**Table 1: Available descriptors in RDKit**

|  |  |
| --- | --- |
| **Descriptor Family** | **Notes** |
| Gasteiger/Marsili Partial Charges | *Tetrahedron* **36**:3219-28 (1980) |
| BalabanJ | *Chem. Phys. Lett.* **89**:399-404 (1982) |
| BertzCT | *J. Am. Chem. Soc.* **103**:3599-601 (1981) |
| Ipc | *J. Chem. Phys.* **67**:4517-33 (1977) |
| HallKierAlpha | *Rev. Comput. Chem.* **2**:367-422 (1991) |
| Kappa1 - Kappa3 | *Rev. Comput. Chem.* **2**:367-422 (1991) |
| Chi0, Chi1 | *Rev. Comput. Chem.* **2**:367-422 (1991) |
| Chi0n - Chi4n | *Rev. Comput. Chem.* **2**:367-422 (1991) |
| Chi0v - Chi4v | *Rev. Comput. Chem.* **2**:367-422 (1991) |
| MolLogP | Wildman and Crippen *JCICS* **39**:868-73 (1999) |
| MolMR | Wildman and Crippen *JCICS* **39**:868-73 (1999) |
| MolWt |  |
| HeavyAtomCount |  |
| HeavyAtomMolWt |  |
| NHOHCount |  |
| NOCount |  |
| NumHAcceptors |  |
| NumHDonors |  |
| NumHeteroatoms |  |
| NumRotatableBonds |  |
| NumValenceElectrons |  |
| RingCount |  |
| TPSA | *J. Med. Chem.* **43**:3714-7, (2000) |
| LabuteASA | *J. Mol. Graph. Mod.* **18**:464-77 (2000) |
| PEOE\_VSA1 - PEOE\_VSA14 | MOE-type descriptors using partial charges and surface area contributions http://www.chemcomp.com/journal/vsadesc.htm ; underlying calculations, calcPEOE\_VSA, is in C++, the assignment to functions for individual descriptors is in Python |
| SMR\_VSA1 - SMR\_VSA10 | [MOE-type descriptors using MR contributions and surface area contributions http://www.chemcomp.com/journal/vsadesc.htm ; underlying calculations, calcSMR\_VSA, is in C++, the assignment to functions for individual descriptors is in Python](http://www.chemcomp.com/journal/vsadesc.htm) |
| SlogP\_VSA1 - SlogP\_VSA12 | [MOE-type descriptors using LogP contributions and surface area contributions http://www.chemcomp.com/journal/vsadesc.htm ; underlying calculations, calcSlogP\_VSA, is in C++, the assignment to functions for individual descriptors is in Python](http://www.chemcomp.com/journal/vsadesc.htm) |
| EState\_VSA1 - EState\_VSA11 | MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper) |
| VSA\_EState1 - VSA\_EState10 | MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper) |
| Topliss fragments | implemented using a set of SMARTS definitions in $(RDBASE)/Data/FragmentDescriptors.csv |

Table X: Results for Random Forest-based classification models.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| accuracy | MCC | precision | recall | f1 | auc | assaytype | threshold | randomseed |
| 0.899±0.016 | 0.211±0.165 | 0.5±0.316 | 0.119±0.079 | 0.193±0.127 | 0.554±0.042 | functional | 1000nM | 1 |
| 0.899±0.012 | 0.271±0.151 | 0.613±0.38 | 0.156±0.089 | 0.237±0.131 | 0.572±0.041 | functional | 1000nM | 2 |
| 0.913±0.009 | 0.275±0.161 | 0.7±0.4 | 0.125±0.079 | 0.209±0.127 | 0.561±0.04 | functional | 1000nM | 3 |
| 0.913±0.019 | 0.389±0.226 | 0.733±0.389 | 0.253±0.153 | 0.361±0.197 | 0.621±0.078 | functional | 1000nM | 4 |
| 0.913±0.014 | 0.229±0.223 | 0.5±0.447 | 0.125±0.137 | 0.194±0.2 | 0.561±0.069 | functional | 1000nM | 5 |
| 0.906±0.009 | 0.107±0.15 | 0.3±0.4 | 0.054±0.066 | 0.089±0.109 | 0.524±0.032 | functional | 1000nM | 6 |
| 0.911±0.016 | 0.179±0.248 | 0.4±0.49 | 0.1±0.146 | 0.154±0.214 | 0.549±0.074 | functional | 1000nM | 7 |
| 0.909±0.01 | 0.265±0.166 | 0.553±0.324 | 0.175±0.127 | 0.249±0.157 | 0.581±0.061 | functional | 1000nM | 8 |
| 0.885±0.034 | 0.239±0.185 | 0.633±0.371 | 0.129±0.079 | 0.211±0.125 | 0.555±0.05 | functional | 1000nM | 9 |
| 0.918±0.005 | 0.239±0.201 | 0.533±0.452 | 0.125±0.112 | 0.197±0.172 | 0.561±0.054 | functional | 1000nM | 10 |
| 0.907±0.014 | 0.24±0.188 | 0.546±0.397 | 0.136±0.107 | 0.209±0.156 | 0.564±0.054 | functional | 1000nM | mean |
| accuracy | MCC | precision | recall | f1 | auc | assaytype | threshold | randomseed |
| 0.68±0.058 | 0.345±0.118 | 0.654±0.078 | 0.575±0.072 | 0.61±0.068 | 0.668±0.057 | functional | 10000nM | 1 |
| 0.69±0.044 | 0.323±0.102 | 0.645±0.084 | 0.457±0.066 | 0.534±0.069 | 0.648±0.047 | functional | 10000nM | 2 |
| 0.74±0.032 | 0.453±0.069 | 0.749±0.058 | 0.553±0.043 | 0.635±0.042 | 0.711±0.032 | functional | 10000nM | 3 |
| 0.685±0.034 | 0.338±0.074 | 0.666±0.082 | 0.497±0.05 | 0.565±0.036 | 0.657±0.029 | functional | 10000nM | 4 |
| 0.7±0.033 | 0.361±0.081 | 0.655±0.037 | 0.535±0.115 | 0.583±0.084 | 0.673±0.045 | functional | 10000nM | 5 |
| 0.7±0.029 | 0.373±0.062 | 0.69±0.055 | 0.526±0.039 | 0.595±0.035 | 0.676±0.028 | functional | 10000nM | 6 |
| 0.699±0.014 | 0.362±0.037 | 0.657±0.024 | 0.535±0.064 | 0.588±0.041 | 0.673±0.021 | functional | 10000nM | 7 |
| 0.642±0.033 | 0.264±0.068 | 0.613±0.041 | 0.508±0.082 | 0.552±0.057 | 0.627±0.034 | functional | 10000nM | 8 |
| 0.683±0.049 | 0.342±0.103 | 0.65±0.066 | 0.554±0.08 | 0.596±0.068 | 0.666±0.051 | functional | 10000nM | 9 |
| 0.704±0.04 | 0.378±0.092 | 0.688±0.083 | 0.529±0.046 | 0.597±0.054 | 0.678±0.041 | functional | 10000nM | 10 |
| 0.692±0.037 | 0.354±0.081 | 0.667±0.061 | 0.527±0.066 | 0.586±0.055 | 0.668±0.038 | functional | 10000nM | mean |
| accuracy | MCC | precision | recall | f1 | auc | assaytype | threshold | randomseed |
| 0.794±0.016 | 0.409±0.055 | 0.715±0.05 | 0.374±0.046 | 0.491±0.049 | 0.66±0.025 | binding | 1000nM | 1 |
| 0.802±0.008 | 0.436±0.025 | 0.723±0.033 | 0.409±0.026 | 0.522±0.023 | 0.676±0.012 | binding | 1000nM | 2 |
| 0.81±0.009 | 0.459±0.03 | 0.76±0.04 | 0.411±0.013 | 0.534±0.019 | 0.682±0.01 | binding | 1000nM | 3 |
| 0.81±0.015 | 0.466±0.048 | 0.764±0.05 | 0.421±0.039 | 0.542±0.042 | 0.686±0.022 | binding | 1000nM | 4 |
| 0.797±0.008 | 0.397±0.031 | 0.72±0.035 | 0.345±0.025 | 0.466±0.028 | 0.649±0.013 | binding | 1000nM | 5 |
| 0.795±0.022 | 0.405±0.074 | 0.7±0.045 | 0.379±0.08 | 0.49±0.074 | 0.661±0.04 | binding | 1000nM | 6 |
| 0.79±0.013 | 0.401±0.04 | 0.714±0.057 | 0.365±0.019 | 0.483±0.026 | 0.655±0.013 | binding | 1000nM | 7 |
| 0.793±0.009 | 0.394±0.029 | 0.687±0.023 | 0.372±0.03 | 0.482±0.029 | 0.656±0.015 | binding | 1000nM | 8 |
| 0.813±0.011 | 0.433±0.037 | 0.736±0.036 | 0.38±0.028 | 0.501±0.032 | 0.668±0.016 | binding | 1000nM | 9 |
| 0.805±0.013 | 0.435±0.043 | 0.736±0.05 | 0.395±0.049 | 0.512±0.043 | 0.672±0.023 | binding | 1000nM | 10 |
| 0.801±0.012 | 0.423±0.041 | 0.726±0.042 | 0.385±0.035 | 0.502±0.037 | 0.666±0.019 | binding | 1000nM | mean |
| accuracy | MCC | precision | recall | f1 | auc | assaytype | threshold | randomseed |
| 0.797±0.01 | 0.42±0.034 | 0.812±0.01 | 0.94±0.006 | 0.872±0.006 | 0.67±0.018 | binding | 10000nM | 1 |
| 0.799±0.02 | 0.41±0.068 | 0.82±0.015 | 0.937±0.016 | 0.874±0.012 | 0.667±0.032 | binding | 10000nM | 2 |
| 0.807±0.017 | 0.431±0.058 | 0.821±0.009 | 0.947±0.016 | 0.88±0.011 | 0.671±0.022 | binding | 10000nM | 3 |
| 0.798±0.025 | 0.403±0.083 | 0.818±0.017 | 0.937±0.014 | 0.873±0.015 | 0.664±0.036 | binding | 10000nM | 4 |
| 0.794±0.012 | 0.404±0.042 | 0.811±0.012 | 0.941±0.006 | 0.871±0.006 | 0.662±0.024 | binding | 10000nM | 5 |
| 0.79±0.014 | 0.372±0.055 | 0.807±0.013 | 0.944±0.012 | 0.87±0.008 | 0.643±0.028 | binding | 10000nM | 6 |
| 0.803±0.01 | 0.397±0.031 | 0.821±0.005 | 0.945±0.013 | 0.879±0.007 | 0.655±0.011 | binding | 10000nM | 7 |
| 0.792±0.012 | 0.402±0.044 | 0.811±0.014 | 0.937±0.015 | 0.869±0.006 | 0.662±0.028 | binding | 10000nM | 8 |
| 0.803±0.018 | 0.439±0.053 | 0.822±0.008 | 0.936±0.02 | 0.875±0.012 | 0.682±0.018 | binding | 10000nM | 9 |
| 0.789±0.013 | 0.373±0.04 | 0.809±0.005 | 0.936±0.014 | 0.868±0.008 | 0.648±0.013 | binding | 10000nM | 10 |
| 0.797±0.015 | 0.405±0.051 | 0.815±0.011 | 0.94±0.013 | 0.873±0.009 | 0.662±0.023 | binding | 10000nM | mean |

**Table 2: Results for Random Forest-based classification models based on Y-scrambled data.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| accuracy | MCC | precision | recall | f1 | auc | assaytype | threshold | randomseed |
| 0.457±0.037 | -0.089±0.075 | 0.439±0.037 | 0.419±0.053 | 0.428±0.045 | 0.456±0.037 | functional | 1000nM | 1 |
| 0.438±0.063 | -0.126±0.124 | 0.433±0.061 | 0.411±0.096 | 0.417±0.073 | 0.438±0.062 | functional | 1000nM | 2 |
| 0.51±0.057 | 0.019±0.116 | 0.5±0.066 | 0.466±0.106 | 0.478±0.081 | 0.509±0.058 | functional | 1000nM | 3 |
| 0.457±0.041 | -0.092±0.08 | 0.434±0.044 | 0.381±0.025 | 0.406±0.033 | 0.455±0.04 | functional | 1000nM | 4 |
| 0.495±0.036 | -0.01±0.072 | 0.49±0.039 | 0.478±0.033 | 0.483±0.028 | 0.495±0.036 | functional | 1000nM | 5 |
| 0.469±0.038 | -0.061±0.076 | 0.476±0.046 | 0.444±0.067 | 0.458±0.054 | 0.47±0.038 | functional | 1000nM | 6 |
| 0.495±0.047 | -0.009±0.094 | 0.498±0.047 | 0.486±0.081 | 0.491±0.062 | 0.495±0.047 | functional | 1000nM | 7 |
| 0.471±0.018 | -0.064±0.035 | 0.491±0.015 | 0.542±0.079 | 0.513±0.042 | 0.469±0.017 | functional | 1000nM | 8 |
| 0.517±0.037 | 0.028±0.078 | 0.502±0.054 | 0.43±0.085 | 0.46±0.063 | 0.514±0.037 | functional | 1000nM | 9 |
| 0.5±0.048 | 0.0±0.097 | 0.495±0.047 | 0.498±0.106 | 0.493±0.074 | 0.5±0.048 | functional | 1000nM | 10 |
| 0.481±0.042 | -0.04±0.085 | 0.476±0.046 | 0.455±0.073 | 0.463±0.055 | 0.48±0.042 | functional | 1000nM | mean |
| accuracy | MCC | precision | recall | f1 | auc | assaytype | threshold | randomseed |
| 0.498±0.054 | -0.01±0.109 | 0.513±0.05 | 0.567±0.056 | 0.539±0.051 | 0.495±0.054 | functional | 10000nM | 1 |
| 0.49±0.056 | -0.023±0.113 | 0.506±0.057 | 0.53±0.114 | 0.513±0.077 | 0.489±0.056 | functional | 10000nM | 2 |
| 0.515±0.041 | 0.029±0.083 | 0.526±0.038 | 0.538±0.072 | 0.529±0.041 | 0.515±0.041 | functional | 10000nM | 3 |
| 0.483±0.018 | -0.033±0.036 | 0.489±0.017 | 0.469±0.066 | 0.477±0.04 | 0.483±0.018 | functional | 10000nM | 4 |
| 0.502±0.032 | 0.005±0.065 | 0.52±0.033 | 0.502±0.024 | 0.511±0.027 | 0.502±0.032 | functional | 10000nM | 5 |
| 0.462±0.057 | -0.08±0.113 | 0.48±0.056 | 0.498±0.08 | 0.487±0.063 | 0.46±0.057 | functional | 10000nM | 6 |
| 0.493±0.024 | -0.015±0.048 | 0.497±0.024 | 0.495±0.046 | 0.496±0.034 | 0.493±0.024 | functional | 10000nM | 7 |
| 0.478±0.039 | -0.043±0.078 | 0.482±0.041 | 0.488±0.059 | 0.484±0.042 | 0.479±0.039 | functional | 10000nM | 8 |
| 0.5±0.023 | -0.015±0.058 | 0.536±0.029 | 0.585±0.059 | 0.556±0.021 | 0.493±0.028 | functional | 10000nM | 9 |
| 0.522±0.018 | 0.043±0.036 | 0.53±0.016 | 0.543±0.057 | 0.535±0.032 | 0.522±0.018 | functional | 10000nM | 10 |
| 0.494±0.036 | -0.014±0.074 | 0.508±0.036 | 0.521±0.063 | 0.513±0.043 | 0.493±0.037 | functional | 10000nM | mean |
| accuracy | MCC | precision | recall | f1 | auc | assaytype | threshold | randomseed |
| 0.483±0.012 | -0.035±0.023 | 0.476±0.012 | 0.452±0.025 | 0.463±0.015 | 0.483±0.012 | binding | 1000nM | 1 |
| 0.477±0.012 | -0.047±0.025 | 0.469±0.014 | 0.452±0.049 | 0.46±0.032 | 0.476±0.012 | binding | 1000nM | 2 |
| 0.489±0.011 | -0.025±0.023 | 0.472±0.012 | 0.448±0.024 | 0.459±0.018 | 0.488±0.011 | binding | 1000nM | 3 |
| 0.489±0.02 | -0.023±0.04 | 0.48±0.02 | 0.465±0.018 | 0.473±0.018 | 0.488±0.02 | binding | 1000nM | 4 |
| 0.497±0.017 | -0.008±0.034 | 0.487±0.018 | 0.454±0.03 | 0.47±0.022 | 0.496±0.017 | binding | 1000nM | 5 |
| 0.484±0.014 | -0.033±0.029 | 0.475±0.016 | 0.445±0.037 | 0.459±0.027 | 0.484±0.014 | binding | 1000nM | 6 |
| 0.514±0.013 | 0.028±0.025 | 0.513±0.012 | 0.493±0.03 | 0.503±0.021 | 0.514±0.013 | binding | 1000nM | 7 |
| 0.508±0.016 | 0.015±0.033 | 0.5±0.019 | 0.471±0.024 | 0.485±0.02 | 0.507±0.016 | binding | 1000nM | 8 |
| 0.493±0.021 | -0.017±0.043 | 0.475±0.023 | 0.442±0.032 | 0.458±0.027 | 0.491±0.021 | binding | 1000nM | 9 |
| 0.476±0.028 | -0.048±0.057 | 0.474±0.03 | 0.446±0.034 | 0.459±0.032 | 0.476±0.028 | binding | 1000nM | 10 |
| 0.491±0.016 | -0.019±0.033 | 0.482±0.018 | 0.457±0.03 | 0.469±0.023 | 0.49±0.016 | binding | 1000nM | mean |
| accuracy | MCC | precision | recall | f1 | auc | assaytype | threshold | randomseed |
| 0.506±0.013 | 0.01±0.027 | 0.491±0.015 | 0.453±0.035 | 0.471±0.023 | 0.505±0.013 | binding | 10000nM | 1 |
| 0.494±0.025 | -0.015±0.049 | 0.477±0.028 | 0.434±0.04 | 0.454±0.031 | 0.492±0.025 | binding | 10000nM | 2 |
| 0.508±0.009 | 0.016±0.018 | 0.506±0.009 | 0.469±0.02 | 0.487±0.014 | 0.508±0.009 | binding | 10000nM | 3 |
| 0.511±0.028 | 0.021±0.056 | 0.515±0.028 | 0.513±0.037 | 0.514±0.032 | 0.511±0.028 | binding | 10000nM | 4 |
| 0.526±0.02 | 0.052±0.041 | 0.52±0.022 | 0.494±0.033 | 0.506±0.027 | 0.526±0.021 | binding | 10000nM | 5 |
| 0.508±0.013 | 0.016±0.027 | 0.51±0.014 | 0.496±0.036 | 0.502±0.019 | 0.508±0.013 | binding | 10000nM | 6 |
| 0.491±0.012 | -0.018±0.024 | 0.489±0.012 | 0.495±0.025 | 0.492±0.018 | 0.491±0.012 | binding | 10000nM | 7 |
| 0.51±0.014 | 0.021±0.028 | 0.513±0.014 | 0.505±0.018 | 0.509±0.015 | 0.51±0.014 | binding | 10000nM | 8 |
| 0.516±0.017 | 0.031±0.034 | 0.516±0.017 | 0.495±0.032 | 0.505±0.023 | 0.516±0.017 | binding | 10000nM | 9 |
| 0.515±0.009 | 0.029±0.019 | 0.511±0.011 | 0.467±0.027 | 0.488±0.018 | 0.514±0.009 | binding | 10000nM | 10 |
| 0.508±0.016 | 0.016±0.032 | 0.505±0.017 | 0.482±0.03 | 0.493±0.022 | 0.508±0.016 | binding | 10000nM | mean |