**1. K-means clustering algorithms method**

To ensure the training and test set compounds are similar regarding their physicochemical, structural, and biological features, we performed a systematic clustering procedure following the methods proposed by Maltarollo et. al. First, the physicochemical properties of molecules including logP, number of hydrogen-bond acceptors (NumHAcceptors) and donors (NumHDonors), topological polar surface area (TPSA), number of rotatable bonds (NumRotatableBonds), fraction of sp3 carbon (FractionCSP3) and molecular weight (MolWt) were calculated by RDKit 2019.03.4.0 and clustered by hierarchical cluster using Euclidean distance and complete linkage type. Secondly, the MACCS fingerprints of molecules were generated by RDKit 2019.03.4.0 and were used for hierarchical cluster analysis using Euclidean distance and complete linkage type. Thirdly, the pIC50 values of molecules were clustered by hierarchical cluster using Euclidean distance and complete linkage type. Then the results of three independent clusters were used to a new hierarchical cluster analysis using Euclidean distance and average increment, and the final assigned clusters were attributed to each molecule. The clustering calculations were all performed by scikit-learn 0.23.2. Finally, a random selection of 20% of molecules of each observed cluster was performed to compose the test set. The clustering results were provided in **Table S1**.

|  |
| --- |
| **Table S1**. K-means clustering algorithms results of 56 CDK4 inhibitors. |
| **Compound** | **Physicochemical** | **Structural** | **Biological** | **Systematic** |
| 1 | 1 | 2 | 0 | 2 |
| 2 | 1 | 2 | 0 | 2 |
| 3 | 1 | 2 | 2 | 0 |
| 4 | 1 | 2 | 2 | 0 |
| 5 | 1 | 2 | 2 | 0 |
| 6 | 1 | 0 | 2 | 0 |
| 7 | 1 | 2 | 2 | 0 |
| 8 | 1 | 0 | 0 | 1 |
| 9 | 1 | 0 | 2 | 0 |
| 10 | 1 | 0 | 0 | 1 |
| 11 | 1 | 0 | 2 | 0 |
| 12 | 0 | 0 | 0 | 1 |
| 13 | 0 | 0 | 0 | 1 |
| 14 | 0 | 2 | 2 | 0 |
| **Table S1**. K-means clustering algorithms results of 56 CDK4 inhibitors. (cont.) |
| **Compound** | **Physicochemical** | **Structural** | **Biological** | **Systematic** |
| 15 | 1 | 2 | 0 | 2 |
| 16 | 0 | 2 | 0 | 2 |
| 17 | 0 | 0 | 0 | 1 |
| 18 | 0 | 0 | 0 | 1 |
| 19 | 0 | 2 | 0 | 2 |
| 20 | 0 | 2 | 0 | 2 |
| 21 | 1 | 2 | 1 | 2 |
| 22 | 1 | 2 | 2 | 0 |
| 23 | 1 | 2 | 2 | 0 |
| 24 | 1 | 2 | 1 | 2 |
| 25 | 1 | 2 | 2 | 0 |
| 26 | 1 | 2 | 1 | 2 |
| 27 | 1 | 2 | 1 | 2 |
| 28 | 1 | 2 | 2 | 0 |
| 29 | 1 | 2 | 1 | 2 |
| 30 | 1 | 2 | 1 | 2 |
| 31 | 1 | 2 | 1 | 2 |
| 32 | 1 | 2 | 1 | 2 |
| 33 | 1 | 2 | 2 | 0 |
| 34 | 1 | 2 | 1 | 2 |
| 35 | 1 | 2 | 2 | 0 |
| 36 | 1 | 2 | 1 | 2 |
| 37 | 1 | 2 | 0 | 2 |
| 38 | 1 | 2 | 2 | 0 |
| 39 | 1 | 2 | 0 | 2 |
| 40 | 1 | 2 | 0 | 2 |
| 41 | 1 | 2 | 0 | 2 |
| 42 | 1 | 2 | 0 | 2 |
| 43 | 1 | 2 | 0 | 2 |
| 44 | 1 | 2 | 0 | 2 |
| 45 | 1 | 2 | 0 | 2 |
| 46 | 1 | 2 | 0 | 2 |
| 47 | 0 | 2 | 0 | 2 |
| 48 | 0 | 2 | 0 | 2 |
| 49 | 1 | 2 | 0 | 2 |
| 50 | 0 | 2 | 0 | 2 |
| 51 | 0 | 3 | 2 | 0 |
| 52 | 0 | 2 | 0 | 2 |
| 53 | 0 | 1 | 0 | 1 |
| 54 | 2 | 1 | 2 | 0 |
| 55 | 0 | 2 | 0 | 2 |
| 56 | 0 | 2 | 0 | 2 |

**2. QSAR applicability domain**

****

**Figure S1.** Williams plot of the current QSAR model. Influential compounds were pointed with high h values higher than the warning value h\*. No outliers are observed which can be judged by their standardized residuals greater than three standard deviation units (3σ). Black squares represent the training set and red triangles represent the test set.