How to predict new compounds

(1) About the models of the prediction system:

The regression and multi-classification models of SCO-R and SAO-R are shown in **Table 1**.

Table 1. Seven models of the multi-layer sweetness prediction system.

Model	SCO-R regression	SAO-R regression	SCO-R multi-classification	SAO-R multi-classification
Descriptor/Fingerprints	MOE2D	MOE2D	MOE2D	toxprint
Algorithm	SVM	SVM	GBT	RF

(2) Descriptor and fingerprints：

The models were built on KNIME (version 4.3.3). The MOE2D descriptors were calculated by MOE (version 2018). The toxprint fingerprints were calculated by ChemoTyper (version 1.0). The details of the selected MOE2D descriptors related to the above models are as follows:

★ SCO-R regression (122):

apol, ast_fraglike, ast_fraglike_ext, ast_violation, ast_violation_ext, a_acc, a_acid, a_aro, a_base, a_don, a_heavy, a_hyd, a_ICM, a_nBr, a_nCl, a_nF, a_nN, a_nO, a_nP, a_nS, balabanJ, BCUT_PEOE_2, BCUT_SLOGP_0, BCUT_SLOGP_1, BCUT_SLOGP_2, BCUT_SMR_0, BCUT_SMR_2, b_1rotN, b_double, b_max1len, b_single, b_triple, chi0v_C, chi1v, chiral, chiral_u, diameter, GCUT_PEOE_3, GCUT_SLOGP_0, h_ema, h_emd, h_emd_C, h_logD, h_logS, h_log_dbo, h_log_pbo, h_pavgQ, h_pKa, h_pKb, h_pstates, h_pstrain, Kier2, Kier3, KierA1, KierA2, KierA3, KierFlex, lip_acc, lip_don, lip_violation, logP(o/w), logS, mutagenic, nmol, opr_brigid, opr_leadlike, opr_violation, PEOE_PC+, PEOE_RPC-, PEOE_VSA+0, PEOE_VSA+1, PEOE_VSA+2, PEOE_VSA+3, PEOE_VSA+4, PEOE_VSA+5, PEOE_VSA+6, PEOE_VSA-0, PEOE_VSA-1, PEOE_VSA-2, PEOE_VSA-3, PEOE_VSA-4, PEOE_VSA-5, PEOE_VSA-6, PEOE_VSA_HYD, PEOE_VSA_NEG, PEOE_VSA_PNEG, PEOE_VSA_POL, PEOE_VSA_POS, PEOE_VSA_PPOS, petitjeanSC, radius, reactive, rsynth, SlogP, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, SlogP_VSA3, SlogP_VSA4, SlogP_VSA5, SlogP_VSA6, SlogP_VSA7, SlogP_VSA8, SlogP_VSA9, SMR_VSA0, SMR_VSA1, SMR_VSA2, SMR_VSA3, SMR_VSA4, SMR_VSA5, SMR_VSA6, SMR_VSA7, TPSA, VAdjMa, VDistEq, vsa_acc, vsa_acid, vsa_base, vsa_don, vsa_other, vsa_pol, weinerPath

★ SAO-R regression (117):

apol, ast_fraglike, ast_fraglike_ext, ast_violation, ast_violation_ext, a_acc, a_acid, a_aro, a_don, a_donacc, a_ICM, a_nBr, a_nCl, a_nF, a_nH, a_nN, a_nO, a_nP, a_nS, balabanJ, BCUT_PEOE_0, BCUT_PEOE_1, BCUT_PEOE_2, BCUT_SLOGP_0, BCUT_SLOGP_1, BCUT_SLOGP_2, BCUT_SMR_1, b_1rotN, b_1rotR, b_double, b_max1len, b_triple, chi1v_C, chiral, chiral_u, diameter, GCUT_PEOE_3, GCUT_SLOGP_0, h_ema, h_emd, h_emd_C, h_logD, h_logS, h_log_dbo, h_pavgQ, h_pKa, h_pKb, h_pstates, h_pstrain, Kier3, KierA1, KierFlex, lip_acc, lip_don, lip_druglike, lip_violation, logP(o/w), logS, mutagenic, nmol, opr_brigid, opr_leadlike, opr_violation, PEOE_RPC-, PEOE_VSA+0, PEOE_VSA+1, PEOE_VSA+2, PEOE_VSA+3, PEOE_VSA+4, PEOE_VSA+5, PEOE_VSA+6, PEOE_VSA-0, PEOE_VSA-1, PEOE_VSA-2, PEOE_VSA-3, PEOE_VSA-4, PEOE_VSA-5, PEOE_VSA-6, PEOE_VSA_NEG, PEOE_VSA_PNEG, PEOE_VSA_POL, PEOE_VSA_POS, PEOE_VSA_PPOS, petitjeanSC, radius, reactive, rsynth, SlogP, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, SlogP_VSA3, SlogP_VSA4, SlogP_VSA5, SlogP_VSA6, SlogP_VSA7, SlogP_VSA8, SlogP_VSA9, SMR, SMR_VSA0, SMR_VSA1, SMR_VSA2, SMR_VSA3, SMR_VSA4, SMR_VSA5, SMR_VSA6, SMR_VSA7, TPSA, VAdjMa, VDistEq, vsa_acc, vsa_don, vsa_hyd, vsa_other, vsa_pol, weinerPath, weinerPol

★ SCO-R multi-classification (122):

apol, ast_fraglike, ast_fraglike_ext, ast_violation, ast_violation_ext, a_acc, a_acid, a_aro, a_base, a_don, a_heavy, a_hyd, a_ICM, a_nBr, a_nCl, a_nF, a_nN, a_nO, a_nP, a_nS, balabanJ, BCUT_PEOE_2, BCUT_SLOGP_0, BCUT_SLOGP_1, BCUT_SLOGP_2, BCUT_SMR_0, BCUT_SMR_2, b_1rotN, b_double, b_max1len, b_single, b_triple, chi0v_C, chi1v, chiral, chiral_u, diameter, GCUT_PEOE_3, GCUT_SLOGP_0, h_ema, h_emd, h_emd_C, h_logD, h_logS, h_log_dbo, h_log_pbo, h_pavgQ, h_pKa, h_pKb, h_pstates, h_pstrain, Kier2, Kier3, KierA1, KierA2, KierA3, KierFlex, lip_acc, lip_don, lip_violation, logP(o/w), logS, mutagenic, nmol, opr_brigid, opr_leadlike, opr_violation, PEOE_PC+, PEOE_RPC-, PEOE_VSA+0, PEOE_VSA+1, PEOE_VSA+2, PEOE_VSA+3, PEOE_VSA+4, PEOE_VSA+5, PEOE_VSA+6, PEOE_VSA-0, PEOE_VSA-1, PEOE_VSA-2, PEOE_VSA-3, PEOE_VSA-4, PEOE_VSA-5, PEOE_VSA-6, PEOE_VSA_HYD, PEOE_VSA_NEG, PEOE_VSA_PNEG, PEOE_VSA_POL, PEOE_VSA_POS, PEOE_VSA_PPOS, petitjeanSC, radius, reactive, rsynth, SlogP, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, SlogP_VSA3, SlogP_VSA4, SlogP_VSA5, SlogP_VSA6, SlogP_VSA7, SlogP_VSA8, SlogP_VSA9, SMR_VSA0, SMR_VSA1, SMR_VSA2, SMR_VSA3, SMR_VSA4, SMR_VSA5, SMR_VSA6, SMR_VSA7, TPSA, VAdjMa, VDistEq, vsa_acc, vsa_acid, vsa_base, vsa_don, vsa_other, vsa_pol, weinerPath

(3) How to form your own prediction pipeline:

The details of constructing your own workflow are shown in **Figure 1** and **Figure2**. Explanation of workflow:

1) The local model file is read by the **Model Reader** node above, while the below reads the model-Normalizer.zip file for normalizer;

2) The node of **File Reader** is used to read the data that needs to be predicted; Please read your data as the examples we provided (example for SCO-R for regression.csv and example for SCO-R multi-classification.csv). If you just want to predict molecules without experimental values or labels, you should ignore the 'lgNOAEL' AND 'label' columns and disconnect the evaluation nodes for example 'scorer'.

3) Convert label to strings using **Number to String** node is required in classification models;

4) Select the corresponding prediction node according to the model read by the model reader;

5) Output for the prediction result. In addition, evaluation nodes can be chosen according to your task.

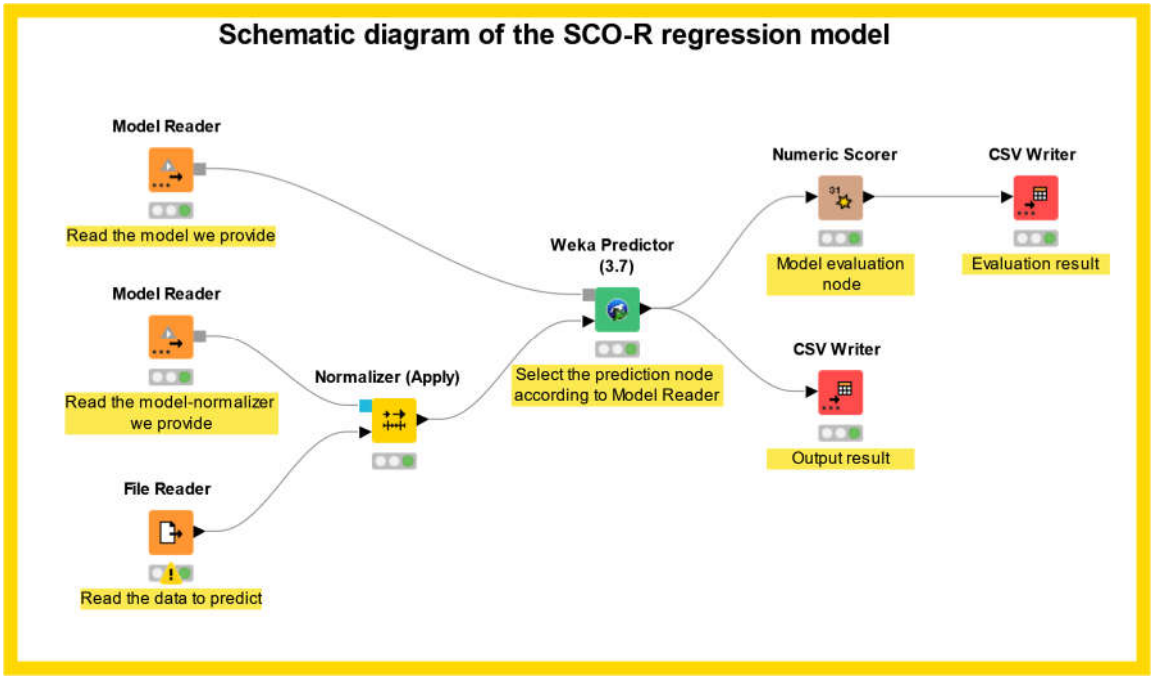

Figure 1. KNIME usage example of regression model.

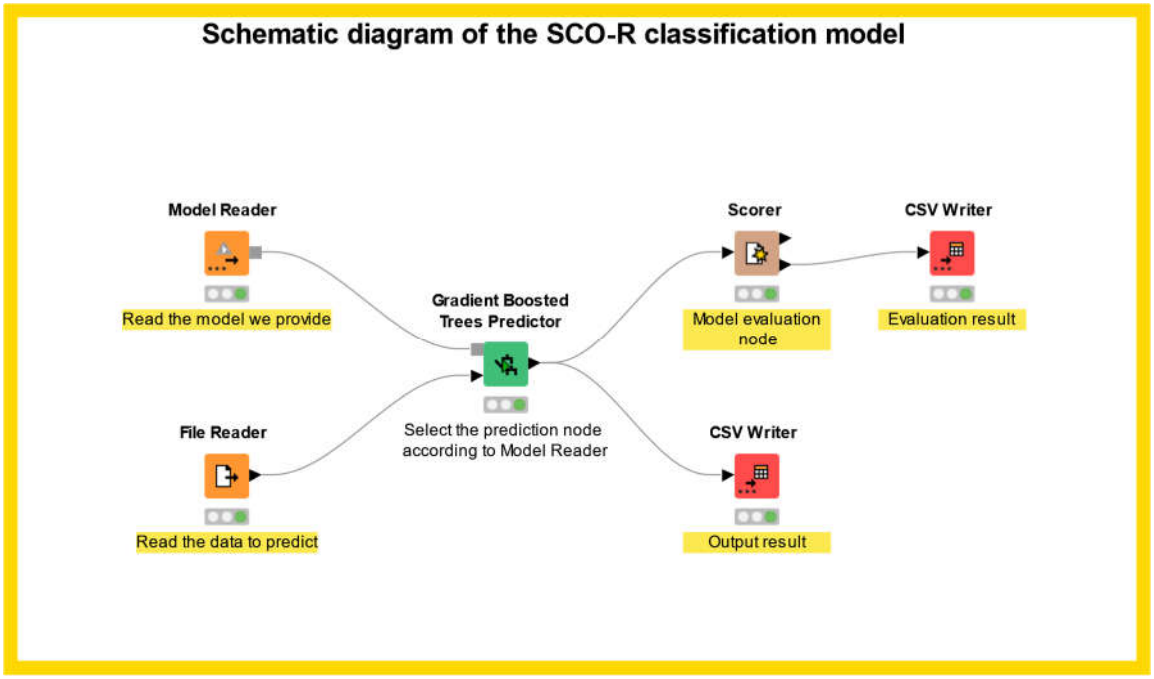

Figure 2. KNIME usage example of classification model.

(4) How to calculate the S indices:

S indices represent the similarity between each molecule and the training dataset by using Tanimoto similarity coefficient and MACCS fingerprints.

Compare the Training_Mean with Query_Mean and Training_Max with Query_Max. We suggest that a molecule having a Query_Mean within 2 fold standard deviation or higher Query_Max is more likely in the application domain.

Table 2. The S indices Training_Mean and Training_Max for four models.

S index	SCO-R regression	SAO-R regression	SCO-R multi-classification	SAO-R multi-classification
Training_Mean	0.212±0.052	0.179±0.042	0.212±0.052	0.182±0.044
Training_Max	0.711±0.139	0.677±0.167	0.716±0.14	0.691±0.163

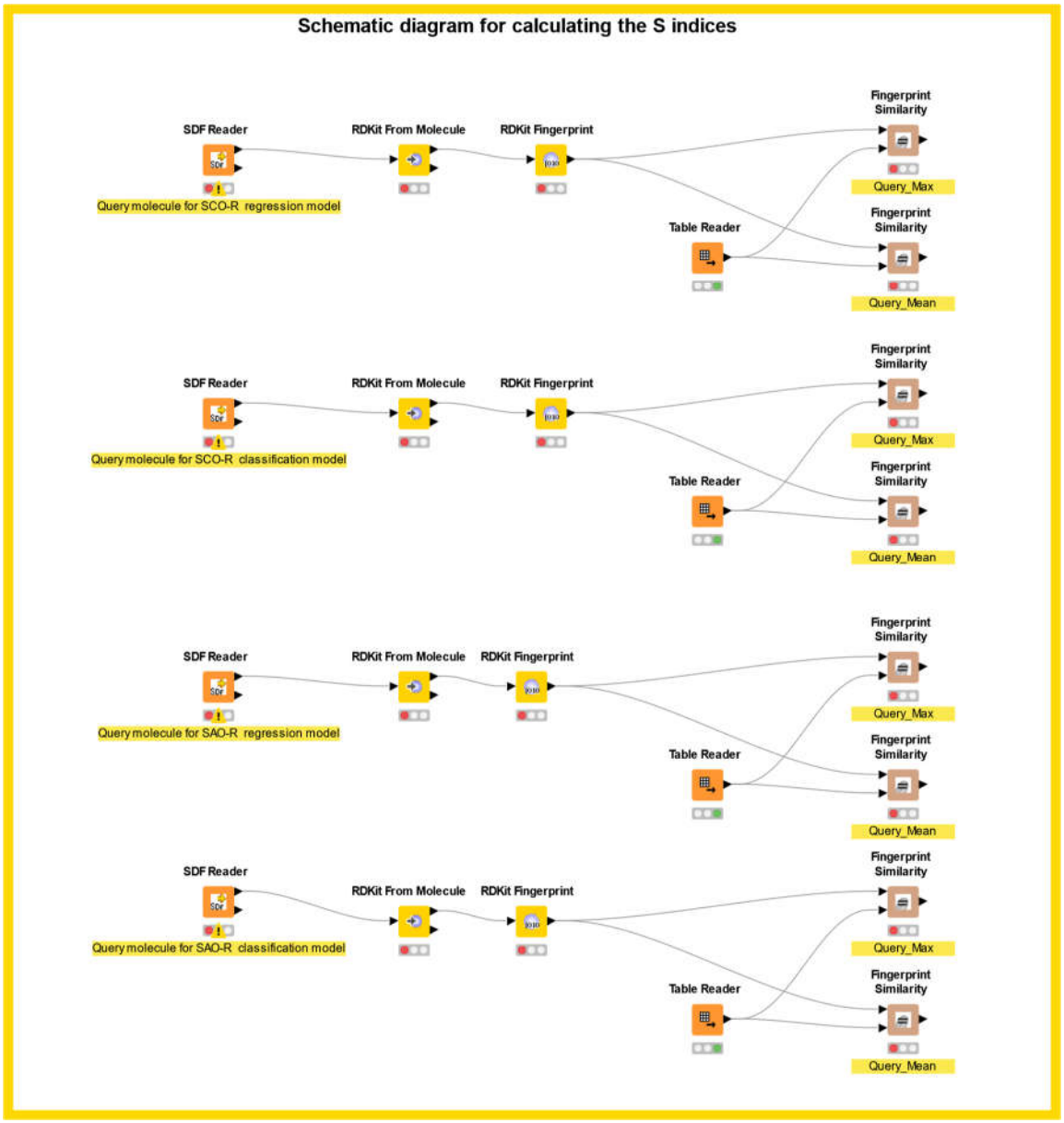

Figure 3. KNIME usage example of calculating the S indices.