**SUPPLEMENTARY MATERIALS**

**An ensemble** **approach to predict binding hotspots in protein-RNA interactions based on SMOTE data balancing and random grouping feature selection** **strategies**

Tong Zhou†, Jie Rong†, Yang Liu, Weikang Gong, Chunhua Li\*

Falcuty of Environmental and Life Sciences, Beijing University of Technology, Beijing 100124, China

\*All correspondence should be addressed to Chunhua Li (E-mail: chunhuali@bjut.edu.cn)

**\_\_\_\_\_\_\_\_\_\_\_\_**

The authors state no conflict of interest.

**†**Tong Zhou and Jie Rong contribute equally to this work.

\*Correspondence to Chunhua Li. E-mail: chunhuali@bjut.edu.cn

**Table S1.** The dataset of 58 protein-RNA complexes

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Training dataset | 1ASY | 1AUD | 1B23 | 1C9S | 1JBS | 1QFQ | 1U0B | 1YVP |
| 2BX2 | 2ERR | 2IX1 | 2M8D | 2PJP | 2Y8W | 2ZI0 | 2ZKO |
| 2ZZN | 3EQT | 3K5Q | 3L25 | 3MOJ | 3QSU | 3RW6 | 3U4M |
| 3VYX | 4ED5 | 4ERD | 4NGD | 4OOG | 4PMW | 4QVC | 4R3I |
| 4RCJ | 4YVI | 5AWH | 5DET | 5EIM | 5ELK | 5H1K | 5IP2 |
| 5UDZ | 5W1H | 5WWX |  |  |  |  |  |
| Independent testing dataset | 1FEU | 1WNE | 1ZDI | 2KXN | 2XB2 | 3AM1 | 3UZS | 3VYY |
| 4CIO | 4G0A | 4JVH | 4NL3 | 5EN1 | 5EV1 | 5HO4 |  |

PDB IDs from Protein Data Bank (PDB)

**Table S2.** The 120 features used in our study

|  |  |  |
| --- | --- | --- |
| **Feature Symbol** | **Description** | **Category** |
| Na | Number of atoms | Physicochemical properties of amino acids |
| Nec | Number of electrostatic charges | Physicochemical properties of amino acids |
| Nphb | Number of potential hydrogen bonds | Physicochemical properties of amino acids |
| Hdrpo | Hydrophobicity | Physicochemical properties of amino acids |
| Hdrpi | Hydrophilicity | Physicochemical properties of amino acids |
| Prop | Propensity | Physicochemical properties of amino acids |
| Isoep | Isoelectric point | Physicochemical properties of amino acids |
| Mass | Mass | Physicochemical properties of amino acids |
| Enc | Expected number of contacts within 14Å sphere | Physicochemical properties of amino acids |
| Eiip | Electron-ion interaction potential | Physicochemical properties of amino acids |
| Phi | Angle phi | Secondary structure of protein |
| Psi | Angle psi | Secondary structure of protein |
| Theta | Angle theta | Secondary structure of protein |
| Tau | Angle tau | Secondary structure of protein |
| probC | Probability of alpha-helix | Secondary structure of protein |
| probH | Probability of beta-strand | Secondary structure of protein |
| probE | Probability of random coil | Secondary structure of protein |
| BmDPXa | Mean of DPX for all atoms of bound state | DPX and CX |
| BsdDPXa | Standard deviation of DPX for all atoms of bound state | DPX and CX |
| BmDPXs | Mean of DPX for side chain of bound state | DPX and CX |
| BsdDPXs | Standard deviation of DPX for side chain of bound state | DPX and CX |
| BmCXa | Mean of CX for all atoms of bound state | DPX and CX |
| BsdCXa | Standard deviation of CX for all atoms of bound state | DPX and CX |
| BmCXs | Mean of CX for side chain of bound state | DPX and CX |
| BsdCXs | Standard deviation of CX for side chain of bound state | DPX and CX |
| UmDPXa | Mean of DPX for all atoms of unbound state | DPX and CX |
| UsdDPXa | Standard deviation of DPX for all atoms of unbound state | DPX and CX |
| UmDPXs | Mean of DPX for side chain of unbound state | DPX and CX |
| UsdDPXs | Standard deviation of DPX for side chain of unbound state | DPX and CX |
| UmCXa | Mean of CX for all atoms of unbound state | DPX and CX |
| UsdCXa | Standard deviation of CX for all atoms of unbound state | DPX and CX |
| UmCXs | Mean of CX for side chain of unbound state | DPX and CX |
| UsdCXs | Standard deviation of CX for side chain of unbound state | DPX and CX |
| DmDPXa | Change of mean of DPX for all atoms between bound and unbound states | DPX and CX |
| DsdDPXa | Change of standard deviation of DPX for all atoms between bound and unbound states | DPX and CX |
| DmDPXs | Change of mean of DPX for side chain between bound and unbound states | DPX and CX |
| DsdDPXs | Change of standard deviation of DPX for side chain between bound and unbound states | DPX and CX |
| DmCXa | Change of mean of CX for all atoms between bound and unbound states | DPX and CX |
| DsdCXa | Change of standard deviation of CX for all atoms between bound and unbound states | DPX and CX |
| DmCXs | Change of mean of CX for side chain between bound and unbound states | DPX and CX |
| DsdCXs | Change of standard deviation of CX for side chain between bound and unbound states | DPX and CX |
| BaSASAaa | Absolute SASA for all atoms of bound state | SASA |
| BrSASAaa | Relative SASA for all atoms of bound state | SASA |
| BaSASAts | Absolute SASA for total side-chain atoms of bound state | SASA |
| BrSASAts | Relative SASA for total side-chain atoms of bound state | SASA |
| BaSASAmc | Absolute SASA for main-chain atoms of bound state | SASA |
| BrSASAmc | Relative SASA for main-chain atoms of bound state | SASA |
| BaSASAnp | Absolute SASA for non-polar atoms of bound state | SASA |
| BrSASAnp | Relative SASA for non-polar atoms of bound state | SASA |
| BaSASAap | Absolute SASA for all polar atoms of bound state | SASA |
| BrSASAap | Relative SASA for all polar atoms of bound state | SASA |
| UaSASAaa | Absolute SASA for all atoms of unbound state | SASA |
| UrSASAaa | Relative SASA for all atoms of unbound state | SASA |
| UaSASAts | Absolute SASA for total side-chain atoms of unbound state | SASA |
| UrSASAts | Relative SASA for total side-chain atoms of unbound state | SASA |
| UaSASAmc | Absolute SASA for main-chain atoms of unbound state | SASA |
| UrSASAmc | Relative SASA for main-chain atoms of unbound state | SASA |
| UaSASAnp | Absolute SASA for non-polar atoms of unbound state | SASA |
| UrSASAnp | Relative SASA for non-polar atoms of unbound state | SASA |
| UaSASAap | Absolute SASA for all polar atoms of unbound state | SASA |
| UrSASAap | Relative SASA for all polar atoms of unbound state | SASA |
| DaSASAaa | Change of absolute SASA for all atoms between bound and unbound states | SASA |
| DrSASAaa | Change of relative SASA for all atoms between bound and unbound states | SASA |
| DaSASAts | Change of absolute SASA for total side-chain atoms between bound and unbound states | SASA |
| DrSASAts | Change of relative SASA for total side-chain atoms between bound and unbound states | SASA |
| DaSASAmc | Change of absolute SASA for main-chain atoms between bound and unbound states | SASA |
| DrSASAmc | Change of relative SASA for main-chain atoms between bound and unbound states | SASA |
| DaSASAnp | Change of absolute SASA for non-polar atoms between bound and unbound states | SASA |
| DrSASAnp | Change of relative SASA for non-polar atoms between bound and unbound states | SASA |
| DaSASAap | Change of absolute SASA for all polar atoms between bound and unbound states | SASA |
| DrSASAap | Change of relative SASA for all polar atoms between bound and unbound states | SASA |
| SaSASAaa | Square root of change of absolute SASA for all atoms between bound and unbound states | SASA |
| SrSASAaa | Square root of change of relative SASA for all atoms between bound and unbound states | SASA |
| SaSASAts | Square root of change of absolute SASA for total side-chain atoms between bound and unbound states | SASA |
| SrSASAts | Square root of change of relative SASA for total side-chain atoms between bound and unbound states | SASA |
| SaSASAmc | Square root of change of absolute SASA for main-chain atoms between bound and unbound states | SASA |
| SrSASAmc | Square root of change of relative SASA for main-chain atoms between bound and unbound states | SASA |
| SaSASAnp | Square root of change of absolute SASA for non-polar atoms between bound and unbound states | SASA |
| SrSASAnp | Square root of change of relative SASA for non-polar atoms between bound and unbound states | SASA |
| SaSASAap | Square root of change of absolute SASA for all polar atoms between bound and unbound states | SASA |
| SrSASAap | Square root of change of relative SASA for all polar atoms between bound and unbound states | SASA |
| PA | The position probability of Alanine | PSSM |
| PG | The position probability of Glycine | PSSM |
| PI | The position probability of Isoleucine | PSSM |
| PL | The position probability of Leucine | PSSM |
| PV | The position probability of Valine | PSSM |
| PM | The position probability of Methionine | PSSM |
| PF | The position probability of Phenylalanine | PSSM |
| PW | The position probability of Tryptophan | PSSM |
| PP | The position probability of Proline | PSSM |
| PC | The position probability of Cystine | PSSM |
| PS | The position probability of Serine | PSSM |
| PT | The position probability of Threonine | PSSM |
| PY | The position probability of Tyrosine | PSSM |
| PN | The position probability of Asparagine | PSSM |
| PQ | The position probability of Glutarnine | PSSM |
| PH | The position probability of Histidine | PSSM |
| PK | The position probability of Lysine | PSSM |
| PR | The position probability of Arginine | PSSM |
| PD | The position probability of Aspartic acid | PSSM |
| PE | The position probability of Glutamic acid | PSSM |
| HSE-up | Number of Cα atoms in upper half-sphere | Solvent exposure |
| HSE-down | Number of Cα atoms in lower half-sphere | Solvent exposure |
| CN | Contact number | Solvent exposure |
| IP | Interface propensity | Interface propensity |
| K | Degree of unweighted AAN | Residue topological features from AAN |
| B | Betweenness centrality of unweighted AAN | Residue topological features from AAN |
| C | Closeness centrality of unweighted AAN | Residue topological features from AAN |
| Kp | Degree of node-weighted AAN based on polarity | Residue topological features from AAN |
| Bp | Betweenness centrality of node-weighted AAN based on polarity | Residue topological features from AAN |
| Cp | Closeness centrality of node-weighted AAN based on polarity | Residue topological features from AAN |
| Km | Degree of node-weighted AAN based on mass | Residue topological features from AAN |
| Bm | Betweenness centrality of node-weighted AAN based on mass | Residue topological features from AAN |
| Cm | Closeness centrality of node-weighted AAN based on mass | Residue topological features from AAN |
| Kh | Degree of node-weighted AAN based on hydrophobicity | Residue topological features from AAN |
| Bh | Betweenness centrality of node-weighted AAN based on hydrophobicity | Residue topological features from AAN |
| Ch | Closeness centrality of node-weighted AAN based on hydrophobicity | Residue topological features from AAN |
| Ks | Degree of node-weighted AAN based on solvent accessibility | Residue topological features from AAN |
| Bs | Betweenness centrality of node-weighted AAN based on solvent accessibility | Residue topological features from AAN |
| Cs | Closeness centrality of node-weighted AAN based on solvent accessibility | Residue topological features from AAN |

**Table S3.** The numerical values of 10 physicochemical properties of the 20 amino acids

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Residue** | **Na** | **Nec** | **Nphb** | **Hdrpo** | **Hdrpi** | **Prop** | **Isoep** | **Mass** | **Enc** | **Eiip** |
| A | 5 | 0 | 2 | 0.25 | -0.5 | -0.17 | 6 | 71.1 | -0.22 | 0.0373 |
| C | 6 | 0 | 2 | 0.04 | -1 | 0.43 | 5.05 | 103.1 | 4.66 | 0.0829 |
| D | 8 | -1 | 4 | -0.72 | 3 | -0.38 | 2.77 | 115.1 | -4.12 | 0.1263 |
| E | 9 | -1 | 4 | -0.62 | 3 | -0.13 | 3.22 | 129.1 | -3.64 | 0.0058 |
| F | 11 | 0 | 2 | 0.61 | -2.5 | 0.82 | 5.48 | 147.2 | 5.27 | 0.0946 |
| G | 4 | 0 | 2 | 0.16 | 0 | -0.07 | 5.97 | 57 | -1.62 | 0.005 |
| H | 10 | 0 | 4 | -0.4 | -0.5 | 0.41 | 7.59 | 137.1 | 1.28 | 0.0242 |
| I | 8 | 0 | 2 | 0.73 | -1.8 | 0.44 | 6.02 | 113.2 | 5.58 | 0 |
| K | 9 | 1 | 2 | -1.1 | 3 | -0.36 | 9.74 | 128.2 | -4.18 | 0.0371 |
| L | 8 | 0 | 2 | 0.53 | -1.8 | 0.4 | 5.98 | 113.2 | 5.01 | 0 |
| M | 8 | 0 | 2 | 0.26 | -1.3 | 0.66 | 5.74 | 131.2 | 3.51 | 0.0823 |
| N | 8 | 0 | 4 | -0.64 | 0.2 | 0.12 | 5.41 | 114.1 | -2.65 | 0.0036 |
| P | 7 | 0 | 2 | -0.07 | 0 | -0.25 | 6.3 | 97.1 | -3.03 | 0.0198 |
| Q | 9 | 0 | 4 | -0.69 | 0.2 | -0.11 | 5.65 | 128.1 | -2.76 | 0.0761 |
| R | 11 | 1 | 4 | -1.76 | 3 | 0.27 | 10.76 | 156.2 | -0.93 | 0.0959 |
| S | 6 | 0 | 4 | -0.26 | 0.3 | -0.33 | 5.68 | 87.1 | -2.84 | 0.0829 |
| T | 7 | 0 | 4 | -0.18 | -0.4 | -0.18 | 5.66 | 101.1 | -1.2 | 0.0941 |
| V | 7 | 0 | 2 | 0.54 | -1.5 | 0.27 | 5.96 | 99.1 | 4.45 | 0.0057 |
| W | 14 | 0 | 3 | 0.37 | -3.4 | 0.83 | 5.89 | 186.2 | 5.2 | 0.0548 |
| Y | 12 | 0 | 3 | 0.02 | -2.3 | 0.66 | 5.66 | 163.2 | 2.15 | 0.0516 |

**Table S4.** The 20×4 residue-nucleotide pairwise propensities

|  |  |
| --- | --- |
| **Animo Acids** | **Nucleotides** |
| **A** | **C** | **G** | **U** | **Average** |
| A | 0.762 | 0.591 | 0.466 | 1.126 | 0.736 |
| R | 2.812 | 2.294 | 1.841 | 3.565 | 2.628 |
| N | 1.289 | 0.83 | 0.789 | 2.218 | 1.282 |
| D | 0.591 | 0.394 | 0.394 | 0.712 | 0.523 |
| C | 2.765 | 0.888 | 1.145 | 1.872 | 1.668 |
| Q | 1.145 | 0.888 | 0.737 | 1.97 | 1.185 |
| E | 0.4 | 0.28 | 0.316 | 0.407 | 0.351 |
| G | 1.034 | 0.888 | 0.737 | 1.312 | 0.993 |
| H | 1.72 | 1.145 | 1.289 | 2.181 | 1.584 |
| I | 1.145 | 0.725 | 0.508 | 1.246 | 0.906 |
| L | 0.737 | 0.394 | 0.362 | 0.776 | 0.567 |
| K | 1.404 | 1.126 | 1.07 | 1.97 | 1.393 |
| M | 1.404 | 0.888 | 0.844 | 1.334 | 1.118 |
| F | 1.691 | 0.689 | 0.859 | 2.673 | 1.478 |
| P | 0.737 | 0.443 | 0.421 | 1.017 | 0.655 |
| S | 1.269 | 0.888 | 0.762 | 1.312 | 1.058 |
| T | 1.289 | 0.737 | 0.601 | 1.502 | 1.032 |
| W | 1.554 | 1.246 | 0.789 | 1.088 | 1.169 |
| Y | 2.004 | 1.107 | 1.107 | 2.673 | 1.723 |
| V | 0.934 | 0.622 | 0.451 | 1.404 | 0.853 |

**Table S5.** Summary of the features in the optimal feature set

|  |  |  |
| --- | --- | --- |
| **Feature Symbol** | **Description** | **Category** |
| Nec | Number of electrostatic charges | Physicochemical properties of amino acids |
| Nphb | Number of potential hydrogen bonds | Physicochemical properties of amino acids |
| Eiip | Electron-ion interaction potential | Physicochemical properties of amino acids |
| PL | Probability of Leucine | PSSM |
| PP | Probability of Proline | PSSM |
| BsdDPXa | Standard deviation of DPX for all atoms of bound state | DPX and CX |
| UsdDPXs | Standard deviation of DPX for side chain of unbound state | DPX and CX |
| DmDPXs | Change of mean of DPX for side chain between bound and unbound states | DPX and CX |
| BmCXs | Mean of CX for side chain of bound state | DPX and CX |
| UrSASAts | Relative SASA for total side-chain atoms of unbound state | SASA |
| BaSASAnp | Absolute SASA for non-polar atoms of bound state | SASA |
| BrSASAnp | Relative SASA for non-polar atoms of bound state | SASA |
| HSE-up | Number of Cα atoms in upper half-sphere | Solvent exposure |
| HSE-down | Number of Cα atoms in lower half-sphere | Solvent exposure |
| CN | Contact number | Solvent exposure |
| IP | Interface propensity | Interface propensity |
| Kh | Degree of node-weighted AAN based on hydrophobicity | Residue topological features from AAN |
| Ks | Degree of node-weighted AAN based on solvent accessibility | Residue topological features from AAN |

**Table S6.** Comparison between different classifiers on training dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **ACC** | **SEN** | **SPE** | **PRE** | **F1** | **MCC** | **AUC** |
| RF | 0.814±0.100 | 0.662±0.190 | 0.852±0.089 | 0.569±0.231 | 0.599±0.192 | 0.491±0.253 | 0.842±0.127 |
| kNN | 0.681±0.150 | 0.875±0.301 | 0.623±0.157 | 0.398±0.187 | 0.536±0.226 | 0.433±0.278 | 0.858±0.145 |
| SVM | 0.677±0.168 | 0.947±0.111 | 0.601±0.211 | 0.467±0.277 | 0.589±0.225 | 0.488±0.273 | 0.844±0.138 |
| Adaboost | 0.726±0.145 | 0.775±0.314 | 0.718±0.168 | 0.424±0.169 | 0.513±0.200 | 0.415±0.245 | 0.846±0.138 |
| GTB | 0.798±0.078 | 0.738±0.233 | 0.817±0.079 | 0.497±0.201 | 0.570±0.197 | 0.479±0.217 | 0.854±0.100 |
| XGBoost | 0.762±0.100 | 0.771±0.310 | 0.750±0.098 | 0.438±0.198 | 0.545±0.219 | 0.439±0.269 | 0.844±0.136 |
| SEC | 0.833±0.091 | 0.800±0.227 | 0.847±0.113 | 0.602±0.231 | 0.646±0.171 | 0.581±0.208 | 0.848±0.145 |

RF: Random Forest. kNN: k-Nearest Neighbor. SVM: Support Vector Machine. Adaboost: Adaptive Boosting. GTB: Gradient Tree Boosting. XGBoost: eXtreme Gradient Boosting. SEC: Stacking Ensemble Classifier.

**Table S7.** Prediction performance of our method on the training dataset

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | ACC | SEN | SPE | PRE | F1 | MCC | AUC |
| Training-50 | 0.818±0.016 | 0.814±0.036 | 0.824±0.022 | 0.569±0.033 | 0.638±0.022 | 0.565±0.023 | 0.859±0.019 |
| Independent testing | 0.766 | 0.900  | 0.792 | 0.474 | 0.621 | 0.557 | 0.829 |

Note: Training-50 represents the average performance across 50 times of 10-fold cross-validation. The measures in table are accuracy (ACC), sensitivity (SEN), specificity (SPE), precision (PRE), F1 score (F1), Matthews correlation coefficient (MCC) and the area under the ROC curve (AUC).



**Figure S1.** Flowchart of the random grouping feature selection strategy combined with the two-step algorithm