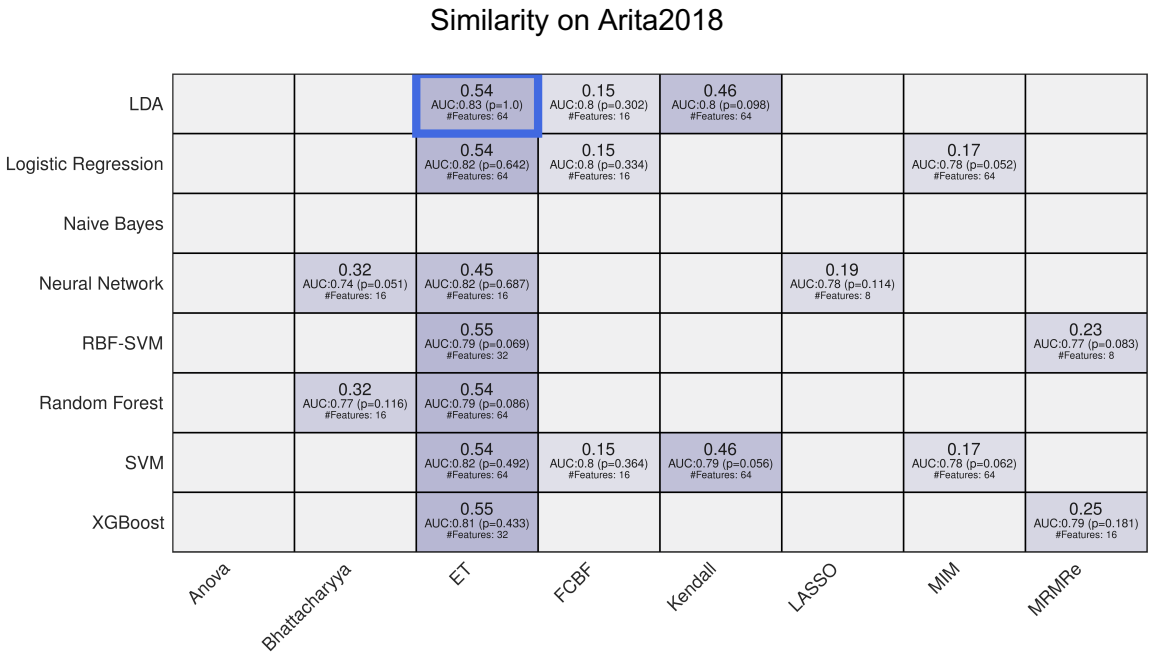

Figure S30: Similarity of the models on dataset Arita2018. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

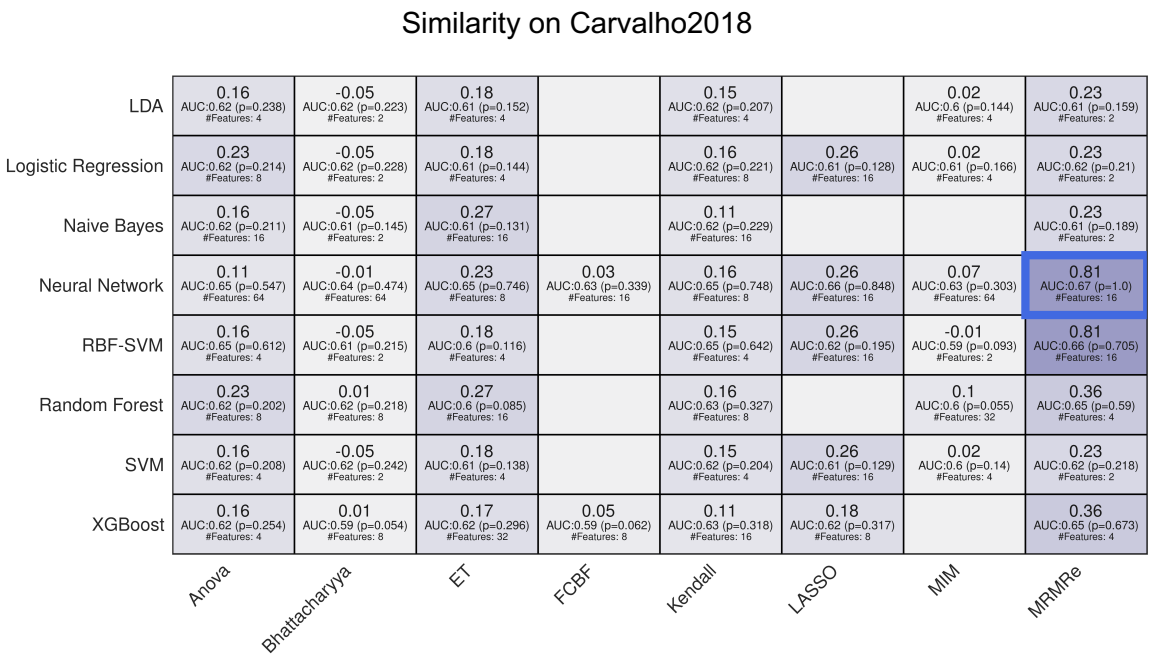

Figure S31: Similarity of the models on dataset Carvalho2018. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

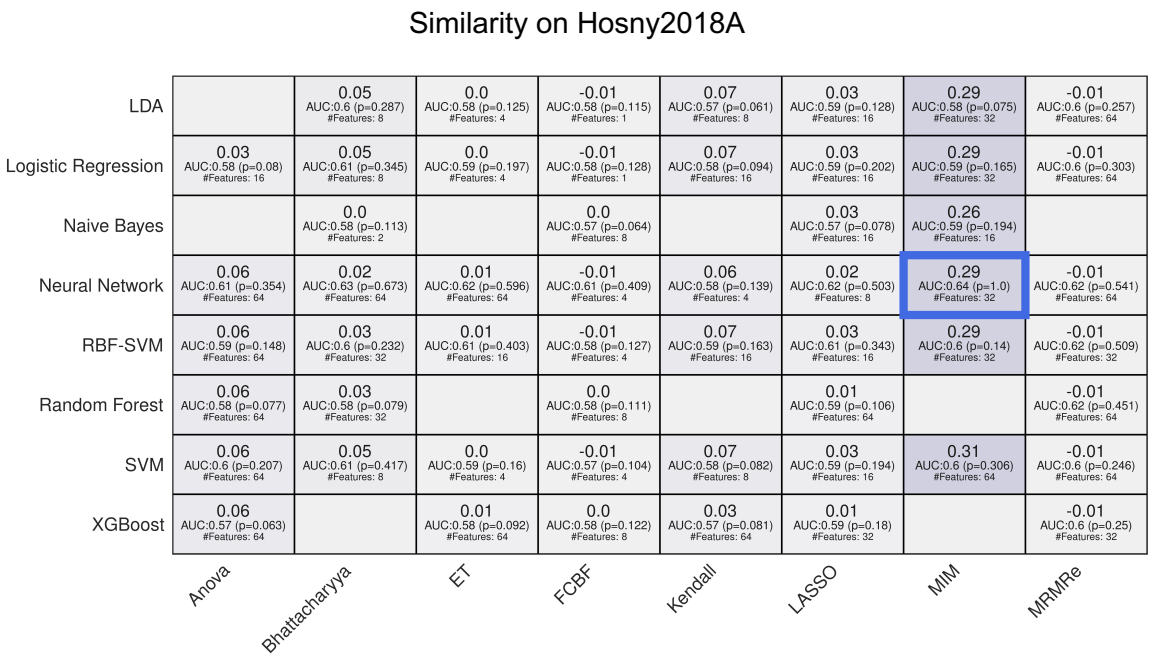

Figure S32: Similarity of the models on dataset Hosny2018A. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

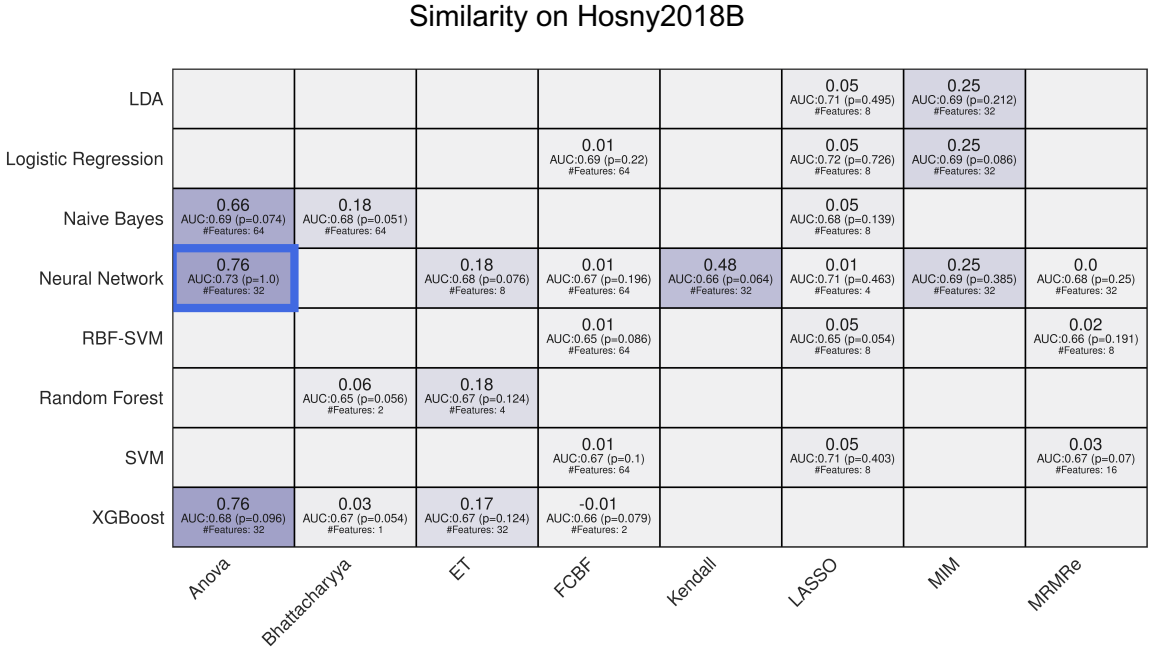

Figure S33: Similarity of the models on dataset Hosny2018B. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

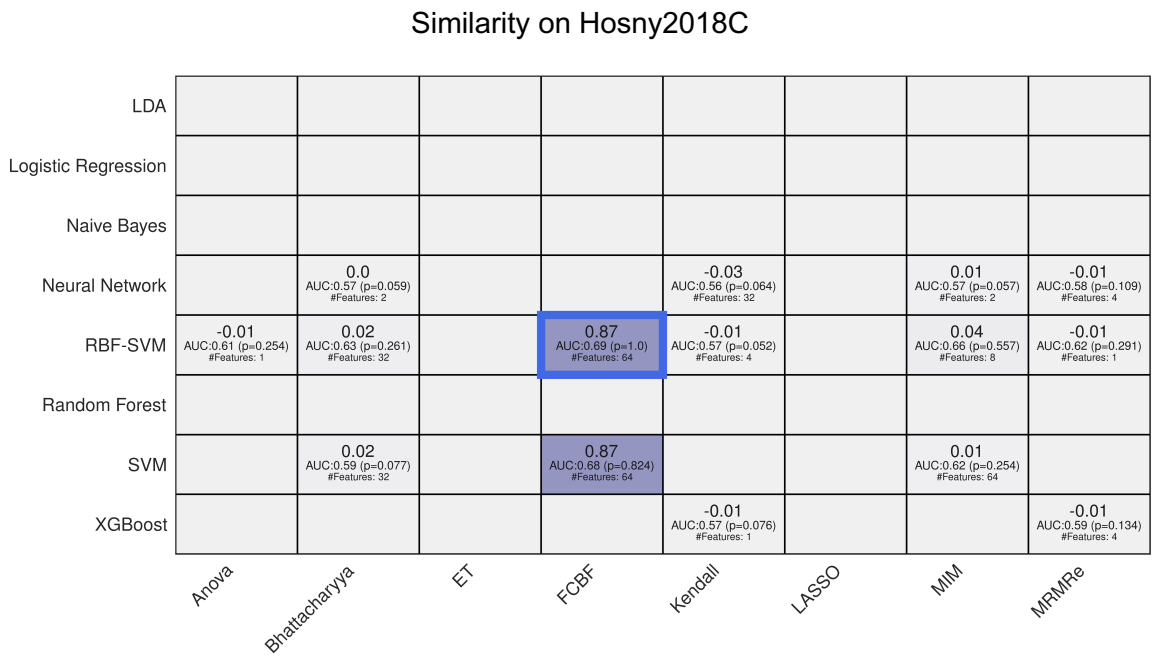

Figure S34: Similarity of the models on dataset Hosny2018C. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

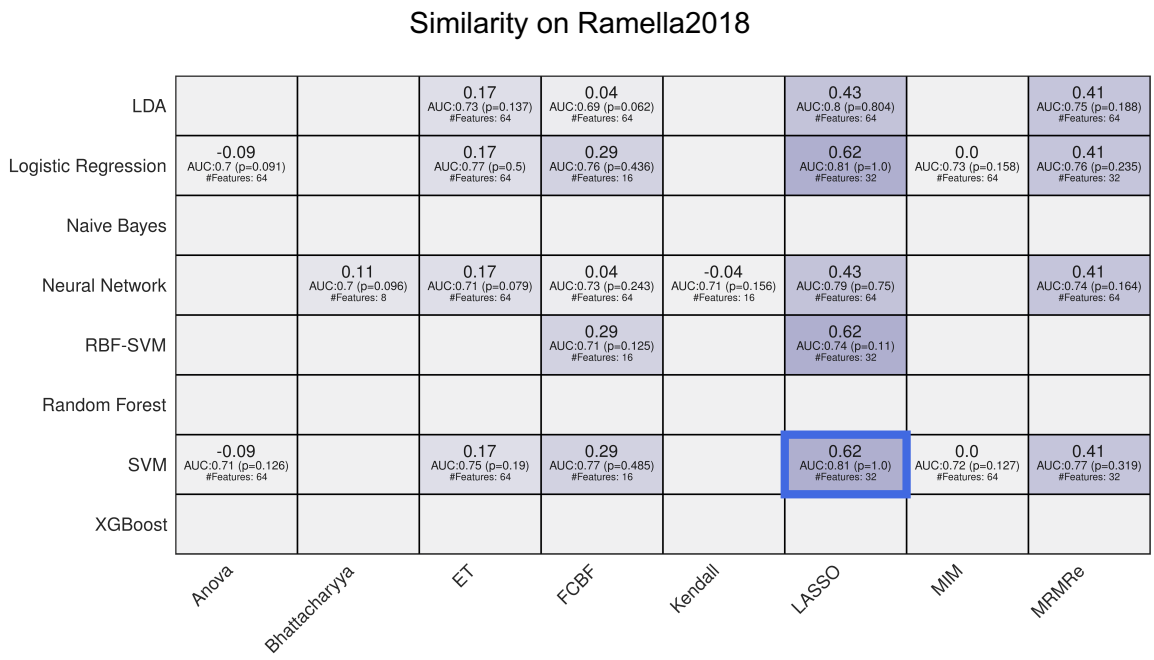

Figure S35: Similarity of the models on dataset Ramella2018. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

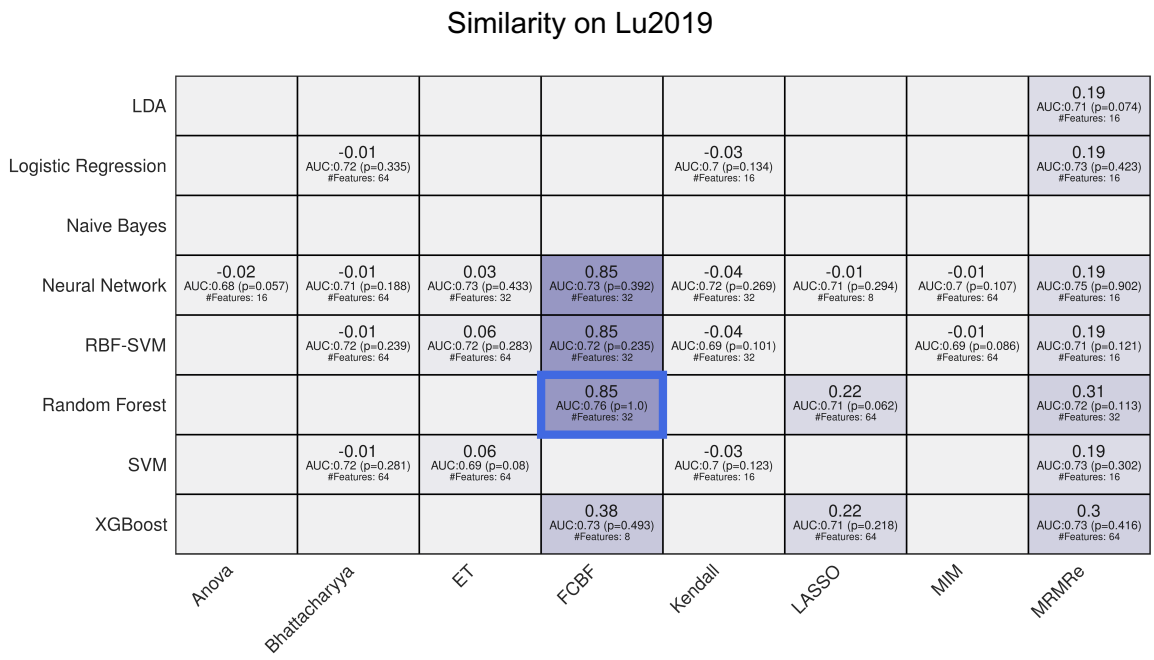

Figure S36: Similarity of the models on dataset Lu2019. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

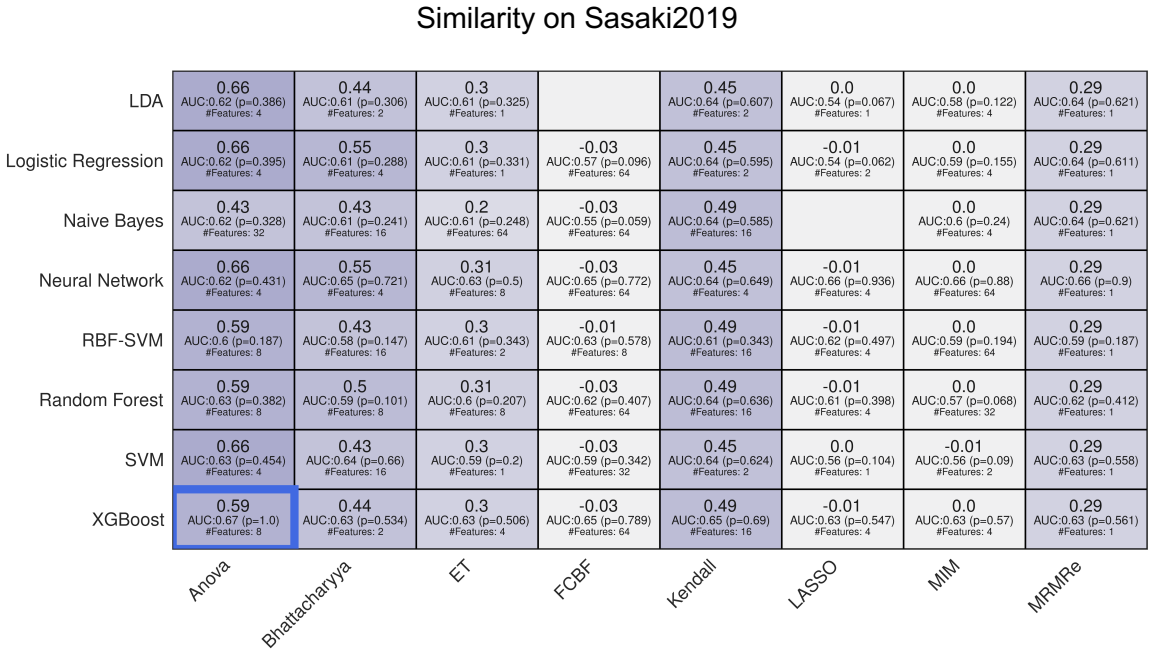

Figure S37: Similarity of the models on dataset Sasaki2019. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

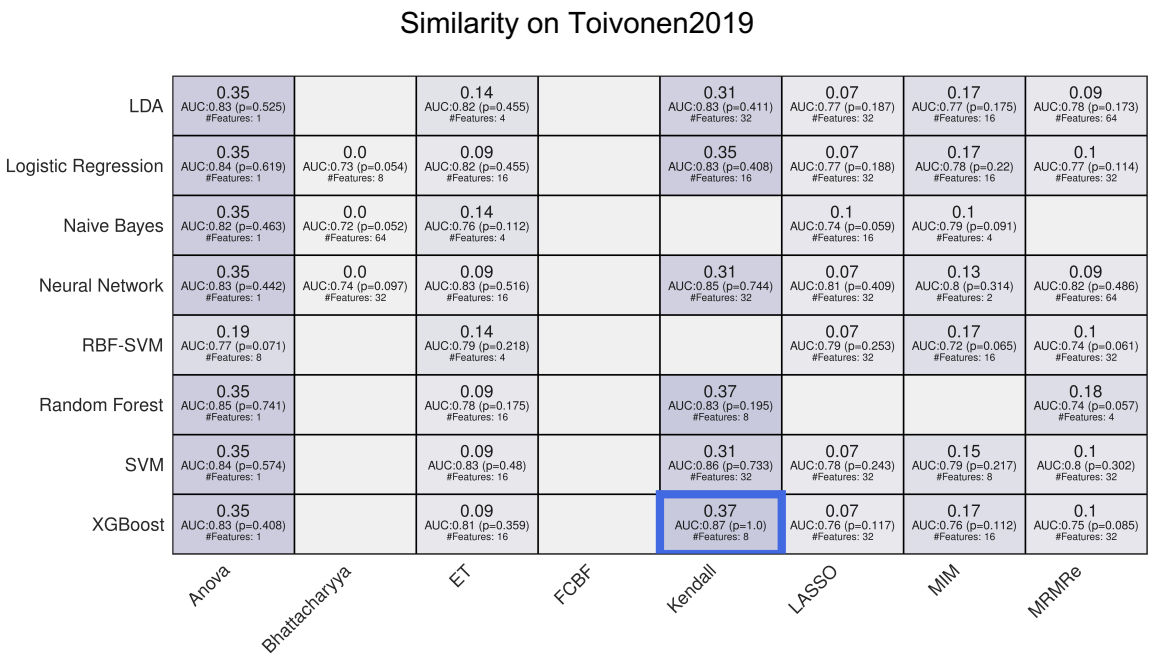

Figure S38: Similarity of the models on dataset Toivonen2019. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

Similarity on Keek2020

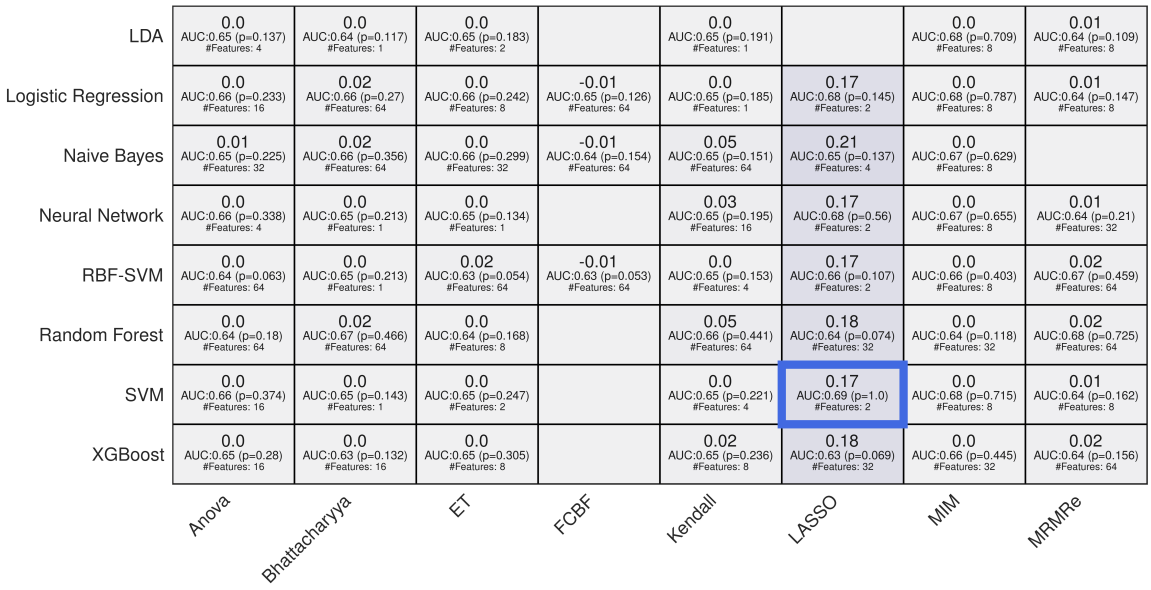

Figure S39: Similarity of the models on dataset Keek2020. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

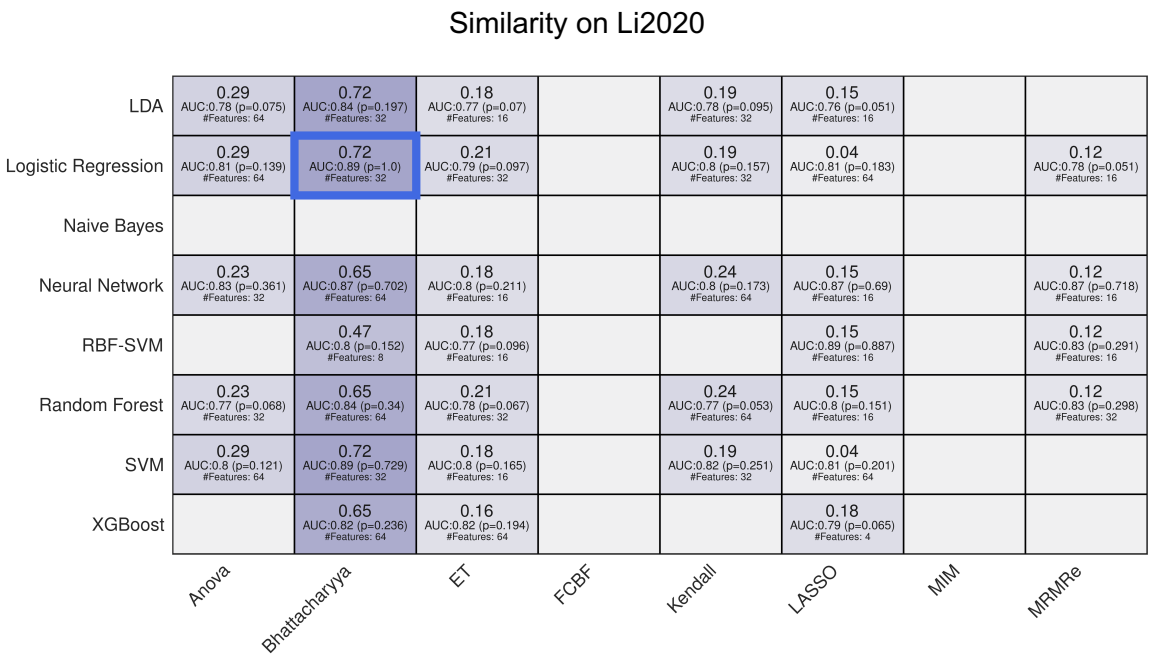

Figure S40: Similarity of the models on dataset Li2020. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

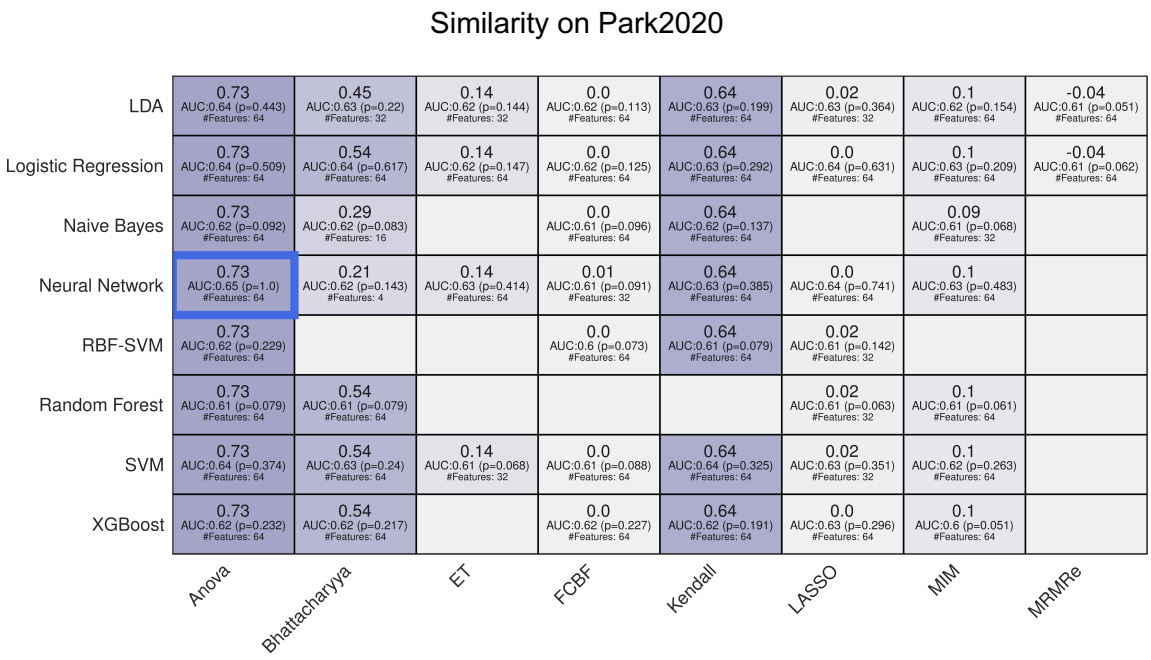

Figure S41: Similarity of the models on dataset Park2020. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

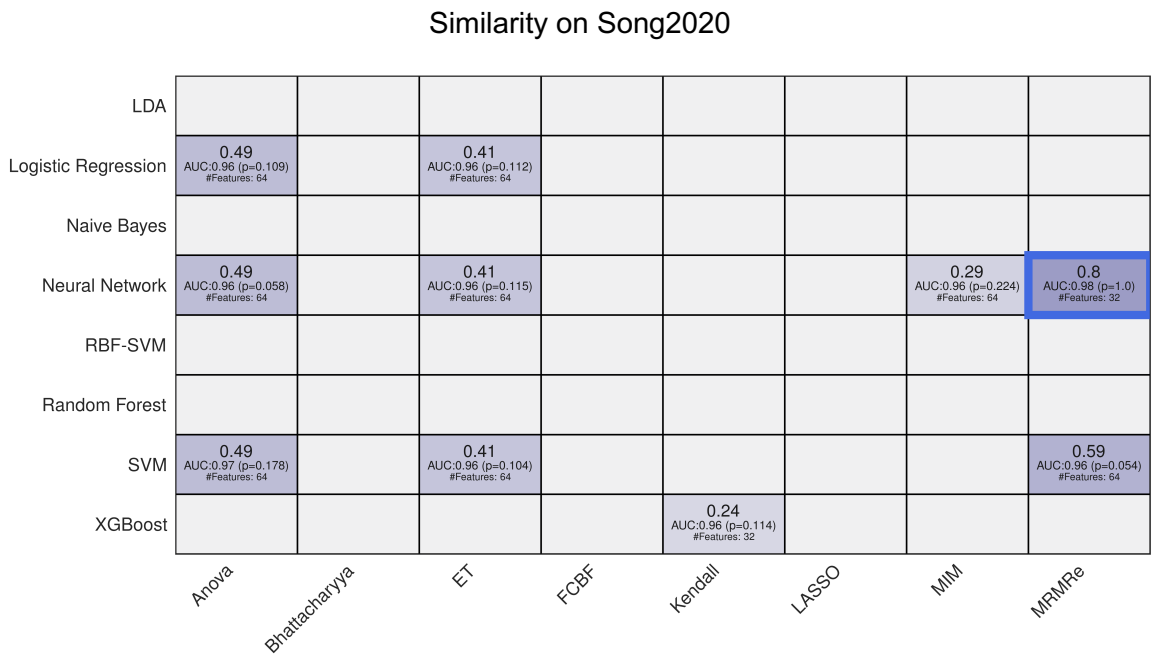

Figure S42: Similarity of the models on dataset Song2020. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.

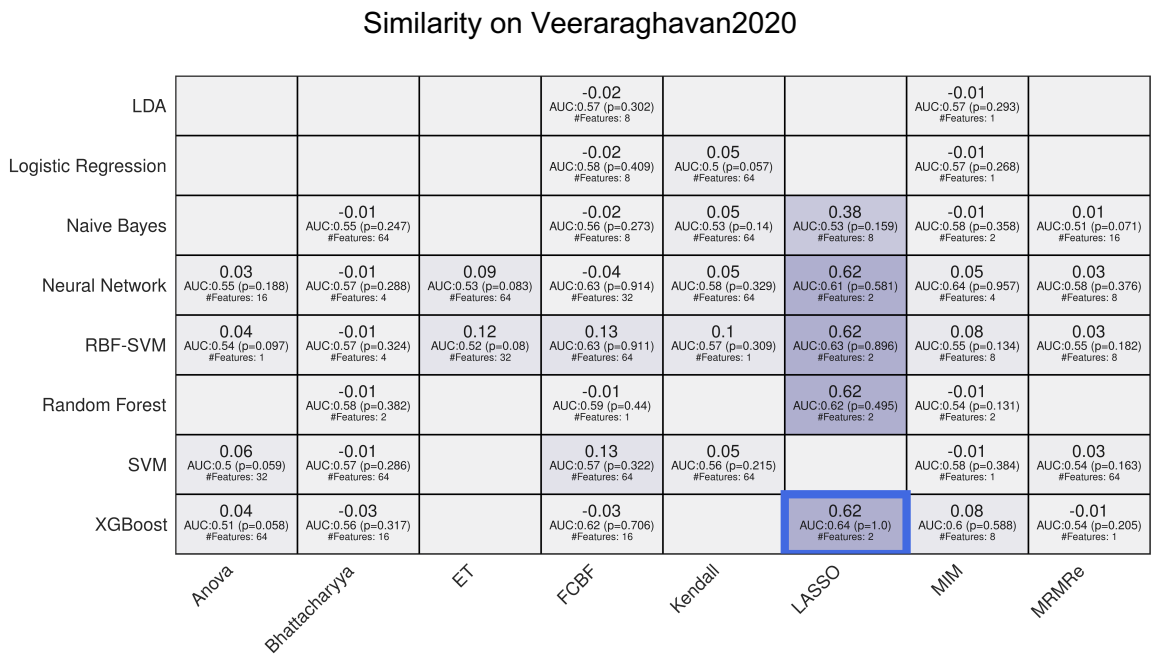

Figure S43: Similarity of the models on dataset Veeraraghavan2020. The best model is framed with a blue border, models that were significantly different to the best model are not shown. Statistical significance was tested using a DeLong test.