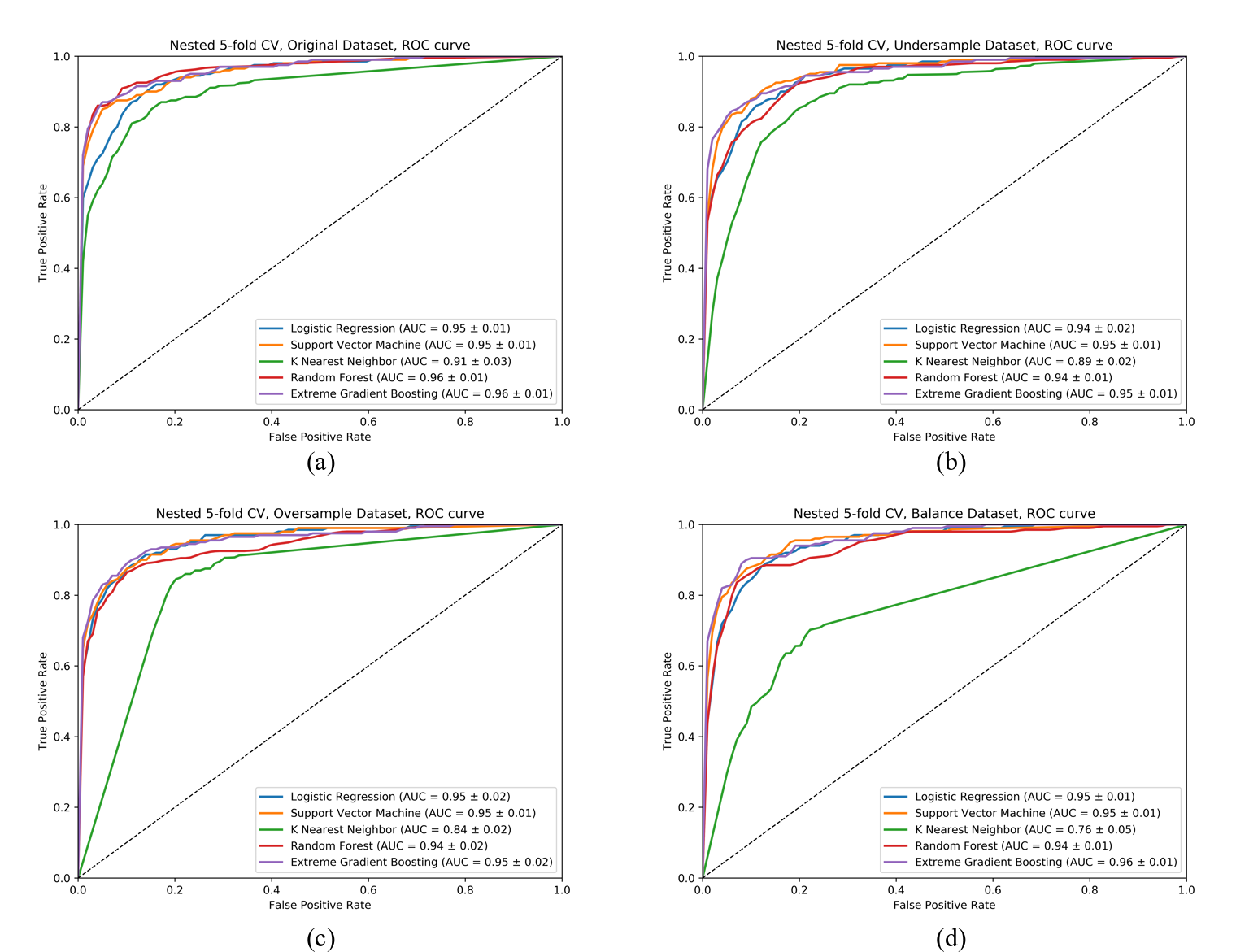
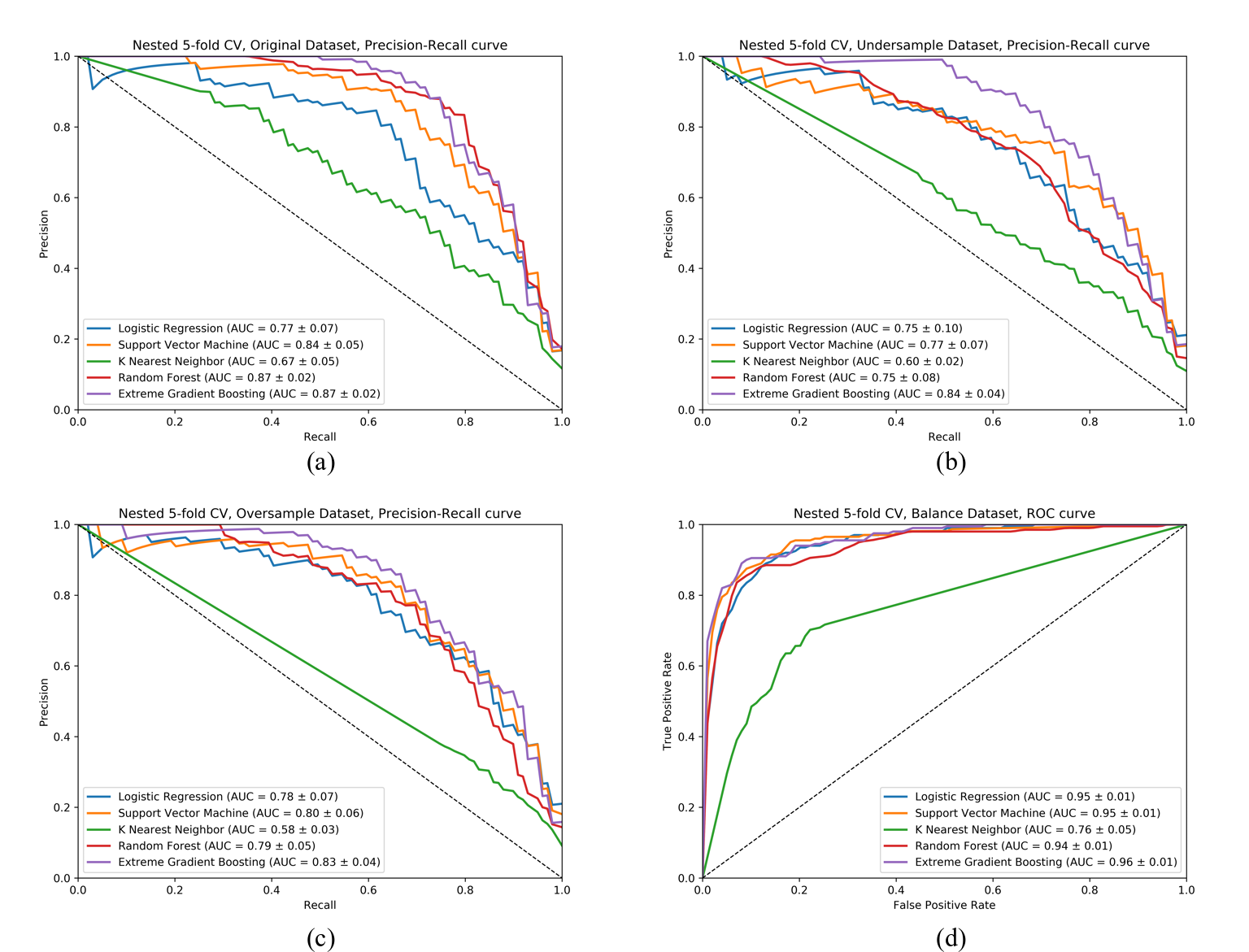


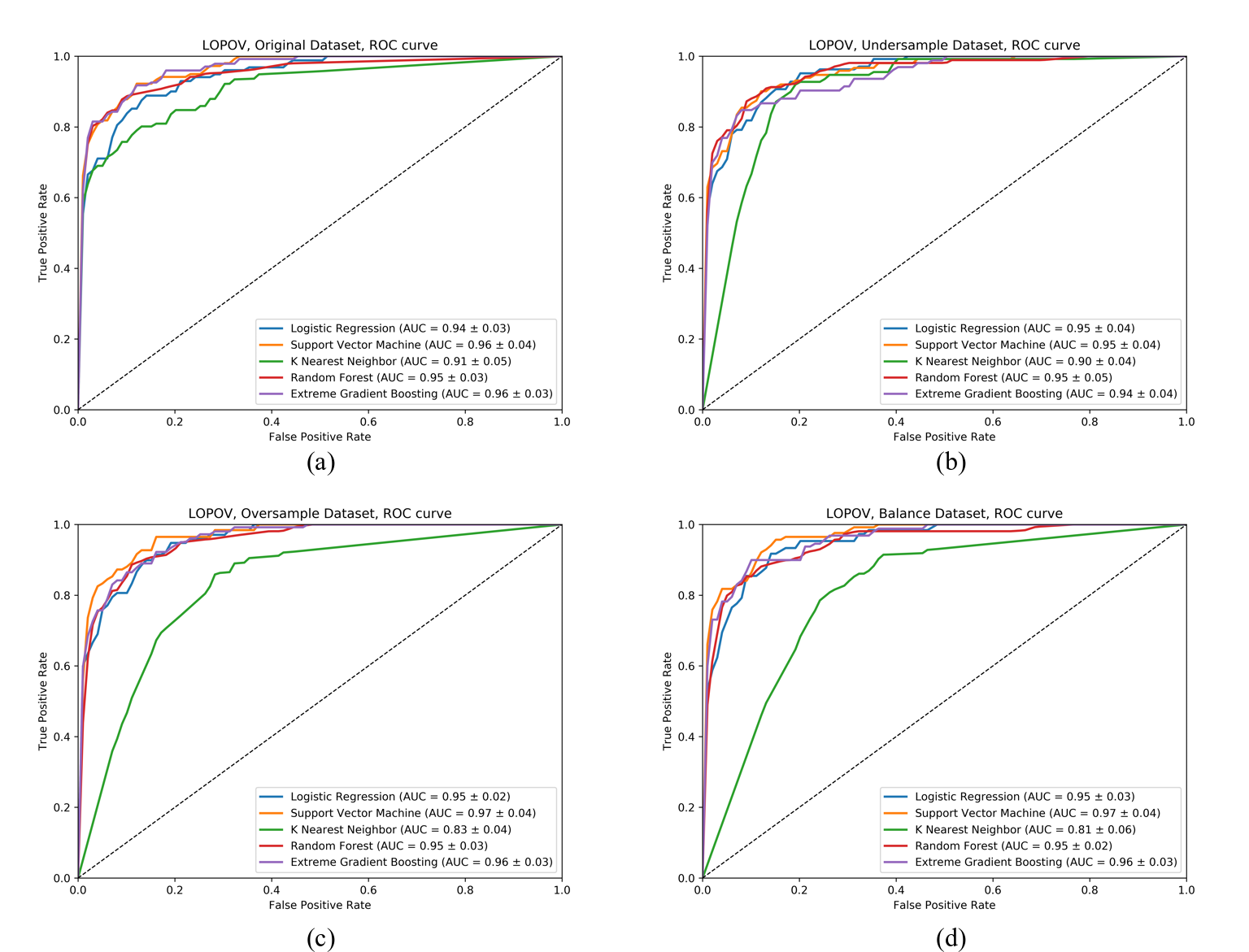
**Supplemental Figure S1.** Workflow of selecting negative samples. The positive proteins (PAgs) were downloaded from the Protegen database and the whole pathogen proteomes were downloaded from the Uniprot database. The negative proteins were selected based on its sequence dis-similarity to the PAgs, as described above. Both positive and negative proteins were tested for homology and proteins with sequence similarity over 30% were removed from the study.



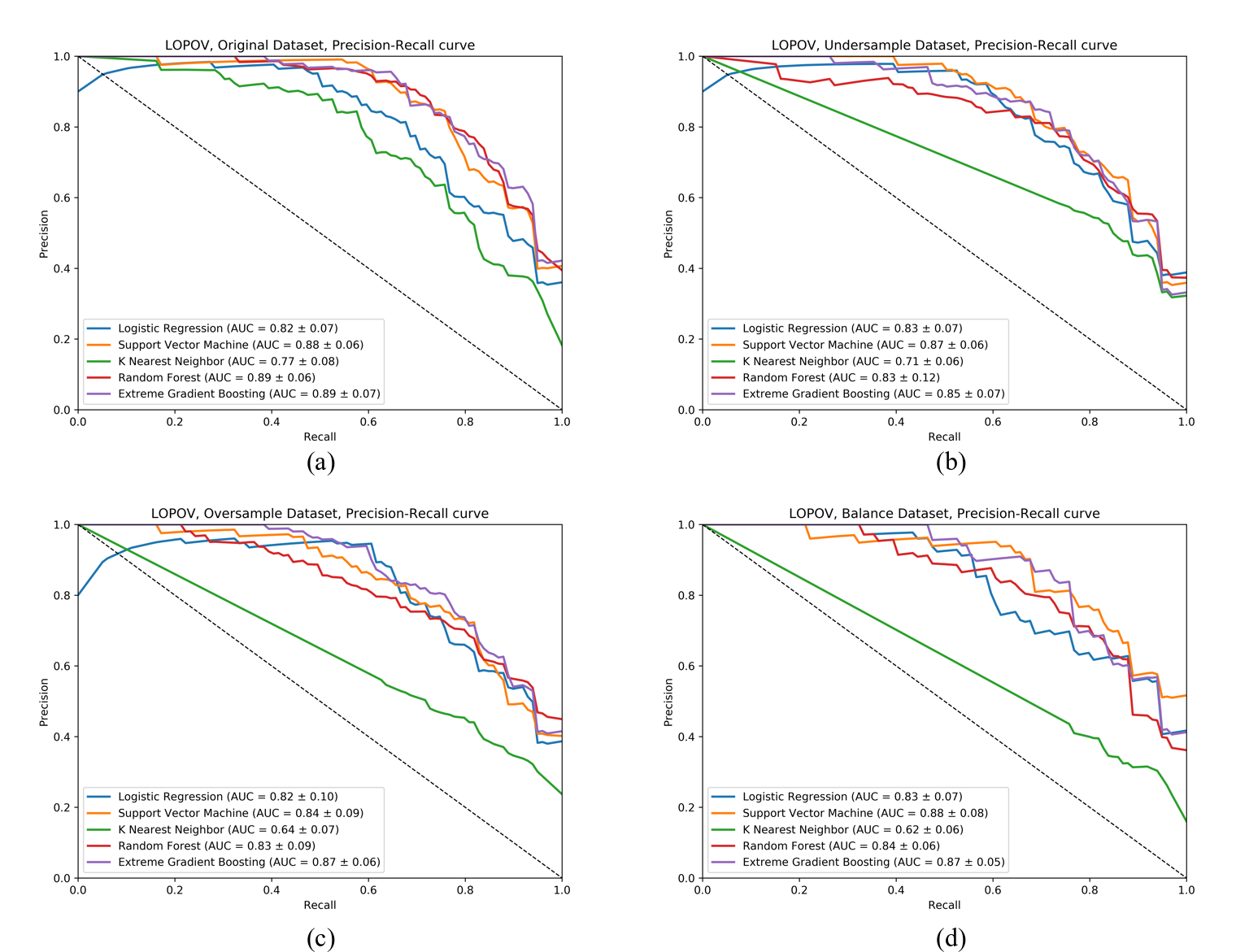
**Supplemental Figure S2.** The average ROC curves in nested five-fold cross validation of five machine learning algorithms (logistic regression, support vector machine, k nearest neighbor, random forest, and extreme gradient boosting) trained on (a) original data; (b) under-sampled data; (c) over-sampled data; and (d) balanced data.



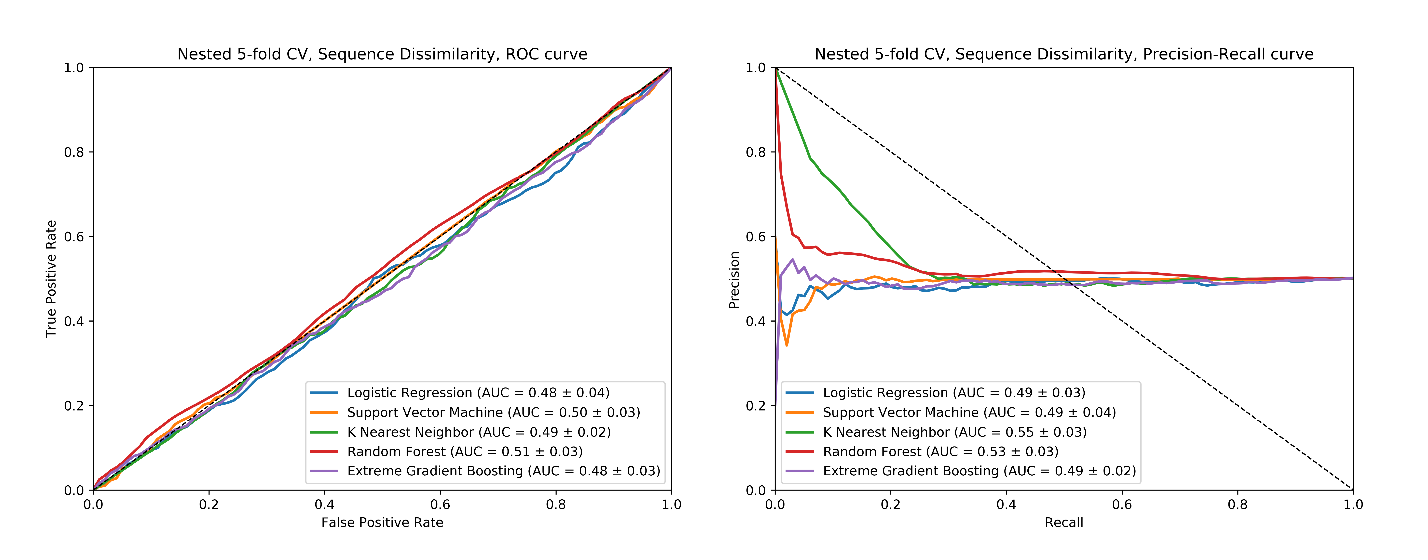
**Supplemental Figure S3.** The average precision-recall curves in nested five-fold cross validation of five machine learning algorithms (logistic regression, support vector machine, k nearest neighbor, random forest, and extreme gradient boosting) trained on (a) original data; (b) under-sampled data; (c) over-sampled data; and (d) balanced data.



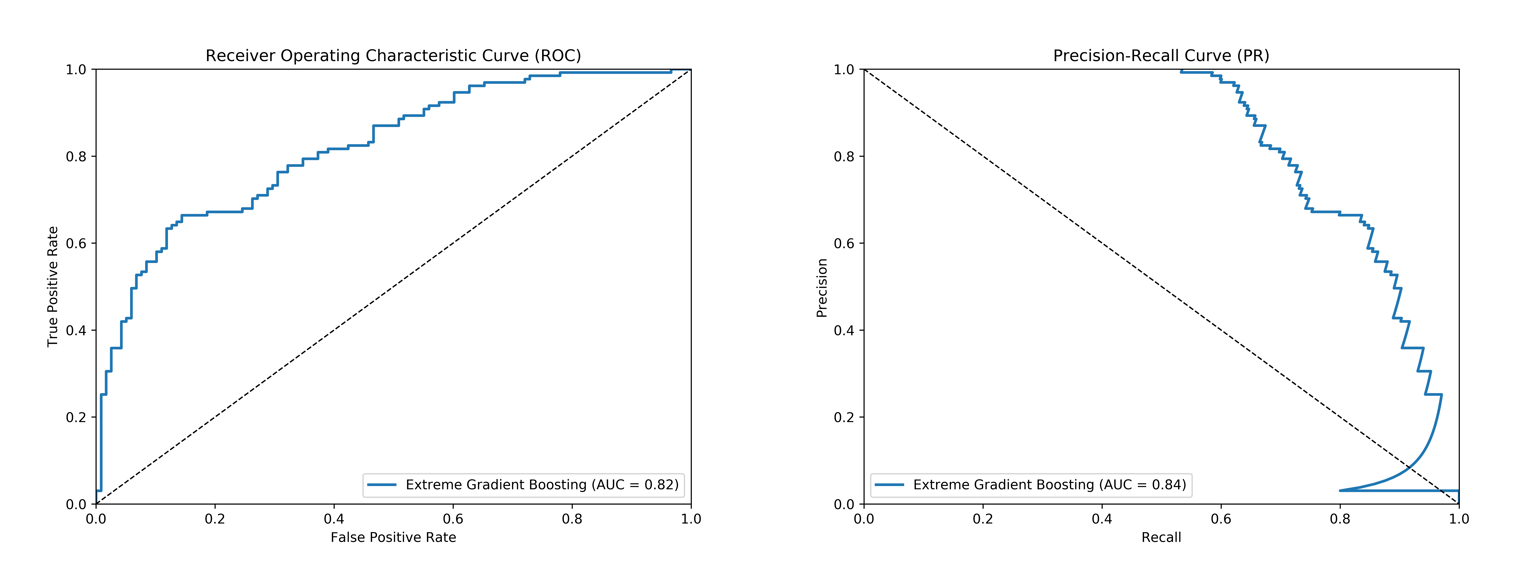
**Supplemental Figure S4.** The average ROC curves in leave-one-pathogen-out validation of five machine learning algorithms (logistic regression, support vector machine, k nearest neighbor, random forest, and extreme gradient boosting) trained on (a) original data; (b) under-sampled data; (c) over-sampled data; and (d) balanced data.



**Supplemental Figure S5.** The average precision-recall curves in leave-one-pathogen-out validation of five machine learning algorithms (logistic regression, support vector machine, k nearest neighbor, random forest, and extreme gradient boosting) trained on (a) original data; (b) under-sampled data; (c) over-sampled data; and (d) balanced data.



**Supplemental Figure S6.** The ROC and precision-recall curves (PRC) of the discrimination between two sets of non-antigen proteins with less than 30% sequence identity in nested five-fold cross validation five machine learning algorithms (extreme gradient boosting trained on the entire training data with feature selection and hyperparameter optimization). The area under ROC and PRC were approximately 0.5, which is equivalent to random prediction. Therefore, the discriminative power of the presented Vaxign-ML was indeed depended on the immunogenic potential of the protective antigens collected in the Protegen database rather than sequence dis-similarity.



**Supplemental Figure S7.** The ROC and precision-recall curves of the Vaxign-ML (extreme gradient boosting trained on the entire training data with feature selection and hyperparameter optimization).

**Supplemental Table S1.** All bioinformatics programs used to compute peptide properties in this study.

|  |  |  |
| --- | --- | --- |
| **Sequence-derived Features** | **Program** | **Download/Website Link** |
| Subcellular localization | psortB(1) | http://www.psort.org/psortb/ |
| Adhesin probability | SPAAN(2) | https://sourceforge.net/projects/adhesin/files/SPAAN/ |
| Signal peptide | SignalP(3) | http://www.cbs.dtu.dk/services/SignalP/ |
| Transmembrane alpha-helix | TMHMM(4) | http://www.cbs.dtu.dk/services/TMHMM/ |
| Immunogenicity | IEDB (5) | http://tools.iedb.org/immunogenicity/ |
| Physicochemical properties | Propy(8) | http://eggnogdb.embl.de/#/app/home |
| **References:** |  |  |
| (1) Yu, N. Y. et al. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26, 1608–1615 (2010).  (2) Sachdeva, G., Kumar, K., Jain, P. & Ramachandran, S. SPAAN: A software program for prediction of adhesins and adhesin-like proteins using neural networks. Bioinformatics 21, 483–491 (2005).  (3) Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785–786 (2011).  (4) Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. 2, 953–971 (2007).  (5) Calis,J.J.A. et al. (2013) Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. PLoS Comput. Biol., 9.  (6) Cao,D.S. et al. (2013) Propy: A tool to generate various modes of Chou’s PseAAC. Bioinformatics, 29, 960–962. | | |

**Supplemental Table S2.** Protegen pathogens’ identifiers, taxonomies and the corresponding Uniprot proteomes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gram | pathogen\_id | taxon\_id | mapped\_taxon\_id | uniprot\_proteome\_id |
| + | 138 | 1280 | 93061 | UP000008816 |
| + | 139 | 1311 | 226185 | UP000001415 |
| + | 47 | 1313 | 171101 | UP000000586 |
| + | 140 | 1314 | 301447 | UP000000750 |
| + | 81 | 1336 | 553483 | UP000001368 |
| + | 7 | 1392 | 1392 | UP000000594 |
| + | 41 | 1485 | 212717 | UP000001412 |
| + | 14 | 1491 | 441771 | UP000001986 |
| + | 136 | 1639 | 169963 | UP000000817 |
| + | 152 | 1648 | 650150 | UP000007944 |
| + | 62 | 1717 | 257309 | UP000002198 |
| + | 290 | 1765 | 83332 | UP000001584 |
| + | 29 | 1773 | 83332 | UP000001584 |
| + | 246 | 1781 | 216594 | UP000001190 |
| - | 135 | 139 | 224326 | UP000001807 |
| - | 141 | 160 | 243276 | UP000000811 |
| - | 195 | 171 | 189518 | UP000001408 |
| - | 24 | 197 | 192222 | UP000000799 |
| - | 134 | 210 | 85962 | UP000000429 |
| - | 8 | 234 | 359391 | UP000002719 |
| - | 12 | 263 | 177416 | UP000001174 |
| - | 137 | 287 | 208964 | UP000002438 |
| - | 46 | 487 | 122586 | UP000000425 |
| - | 109 | 520 | 520 | UP000047656 |
| - | 42 | 590 | 83333 | UP000000625 |
| - | 43 | 620 | 300267 | UP000002716 |
| - | 11 | 632 | 632 | UP000000815 |
| - | 111 | 633 | 632 | UP000000815 |
| - | 209 | 636 | 634503 | UP000001485 |
| - | 79 | 645 | 380703 | UP000000756 |
| - | 102 | 666 | 666 | UP000036184 |
| - | 301 | 672 | 672 | UP000002675 |
| - | 86 | 715 | 416269 | UP000001432 |
| - | 31 | 727 | 71421 | UP000000579 |
| - | 178 | 738 | 557723 | UP000006743 |
| - | 161 | 747 | 272843 | UP000000809 |
| - | 37 | 777 | 227377 | UP000002671 |
| - | 108 | 780 | 272947 | UP000002480 |
| - | 302 | 783 | 272947 | UP000002480 |
| - | 208 | 813 | 272561 | UP000000431 |
| - | 84 | 2096 | 710127 | UP000001418 |
| - | 200 | 2099 | 262719 | UP000000548 |
| - | 121 | 28450 | 272560 | UP000000605 |
| - | 80 | 55601 | 882102 | UP000006800 |
| - | 25 | 83334 | 83333 | UP000000625 |
| - | 120 | 83554 | 1112253 | UP000014824 |
| - | 82 | 83555 | 272561 | UP000000431 |
| - | 112 | 83558 | 272561 | UP000000431 |
| - | 113 | 83560 | 272561 | UP000000431 |
| - | 130 | 393305 | 632 | UP000000815 |

|  |
| --- |
|  |

|  |
| --- |
|  |
| **Supplemental Table S3.** All proteins (protective and non-protective) in the external independent dataset, 200BPA.  [Please see the attached Table\_S3\_Independent\_Curated.xlsx.] |

**Supplemental Table S4.** Leave-one-pathogen-out evaluation metrics of five machine learning algorithms trained using four data re-sampling methods.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Original Data | | | | Under-sampled | | | | Over-sampled | | | | Balanced | | | |
| AUROC | AUPRC | WF1 | MCC | AUROC | AUPRC | WF1 | MCC | AUROC | AUPRC | WF1 | MCC | AUROC | AUPRC | WF1 | MCC |
| **Logistic Regression (LR)** | | | | | | | | | | | | | | | |
| 0.94 | 0.82 | 0.86 | 0.58 | 0.95 | 0.83 | 0..80 | 0.56 | 0.94 | 0.82 | 0.84 | 0.59 | 0.95 | 0.83 | 0.77 | 0.51 |
| **Support Vector Machine (SVM)** | | | | | | | | | | | | | | | |
| 0.96 | 0.88 | 0.91 | 0.70 | 0.95 | 0.87 | 0.88 | 0.64 | 0.96 | 0.85 | 0.90 | 0.67 | 0.97 | 0.88 | 0.91 | 0.71 |
| **K-Nearest Neighbor (KNN)** | | | | | | | | | | | | | | | |
| 0.91 | 0.77 | 0.90 | 0.65 | 0.9 | 0.71 | 0.71 | 0.43 | 0.84 | 0.67 | 0.74 | 0.44 | 0.81 | 0.62 | 0.76 | 0.45 |
| **Random Forest (RF)** | | | | | | | | | | | | | | | |
| 0.95 | 0.89 | 0.93 | 0.75 | 0.95 | 0.83 | 0.90 | 0.67 | 0.96 | 0.84 | 0.91 | 0.68 | 0.95 | 0.84 | 0.91 | 0.69 |
| **Extreme Gradient Boosting (XGB)** | | | | | | | | | | | | | | | |
| 0.96 | 0.89 | 0.94 | 0.77 | 0.94 | 0.85 | 0.89 | 0.67 | 0.96 | 0.87 | 0.93 | 0.75 | 0.96 | 0.87 | 0.92 | 0.73 |

Abbreviations: AUROC=area under ROC curve; AUPRC=area under precision recall curve; WF1: weighted F1 score; MCC=Matthew’s correlation coefficient.  
The underlined denotes the highest values for each metric.

**Supplemental Table S5.** Nested five-fold cross validation evaluation metrics of five machine learning algorithms trained using two sets of randomly selected non-antigen proteins with less than 30% sequence identity.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Precision | Recall | Weighted\_F1 | MCC |
| Logistic Regression | 0.549359 | 0.625136 | 0.354529 | 0.002837 |
| Support Vector Machine | 0.477664 | 0.468337 | 0.476797 | -0.04588 |
| K Nearest Neighbor | 0.486569 | 0.500497 | 0.485354 | -0.0288 |
| Random Forest | 0.511025 | 0.495523 | 0.51055 | 0.021724 |
| Extreme Gradient Boosting | 0.481015 | 0.480422 | 0.481054 | -0.03777 |

**Supplemental Table S6.** Selected 180 features and its importance in the Vaxign-ML model.

|  |  |  |
| --- | --- | --- |
| Extracellular\_Probability | 0.0162 | subcellular localization |
| CytoplasmicMembrane\_Probability | 0.0232 | subcellular localization |
| Periplasmic\_Probability | 0.0064 | subcellular localization |
| OuterMembrane\_Probability | 0.0056 | subcellular localization |
| SPAAN\_Score | 0.0357 | adhesin probability |
| Total\_N-in\_prob | 0.0047 | transmembrane helix |
| Immunogenicity\_Score | 0.0226 | immunogenicity |
| R | 0.0156 | amino acid composition |
| N | 0.0114 | amino acid composition |
| D | 0.0056 | amino acid composition |
| C | 0.0120 | amino acid composition |
| Q | 0.0056 | amino acid composition |
| H | 0.0117 | amino acid composition |
| I | 0.0047 | amino acid composition |
| L | 0.0089 | amino acid composition |
| K | 0.0265 | amino acid composition |
| M | 0.0092 | amino acid composition |
| F | 0.0053 | amino acid composition |
| W | 0.0045 | amino acid composition |
| Y | 0.0050 | amino acid composition |
| V | 0.0047 | amino acid composition |
| \_PolarizabilityC1 | 0.0114 | polarizability |
| \_SecondaryStrC3 | 0.0045 | secondary structure |
| \_ChargeC3 | 0.0073 | charge |
| \_PolarityC2 | 0.0056 | polarity |
| \_NormalizedVDWVC2 | 0.0022 | intermolecular force |
| \_NormalizedVDWVC3 | 0.0087 | intermolecular force |
| \_HydrophobicityC3 | 0.0142 | hydrophobicity |
| \_PolarizabilityT12 | 0.0092 | polarizability |
| \_PolarizabilityT23 | 0.0028 | polarizability |
| \_SolventAccessibilityT12 | 0.0053 | solvent accessibility |
| \_SolventAccessibilityT23 | 0.0056 | solvent accessibility |
| \_SecondaryStrT12 | 0.0075 | secondary structure |
| \_SecondaryStrT13 | 0.0008 | secondary structure |
| \_ChargeT12 | 0.0028 | charge |
| \_ChargeT23 | 0.0061 | charge |
| \_PolarityT12 | 0.0053 | polarity |
| \_NormalizedVDWVT12 | 0.0101 | intermolecular force |
| \_PolarizabilityD1050 | 0.0036 | polarity |
| \_SolventAccessibilityD1001 | 0.0248 | solvent accessibility |
| \_SolventAccessibilityD1050 | 0.0045 | solvent accessibility |
| \_SolventAccessibilityD1100 | 0.0050 | solvent accessibility |
| \_SolventAccessibilityD2025 | 0.0050 | solvent accessibility |
| \_SolventAccessibilityD2050 | 0.0011 | solvent accessibility |
| \_SolventAccessibilityD3025 | 0.0081 | solvent accessibility |
| \_SecondaryStrD1100 | 0.0008 | secondary structure |
| \_SecondaryStrD2025 | 0.0061 | secondary structure |
| \_SecondaryStrD2100 | 0.0056 | secondary structure |
| \_ChargeD1001 | 0.0240 | charge |
| \_ChargeD1050 | 0.0031 | charge |
| \_ChargeD1075 | 0.0020 | charge |
| \_ChargeD2001 | 0.0600 | charge |
| \_ChargeD2100 | 0.0003 | charge |
| \_ChargeD3001 | 0.0050 | charge |
| \_ChargeD3025 | 0.0053 | charge |
| \_ChargeD3050 | 0.0017 | charge |
| \_ChargeD3075 | 0.0056 | charge |
| \_PolarityD2025 | 0.0070 | polarity |
| \_PolarityD2050 | 0.0020 | polarity |
| \_PolarityD2075 | 0.0017 | polarity |
| \_PolarityD3050 | 0.0056 | polarity |
| \_NormalizedVDWVD1025 | 0.0017 | intermolecular force |
| \_NormalizedVDWVD1075 | 0.0039 | intermolecular force |
| \_NormalizedVDWVD2025 | 0.0050 | intermolecular force |
| \_NormalizedVDWVD2050 | 0.0036 | intermolecular force |
| \_HydrophobicityD1075 | 0.0036 | hydrophobicity |
| \_HydrophobicityD2050 | 0.0031 | hydrophobicity |
| \_HydrophobicityD2100 | 0.0014 | hydrophobicity |
| QSOSW24 | 0.0075 | sequence ordering |
| QSOSW25 | 0.0025 | sequence ordering |
| QSOSW29 | 0.0045 | sequence ordering |
| QSOSW36 | 0.0059 | sequence ordering |
| QSOgrant24 | 0.0000 | sequence ordering |
| QSOgrant26 | 0.0031 | sequence ordering |
| MoreauBrotoAuto\_Hydrophobicity16 | 0.0028 | hydrophobicity |
| MoreauBrotoAuto\_AvFlexibility3 | 0.0059 | flexibility |
| MoreauBrotoAuto\_Polarizability2 | 0.0042 | polarity |
| MoreauBrotoAuto\_FreeEnergy13 | 0.0042 | free energy |
| MoreauBrotoAuto\_FreeEnergy16 | 0.0050 | free energy |
| MoreauBrotoAuto\_Steric13 | 0.0020 | steric |
| GearyAuto\_Hydrophobicity4 | 0.0120 | hydrophobicity |
| GearyAuto\_Hydrophobicity5 | 0.0081 | hydrophobicity |
| GearyAuto\_Hydrophobicity6 | 0.0017 | hydrophobicity |
| GearyAuto\_Hydrophobicity8 | 0.0036 | hydrophobicity |
| GearyAuto\_Hydrophobicity9 | 0.0087 | hydrophobicity |
| GearyAuto\_Hydrophobicity10 | 0.0053 | hydrophobicity |
| GearyAuto\_Hydrophobicity11 | 0.0050 | hydrophobicity |
| GearyAuto\_Hydrophobicity12 | 0.0028 | hydrophobicity |
| GearyAuto\_Hydrophobicity13 | 0.0042 | hydrophobicity |
| GearyAuto\_Hydrophobicity14 | 0.0022 | hydrophobicity |
| GearyAuto\_Hydrophobicity15 | 0.0047 | hydrophobicity |
| GearyAuto\_Hydrophobicity16 | 0.0039 | hydrophobicity |
| GearyAuto\_AvFlexibility2 | 0.0112 | flexibility |
| GearyAuto\_AvFlexibility5 | 0.0056 | flexibility |
| GearyAuto\_AvFlexibility6 | 0.0034 | flexibility |
| GearyAuto\_AvFlexibility7 | 0.0064 | flexibility |
| GearyAuto\_AvFlexibility8 | 0.0042 | flexibility |
| GearyAuto\_AvFlexibility10 | 0.0025 | flexibility |
| GearyAuto\_AvFlexibility11 | 0.0087 | flexibility |
| GearyAuto\_AvFlexibility12 | 0.0039 | flexibility |
| GearyAuto\_AvFlexibility13 | 0.0025 | flexibility |
| GearyAuto\_AvFlexibility15 | 0.0042 | flexibility |
| GearyAuto\_Polarizability1 | 0.0047 | polarity |
| GearyAuto\_Polarizability2 | 0.0045 | polarity |
| GearyAuto\_Polarizability3 | 0.0034 | polarity |
| GearyAuto\_Polarizability4 | 0.0020 | polarity |
| GearyAuto\_Polarizability6 | 0.0036 | polarity |
| GearyAuto\_Polarizability7 | 0.0034 | polarity |
| GearyAuto\_Polarizability8 | 0.0034 | polarity |
| GearyAuto\_Polarizability9 | 0.0028 | polarity |
| GearyAuto\_Polarizability10 | 0.0050 | polarity |
| GearyAuto\_Polarizability14 | 0.0059 | polarity |
| GearyAuto\_Polarizability15 | 0.0017 | polarity |
| GearyAuto\_FreeEnergy1 | 0.0050 | free energy |
| GearyAuto\_FreeEnergy2 | 0.0042 | free energy |
| GearyAuto\_FreeEnergy3 | 0.0106 | free energy |
| GearyAuto\_FreeEnergy4 | 0.0050 | free energy |
| GearyAuto\_FreeEnergy5 | 0.0042 | free energy |
| GearyAuto\_FreeEnergy6 | 0.0028 | free energy |
| GearyAuto\_FreeEnergy7 | 0.0036 | free energy |
| GearyAuto\_FreeEnergy8 | 0.0045 | free energy |
| GearyAuto\_FreeEnergy9 | 0.0028 | free energy |
| GearyAuto\_FreeEnergy10 | 0.0047 | free energy |
| GearyAuto\_FreeEnergy11 | 0.0042 | free energy |
| GearyAuto\_FreeEnergy12 | 0.0031 | free energy |
| GearyAuto\_FreeEnergy13 | 0.0061 | free energy |
| GearyAuto\_FreeEnergy14 | 0.0034 | free energy |
| GearyAuto\_FreeEnergy15 | 0.0028 | free energy |
| GearyAuto\_ResidueASA1 | 0.0017 | surface area |
| GearyAuto\_ResidueASA2 | 0.0028 | surface area |
| GearyAuto\_ResidueASA3 | 0.0034 | surface area |
| GearyAuto\_ResidueASA4 | 0.0034 | surface area |
| GearyAuto\_ResidueASA5 | 0.0020 | surface area |
| GearyAuto\_ResidueASA6 | 0.0014 | surface area |
| GearyAuto\_ResidueASA7 | 0.0028 | surface area |
| GearyAuto\_ResidueASA8 | 0.0022 | surface area |
| GearyAuto\_ResidueASA9 | 0.0025 | surface area |
| GearyAuto\_ResidueASA11 | 0.0078 | surface area |
| GearyAuto\_ResidueASA12 | 0.0053 | surface area |
| GearyAuto\_ResidueASA15 | 0.0031 | surface area |
| GearyAuto\_ResidueASA16 | 0.0017 | surface area |
| GearyAuto\_ResidueVol3 | 0.0014 | residue volume |
| GearyAuto\_ResidueVol4 | 0.0034 | residue volume |
| GearyAuto\_ResidueVol5 | 0.0089 | residue volume |
| GearyAuto\_ResidueVol6 | 0.0017 | residue volume |
| GearyAuto\_ResidueVol9 | 0.0003 | residue volume |
| GearyAuto\_ResidueVol10 | 0.0025 | residue volume |
| GearyAuto\_ResidueVol12 | 0.0017 | residue volume |
| GearyAuto\_ResidueVol13 | 0.0014 | residue volume |
| GearyAuto\_Steric1 | 0.0014 | steric |
| GearyAuto\_Steric2 | 0.0022 | steric |
| GearyAuto\_Steric3 | 0.0017 | steric |
| GearyAuto\_Steric4 | 0.0045 | steric |
| GearyAuto\_Steric5 | 0.0092 | steric |
| GearyAuto\_Steric6 | 0.0034 | steric |
| GearyAuto\_Steric7 | 0.0042 | steric |
| GearyAuto\_Steric8 | 0.0017 | steric |
| GearyAuto\_Steric9 | 0.0028 | steric |
| GearyAuto\_Steric10 | 0.0025 | steric |
| GearyAuto\_Steric11 | 0.0039 | steric |
| GearyAuto\_Steric12 | 0.0020 | steric |
| GearyAuto\_Steric13 | 0.0031 | steric |
| GearyAuto\_Steric15 | 0.0022 | steric |
| GearyAuto\_Steric16 | 0.0059 | steric |
| GearyAuto\_Mutability1 | 0.0045 | mutability |
| GearyAuto\_Mutability2 | 0.0067 | mutability |
| GearyAuto\_Mutability3 | 0.0011 | mutability |
| GearyAuto\_Mutability4 | 0.0031 | mutability |
| GearyAuto\_Mutability5 | 0.0017 | mutability |
| GearyAuto\_Mutability6 | 0.0022 | mutability |
| GearyAuto\_Mutability7 | 0.0022 | mutability |
| GearyAuto\_Mutability8 | 0.0028 | mutability |
| GearyAuto\_Mutability9 | 0.0011 | mutability |
| GearyAuto\_Mutability10 | 0.0017 | mutability |
| GearyAuto\_Mutability11 | 0.0036 | mutability |
| GearyAuto\_Mutability12 | 0.0045 | mutability |
| GearyAuto\_Mutability13 | 0.0022 | mutability |
| GearyAuto\_Mutability14 | 0.0039 | mutability |
| GearyAuto\_Mutability15 | 0.0101 | mutability |
| GearyAuto\_Mutability16 | 0.0025 | mutability |