UNIVERSITY OF COPENHAGEN

A Physicist's Approach To Machine Learning

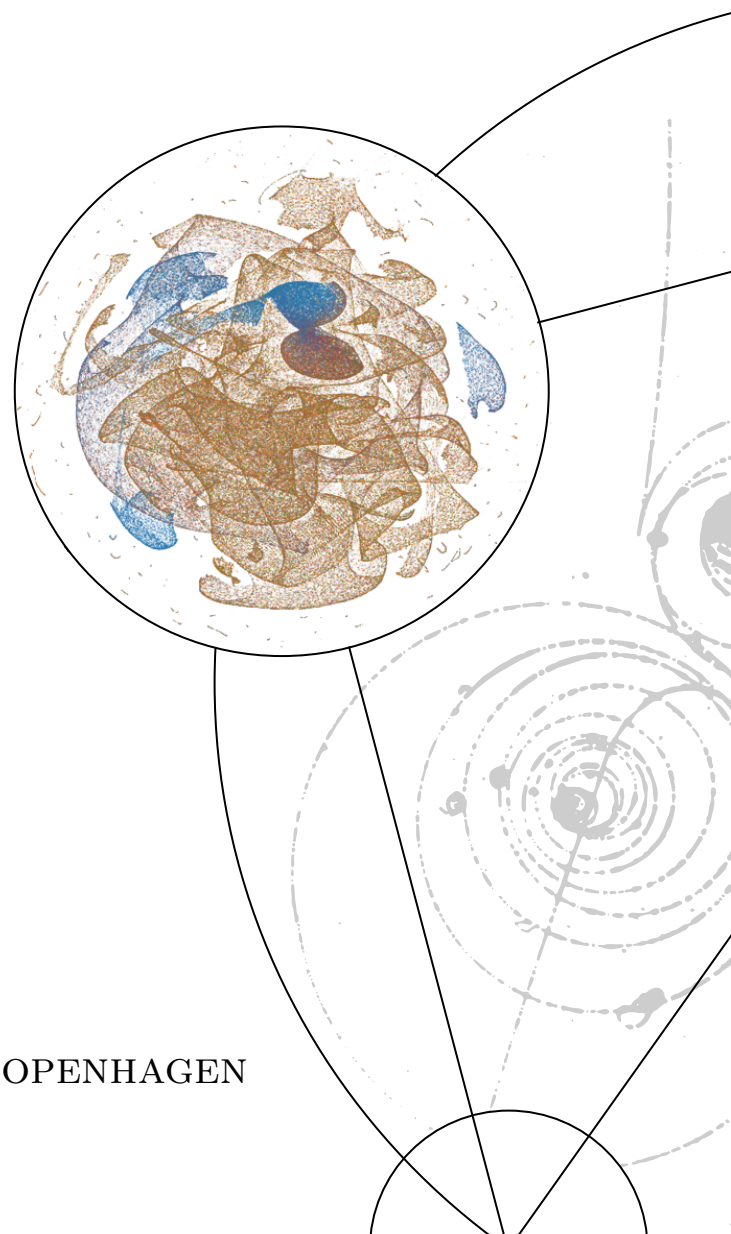
Understanding The Basic Bricks

Christian Michelsen

Master's Thesis (Cand. Scient.)

January 3rd, 2020

Supervised by
Troels Petersen



University of Copenhagen

https://github.com/christianmichelsen

This thesis was inspired by the works of Edward R. Tufte and is based on the Tufte-LaTeX package.

First printing, January 2020

Abstract

Here will be a decent abstract at some point^{TM}.

Contents

Part I

Part I of this thesis covers the introductory theory of machine learning in chapter 2 along with some extra technical aspects of it. In chapter 3 machine learning is applied to estimate Danish housing prices as precisely and accurately as possible.

Part II

Part II of this thesis deals with particle physics and the discriminatory power of machine learning for quark-gluon identification and subsequent analysis. In chapter 4 the theory of the Standard Model is introduced together with a description of the ALEPH detector. Theory is applied in chapter 5 where the types of jets and events in each collision is analysed using machine learning to improve the understanding of how gluon jets hadronizes and splits: simply said how they look and behave.

A. Housing Prices Appendix

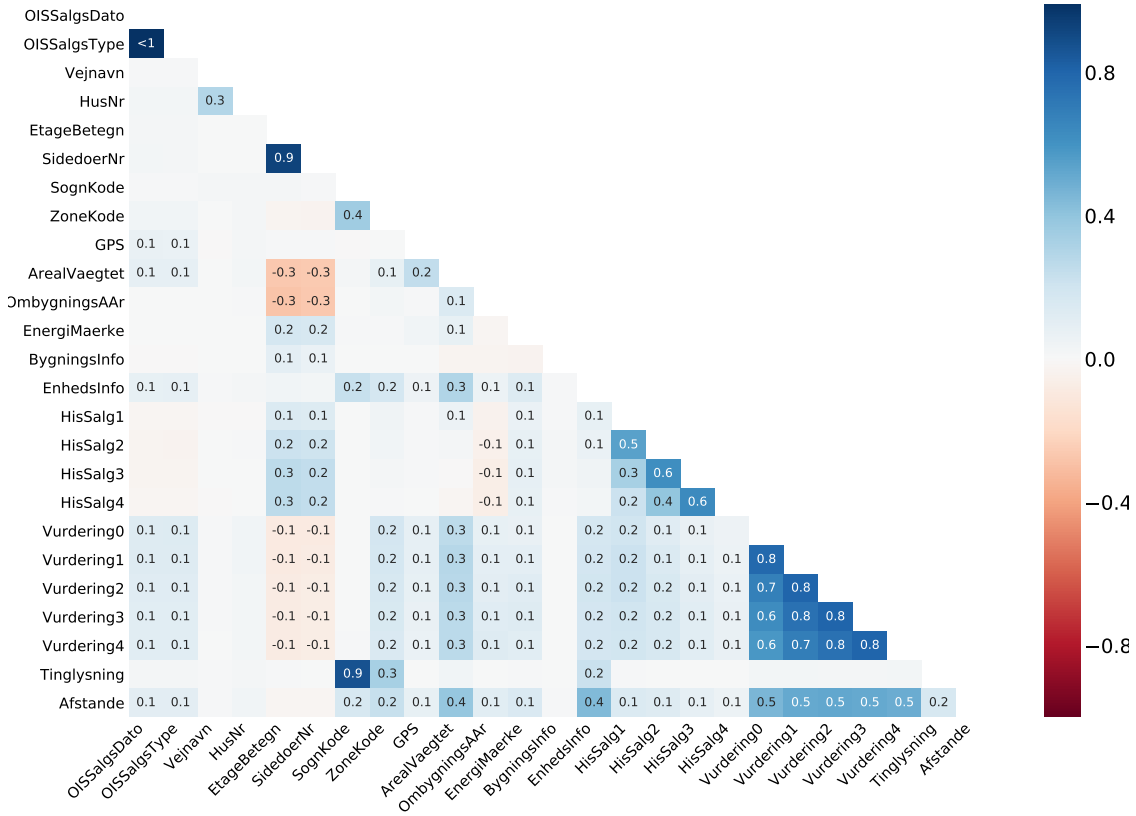


Figure A.1: XXXX **TODO!**.

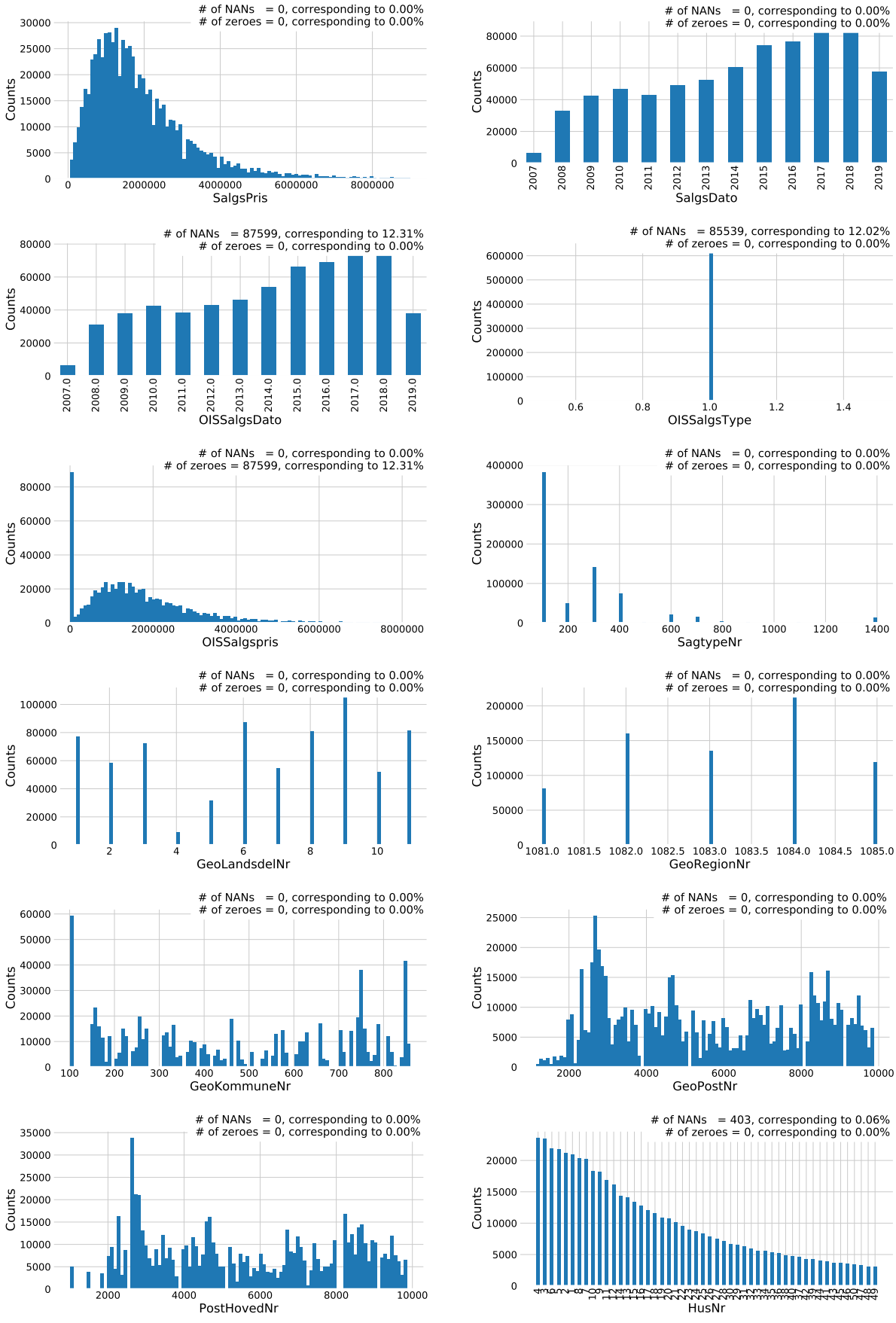


Figure A.2: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

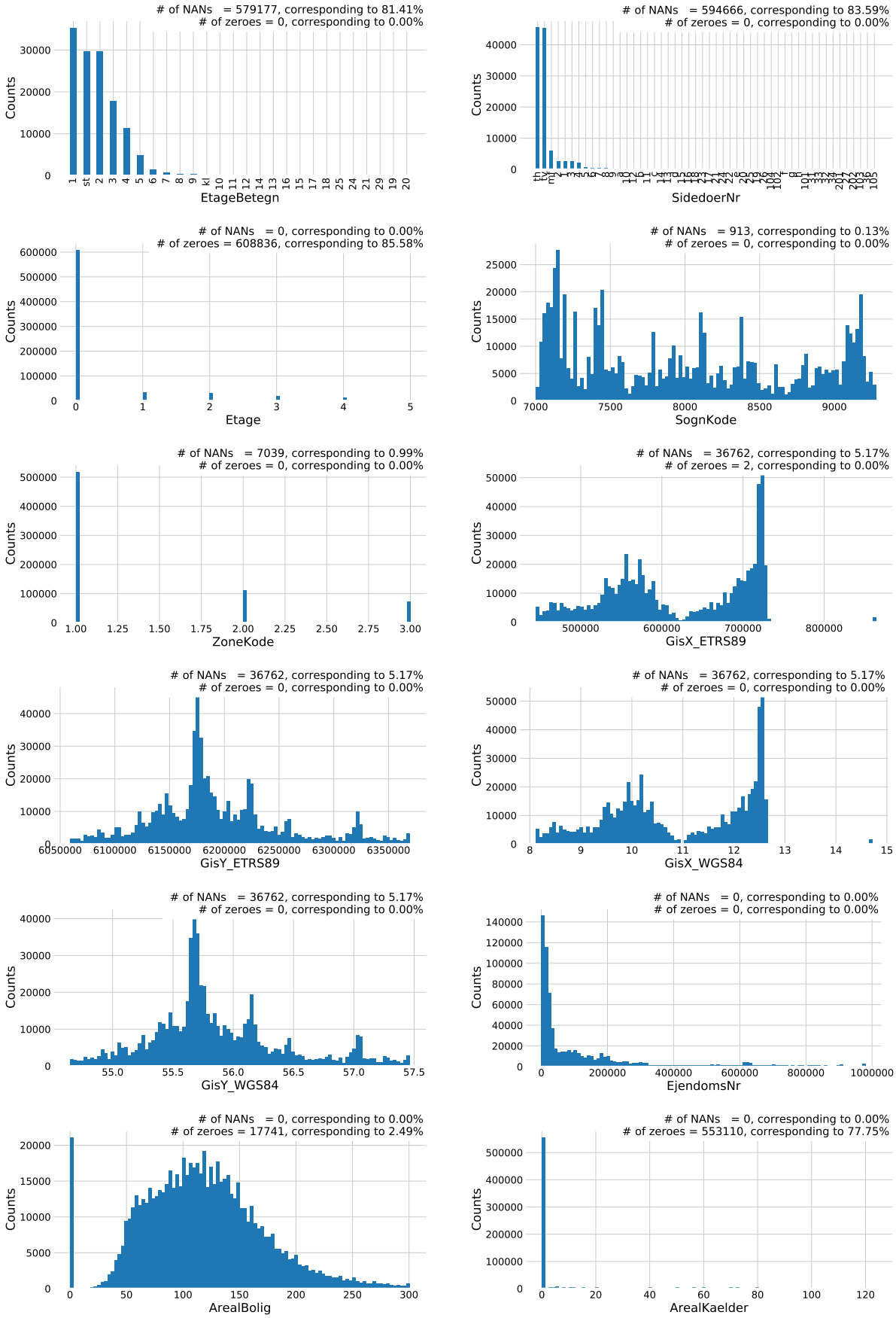


Figure A.3: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

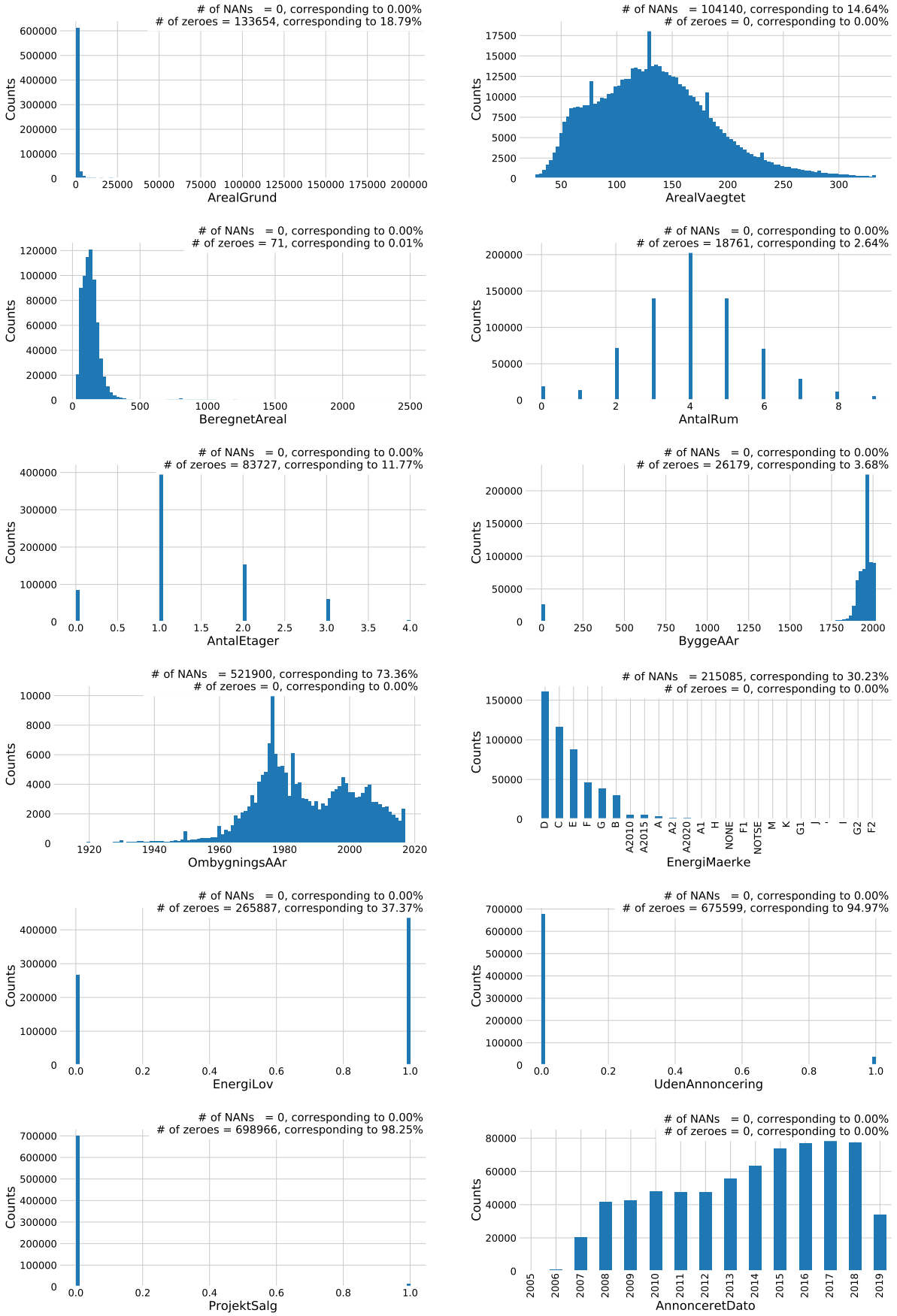


Figure A.4: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

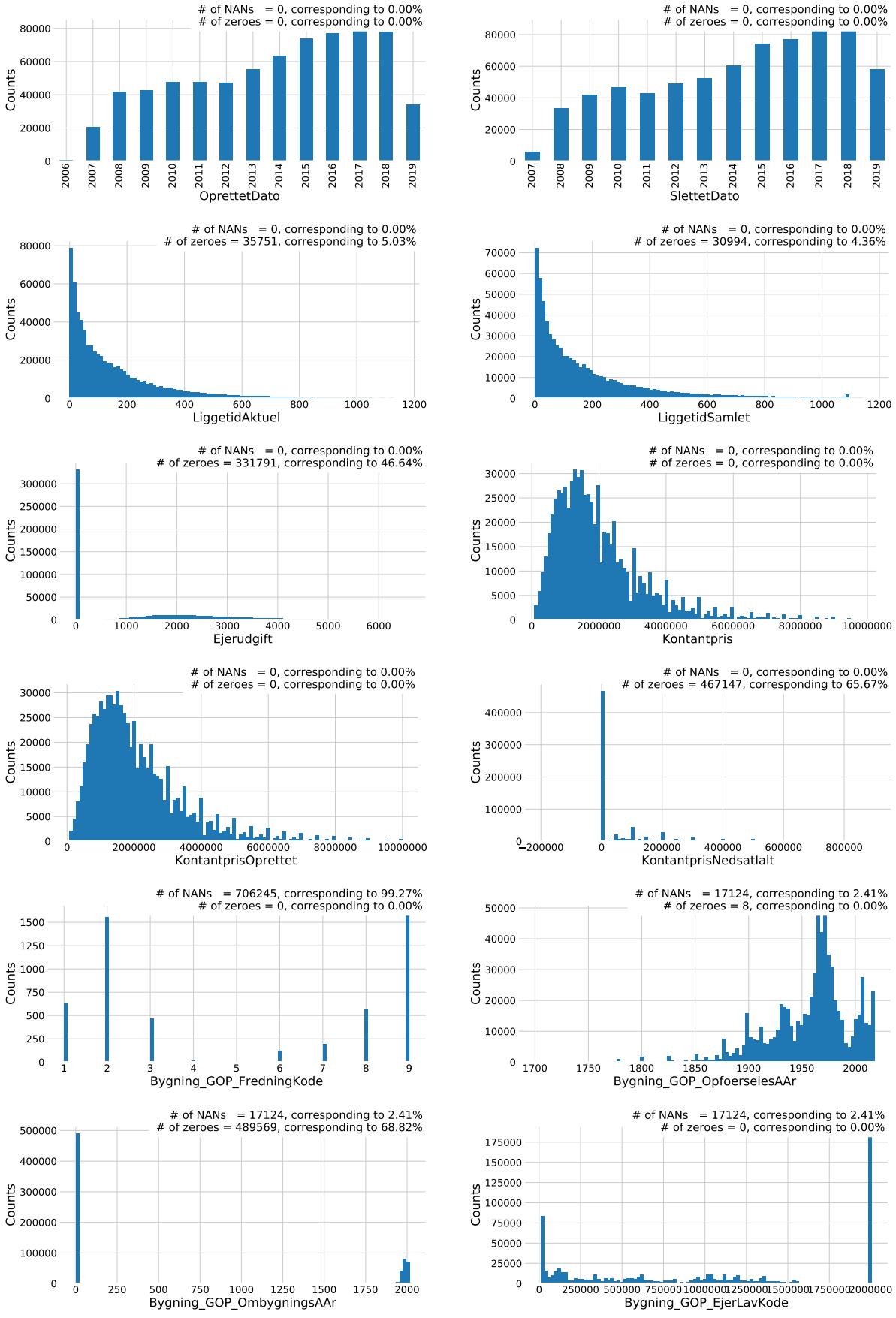


Figure A.5: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

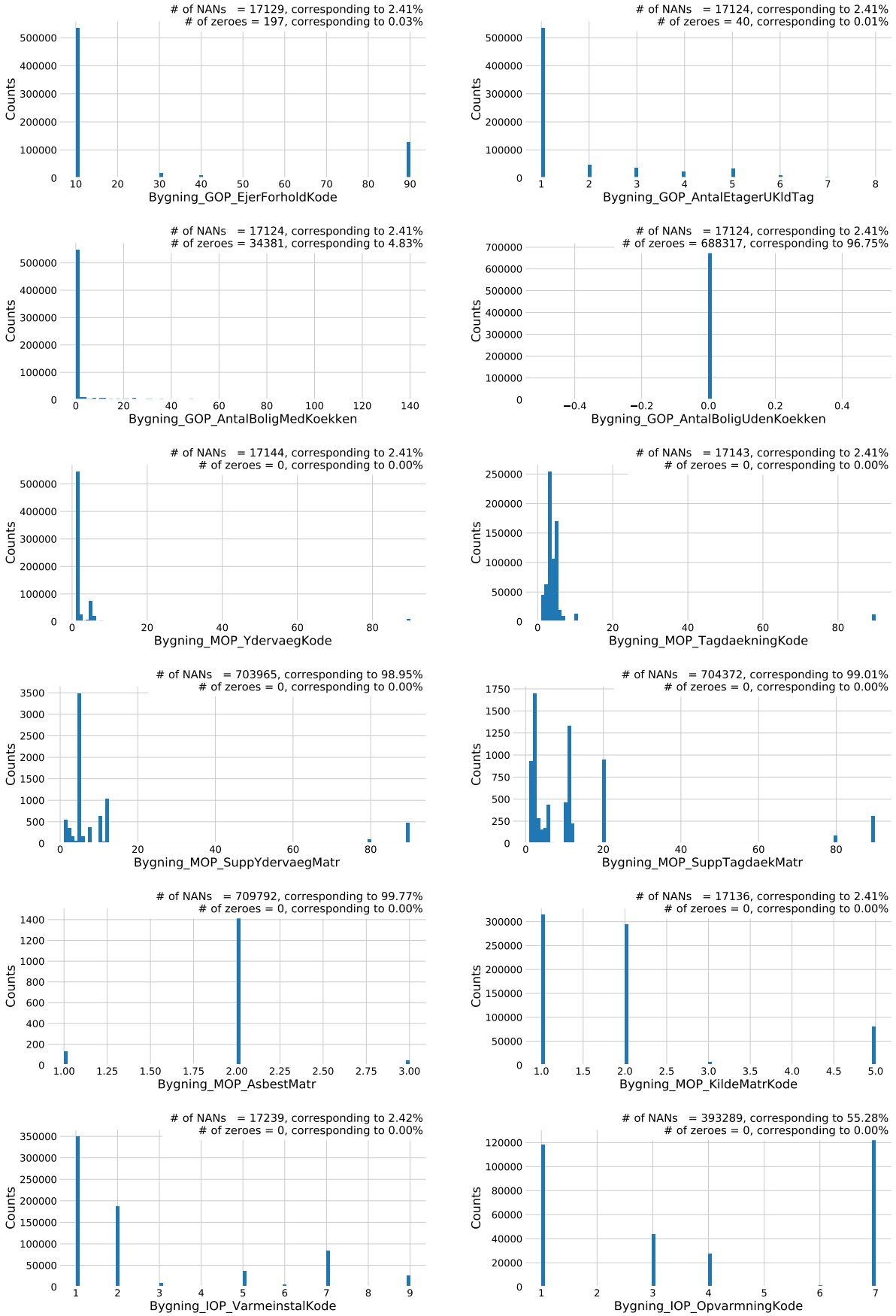


Figure A.6: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

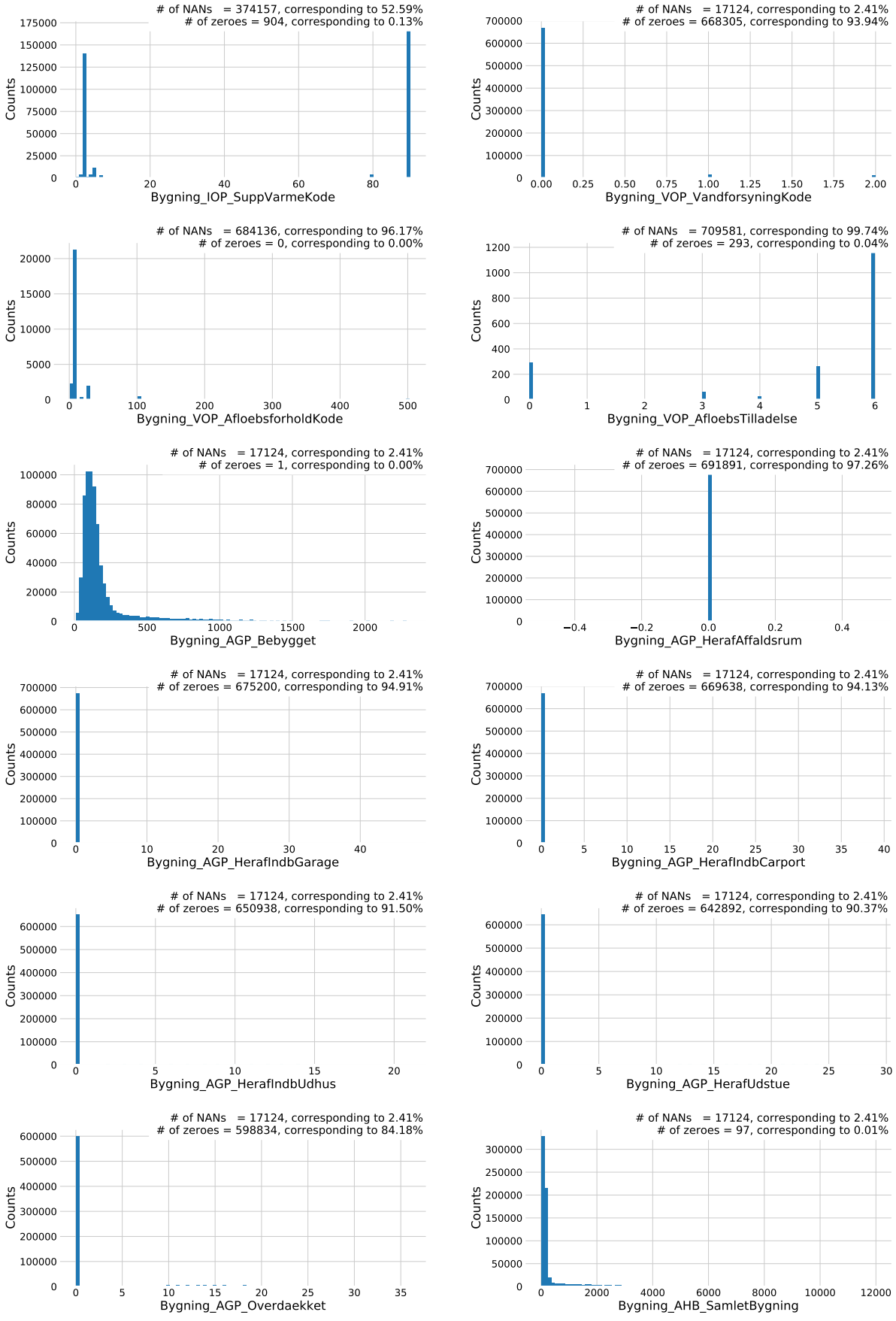


Figure A.7: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

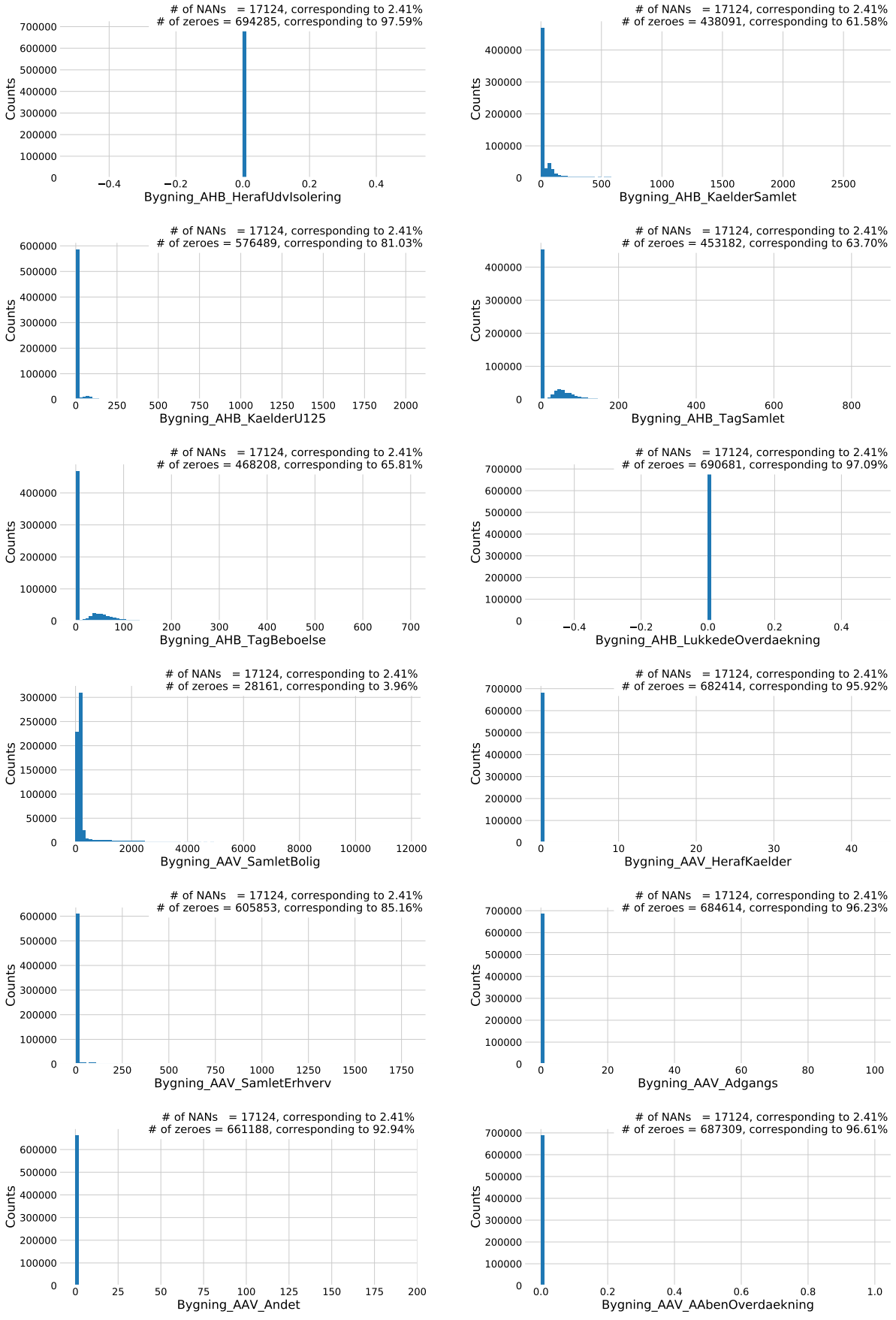


Figure A.8: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

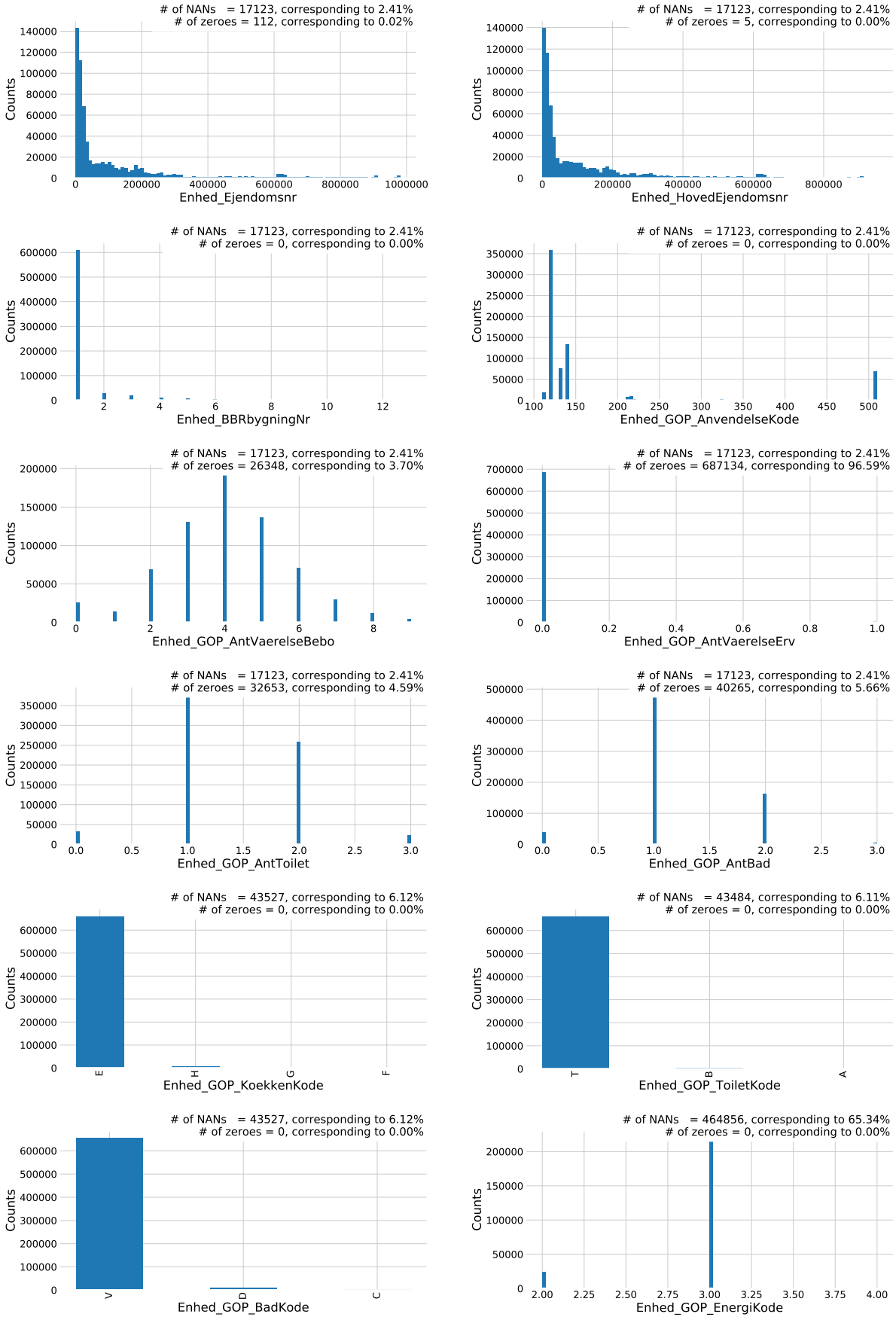


Figure A.9: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

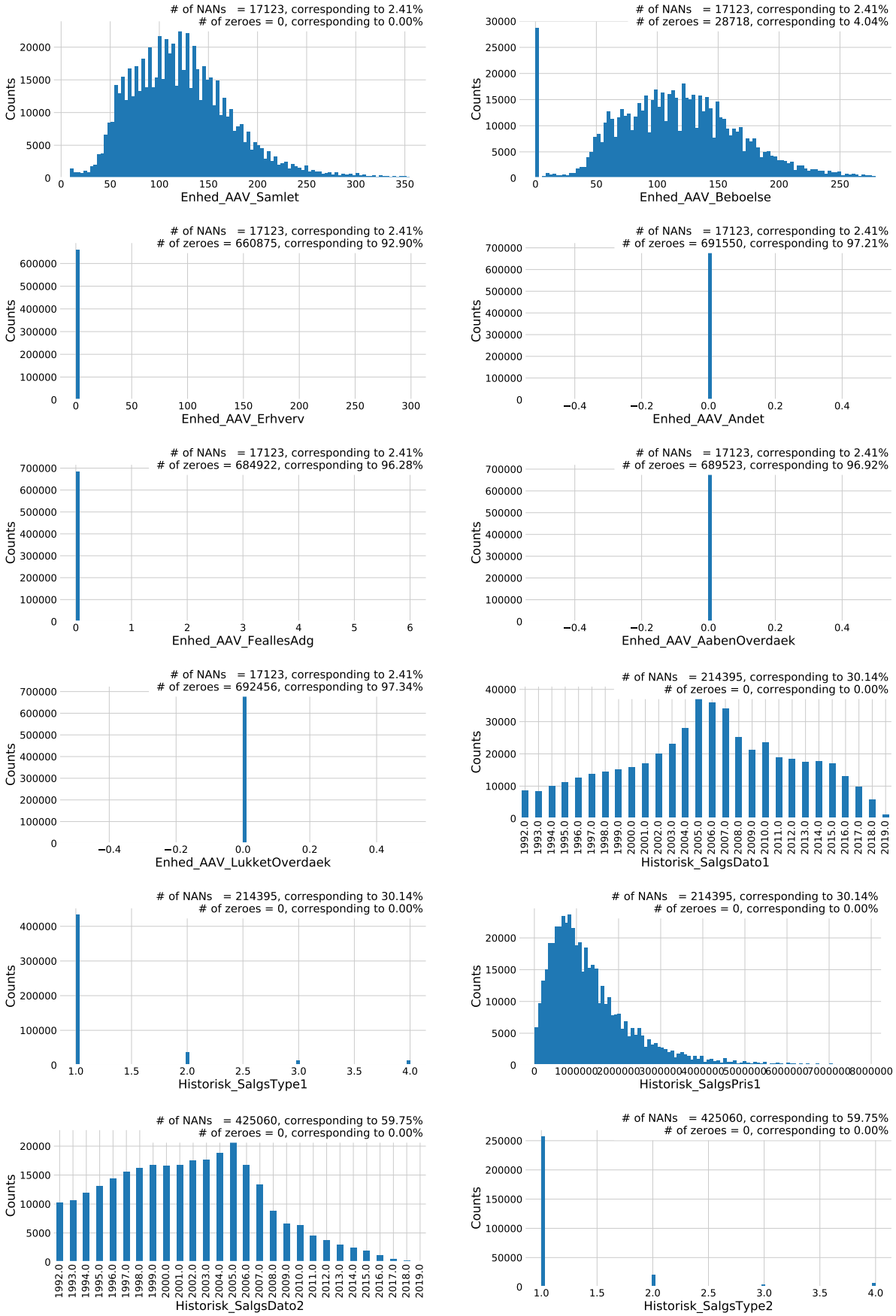


Figure A.10: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

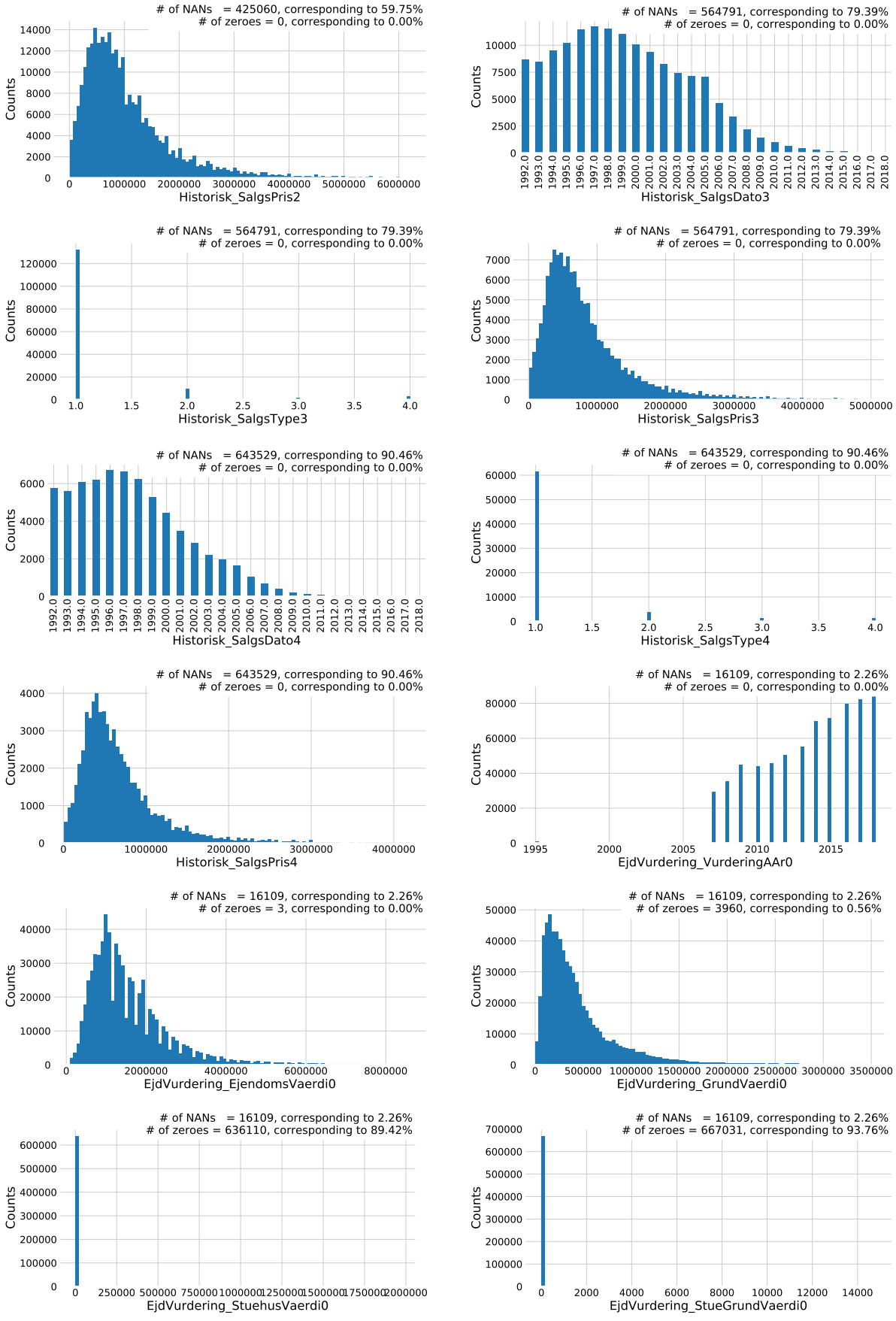


Figure A.11: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

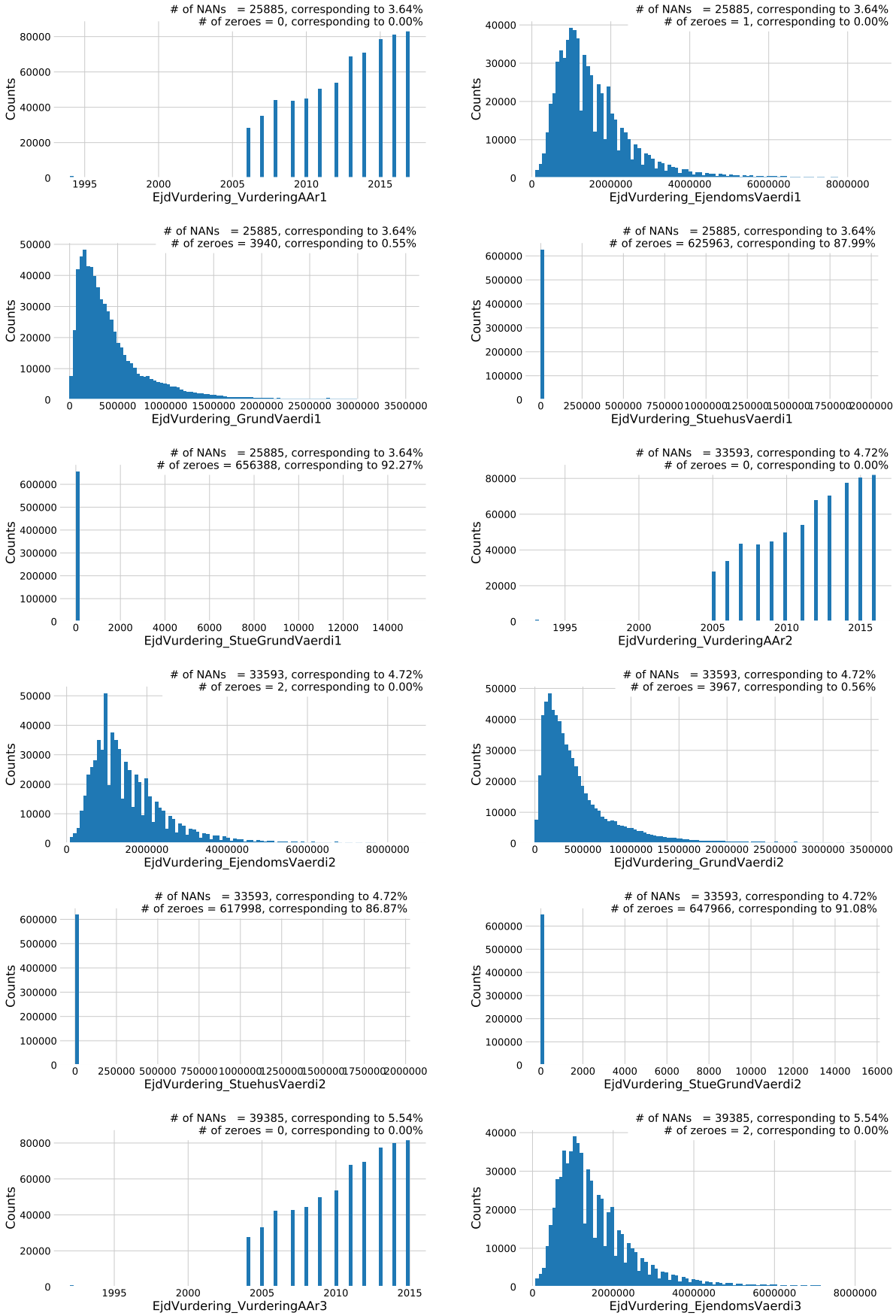


Figure A.12: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

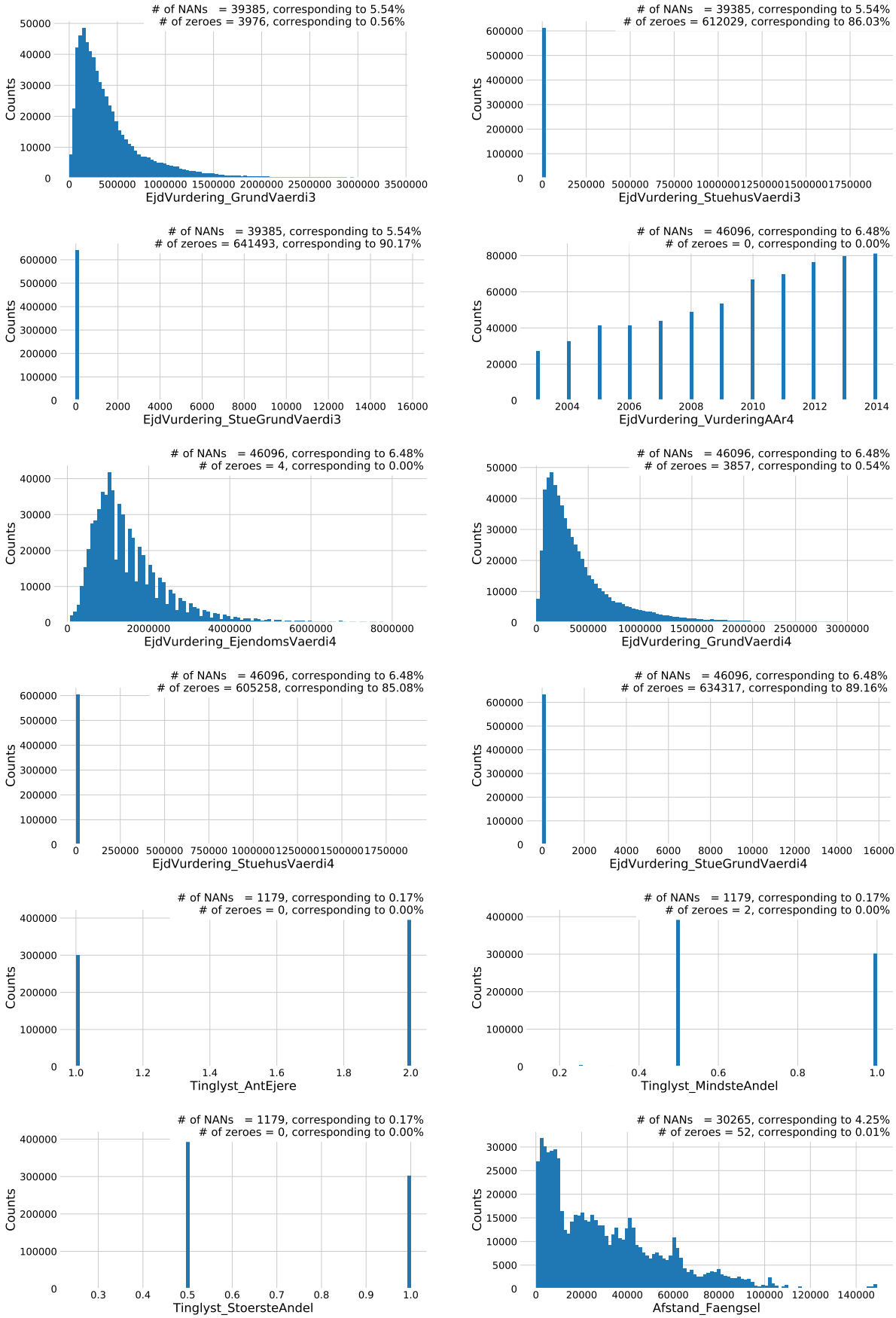


Figure A.13: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

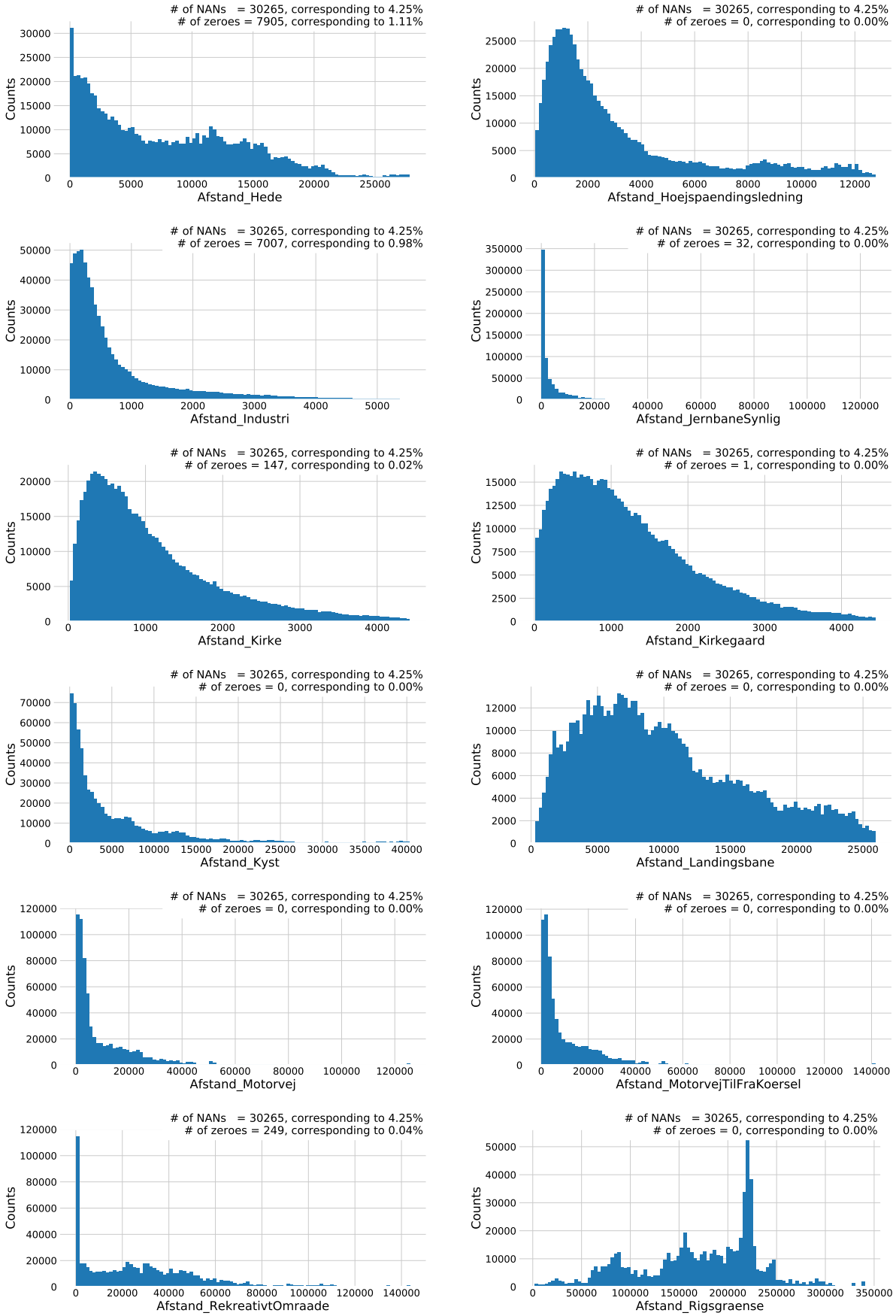


Figure A.14: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

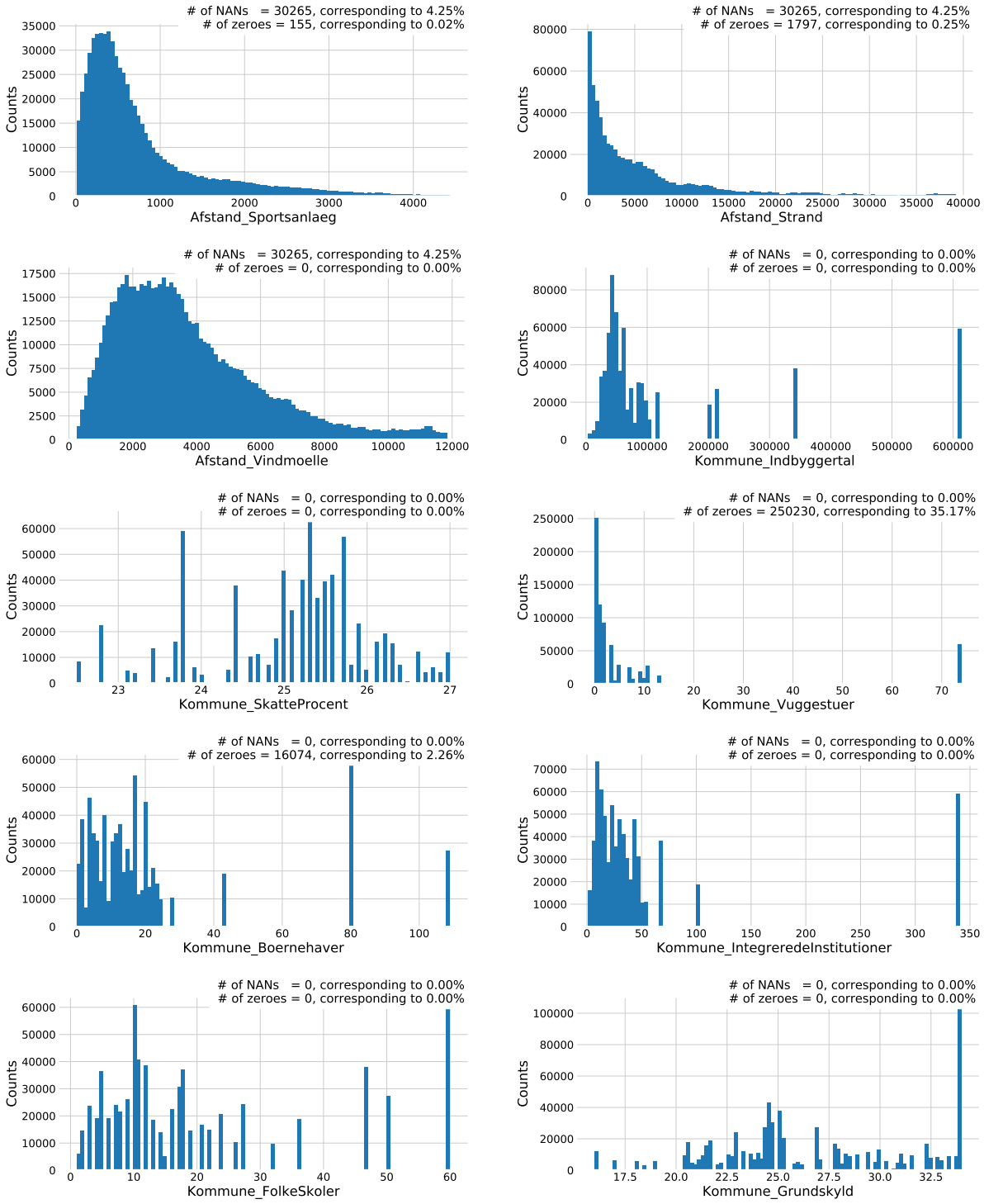


Figure A.15: Distributions the 168 input variables (excluding `ID` and `Vejnavn`).

OISSalgsType	GeoLandsdelNr	GeoRegionNr
GeoKommuneNr	GeoPostNr	PostHovedNr
Etage	SognKode	ZoneKode
GisX_WGS84	GisY_WGS84	EjendomsNr
ArealBolig	ArealKaelder	ArealGrund
BeregnetAreal	AntalRum	AntalEtager
ByggeAAr	OmbygningsAAr	EnergiLov
UdenAnnoncering	ProjektSalg	LiggetidAktuel
LiggetidSamlet	Ejerudgift	Bygning_GOP_OpfoerselesAAr
Bygning_GOP_OmbygningsAAr	Bygning_GOP_EjerLavKode	Bygning_GOP_EjerForholdKode
Bygning_GOP_AntalEtagerUKldTag	Bygning_GOP_AntalBoligMedKoekken	Bygning_GOP_AntalBoligUdenKoekken
Bygning_MOP_YdervaegKode	Bygning_MOP_TagdaekningKode	Bygning_MOP_KildeMatrKode
Bygning_IOP_VarmeinstalKode	Bygning_IOP_OpvarmningKode	Bygning_IOP_SuppVarmeKode
Bygning_VOP_VandforsyningKode	Bygning_AGP_Bebygget	Bygning_AGP_HerafAffaldsrum
Bygning_AGP_HerafIndbGarage	Bygning_AGP_HerafIndbCarport	Bygning_AGP_HerafIndbUdhus
Bygning_AGP_HerafUdstue	Bygning_AGP_Overdaekket	Bygning_AHB_SamletBygning
Bygning_AHB_HerafUdvIsolering	Bygning_AHB_KaelderSamlet	Bygning_AHB_KaelderU125
Bygning_AHB_TagSamlet	Bygning_AHB_TagBeboelse	Bygning_AHB_LukkedeOverdaekning
Bygning_AAV_SamletBolig	Bygning_AAV_HerafKaelder	Bygning_AAV_Adgangs
Bygning_AAV_Andet	Bygning_AAV_AAbenOverdaekning	Enhed_Ejendomsnr
Enhed_GOP_AnvendelseKode	Enhed_GOP_AntVaerelseErv	Enhed_GOP_AntToilet
Enhed_GOP_AntBad	Enhed_GOP_EnergiKode	Enhed_AAV_Erhverv
Enhed_AAV_Andet	Enhed_AAV_FeallesAdg	Enhed_AAV_AabenOverdaek
Enhed_AAV_LukketOverdaek	Historisk_SalgsType1	Historisk_SalgsPris1
Historisk_SalgsType2	Historisk_SalgsPris2	Historisk_SalgsType3
Historisk_SalgsPris3	EjdVurdering_VurderingAAr0	EjdVurdering_EjendomsVaerdi0
EjdVurdering_GrundVaerdi0	EjdVurdering_StuehusVaerdi0	EjdVurdering_StueGrundVaerdi0
EjdVurdering_VurderingAAr1	EjdVurdering_EjendomsVaerdi1	EjdVurdering_GrundVaerdi1
EjdVurdering_StuehusVaerdi1	EjdVurdering_StueGrundVaerdi1	EjdVurdering_VurderingAAr2
EjdVurdering_EjendomsVaerdi2	EjdVurdering_GrundVaerdi2	EjdVurdering_StuehusVaerdi2
EjdVurdering_StueGrundVaerdi2	EjdVurdering_VurderingAAr3	EjdVurdering_EjendomsVaerdi3
EjdVurdering_GrundVaerdi3	EjdVurdering_StuehusVaerdi3	EjdVurdering_StueGrundVaerdi3
EjdVurdering_VurderingAAr4	EjdVurdering_EjendomsVaerdi4	EjdVurdering_GrundVaerdi4
EjdVurdering_StuehusVaerdi4	EjdVurdering_StueGrundVaerdi4	Tinglyst_AntEjere
Tinglyst_MindsteAndel	Tinglyst_StoersteAndel	Afstand_Faengsel
Afstand_Hede	Afstand_Hoejspaendingsledning	Afstand_Industri
Afstand_JernbaneSynlig	Afstand_Kirke	Afstand_Kirkegaard
Afstand_Kyst	Afstand_Landingsbane	Afstand_Motorvej
Afstand_MotorvejTilFraKoersel	Afstand_RekreativtOmraade	Afstand_Rigsgraense
Afstand_Sportsanlaeg	Afstand_Strand	Afstand_Vindmoelle
Kommune_Indbyggertal	Kommune_SkatteProcent	Kommune_Vuggestuer
Kommune_Boernehaver	Kommune_IntegreredeInstitutioner	Kommune_FolkeSkoler
Kommune_Grundskyld	dag_i_maaned	maaned
aar	SalgsDato_siden0	Historisk_SalgsDato1_siden0
Historisk_SalgsDato2_siden0	Historisk_SalgsDato3_siden0	HusNr_tal
HusNr_bogstav	SidedoerNummer	Vej
ArealVaegtet_same_as_BeregnetAreal	ByggeAAr_diff	OmbygningsAAr_diff
Energi	Prophet_index	

Table A.1: XXX **TODO!**.

Figure A.16: XXXX TODO!.

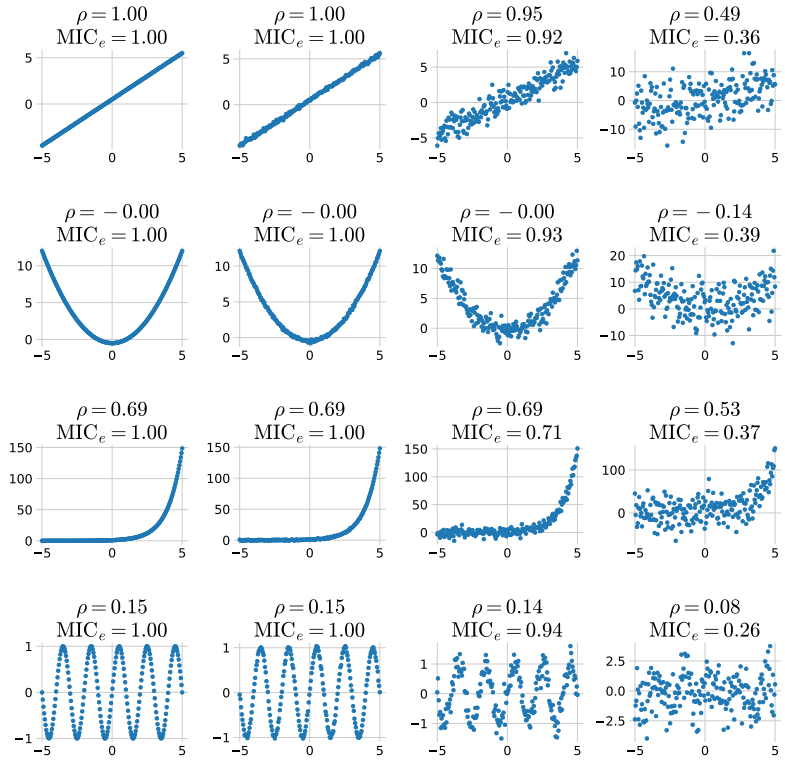


Figure A.17: MIC non-linear correlation.

Energy rating label	Code
A2020	2
A2015	4
A2010	6
A2	8
A1	10
A	12
B	20
C	30
D	40
E	50
F2	60
F1	62
F	64
G2	70
G1	72
G	74
H, I, J, K, M	90
NAN	100

Table A.2: Energy rating mapping. If the energy rating is e.g. "A2" this gets the code 8.

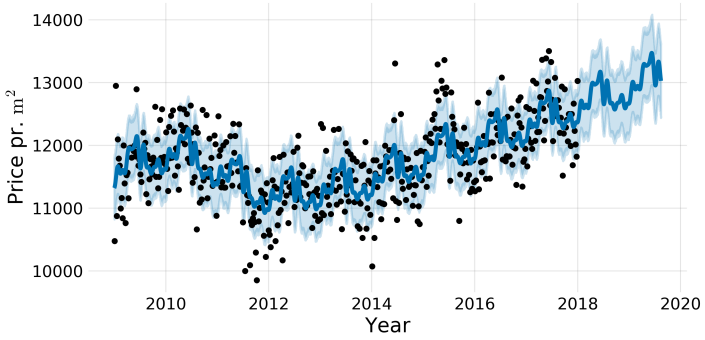


Figure A.18: The predictions of the Facebook Prophet model trained on square meter prices for owner-occupied apartments sold before January 1st, 2018. The data is down-sampled to weekly bins where the median of each week is used as in input to the Prophet model. This can be seen as black dots in the figure. The model's forecasts for 2018 and 2019 are shown in blue with a light blue error band showing the $1-\sigma$ confidence interval.

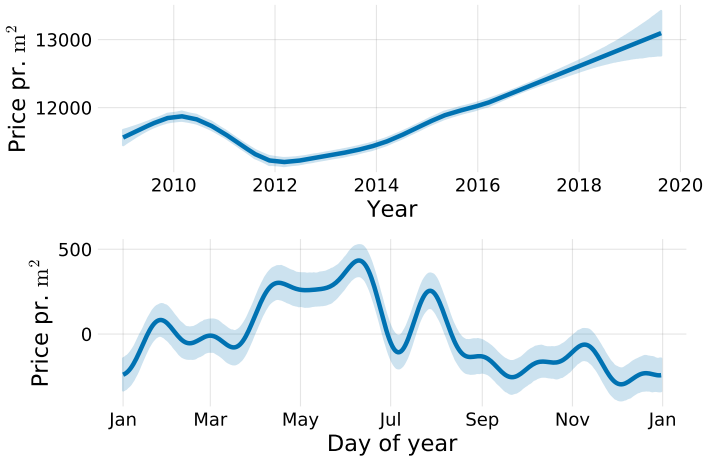


Figure A.19: The trends of the Facebook Prophet model trained on square meter prices for owner-occupied apartments sold before January 1st, 2018. In the top plot is the overall trend as a function of year and in the bottom plot is the yearly variation as a function of day of year. It can be seen that the square meter price is higher during the Summer months compared to the Winter months, however, compared to the overall trend this effect is minor ($< 10\%$).

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	226	141	0.1664
2.5	False	201	115	0.1770
5	True	301	90	0.1623
5	False	375	82	0.1786
10	True	318	97	0.1618
10	False	226	56	0.1893
20	True	265	81	0.1626
20	False	687	124	0.1799
∞	**True**	**405**	**110**	**0.1600**
∞	False	94	32	0.2036

Table A.3: Rmse-ejerlejlighed-appendix.

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	333	75	0.1595
2.5	False	496	57	0.1523
5	True	280	66	0.1606
5	**False**	**734**	**96**	**0.1513**
10	True	367	83	0.1618
10	False	351	52	0.1590
20	True	269	62	0.1609
20	False	333	49	0.1587
∞	True	388	83	0.1595
∞	False	268	42	0.1648

Table A.4: Logcosh-ejerlejlighed-appendix.

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	293	56	0.1598
2.5	False	814	101	0.1466
5	True	304	68	0.1610
5	False	923	110	0.1468
10	True	266	62	0.1610
10	**False**	**770**	**97**	**0.1450**
20	True	288	65	0.1613
20	False	967	117	0.1467
∞	True	340	72	0.1601
∞	False	807	99	0.1480

Table A.5: Cauchy-ejerlejlighed-appendix.

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	285	64	0.1628
2.5	False	718	90	0.1517
5	True	257	62	0.1600
5	False	702	91	0.1499
10	True	272	62	0.1601
10	False	771	99	0.1466
20	True	260	61	0.1603
20	False	876	107	0.1486
∞	True	310	69	0.1584
∞	**False**	**973**	**115**	**0.1459**

Table A.6: Welsch-ejerlejlighed-appendix.

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	229	54	0.1601
2.5	False	304	45	0.1577
5	True	205	54	0.1629
5	False	343	51	0.1549
10	True	257	61	0.1596
10	False	332	47	0.1573
20	True	272	62	0.1608
20	False	403	56	0.1537
∞	True	344	74	0.1578
∞	**False**	**453**	**59**	**0.1527**

Table A.7: Fair-ejerlejlighed-appendix.

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	458	339	0.1983
2.5	False	844	439	0.1913
5	True	733	478	0.1968
5	False	1126	541	0.1888
10	True	434	310	0.1999
10	False	917	444	0.1884
20	True	398	286	0.2013
20	**False**	**1206**	**575**	**0.1867**
∞	True	730	505	0.1977
∞	False	1264	625	0.1876

Table A.8: Rmse-villa-appendix.

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	346	223	0.2018
2.5	False	1095	415	0.1877
5	True	618	331	0.1976
5	False	1601	546	0.1847
10	True	506	280	0.1990
10	False	1160	400	0.1873
20	True	445	269	0.2011
20	False	1313	497	0.1876
∞	True	432	258	0.1982
∞	**False**	**2151**	**739**	**0.1842**

Table A.9: Logcosh-villa-appendix.

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	434	244	0.1991
2.5	False	1007	356	0.1872
5	True	350	208	0.1999
5	False	1130	389	0.1858
10	True	436	240	0.1992
10	False	1183	394	0.1850
20	True	397	242	0.2003
20	**False**	**1514**	**542**	**0.1833**
∞	True	449	257	0.1992
∞	False	1351	470	0.1844

Table A.10: Cauchy-villa-appendix.

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	867	440	0.1960
2.5	False	835	300	0.1897
5	True	301	184	0.2035
5	False	893	312	0.1878
10	True	341	200	0.2014
10	False	1113	390	0.1869
20	True	338	209	0.2022
20	False	1212	424	0.1875
∞	True	579	321	0.1970
∞	**False**	**1497**	**509**	**0.1837**

Table A.11: Welsch-villa-appendix.

Half-life	$\log_{10}$	N_{trees}	Time [s]	f_{eval}
2.5	True	508	278	0.1956
2.5	False	862	301	0.1882
5	True	506	278	0.1957
5	**False**	**1357**	**462**	**0.1835**
10	True	875	436	0.1946
10	False	954	325	0.1861
20	True	763	402	0.1943
20	False	1256	435	0.1840
∞	True	535	303	0.1973
∞	False	1337	456	0.1844

Table A.12: Fair-villa-appendix.

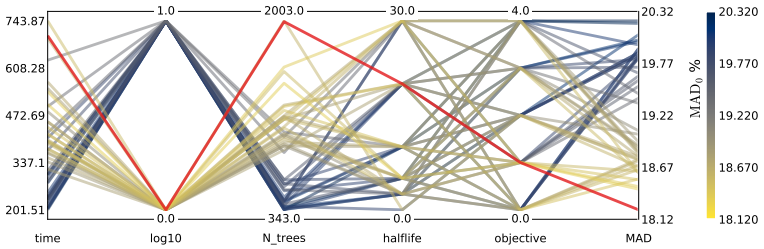


Figure A.20: Hyperparameter optimization results of the housing model for houses. The results are shown as parallel coordinates with each hyperparameter along the x-axis and the value of that parameter on the y-axis. Each line is an event in the 4-dimensional space colored according to the performance of that hyperparameter as measured by MAD_0 from highest MAD in dark blue to lowest AUC in yellow. The single best hyperparameter is shown in red. For the hyperparameter `log10` `0` means False and `1` means True, for `Halftime` ∞ is mapped to `30`, and for `objektive` the functions Cauchy (0), Fair (1), LogCosh (2) SquaredError (3), and Welsch (4) are mapped to the integers in the parentheses.

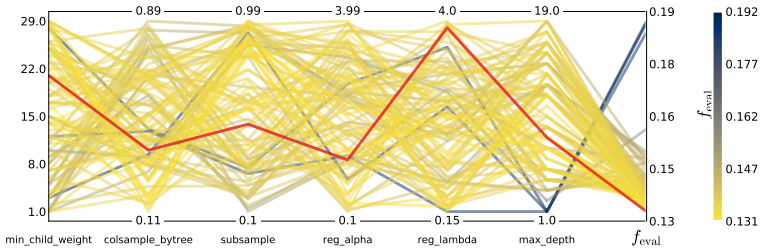


Figure A.21: Hyperparameter optimization results of XGBoost parameters of the housing model for apartments shown as parallel coordinates. Here shown for random search as hyperparameter optimization.

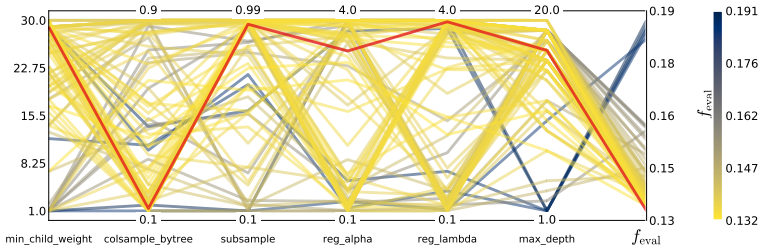


Figure A.22: Hyperparameter optimization results of XGBoost parameters of the housing model for apartments shown as parallel coordinates. Here shown for Bayesian optimization as hyperparameter optimization.

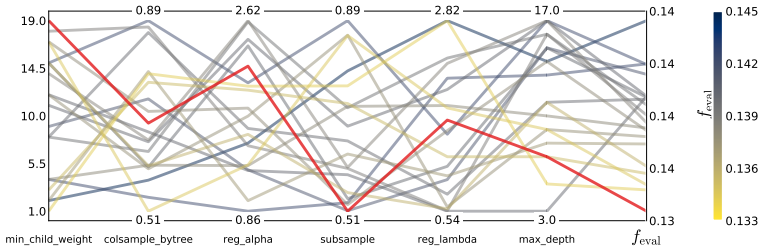


Figure A.23: Hyperparameter optimization results of XGBoost parameters of the housing model for houses shown as parallel coordinates. Here shown for random search as hyperparameter optimization.

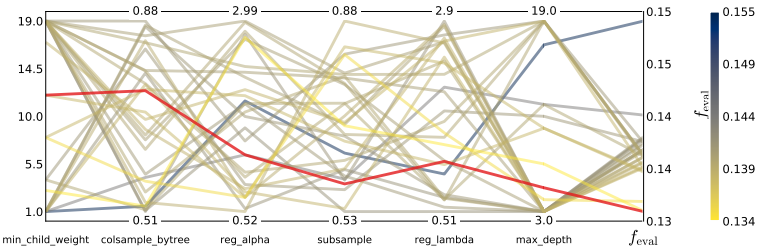


Figure A.24: Hyperparameter optimization results of XGBoost parameters of the housing model for houses shown as parallel coordinates. Here shown for Bayesian optimization as hyperparameter optimization.

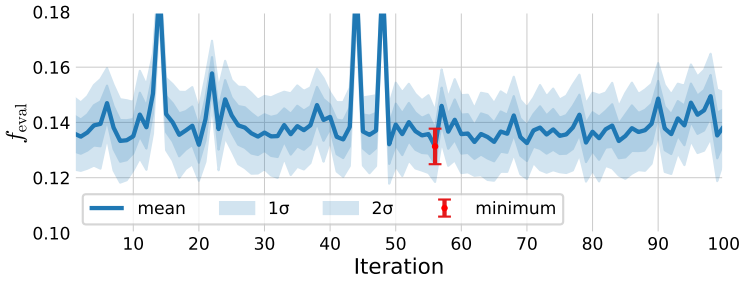


Figure A.25: XXX of the housing model for apartments shown as parallel coordinates. Here shown for random search as hyperparameter optimization.

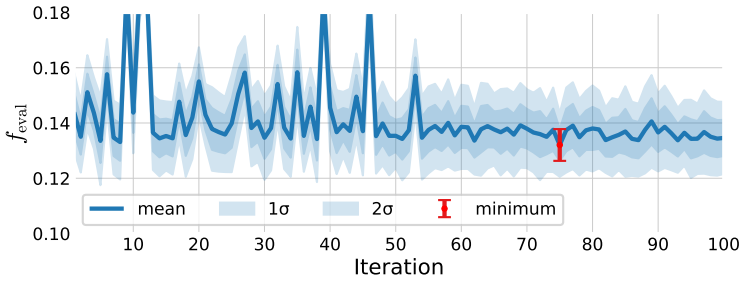


Figure A.26: XXX of the housing model for apartments shown as parallel coordinates. Here shown for Bayesian optimization as hyperparameter optimization.

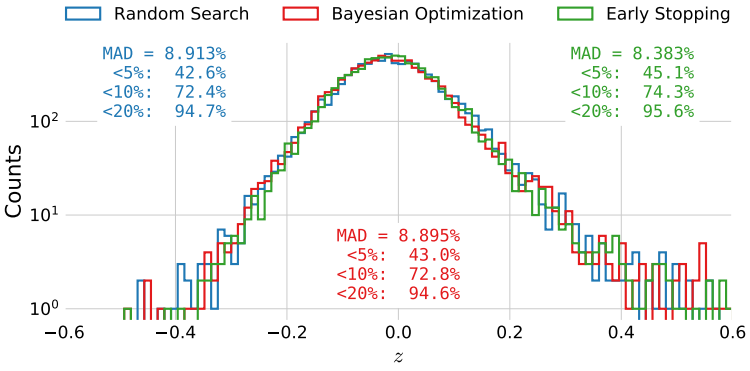


Figure A.27: Histogram of z-values of the XGB-model trained on apartments. The performance after hyperparameter optimization (HPO) using Random Search (RS) is shown in blue, for Bayesian Optimization (BO) in red. After finding the best model, BO in this case, the model is retrained using early stopping, the performance of which is shown in green.

	MAD (%)	$\leq 5\%(\%)$	$\leq 10\%(\%)$	$\leq 20\%(\%)$	μ
Train	6.35	56.22	83.41	97.08	0.00902 ± 0.00068
Test	8.38	45.06	74.32	95.58	-0.00820 ± 0.00115
2019	9.12	42.63	71.36	93.65	0.00297 ± 0.00235

Table A.13: XXX ejer tight

	MAD (%)	$\leq 5\%(\%)$	$\leq 10\%(\%)$	$\leq 20\%(\%)$	μ
Train	15.63	25.65	47.89	75.82	0.04543 ± 0.00080
Test	16.49	24.30	45.77	75.19	0.01686 ± 0.00194
2019	17.17	23.67	44.25	73.54	0.02056 ± 0.00279

Table A.14: XXX villa tight

B. Quarks vs. Gluons Appendix

	b	*c*	*uds*	*g*	non-*q*-matched
2	37.2 %	12.9 %	29.1 %	0.0 %	20.7 %
3	22.6 %	8.9 %	19.7 %	31.2 %	17.5 %
4	14.6 %	7.0 %	15.0 %	45.1 %	18.3 %
5	10.0 %	5.7 %	12.2 %	52.5 %	19.6 %
6	7.1 %	4.4 %	8.8 %	54.4 %	25.2 %

Table B.1: Number of different types of jets for MC and MCb written in relative numbers such that each row sum to 100 %. See also Table 5.3.

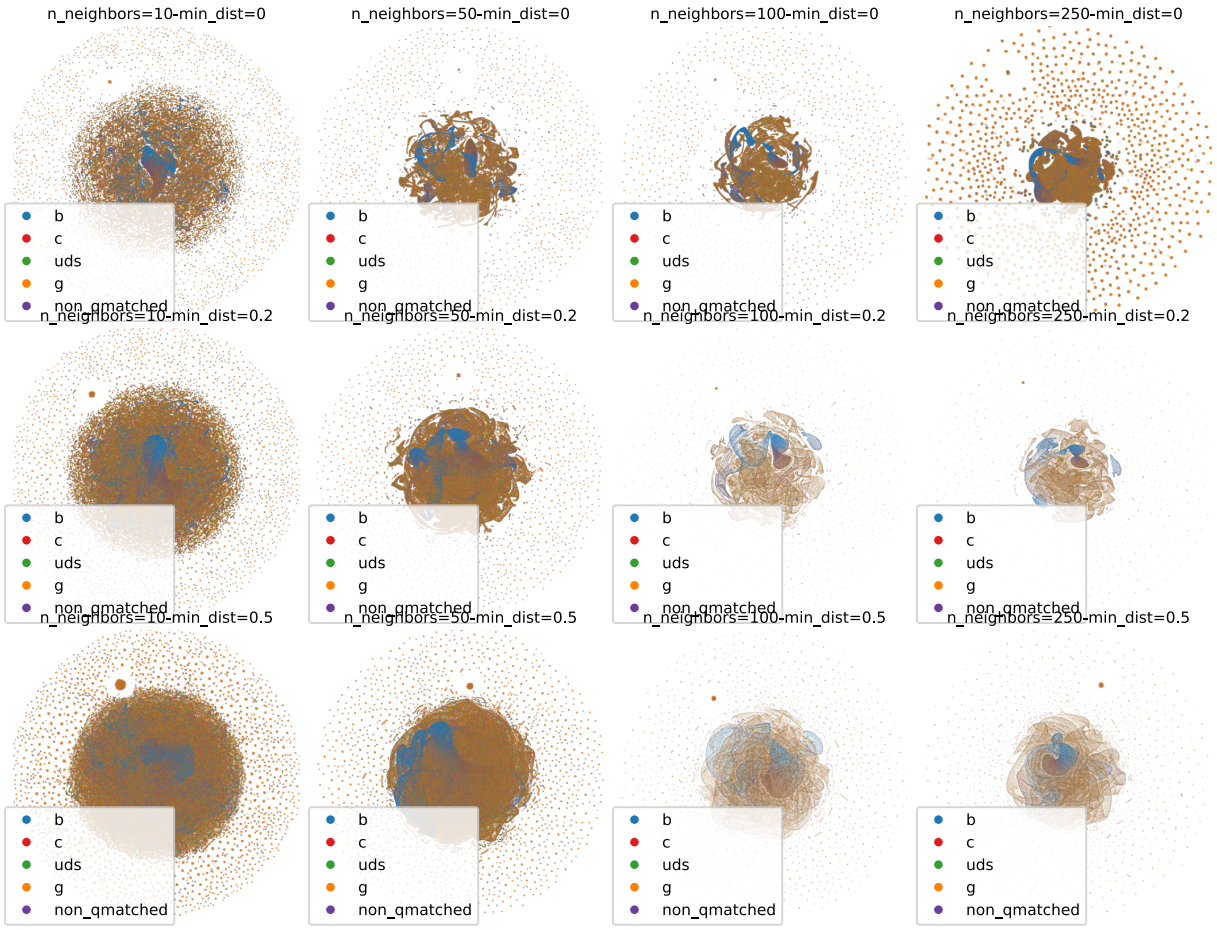


Figure B.1: Grid search of the two parameters `n_neighbors` and `min_dist` for the UMAP algorithm run on 4-jet events. For an explanation of these, see section 5.2.

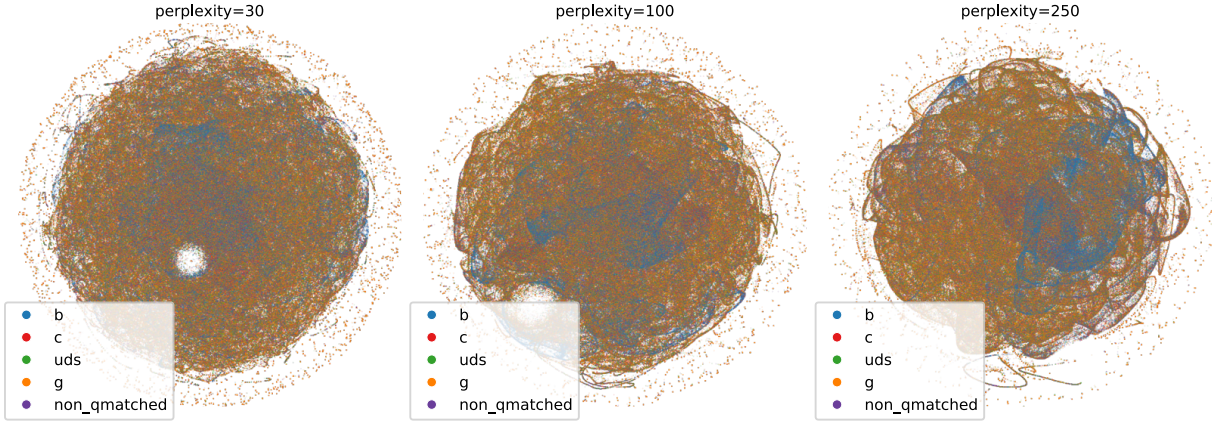


Figure B.2: Visualization of the t-SNE algorithm as a function of the `perplexity` parameters for 4-jet events.

Hyperparameter	Range
`subsample`	$\mathcal{U}(0.4, 1)$
`colsample_bytree`	$\mathcal{U}_{\text{trunc}}(0.4, 1, 2)$
`max_depth`	$\mathcal{U}_{\text{int}}(1, 20)$
`min_child_weight`	$\mathcal{U}_{\text{int}}(0, 10)$

Table B.2: Probability Density Functions for the random search hyperparameter optimization process for the XGBoost model.

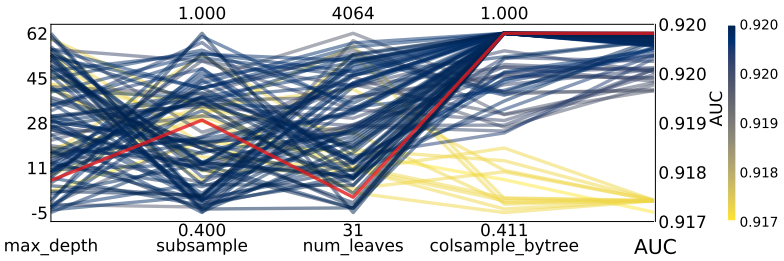


Figure B.3: Hyperparameter optimization results of b-tagging for 3-jet events. The results are shown as parallel coordinates with each hyperparameter along the x-axis and the value of that parameter on the y-axis. Each line is an event in the 4-dimensional space colored according to the performance of that hyperparameter as measured by AUC from highest AUC in dark blue to lowest AUC in yellow. The single best hyperparameter is shown in red.

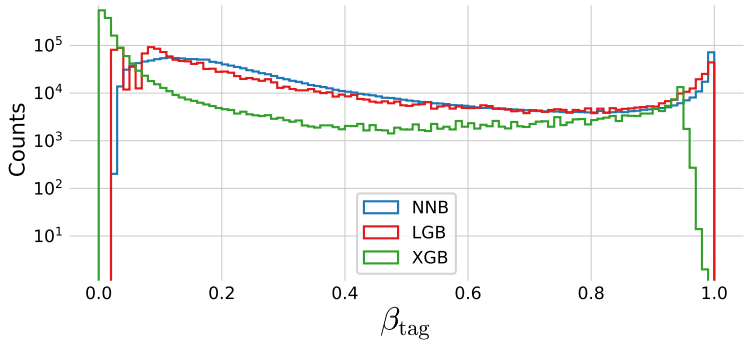


Figure B.4: Histogram of b-tag scores β_{tag} in 3-jet events for NNB (the neural network pre-trained by ALEPH, also called `nnbjet`) in blue, LGB in red, and XGB in green.

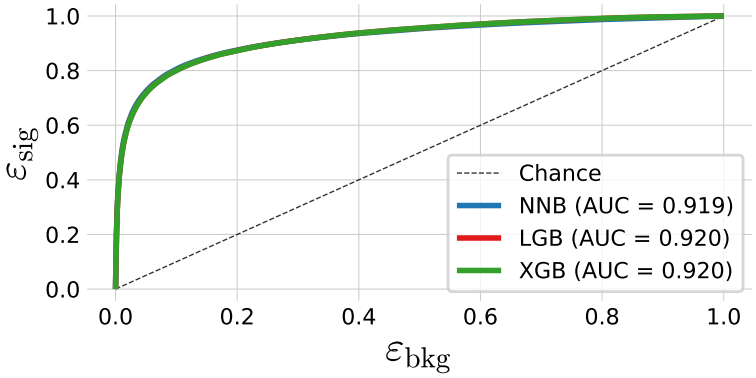


Figure B.5: ROC curve of the three b-tag models in 3-jet events for NNB (the pre-trained neural network trained by ALEPH, also called `nnbjet`) in blue, LGB in red, and XGB in green. In the legend the area under curve (AUC) is also shown. Notice that the LGB and XGB models share performance and it is thus due to overplotting that only the green line for XGB can be seen. In the machine learning community the background efficiency ε_{bkg} is sometimes know as the false positive rate (FPR) and the signal efficiency ε_{sig} as the true positive rate (TPR).

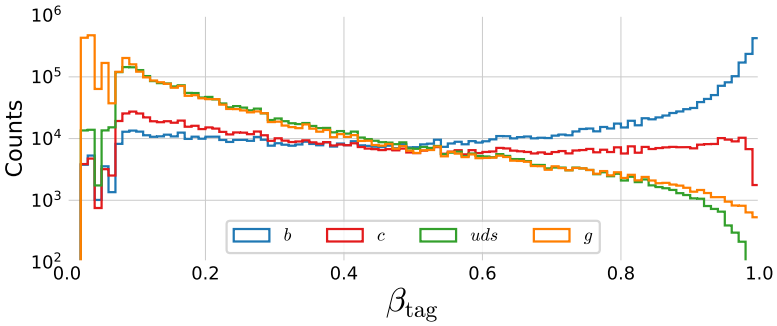


Figure B.6: Distribution of b-tags in 3-jet events for b-jets in blue, c-jets in red, uds in green and g in orange.

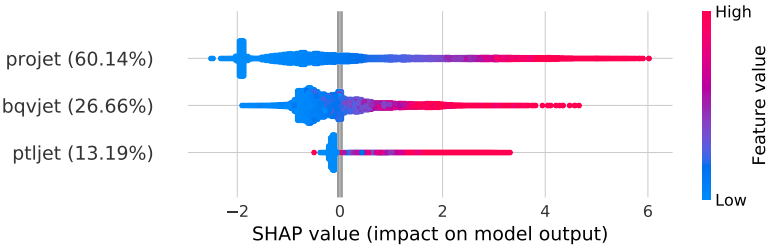


Figure B.7: Global feature importances for the LGB b-tagging algorithm on 3-jet events. The normalized feature importance is shown in the parenthesis and the each dot is an observation showing the dependance between the SHAP value and the feature's value.

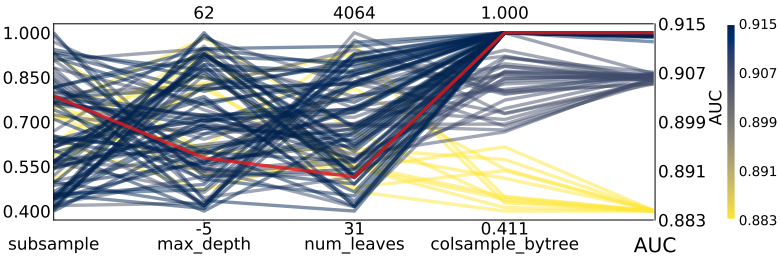


Figure B.8: Hyperparameter optimization results of g-tagging for 3-jet events for energy ordered jets.

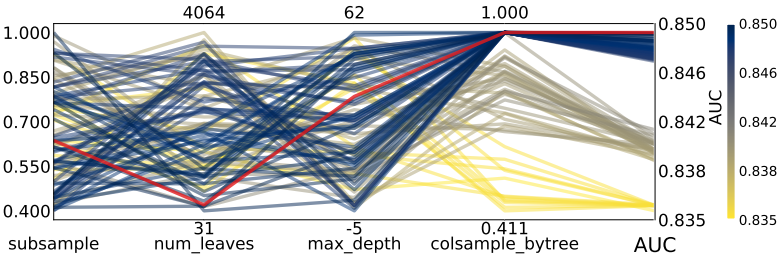


Figure B.9: Hyperparameter optimization results of g-tagging for 3-jet events for (row) shuffled jets.

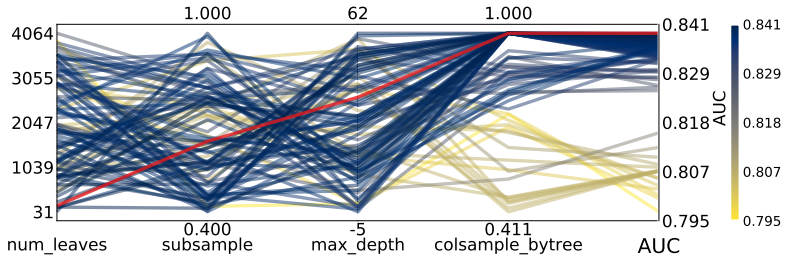


Figure B.10: Hyperparameter optimization results of g-tagging for 4-jet events for energy ordered jets.

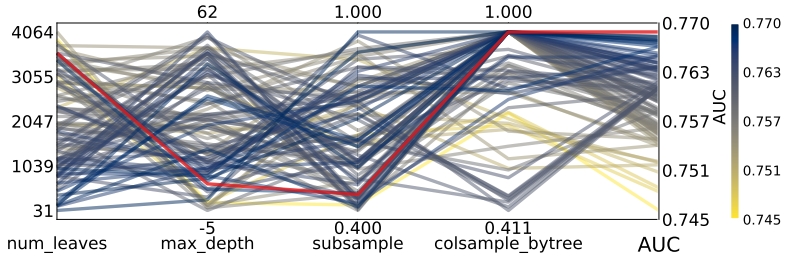


Figure B.11: Hyperparameter optimization results of g-tagging for 4-jet events for (row) shuffled jets.

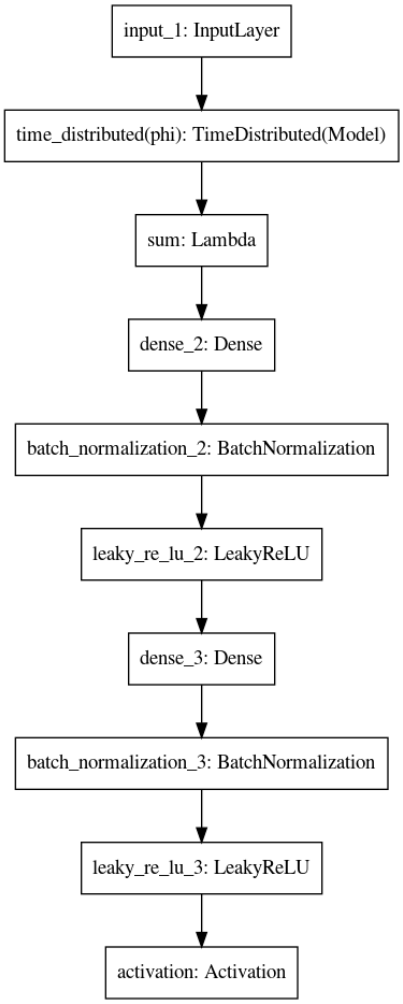


Figure B.12: Architecture of the PermNet neural network.

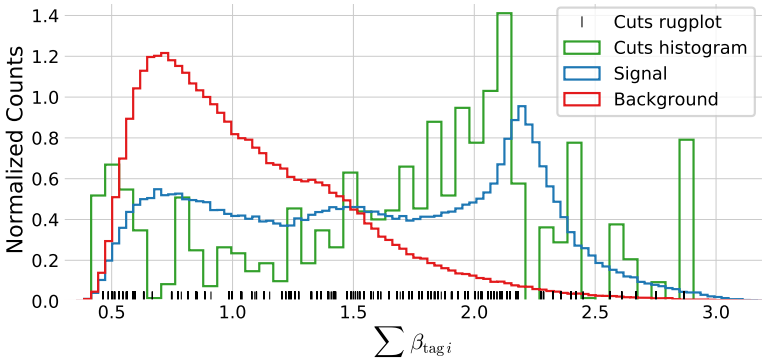


Figure B.13: Histogram of the distribution of signal in blue and background in red for the 1-dimensional sum of b-tags for 4-jet events. A histogram of the cut values from the LGB model trained on this data is shown in green together with a rug plot of the cut values in black. Notice how most of the cuts match up with the signal peak at around a $\sum \beta_i \sim 2.1$.

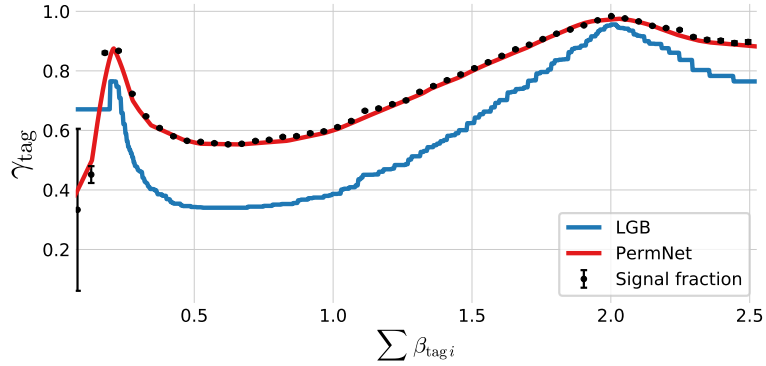


Figure B.14: Plot of the (1D) g-tag scores for 3-jet events as a function of $\sum \beta_i$ for the LGB model in blue and the PermNet model in red. The signal fraction (based on the signal and background histograms in Figure B.15) is plotted as black error bars where the size of the error bars is based on the propagated uncertainties of the signal and background histogram assuming Poissonian statistics.

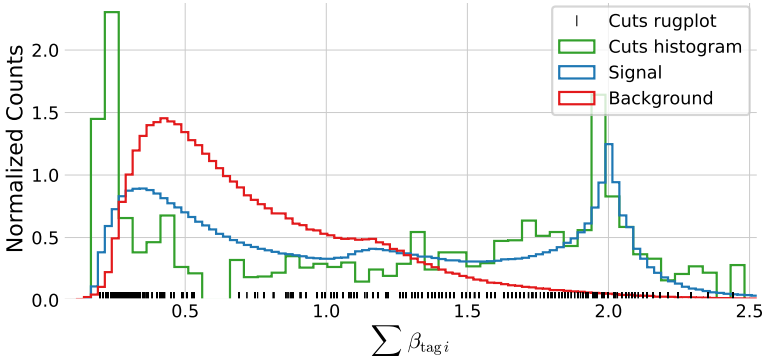


Figure B.15: Histogram of the distribution of signal in blue and background in red for the 1-dimensional sum of b-tags for 3-jet events. A histogram of the cut values from the LGB model trained on this data is shown in green together with a rug plot of the cut values in black. Notice how most of the cuts match up with the signal peak at around a $\sum \beta_i \sim 2.1$.

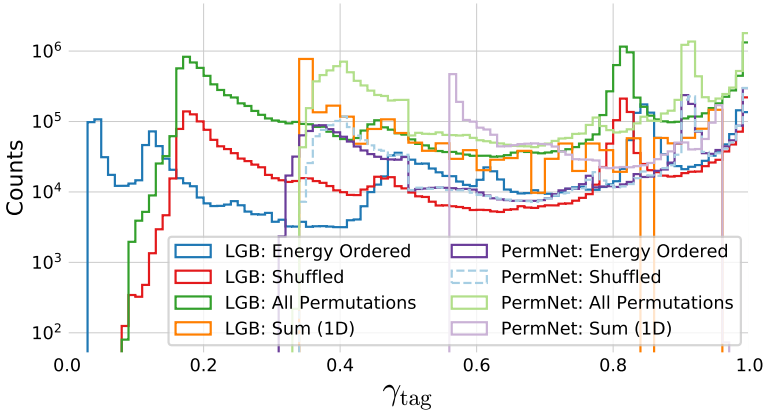


Figure B.16: Distribution of g-tag scores in 3-jet events shown with a logarithmic y-scale for LGB: Energy Ordered in blue, LGB: Shuffled in red, LGB: All Permutations in green, LGB: Sum 1D in orange, PermNet: Energy Ordered in purple, PermNet: Shuffled in light-blue, PermNet: All Permutations in light-green, PermNet: Sum 1D in light-purple. Here LGB and PermNet are the two different type of models and "Energy Ordered", "Shuffled", "All Permutations", and "Sum 1D" are the different methods used for making the input data permutation invariant (except energy ordered).

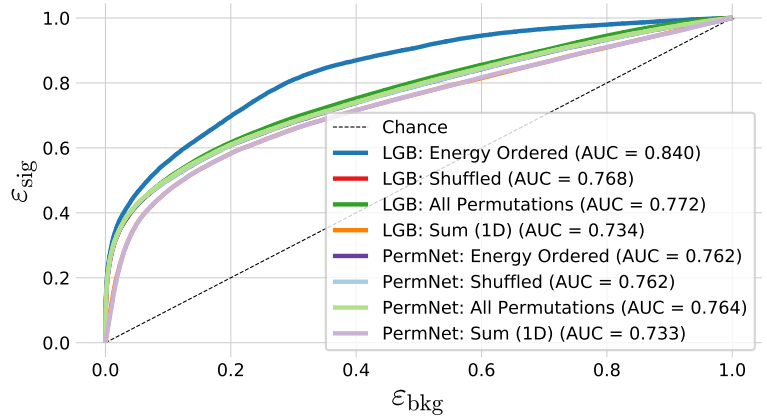


Figure B.17: ROC curve of the eight g-tag models in 4-jet events. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 5.20 and in the legend also the Area Under the ROC curve (AUC) is shown. Notice that the XGB model which uses the energy ordered data produced the best model, however, this model is not permutation invariant. Of the permutation invariant models (the rest), the XGB model trained on all permutations of the b-tags performs highest. The lowest performing models are the two models trained only on the 1-dimensional sum of b-tags, as expected, however, still with a better performance than expected by the author.

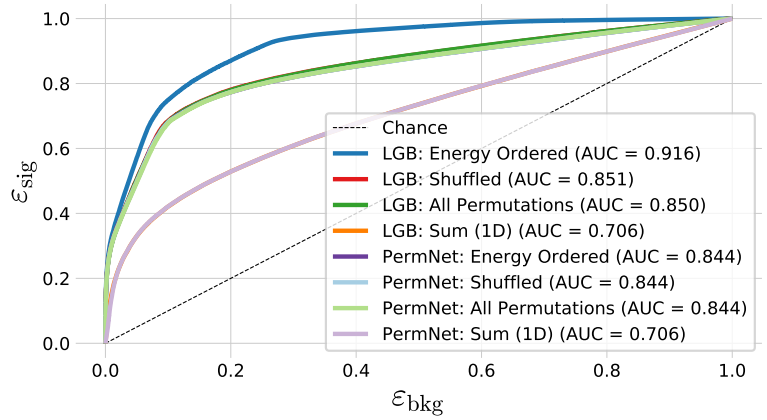


Figure B.18: ROC curve of the eight g-tag models in 3-jet. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 5.20 and in the legend also the Area Under the ROC curve (AUC) is shown.

β_{tag_i}	Energy Ordered	Shuffled	All Permutations
1	0.827 ± 0.006	0.924 ± 0.006	0.923 ± 0.006
2	0.749 ± 0.006	0.909 ± 0.006	0.918 ± 0.005
3	1.198 ± 0.006	0.878 ± 0.005	0.906 ± 0.005

Table B.3: Global SHAP feature importances $\phi^{\text{tot}}_{\beta_i}$ for the three g-Tagging Models in 3-Jet Events. Each $\phi^{\text{tot}}_{\beta_i}$ is shown for the three methods in the columns and the three b-tags as variables in the rows.

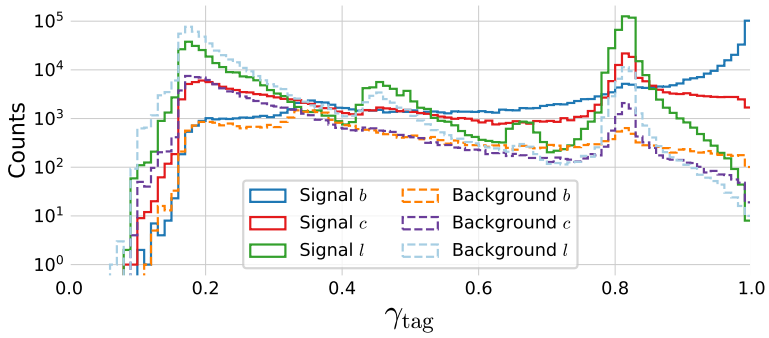


Figure B.19: Histogram of g-tag scores from the LGB-model in 3-jet events for b signal in blue, c signal in red, l (uds) signal in green, b background in orange, c background in purple, l (uds) background in light-blue.

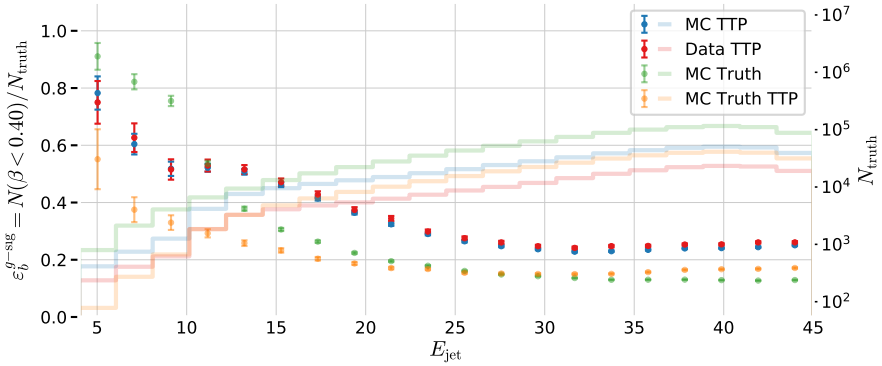


Figure B.20: b-tag efficiency for b-jets in the g-signal region for 3-jet events, $\varepsilon_b^{g\text{-sig}}$, as a function of jet energy E_{jet}. In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth TTP in green, and MC Truth TTP in orange. The efficiencies (the errorbars) can be read off on the left y-axis and the counts (histograms) on the right y-axis.

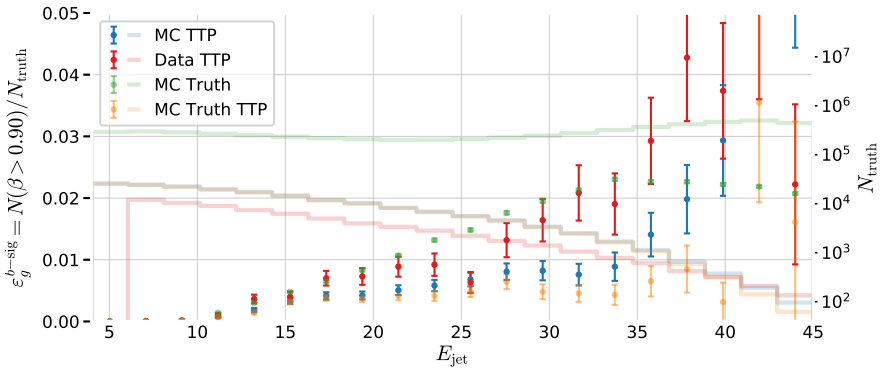


Figure B.21: b-tag efficiency for g-jets in the b-signal region for 3-jet events, $\varepsilon_g^{b\text{-sig}}$, as a function of jet energy E_{jet}. In the plot the efficiencies are shown for MC TTP in blue, Data TTP in red, MC Truth TTP in green, and MC Truth TTP in orange. The efficiencies (the errorbars) can be read off on the left y-axis and the counts (histograms) on the right y-axis.

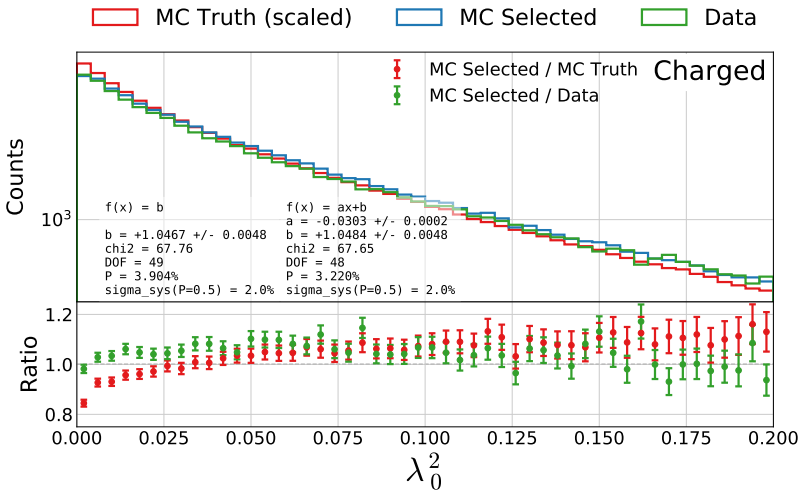


Figure B.22: Distribution of the generalized angularity λ_0^2 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

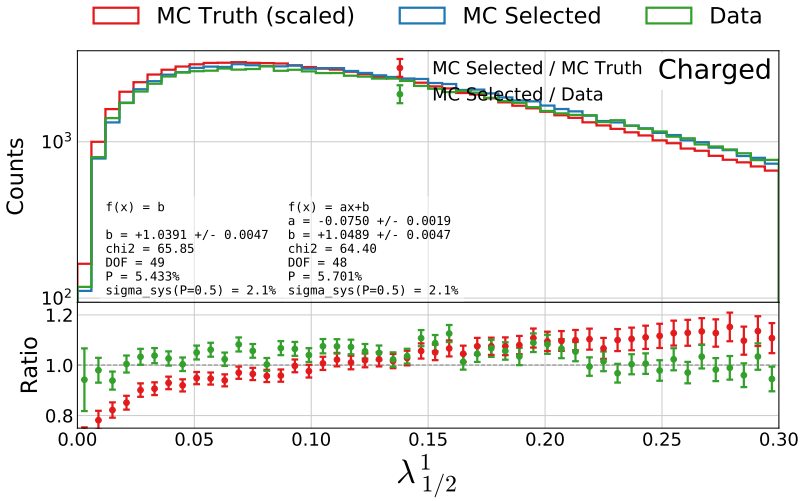


Figure B.23: Distribution of the generalized angularity $\lambda_{\frac{1}{2}}^1$ for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

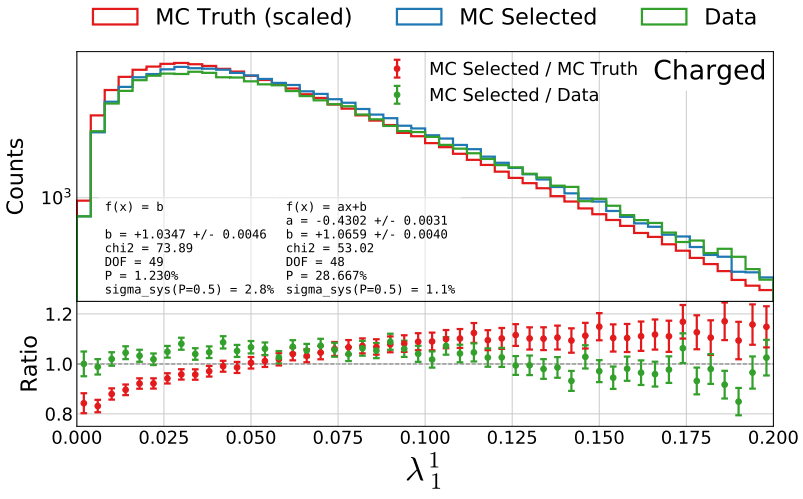


Figure B.24: Distribution of the generalized angularity λ_1^1 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

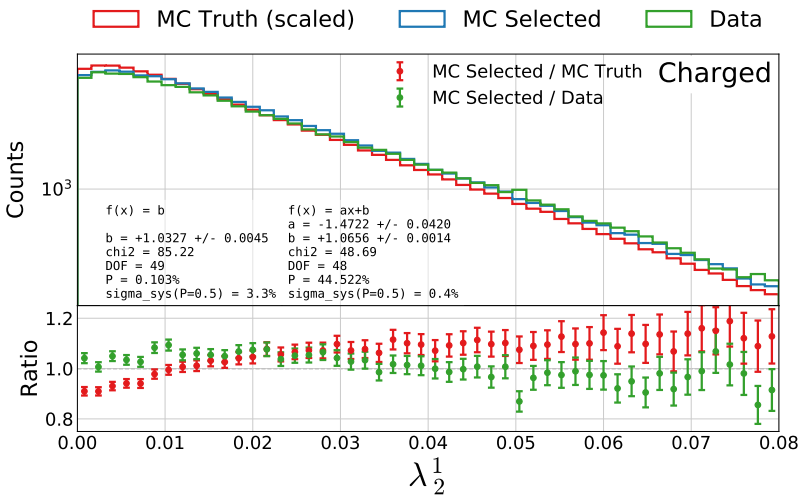


Figure B.25: Distribution of the generalized angularity λ_1^2 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

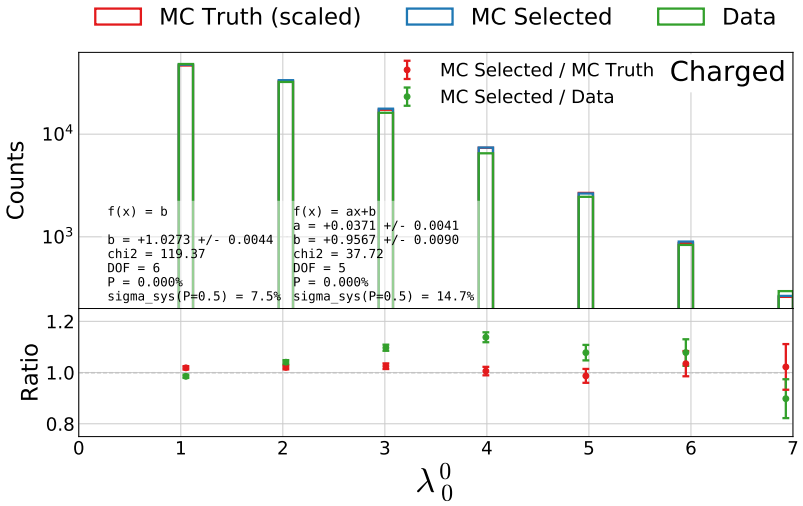


Figure B.26: Distribution of the generalized angularity λ_0^0 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

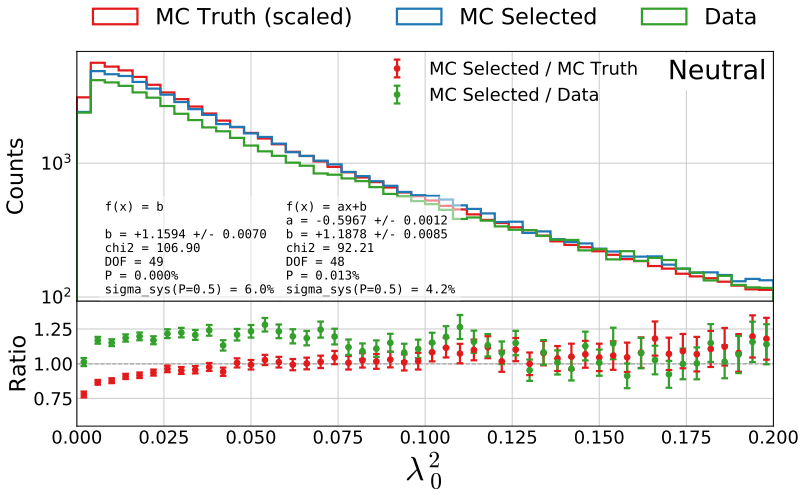


Figure B.27: Distribution of the generalized angularity λ_0^2 for charged gluons jets in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

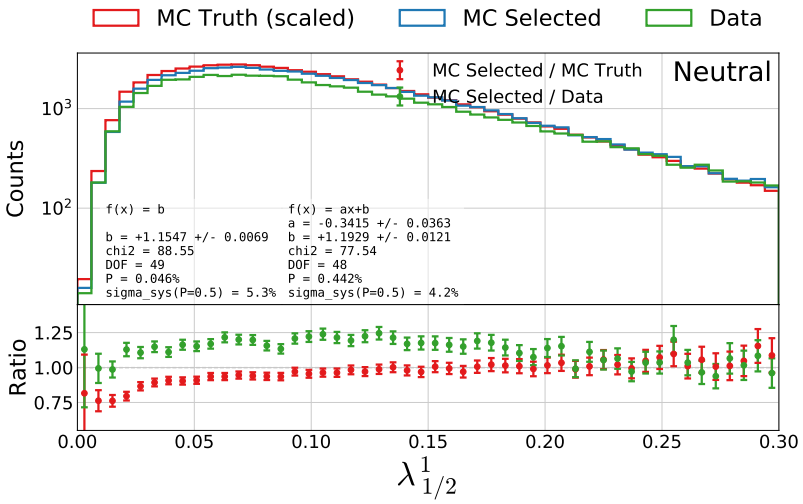


Figure B.28: Distribution of the generalized angularity $\lambda^1_{\frac{1}{2}}$ for neutral gluons clusters in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

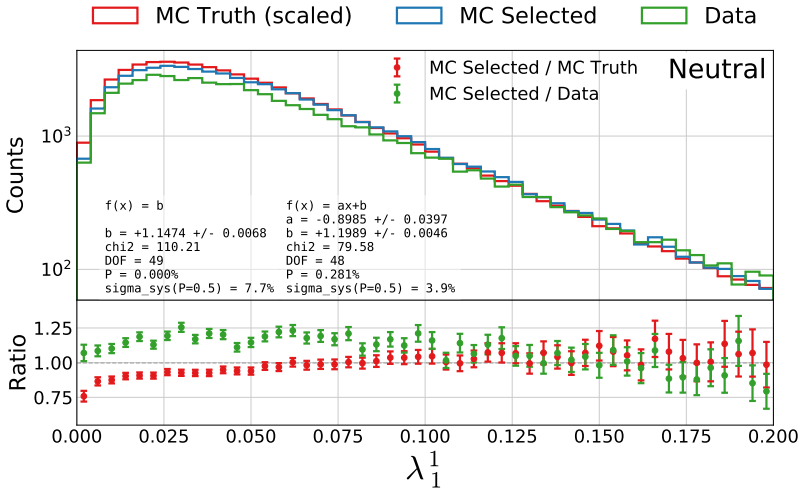


Figure B.29: Distribution of the generalized angularity λ^1_1 for neutral gluons clusters in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

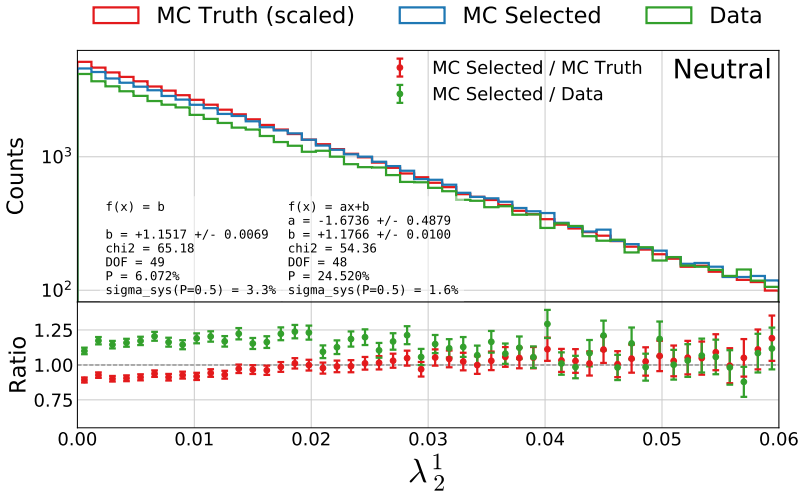


Figure B.30: Distribution of the generalized angularity λ^2_1 for neutral gluons clusters in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

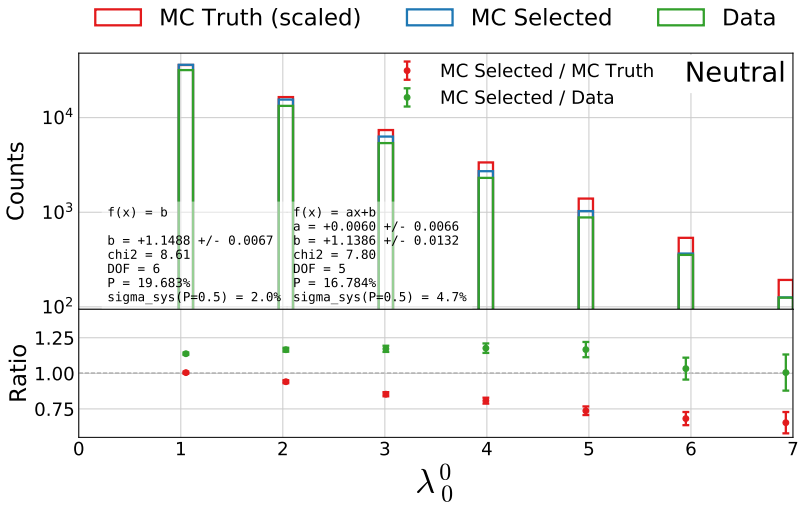


Figure B.31: Distribution of the generalized angularity λ_0^0 for neutral gluons clusters in 3-jet events. The distributions for MC Truth is shown in red, MC Truth in blue, and Data in green in the top plot and in the bottom plot the ratio between MC Selected and MC Truth is shown in red and between MC Selected and Data in green.

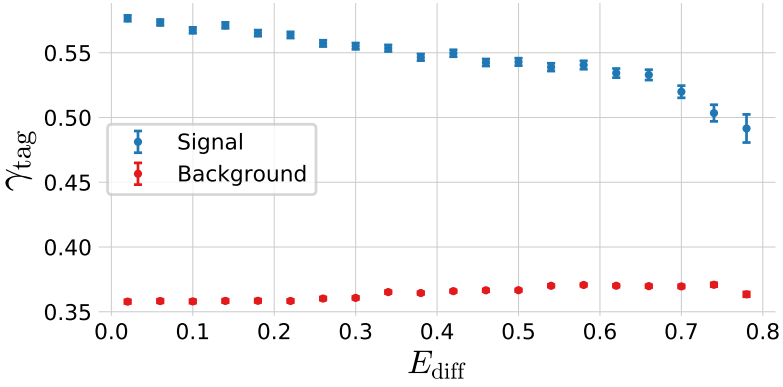


Figure B.32: Relationship between the g-tag value γ_{tag} and the gluon splitting variable E_{diff}. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

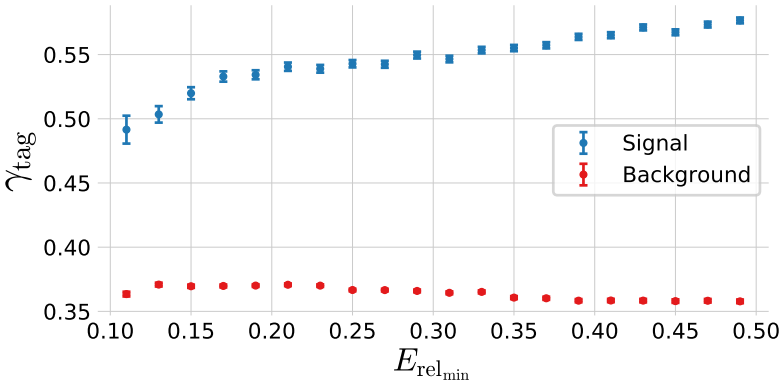


Figure B.33: Relationship between the g-tag value γ_{tag} and the gluon splitting variable $E_{\text{rel}_{\text{min}}}$. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

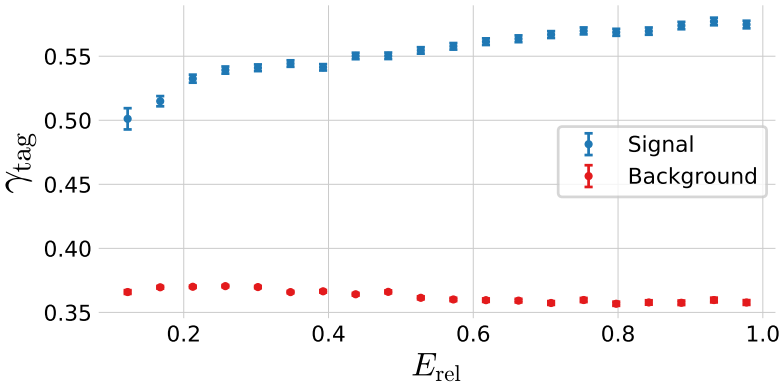


Figure B.34: Relationship between the g-tag value γ_{tag} and the gluon splitting variable E_{rel}. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

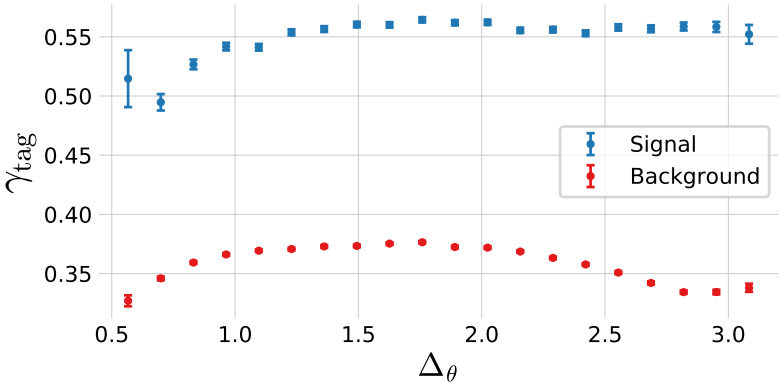


Figure B.35: Relationship between the g-tag value γ_{tag} and the gluon splitting variable Δ_θ. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

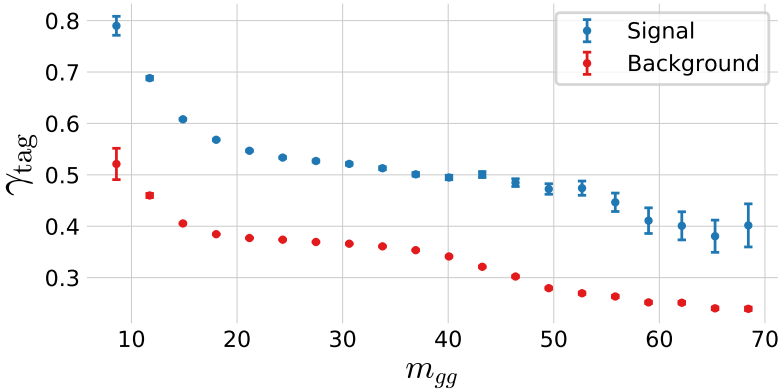


Figure B.36: Relationship between the g-tag value γ_{tag} and the gluon splitting variable m_{gg}. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

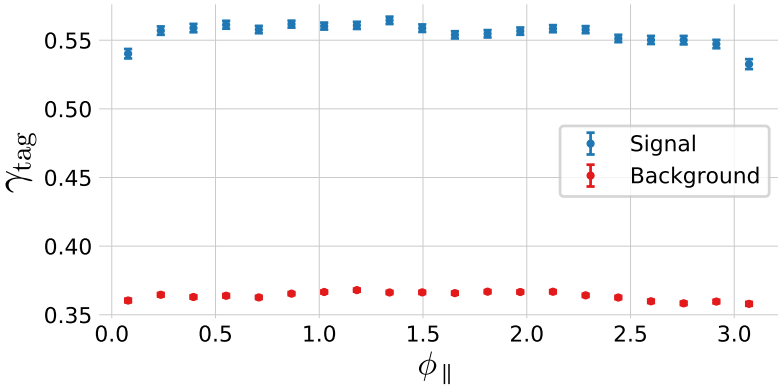


Figure B.37: Relationship between the g-tag value γ_{tag} and the gluon splitting variable $\phi_\parallel$. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

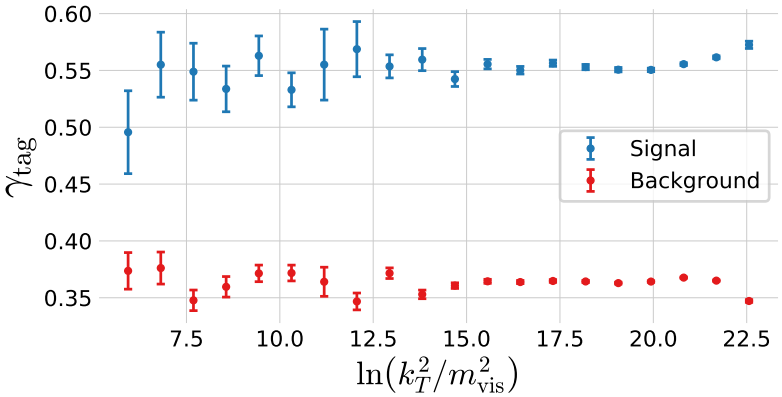


Figure B.38: Relationship between the g-tag value γ_{tag} and the gluon splitting variable $\ln\left(k_t^2/m_{\text{vis}}^2\right)$. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

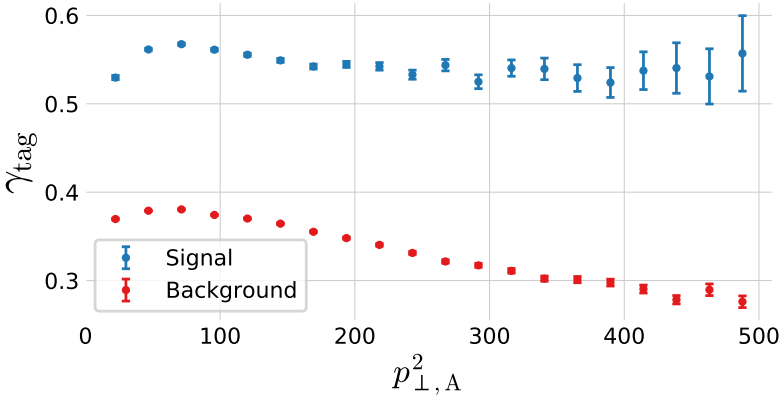


Figure B.39: Relationship between the g-tag value γ_{tag} and the gluon splitting variable $p^2_{\perp,\text{A}}$. The signal events (according to MC Truth) are plotted in blue and background events (according to MC Truth) in red.

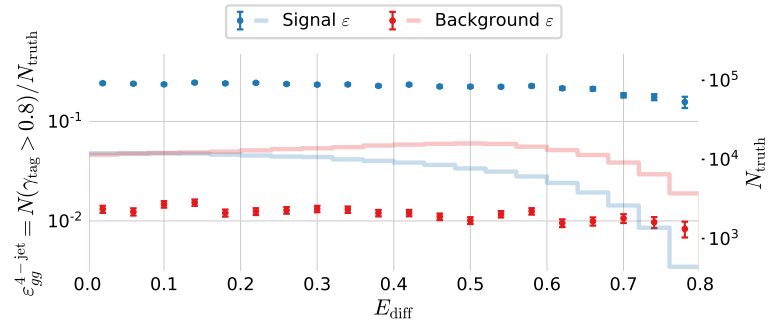


Figure B.40: Efficiency of the g-tagging algorithm for 4-jet events as a function of normalized gluon-gluon jet energy difference (asymmetry) E_{diff} in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

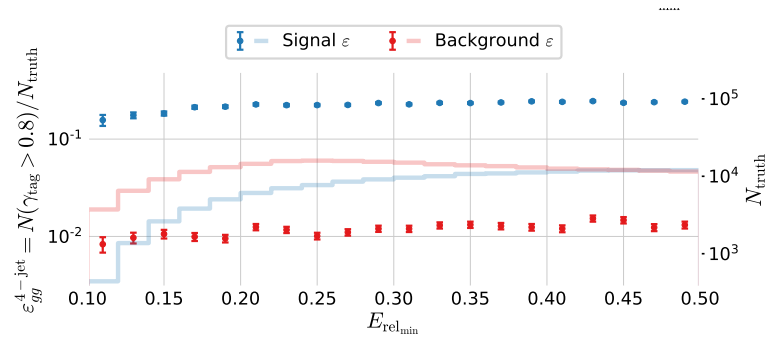


Figure B.41: Efficiency of the g-tagging algorithm for 4-jet events as a function of $E_{\text{rel}_{\text{min}}}$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

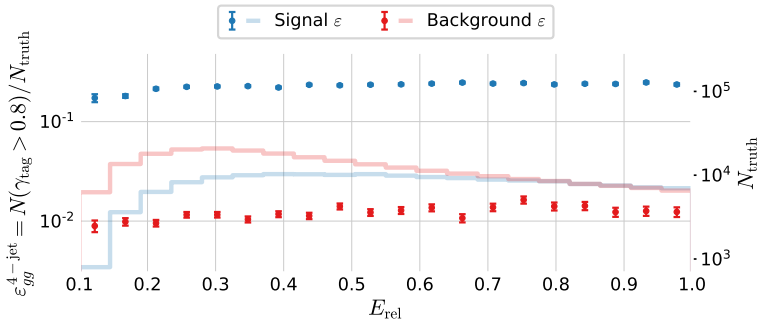


Figure B.42: Efficiency of the g-tagging algorithm for 4-jet events as a function of E_{rel} in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

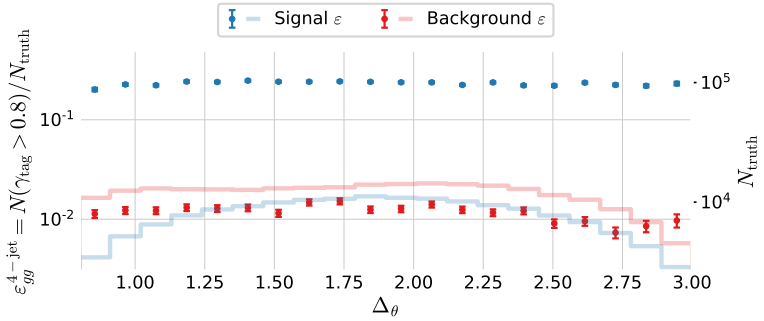


Figure B.43: Efficiency of the g-tagging algorithm for 4-jet events as a function of Δ_θ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

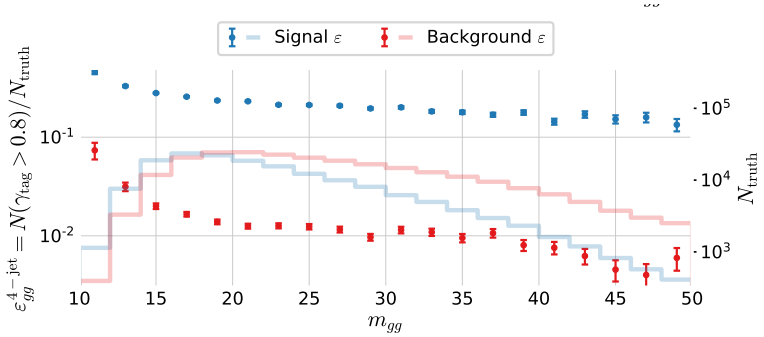


Figure B.44: Efficiency of the g-tagging algorithm for 4-jet events as a function of m_{gg} in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

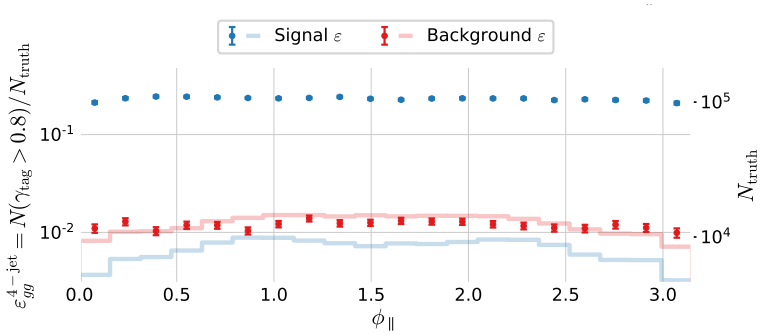


Figure B.45: Efficiency of the g-tagging algorithm for 4-jet events as a function of $\phi_{\parallel}$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

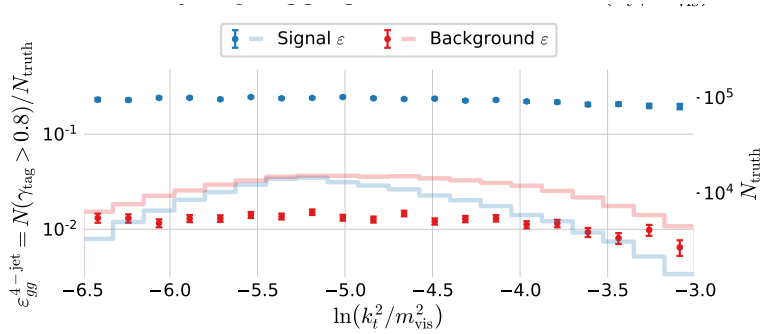


Figure B.46: Efficiency of the g-tagging algorithm for 4-jet events as a function of $\ln\left(k_t^2/m_{\text{vis}}^2\right)$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

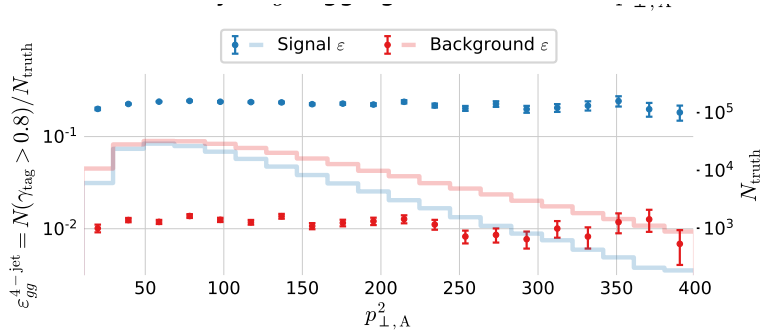


Figure B.47: Efficiency of the g-tagging algorithm for 4-jet events as a function of $p_{\perp,\text{A}}^2$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

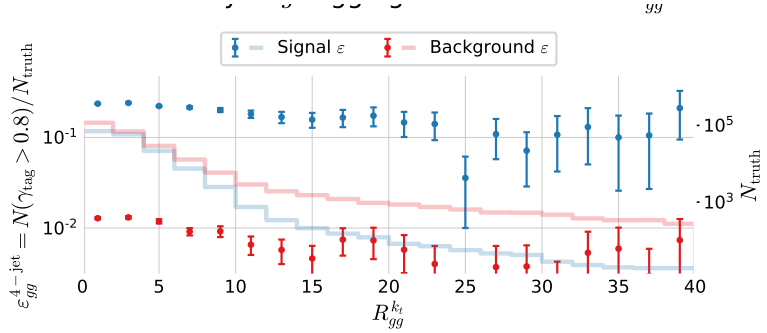


Figure B.48: Efficiency of the g-tagging algorithm for 4-jet events as a function of $R_{gg}^{k_t}$ in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

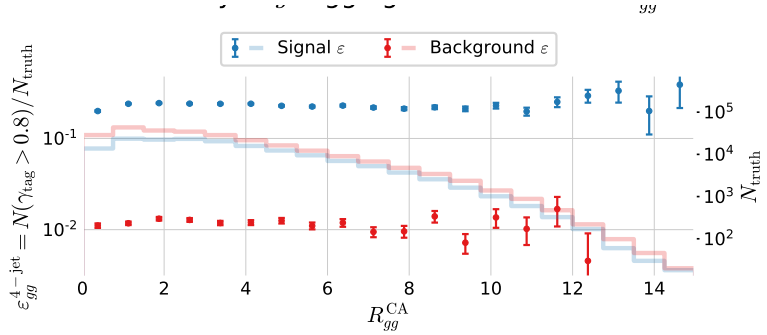


Figure B.49: Efficiency of the g-tagging algorithm for 4-jet events as a function of E in MC. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

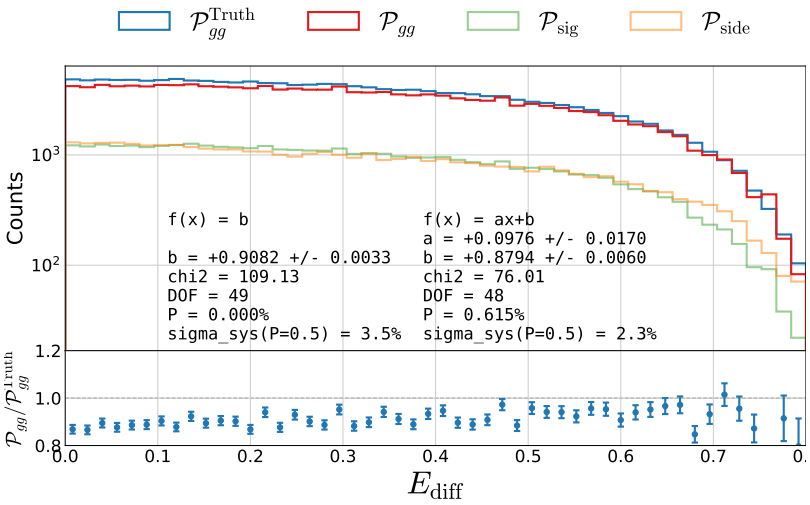


Figure B.50: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for E_{diff}. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal region $\mathcal{P}_{\text{sig}}$ in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between $\mathcal{P}_{gg}$ and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

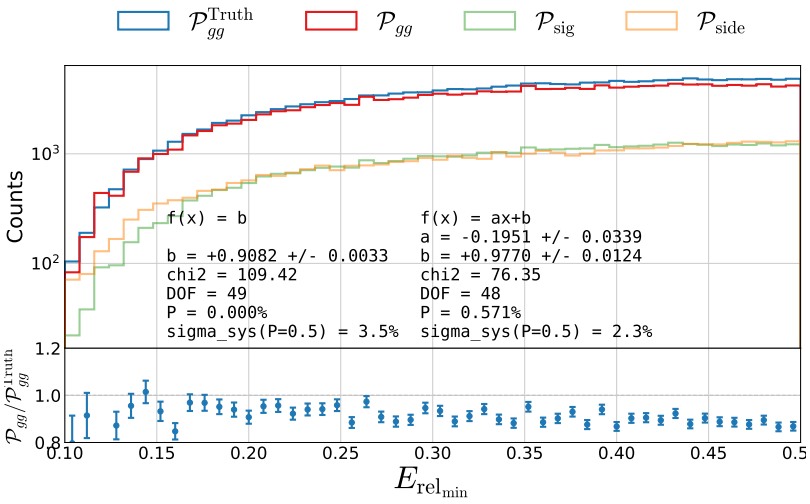


Figure B.51: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for $E_{\text{rel}_{\text{min}}}$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal region $\mathcal{P}_{\text{sig}}$ in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between $\mathcal{P}_{gg}$ and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

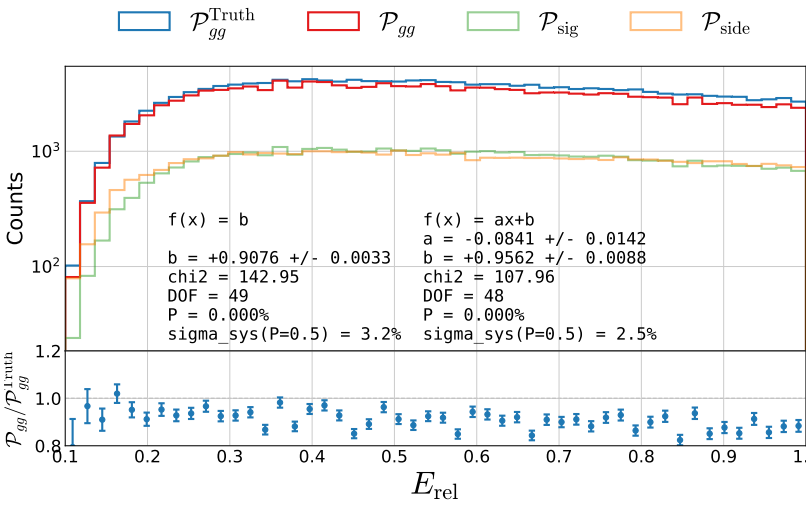


Figure B.52: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for E_{rel}. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal region $\mathcal{P}_{\text{sig}}$ in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between $\mathcal{P}_{gg}$ and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

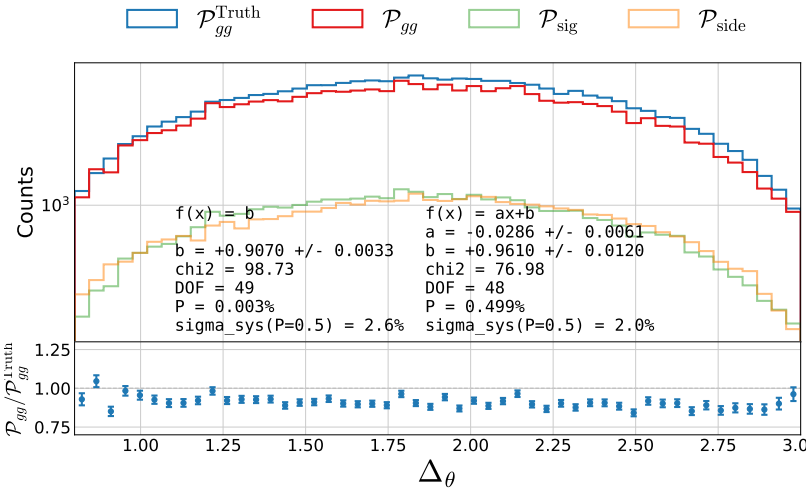


Figure B.53: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for Δ_θ. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal region $\mathcal{P}_{\text{sig}}$ in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between $\mathcal{P}_{gg}$ and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

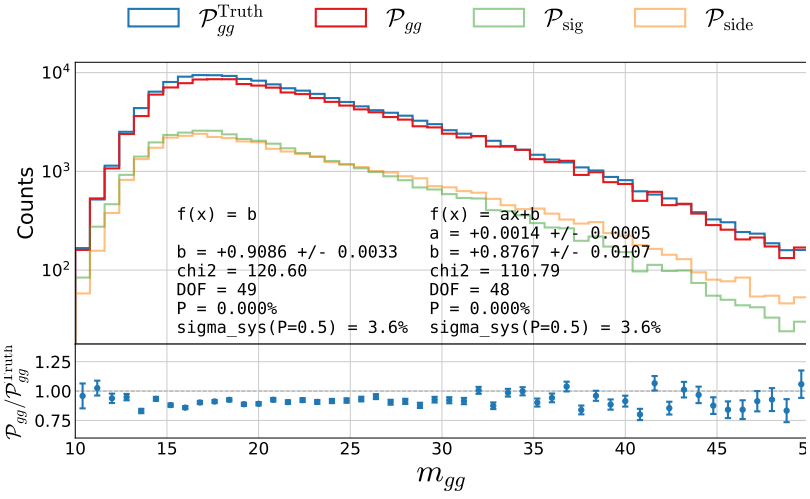


Figure B.54: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for m_{gg}. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal region $\mathcal{P}_{\text{sig}}$ in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between $\mathcal{P}_{gg}$ and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

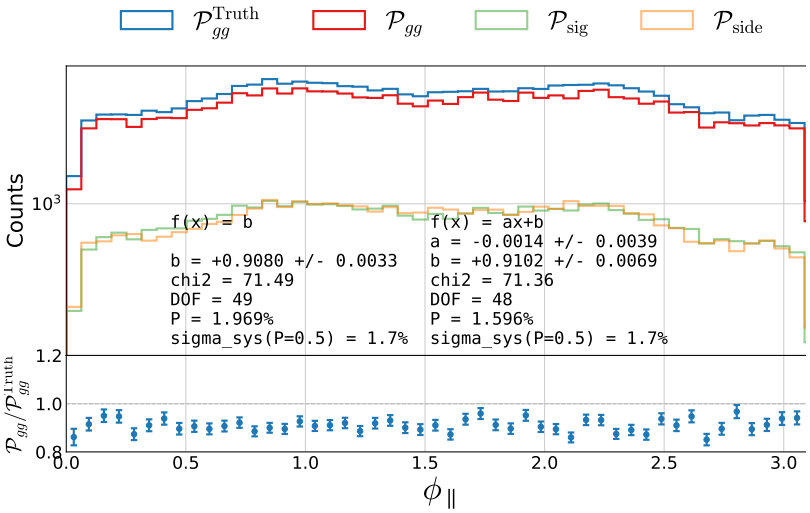


Figure B.55: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for $\phi_{\parallel}$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal region $\mathcal{P}_{\text{sig}}$ in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between $\mathcal{P}_{gg}$ and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

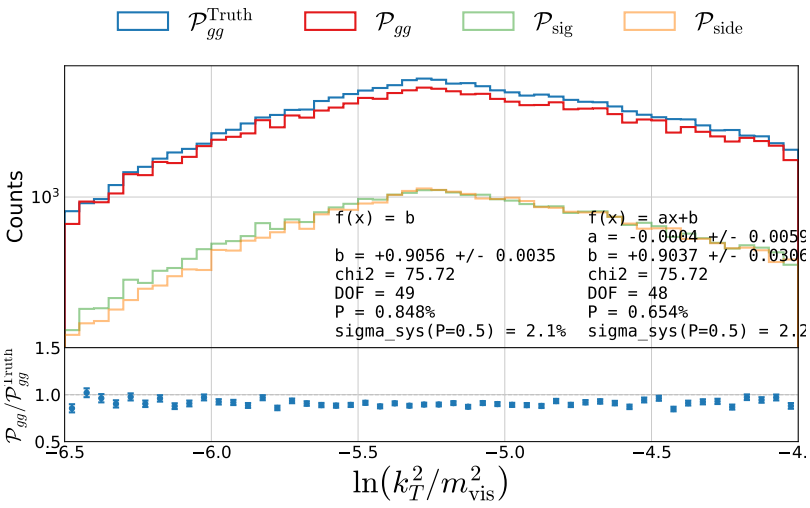


Figure B.56: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for $\ln\left(k_t^2/m_{\text{vis}}^2\right)$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal region $\mathcal{P}_{\text{sig}}$ in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between $\mathcal{P}_{gg}$ and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

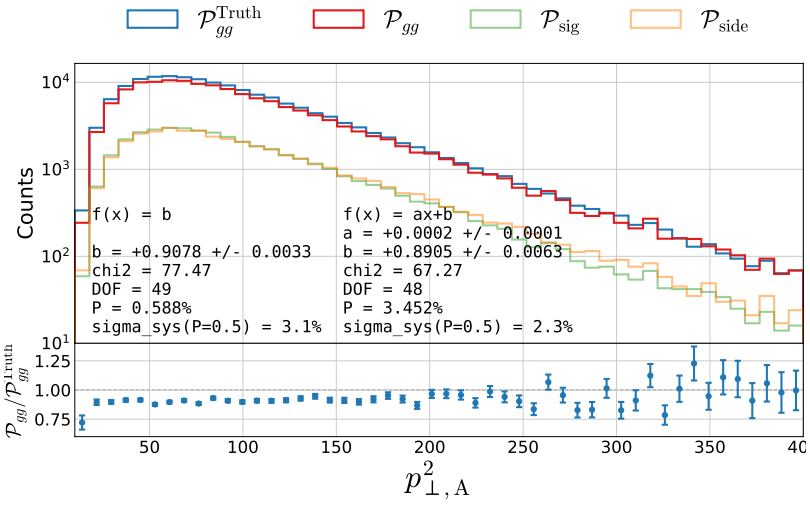


Figure B.57: Closure plot comparing MC Truth and the efficiency corrected g-tagging model in 4-jet events for $p_{\perp,\text{A}}^2$. In the top part of the plot $\mathcal{P}_{gg}^{\text{Truth}}$ based on MC Truth is shown in blue, the $\mathcal{P}_{gg}$ based on MC but without Truth in red, the distribution in the signal region $\mathcal{P}_{\text{sig}}$ in light green and the distribution in the sideband region $\mathcal{P}_{\text{side}}$ in light orange. In the bottom part of the plot the ratio between $\mathcal{P}_{gg}$ and $\mathcal{P}_{gg}^{\text{Truth}}$ is shown.

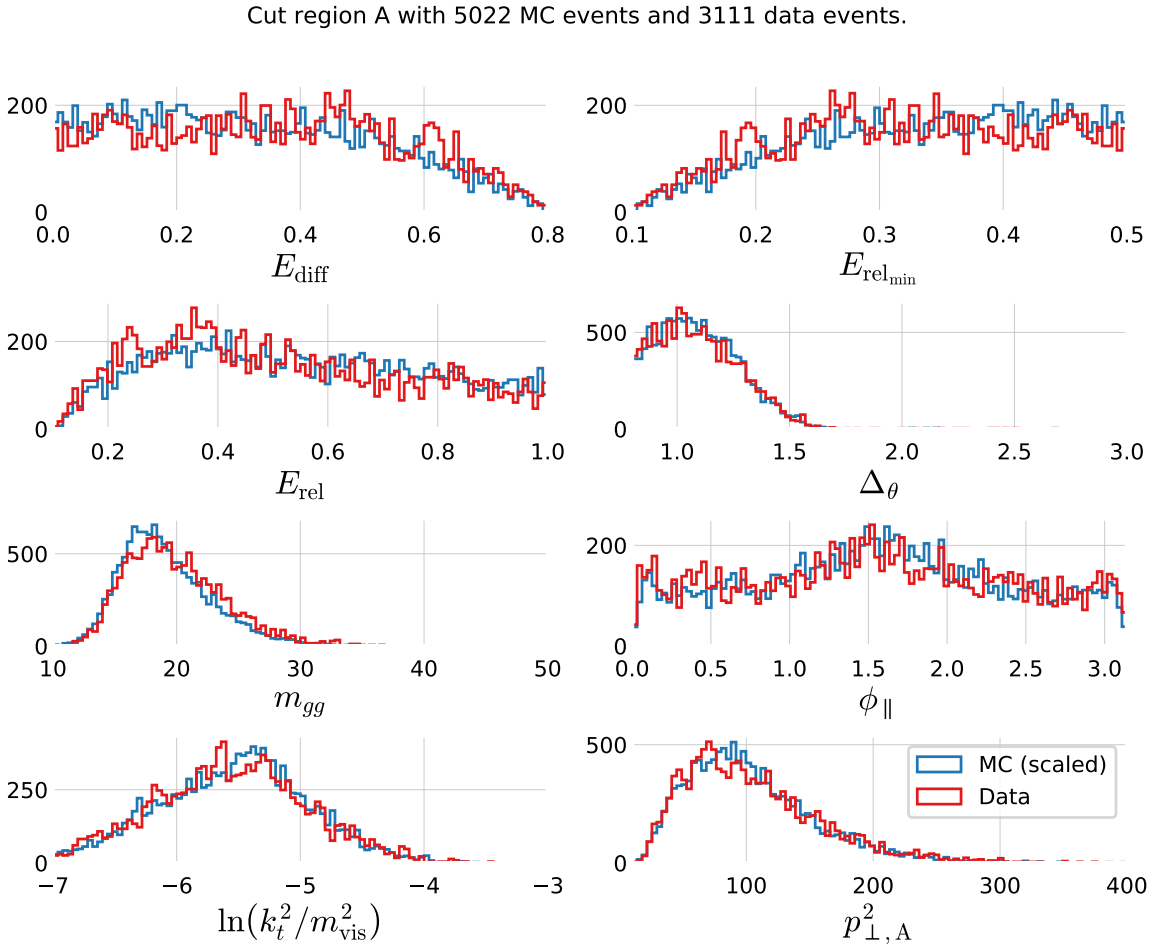


Figure B.58: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area A, see Table **??**. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area A which has 5022 events in the MC sample and 3111 in the Data sample.

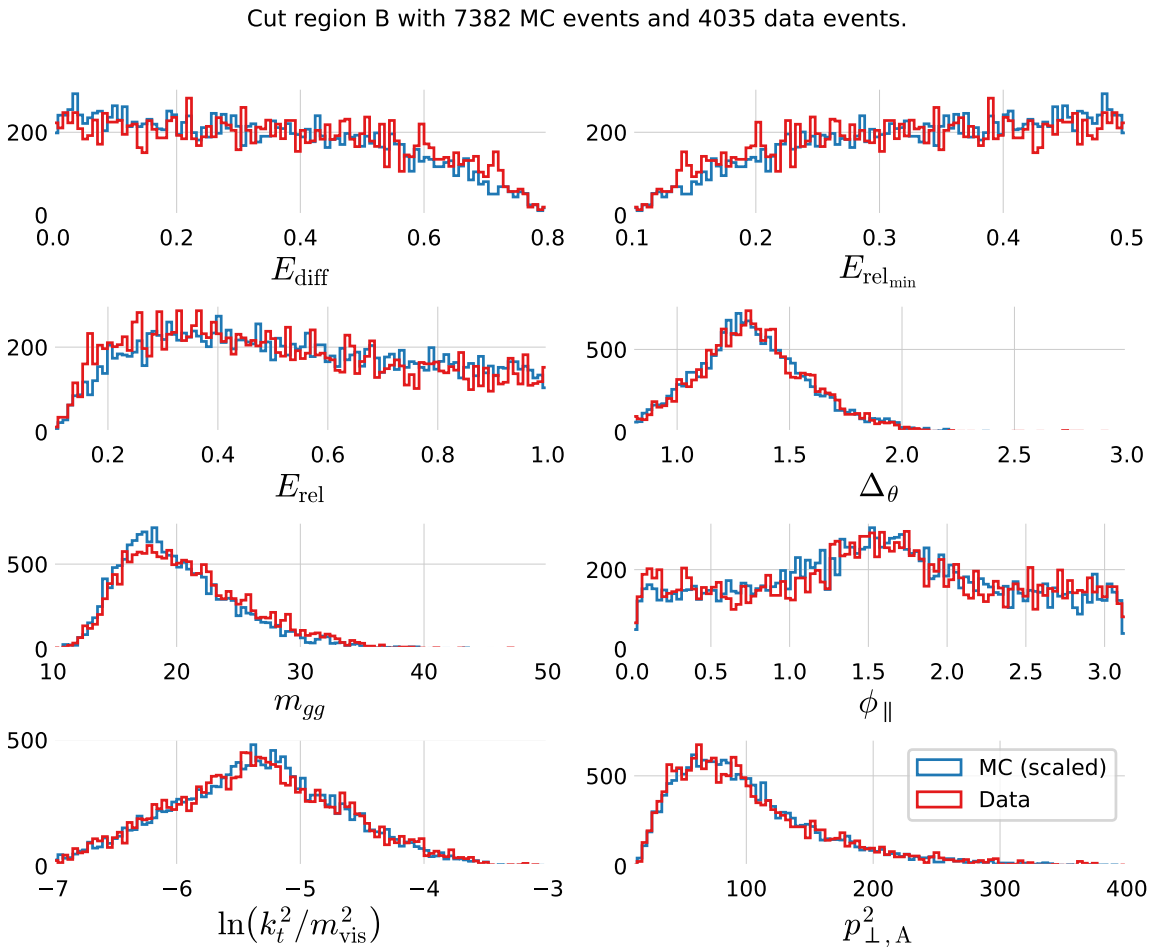


Figure B.59: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area B, see Table **??**. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area B which has 7382 events in the MC sample and 4035 in the Data sample.

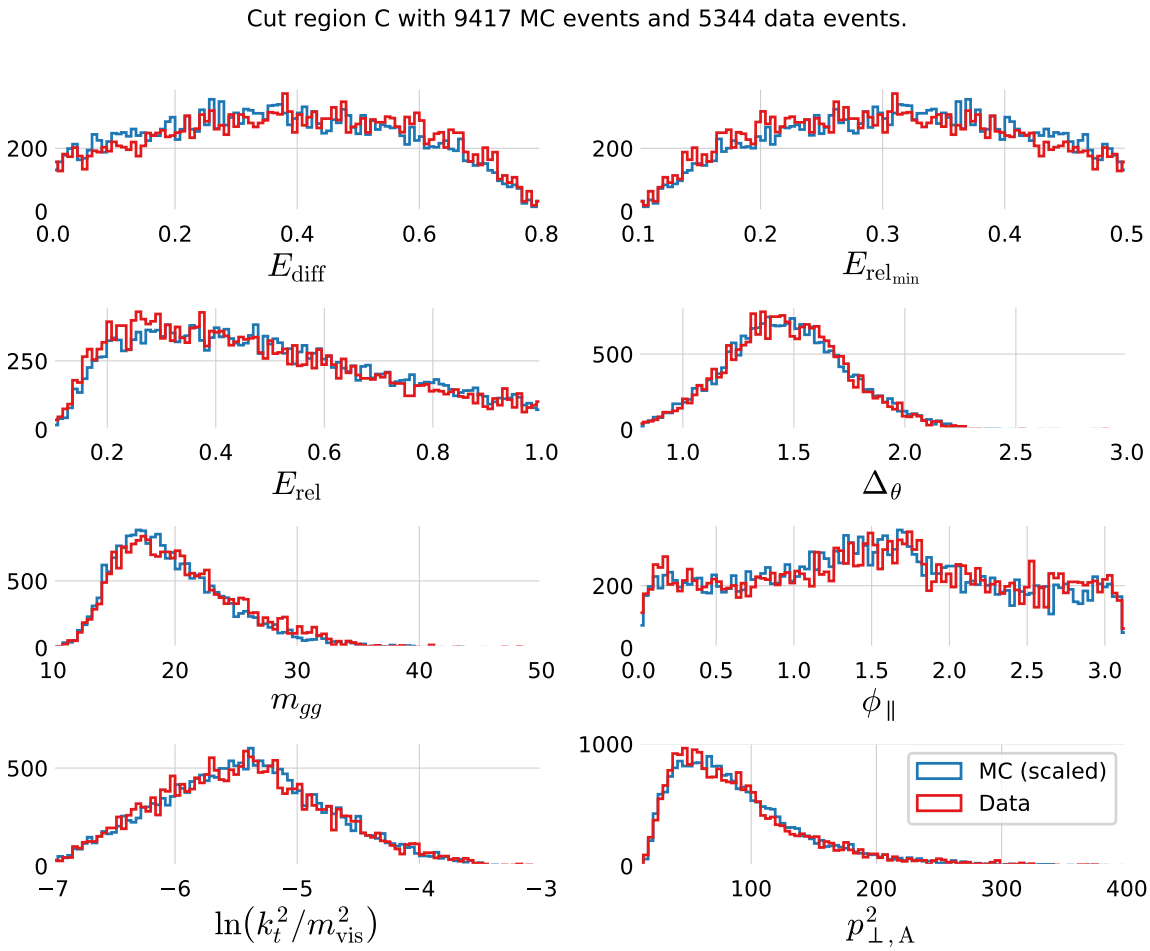


Figure B.6o: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area C, see Table **??**. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area C which has 9417 events in the MC sample and 5344 in the Data sample.

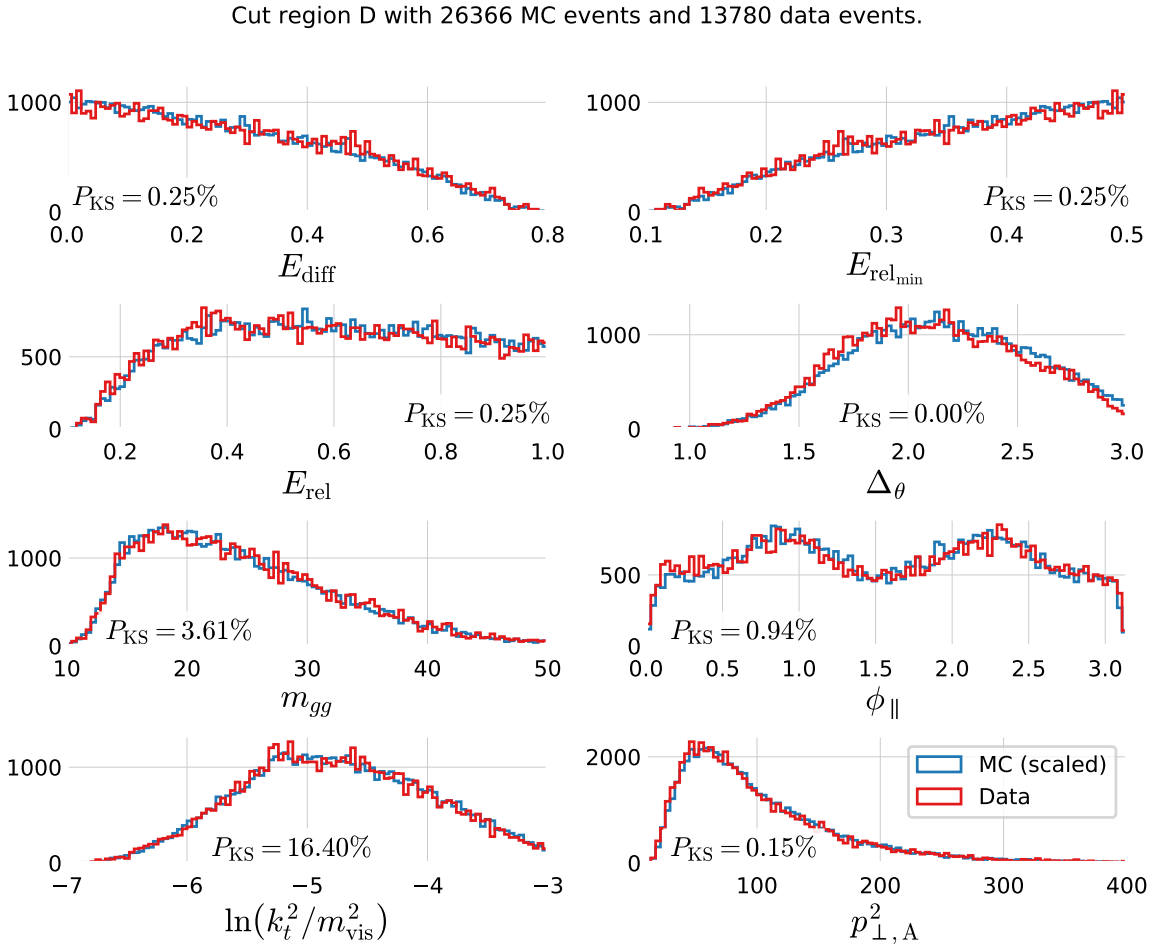


Figure B.61: Comparison of the gluon splitting distributions in MC and Data for $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area D, see Table **??**. The distribution for MC (scaled to Data) is shown in blue and for Data in red. These eight distributions are for the $R_{gg}^{k_t}$-R_{gg}^{CA} Phase Space Area D which has 26 366 events in the MC sample and 13 780 in the Data sample.

List of Figures

List of Tables

Bibliography

[1] Advanced Topics in Machine Learning (ATML). URL https://kurser.ku.dk/course/ndak15014u.

[2] Allstate Claims Severity - Fair Loss. URL https://kaggle.com/c/allstate-claims-severity.

[3] Dmlc/xgboost. URL https://github.com/dmlc/xgboost.

[4] HEP meets ML award | The Higgs Machine Learning Challenge. URL https://higgsml.lal.in2p3.fr/prizes-and-award/award/.

[5] The Large Electron-Positron Collider | CERN. URL https://home.cern/science/accelerators/large-electron-positron-collider.

[6] Microsoft/LightGBM. URL https://github.com/microsoft/LightGBM/blob/b3975555d7023fd05f8e56326905fe7b185109de/src/treelearner/serial_tree_learner.cpp#L282.

[7] Scikit-hep/uproot. URL https://github.com/scikit-hep/uproot.

[8] Datashader: Revealing the Structure of Genuinely Big Data. URL https://github.com/holoviz/datashader.

[9] O. . Www.OIS.dk - Din genvej til ejendomsdata. URL https://www.ois.dk/.

[10] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems. URL http://tensorflow.org/.

[11] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning From Data*. AMLBook. ISBN 978-1-60049-006-4.

[12] D. Albanese, S. Riccadonna, C. Donati, and P. Franceschi. A practical tool for maximal information coefficient analysis. 7. ISSN 2047-217X. doi: 10.1093/gigascience/giy032. URL https://doi.org/10.1093/gigascience/giy032.

[13] H. Albrecht, H. Ehrlichmann, T. Hamacher, R. P. Hofmann, T. Kirchhoff, A. Nau, S. Nowak, H. Schröder, H. D. Schulz, M. Walter, R. Wurth, C. Hast, H. Kolanoski, A. Kosche,

A. Lange, A. Lindner, R. Mankel, M. Schieber, T. Siegmund, B. Spaan, H. Thurn, D. Töpfer, D. Wegener, M. Bittner, P. Eckstein, M. Paulini, K. Reim, H. Wegener, R. Eckmann, R. Mundt, T. Oest, R. Reiner, W. Schmidt-Parzefall, W. Funk, J. Stiewe, S. Werner, K. Ehret, W. Hofmann, A. Hüpper, S. Khan, K. T. Knöpfle, M. Seeger, J. Spengler, D. I. Britton, C. E. K. Charlesworth, K. W. Edwards, E. R. F. Hyatt, H. Kapitza, P. Krieger, D. B. MacFarlane, P. M. Patel, J. D. Prentice, P. R. B. Saull, K. Tzamariudaki, R. G. Van de Water, T. S. Yoon, D. Reßing, M. Schmidtler, M. Schneider, K. R. Schubert, K. Strahl, R. Waldi, S. Weseler, G. Kernel, P. Križnič, T. Podobnik, T. Živko, V. Balagura, I. Belyaev, S. Chechelnitsky, M. Danilov, A. Droutskoy, Y. Gershtein, A. Golutvin, G. Kostina, D. Litvintsev, V. Lubimov, P. Pakhlov, F. Ratnikov, S. Semenov, A. Snizhko, V. Soloshenko, I. Tichomirov, and Y. Zaitsev. A model-independent determination of the inclusive semileptonic decay fraction of B mesons. 318(2): 397–404. ISSN 0370-2693. doi: 10.1016/0370-2693(93)90146-9. URL http://www.sciencedirect.com/science/article/pii/0370269393901469.

[14] E. Anderson. The Species Problem in Iris. 23(3):457–509. ISSN 00266493. doi: 10.2307/2394164. URL www.jstor.org/stable/2394164.

[15] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand. Parton fragmentation and string dynamics. 97(2): 31–145. ISSN 0370-1573. doi: 10.1016/0370-1573(83)90080-7. URL http://www.sciencedirect.com/science/article/pii/0370157383900807.

[16] S. R. Armstrong. A Search for the standard model Higgs boson in four jet final states at center-of-mass energies near 183-GeV with the ALEPH detector at LEP. URL http://wwwlib.umi.com/dissertations/fullcit?p9910371.

[17] M. Awad and R. Khanna. Support Vector Regression. In M. Awad and R. Khanna, editors, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pages 67–80. Apress. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_4. URL https://doi.org/10.1007/978-1-4302-5990-9_4.

[18] R. J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series)*. WileyBlackwell, reprint edition. ISBN 0-471-92295-1. URL http://www.amazon.co.uk/Statistics-Statistical-Physical-Sciences-Manchester/dp/0471922951%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0471922951.

[19] J. T. Barron. A General and Adaptive Robust Loss Function. URL http://arxiv.org/abs/1701.03077.

[20] W. Bartel et al. Experimental study of jets in electron-positron annihilation. 101(1):129–134. ISSN 0370-2693. doi: 10.1016/0370-2693(81)90505-0. URL http://www.sciencedirect.com/science/article/pii/0370269381905050.

[21] E. Becht, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Evaluation of UMAP as an alternative to t-SNE for single-cell data. page 298430, . doi: 10.1101/298430. URL https://www.biorxiv.org/content/10.1101/298430v1.

[22] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. 37 (1):38–44, . ISSN 1546-1696. doi: 10.1038/nbt.4314. URL https://www.nature.com/articles/nbt.4314.

[23] J. Bergstra and Y. Bengio. Random Search for Hyper-parameter Optimization. 13:281–305. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2188385.2188395.

[24] C. Bierlich. Rope hadronization, geometry and particle production in pp and pA Collisions. URL https://lup.lub.lu.se/search/ws/files/18474576/thesis.pdf.

[25] Bolighed. Bolighed - usikkerhed i data-vurderingen. URL https://bolighed.dk/om-bolighed/spoergsmaal-og-svar/#boligvaerdi.

[26] L. Breiman. Random Forests. 45(1):5–32. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

[27] R. P. Brent. *Algorithms for Minimization without Derivatives*. Prentice-Hall Series in Automatic Computation. Prentice-Hall. URL https://cds.cern.ch/record/113464.

[28] E. Brochu, V. M. Cora, and N. de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. URL http://arxiv.org/abs/1012.2599.

[29] R. Brun and F. Rademakers. ROOT — An object oriented data analysis framework. 389(1):81–86. ISSN 0168-9002. doi: 10.1016/S0168-9002(97)00048-X. URL http://www.sciencedirect.com/science/article/pii/S016890029700048X.

[30] A. Buckley, J. Butterworth, S. Gieseke, D. Grellscheid, S. Hoche, H. Hoeth, F. Krauss, L. Lonnblad, E. Nurse, P. Richardson, S. Schumann, M. H. Seymour, T. Sjostrand,

P. Skands, and B. Webber. General-purpose event generators for LHC physics. 504(5):145–233. ISSN 03701573. doi: 10.1016/j.physrep.2011.03.005. URL http://arxiv.org/abs/1101.2599.

[31] C. Burgard. Standard model of physics | TikZ example. URL http://www.texample.net/tikz/examples/model-physics/.

[32] D. Buskulic et al. An investigation of Bdo and Bso oscillation. 322(4):441–458. ISSN 0370-2693. doi: 10.1016/0370-2693(94)91177-0. URL http://www.sciencedirect.com/science/article/pii/0370269394911770.

[33] M. Cacciari, G. P. Salam, and G. Soyez. FastJet user manual. 72(3):1896. ISSN 1434-6044, 1434-6052. doi: 10.1140/epjc/s10052-012-1896-2. URL http://arxiv.org/abs/1111.6097.

[34] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber. Longitudinally-invariant kt-clustering algorithms for hadron-hadron collisions. 406(1):187–224, . ISSN 0550-3213. doi: 10.1016/0550-3213(93)90166-M. URL http://www.sciencedirect.com/science/article/pii/055032139390166M.

[35] S. Catani, G. Turnock, and B. R. Webber. Jet broadening measures in e+ e- annihilation. B295:269–276, . doi: 10.1016/0370-2693(92)91565-Q.

[36] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. pages 785–794. doi: 10.1145/2939672.2939785. URL http://arxiv.org/abs/1603.02754.

[37] A. Collaboration. Electron efficiency measurements with the ATLAS detector using 2012 LHC proton-proton collision data. 77(3):195, . ISSN 1434-6044, 1434-6052. doi: 10.1140/epjc/s10052-017-4756-2. URL http://arxiv.org/abs/1612.01456.

[38] C. Collaboration. Search for a Higgs boson in the decay channel H to ZZ(*) to q qbar l-l+ in pp collisions at sqrt(s) = 7 TeV. 2012(4):36, . ISSN 1029-8479. doi: 10.1007/JHEP04(2012)036. URL http://arxiv.org/abs/1202.1416.

[39] T. A. Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. 716(1):1–29, . ISSN 03702693. doi: 10.1016/j.physletb.2012.08.020. URL http://arxiv.org/abs/1207.7214.

[40] T. C. Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. 716(1):30–61, . ISSN 03702693. doi: 10.1016/j.physletb.2012.08.021. URL http://arxiv.org/abs/1207.7235.

[41] M. Dam. An upper limit for Br(Z->ggg) from symmetric 3–jet zo hadronic decays. 389(2):405–415. ISSN 0370-2693. doi: 10.1016/S0370-2693(96)01450-5.

[42] A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, and S. Gravel. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. 15(11). ISSN 1553-7390. doi: 10.1371/journal.pgen.1008432. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6853336/.

[43] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber. Better Jet Clustering Algorithms. 1997(08):001–001. ISSN 1029-8479. doi: 10.1088/1126-6708/1997/08/001. URL http://arxiv.org/abs/hep-ph/9707323.

[44] S. D. DST. Price Index (EJ14) - Statistics Denmark. URL https://www.dst.dk/en/Statistik/emner/priser-og-forbrug/ejendomme.

[45] S. D. Ellis and D. E. Soper. Successive combination jet algorithm for hadron collisions. 48(7):3160–3166. doi: 10.1103/PhysRevD.48.3160. URL https://link.aps.org/doi/10.1103/PhysRevD.48.3160.

[46] D. et al. Buskulic. A precise measurement of hadrons. 313(3): 535–548. ISSN 0370-2693. doi: 10.1016/0370-2693(93)90028-G. URL http://www.sciencedirect.com/science/article/pii/037026939390028G.

[47] F. Faye. Frederik Faye / deepcalo. URL https://gitlab.com/ffaye/deepcalo.

[48] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. 7(2):179–188. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x.

[49] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In P. Vitányi, editor, *Computational Learning Theory*, pages 23–37. Springer Berlin Heidelberg. ISBN 978-3-540-49195-8. AdaBoost.

[50] S. L. Glashow. Partial-symmetries of weak interactions. 22(4): 579–588. ISSN 0029-5582. doi: 10.1016/0029-5582(61)90469-2. URL http://www.sciencedirect.com/science/article/pii/0029558261904692.

[51] P. Gras, S. Höche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer, A. Siódmok, P. Skands, G. Soyez, and J. Thaler. Systematics of quark/gluon tagging. 2017(7):91. ISSN 1029-8479. doi: 10.1007/JHEP07(2017)091. URL http://arxiv.org/abs/1704.03878.

[52] N. Guttenberg, N. Virgo, O. Witkowski, H. Aoki, and R. Kanai. Permutation-equivariant neural networks applied to dynamics prediction. URL http://arxiv.org/abs/1612.04530.

[53] A. E. Harvey and S. Peters. Estimation Procedures for Structural Time Series Models. doi: 10.1002/for.3980090203.

[54] T. Hastie and R. Tibshirani. Generalized Additive Models: Some Applications. 82(398):371–386. ISSN 01621459. doi: 10.2307/2289439. URL www.jstor.org/stable/2289439.

[55] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag, 2 edition. ISBN 978-0-387-84857-0. URL //www.springer.com/la/book/9780387848570.

[56] K. Hornik. Approximation capabilities of multilayer feedforward networks. 4(2):251–257. ISSN 0893-6080. doi: 10.1016/0893-6080(91)90009-T. URL http://www.sciencedirect.com/science/article/pii/089360809190009T.

[57] P. Huber and E. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley. ISBN 978-1-118-21033-8. URL https://books.google.dk/books?id=j1OhquR_j88C.

[58] S. Hviid, Juul. Working Paper: A regional model of the Danish housing market. URL http://www.nationalbanken.dk/en/publications/Pages/2017/11/Working-Paper-A-regional-model-of-the-Danish-housing-market.aspx.

[59] F. James and M. Roos. Minuit – a system for function minimization and analysis of the parameter errors and correlations. 10:343–367. doi: 10.1016/0010-4655(75)90039-9.

[60] R. E. Kalman. A new approach to linear filtering and prediction problems. 82:35–45.

[61] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc. URL http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

[62] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. URL http://arxiv.org/abs/1412.6980.

[63] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-Normalizing Neural Networks. URL http://arxiv.org/abs/1706.02515.

[64] A. J. Larkoski, J. Thaler, and W. J. Waalewijn. Gaining (Mutual) Information about Quark/Gluon Discrimination. 2014 (11). ISSN 1029-8479. doi: 10.1007/JHEP11(2014)129. URL http://arxiv.org/abs/1408.3122.

[65] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. 49(4): 764–766. ISSN 0022-1031. doi: 10.1016/j.jesp.2013.03.013. URL http://www.sciencedirect.com/science/article/pii/S0022103113000668.

[66] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL http://dl.acm.org/citation.cfm?id=3295222.3295230.

[67] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles - SHAP. . URL http://arxiv.org/abs/1802.03888.

[68] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. . URL http://arxiv.org/abs/1802.03888.

[69] A. L. Maas. Rectifier nonlinearities improve neural network acoustic models.

[70] L. McInnes and J. Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. URL http://arxiv.org/abs/1802.03426.

[71] T. C. Mills. *Time Series Techniques for Economists / Terence c. Mills*. Cambridge University Press Cambridge [England] ; New York. ISBN 0-521-34339-9 0-521-40574-2. URL http://www.loc.gov/catdir/toc/cam031/89007187.html.

[72] N. Mohd Razali and B. Yap. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. 2.

[73] I. Mulalic, H. Rasmussen, J. Rouwendal, and H. H. Woltmann. The Financial Crisis and Diverging House Prices: Evidence from the Copenhagen Metropolitan Area. ISSN 1556-5068. doi: 10.2139/ssrn.3041272. URL https://www.ssrn.com/abstract=3041272.

[74] Particle Data Group et al. Review of Particle Physics. 98 (3):030001. doi: 10.1103/PhysRevD.98.030001. URL https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. 12:2825–2830.

[76] E. Polley and M. van der Laan. Super Learner In Prediction. URL https://biostats.bepress.com/ucbbiostat/paper266.

[77] J. Proriol, J. Jousset, C. Guicheney, A. Falvard, P. Henrard, D. Pallin, P. Perret, and B. Brandl. TAGGING B QUARK EVENTS IN ALEPH WITH NEURAL NETWORKS (comparison of different methods : Neural Networks and Discriminant Analysis). page 27.

[78] A. Purcell. Go on a particle quest at the first CERN webfest. URL https://cds.cern.ch/record/1473657.

[79] S. Ravanbakhsh, J. Schneider, and B. Poczos. Deep Learning with Sets and Point Clouds. URL http://arxiv.org/abs/1611.04500.

[80] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting Novel Associations in Large Data Sets. 334(6062):1518–1524. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1205438. URL http://science.sciencemag.org/content/334/6062/1518.

[81] J. W. Rohlf. *Modern Physics from A to Z*. John Wiley and Sons. ISBN 978-0-471-57270-1.

[82] P. J. Rousseeuw and C. Croux. Alternatives to the Median Absolute Deviation. 88(424):1273–1283. ISSN 0162-1459. doi: 10.1080/01621459.1993.10476408. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476408.

[83] A. Salam. Weak and electromagnetic interactions. In *Selected Papers of Abdus Salam*, volume Volume 5 of *World Scientific Series in 20th Century Physics*, pages 244–254. WORLD SCIENTIFIC. ISBN 978-981-02-1662-7. doi: 10.1142/9789812795915_0034. URL https://www.worldscientific.com/doi/abs/10.1142/9789812795915_0034.

[84] S. Schmitt. Data Unfolding Methods in High Energy Physics. 137:11008. ISSN 2100-014X. doi: 10.1051/epjconf/201713711008. URL http://arxiv.org/abs/1611.01927.

[85] L. Scodellaro. B tagging in ATLAS and CMS. URL http://arxiv.org/abs/1709.01290.

[86] L. Shapley. A value for n-person games. In *The Shapley Value*, volume 28 of *Annals of Math Studies*, pages 307–317. doi: 10.1017/CBO9780511528446.003.

[87] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands. An Introduction to PYTHIA 8.2. 191:159–177. ISSN 00104655. doi: 10.1016/j.cpc.2015.01.024. URL http://arxiv.org/abs/1410.3012.

[88] P. Skands. Peter Skands.

[89] S. J. Taylor and B. Letham. Forecasting at scale. doi: 10.7287/peerj.preprints.3190v2. URL https://peerj.com/preprints/3190.

[90] i. team. Iminuit – A python interface to minuit. URL https://github.com/scikit-hep/iminuit.

[91] J. Thaler. Report of the Les Houches Quark/Gluon Subgroup. (1):28.

[92] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. 58(1):267–288. ISSN 0035-9246. URL www.jstor.org/stable/2346178.

[93] A. Tikhonov. *On the Stability of Inverse Problems*, volume vol. 39 of *Doklady Akademii Nauk SSSR*.

[94] S. Toghi Eshghi, A. Au-Yeung, C. Takahashi, C. R. Bolen, M. N. Nyachienga, S. P. Lear, C. Green, W. R. Mathews, and W. E. O'Gorman. Quantitative Comparison of Conventional and t-SNE-guided Gating Analyses. 10. ISSN 1664-3224. doi: 10.3389/fimmu.2019.01194. URL https://www.frontiersin.org/articles/10.3389/fimmu.2019.01194/full.

[95] d. L. M. J. van, E. C. Polley, and A. E. Hubbard. Super Learner. 6(1). ISSN 1544-6115. doi: 10.2202/1544-6115.1309. URL https://www.degruyter.com/view/j/sagmb.2007.6.issue-1/sagmb.2007.6.1.1309/sagmb.2007.6.1.1309.xml.

[96] J. J. van der Bij and E. W. N. Glover. Z boson production and decay via gluons. 313(2):237–257. ISSN 0550-3213. doi: 10.1016/0550-3213(89)90317-9. URL http://www.sciencedirect.com/science/article/pii/0550321389903179.

[97] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. 9:2579–2605. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[98] F. van Veen. The Neural Network Zoo. URL http://www.asimovinstitute.org/neural-network-zoo/.

[99] V. Vapnik. Principles of Risk Minimization for Learning Theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, pages 831–838. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-222-9. URL http://dl.acm.org/citation.cfm?id=2986916.2987018.

[100] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore,

J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. page arXiv:1907.10121. URL https://ui.adsabs.harvard.edu/abs/2019arXiv190710121V/abstract.

[101] I. Wallach and R. Lilien. The protein–small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. 25(5):615–620. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp035. URL https://academic.oup.com/bioinformatics/article/25/5/615/183421.

[102] S. Weinberg. A Model of Leptons. 19(21):1264–1266. doi: 10.1103/PhysRevLett.19.1264. URL https://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[103] H. Wickham. Tidy data. 59(10):1–23. ISSN 1548-7660. doi: 10.18637/jss.v059.i10. URL https://www.jstatsoft.org/v059/i10.

[104] M. Wobisch and T. Wengler. Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering. URL http://arxiv.org/abs/hep-ph/9907280.

[105] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola. Deep Sets. URL http://arxiv.org/abs/1703.06114.