Supplementary Information

In silico estimation of chemical aquatic toxicity on crustacean using chemical category methods

Qianqian Cao, Lin Liu, Hongbin Yang, Yingchun Cai, Weihua Li, Guixia Liu, Philip W. Lee, Yun Tang*

Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China
### Table S1 Performance of binary classification models of all crustacean using different fingerprints and modeling methods

<table>
<thead>
<tr>
<th>Model</th>
<th>10-fold cross validation on training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>CA</td>
</tr>
<tr>
<td>CDK-NN</td>
<td>0.79</td>
<td>0.72</td>
</tr>
<tr>
<td>CDK-CT</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>CDK-KNN</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>CDK-NB</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>CDK-RF</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td>CDK-SVM</td>
<td>0.82</td>
<td>0.75</td>
</tr>
<tr>
<td>Est-NN</td>
<td>0.79</td>
<td>0.74</td>
</tr>
<tr>
<td>Est-CT</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>Est-KNN</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>Est-NB</td>
<td>0.77</td>
<td>0.71</td>
</tr>
<tr>
<td>Est-RF</td>
<td>0.80</td>
<td>0.72</td>
</tr>
<tr>
<td>Est-SVM</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>Ext-NN</td>
<td>0.79</td>
<td>0.73</td>
</tr>
<tr>
<td>Ext-CT</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>Ext-KNN</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>Ext-NB</td>
<td>0.81</td>
<td>0.74</td>
</tr>
<tr>
<td>Ext-RF</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td>Ext-SVM</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>Gra-NN</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>Gra-CT</td>
<td>0.69</td>
<td>0.7</td>
</tr>
<tr>
<td>Gra-KNN</td>
<td>0.79</td>
<td>0.72</td>
</tr>
<tr>
<td>Gra-NB</td>
<td>0.75</td>
<td>0.67</td>
</tr>
<tr>
<td>Gra-RF</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>Gra-SVM</td>
<td>0.83</td>
<td>0.75</td>
</tr>
<tr>
<td>Mac-ANN</td>
<td>0.79</td>
<td>0.73</td>
</tr>
<tr>
<td>Mac-CT</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Mac-KNN</td>
<td>0.80</td>
<td>0.72</td>
</tr>
<tr>
<td>Mac-NB</td>
<td>0.74</td>
<td>0.66</td>
</tr>
<tr>
<td>Mac-RF</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td>Mac-SVM</td>
<td>0.83</td>
<td>0.76</td>
</tr>
<tr>
<td>Pub-ANN</td>
<td>0.80</td>
<td>0.74</td>
</tr>
<tr>
<td>Pub-CT</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>Pub-KNN</td>
<td>0.78</td>
<td>0.72</td>
</tr>
<tr>
<td>Pub-NB</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>Pub-RF</td>
<td>0.82</td>
<td>0.75</td>
</tr>
<tr>
<td>Pub-SVM</td>
<td>0.81</td>
<td>0.75</td>
</tr>
<tr>
<td>Sub-NN</td>
<td>0.82</td>
<td>0.75</td>
</tr>
<tr>
<td>Sub-CT</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td>Method</td>
<td>0.78</td>
<td>0.71</td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Sub-KNN</td>
<td>0.78</td>
<td>0.72</td>
</tr>
<tr>
<td>Sub-NB</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>Sub-RF</td>
<td>0.80</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Table S2 The parameters settings of machine learning methods for models building

<table>
<thead>
<tr>
<th>FPName</th>
<th>RF(trees)</th>
<th>kNN(k)</th>
<th>NN(n_mid)</th>
<th>SVM(c)</th>
<th>SVM(g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK</td>
<td>50</td>
<td>13</td>
<td>40</td>
<td>2.0</td>
<td>0.0078125</td>
</tr>
<tr>
<td>Est</td>
<td>70</td>
<td>13</td>
<td>35</td>
<td>2048</td>
<td>0.00195</td>
</tr>
<tr>
<td>Ext</td>
<td>80</td>
<td>11</td>
<td>20</td>
<td>2.0</td>
<td>0.0078125</td>
</tr>
<tr>
<td>Local models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gra</td>
<td>30</td>
<td>13</td>
<td>5</td>
<td>2.0</td>
<td>0.03125</td>
</tr>
<tr>
<td>Mac</td>
<td>80</td>
<td>9</td>
<td>30</td>
<td>0.5</td>
<td>0.125</td>
</tr>
<tr>
<td>Pub</td>
<td>90</td>
<td>13</td>
<td>15</td>
<td>128</td>
<td>0.000122</td>
</tr>
<tr>
<td>Sub</td>
<td>90</td>
<td>11</td>
<td>25</td>
<td>2048</td>
<td>0.00195</td>
</tr>
<tr>
<td>CDK</td>
<td>90</td>
<td>9</td>
<td>20</td>
<td>8.0</td>
<td>0.00195</td>
</tr>
<tr>
<td>Est</td>
<td>40</td>
<td>13</td>
<td>15</td>
<td>2.0</td>
<td>0.125</td>
</tr>
<tr>
<td>Ext</td>
<td>70</td>
<td>13</td>
<td>15</td>
<td>32</td>
<td>0.000122</td>
</tr>
<tr>
<td>Gra</td>
<td>70</td>
<td>7</td>
<td>20</td>
<td>8.0</td>
<td>0.00195</td>
</tr>
<tr>
<td>Global models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mac</td>
<td>40</td>
<td>11</td>
<td>10</td>
<td>2.0</td>
<td>0.125</td>
</tr>
<tr>
<td>Pub</td>
<td>90</td>
<td>13</td>
<td>25</td>
<td>2.0</td>
<td>0.3125</td>
</tr>
<tr>
<td>Sub</td>
<td>80</td>
<td>13</td>
<td>5</td>
<td>0.5</td>
<td>0.125</td>
</tr>
</tbody>
</table>

Table S3 The AD parameters and outlier counts for test set and external validation set

<table>
<thead>
<tr>
<th>Variable</th>
<th>Test set</th>
<th>External validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>Z</td>
<td>N_{OD}</td>
</tr>
<tr>
<td>3</td>
<td>0.8</td>
<td>8</td>
</tr>
</tbody>
</table>
**Figure Legends**

**Figure S1.** Tanimoto similarity index for data sets in local and global models. A: x-axis and y-axis were represented the number of 709 compounds, respectively; B: x-axis and y-axis were represented the number of 824 compounds, respectively.

![Figure S1](image)

**Figure S2.** Workflow of model building for chemical acute aquatic toxicity.