How to predict new compounds

(1) About the models of the prediction system:

Table 1 shows the five models of the permeability prediction system.

Table 1. five models of the permeability prediction system.

Dataset	RRCK-C	PAMPA-C	Caco2-L	Caco2-C	Caco2-A
Model	MOE2D-SVR	MOE2D-LightGBM	MOE2D-RF	MOE2D-RF	MOE2D-RF

(2) Descriptors：

The models were built on KNIME platform (version 4.7.0). The MOE2D descriptors were calculated by MOE (version 2019). The details of the selected MOE2D descriptors related to the above five models are as follows:

(1) RRCK-C (103):

['apol', 'ast_violation', 'a_acc', 'a_aro', 'a_don', 'a_donacc', 'a_heavy', 'a_hyd', 'a_IC', 'a_nC', 'a_nF', 'a_nH', 'a_nN', 'a_nO', 'a_nS', 'balabanJ', 'BCUT_PEOE_1', 'BCUT_PEOE_2', 'BCUT_SLOGP_1', 'BCUT_SLOGP_2', 'BCUT_SMR_1', 'bpol', 'b_1rotN', 'b_double', 'b_max1len', 'chi1v', 'chi1v_C', 'chi1_C', 'chiral', 'chiral_u', 'diameter', 'h_ema', 'h_emd', 'h_emd_C', 'h_logD', 'h_logP', 'h_logS', 'h_log_pbo', 'h_pavgQ', 'h_pKa', 'h_pKb', 'h_pstrain', 'Kier2', 'Kier3', 'KierA2', 'KierFlex', 'lip_acc', 'lip_violation', 'logP(o/w)', 'logS', 'opr_brigid', 'opr_nring', 'opr_nrot', 'opr_violation', 'PEOE_PC+', 'PEOE_PC-', '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_VSA1', 'SMR_VSA2', 'SMR_VSA3', 'SMR_VSA4', 'SMR_VSA5', 'SMR_VSA6', 'SMR_VSA7', 'TPSA', 'VDistEq', 'vsa_don', 'weinerPath', 'weinerPol']

(2) PAMPA-C (71):

['apol', 'a_aro', 'a_don', 'a_donacc', 'a_hyd', 'a_nO', 'balabanJ', 'BCUT_SLOGP_2', 'b_1rotN', 'b_max1len', 'chi1v_C', 'chi1_C', 'chiral', 'diameter', 'GCUT_PEOE_3', 'h_emd', 'h_emd_C', 'h_logD', 'h_logP', 'h_logS', 'h_log_pbo', 'h_pavgQ', 'h_pKa', 'h_pKb', 'h_pstates', 'h_pstrain', 'Kier3', 'KierFlex', 'lip_violation', 'logP(o/w)', 'logS', 'opr_brigid', 'opr_nring', 'opr_nrot', 'opr_violation', 'PEOE_PC-', 'PEOE_VSA+0', 'PEOE_VSA+1', 'PEOE_VSA+2', 'PEOE_VSA+3', 'PEOE_VSA+4', 'PEOE_VSA+6', 'PEOE_VSA-0', 'PEOE_VSA-2', 'PEOE_VSA-4', 'PEOE_VSA-6', 'PEOE_VSA_NEG', 'PEOE_VSA_PNEG', 'PEOE_VSA_POS', 'petitjeanSC', 'radius', 'reactive', 'SlogP', 'SlogP_VSA0', 'SlogP_VSA1', 'SlogP_VSA3', 'SlogP_VSA4', 'SlogP_VSA5', 'SlogP_VSA8', 'SlogP_VSA9', 'SMR_VSA1', 'SMR_VSA2', 'SMR_VSA3', 'SMR_VSA4', 'SMR_VSA5', 'SMR_VSA6', 'SMR_VSA7', 'VAdjMa', 'VDistEq', 'vsa_don', 'weinerPath']

(3) Caco2-L (82):

['apol', 'ast_fraglike', 'ast_violation', 'ast_violation_ext', 'a_acc', 'a_aro', 'a_don', 'a_hyd', 'a_nN', 'a_nS', 'balabanJ', 'BCUT_PEOE_2', 'BCUT_SLOGP_2', 'BCUT_SLOGP_3', 'BCUT_SMR_2', 'b_double', 'b_max1len', 'b_rotR', 'chiral', 'chiral_u', 'diameter', 'GCUT_PEOE_3', 'GCUT_SLOGP_3', 'h_emd_C', 'h_logD', 'h_logP', 'h_logS', 'h_log_pbo', 'h_pavgQ', 'h_pKa', 'h_pKb', 'h_pstates', 'h_pstrain', 'lip_don', 'lip_druglike', 'lip_violation', 'logP(o/w)', 'logS', 'mutagenic', 'opr_brigid', 'opr_leadlike', 'opr_violation', 'PEOE_PC-', '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-6', 'PEOE_VSA_FNEG', 'PEOE_VSA_FPOS', 'PEOE_VSA_NEG', 'radius', 'reactive', 'rsynth', 'SlogP', 'SlogP_VSA1', 'SlogP_VSA3', 'SlogP_VSA4', 'SlogP_VSA5', 'SlogP_VSA6', 'SlogP_VSA7', 'SlogP_VSA8', 'SlogP_VSA9', 'SMR_VSA1', 'SMR_VSA3', 'SMR_VSA4', 'SMR_VSA5', 'SMR_VSA6', 'SMR_VSA7', 'VAdjEq', 'VAdjMa', 'VDistEq', 'vsa_don', 'weinerPath']

(4) Caco2-C (79):

['apol', 'ast_violation', 'ast_violation_ext', 'a_acc', 'a_aro', 'a_don', 'a_donacc', 'a_nCl', 'a_nF', 'a_nN', 'a_nS', 'balabanJ', 'b_1rotN', 'b_double', 'b_max1len', 'chiral', 'chiral_u', 'diameter', 'GCUT_PEOE_3', 'h_emd', 'h_emd_C', 'h_logD', 'h_logP', 'h_logS', 'h_log_dbo', 'h_log_pbo', 'h_pavgQ', 'h_pKa', 'h_pKb', 'h_pstates', 'h_pstrain', 'lip_druglike', 'lip_violation', 'logP(o/w)', 'logS', 'opr_brigid', 'opr_leadlike', 'opr_nring', 'opr_violation', 'PEOE_PC-', '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-6', 'PEOE_VSA_NEG', 'petitjeanSC', 'radius', 'reactive', 'SlogP', 'SlogP_VSA0', 'SlogP_VSA1', 'SlogP_VSA3', 'SlogP_VSA4', 'SlogP_VSA5', 'SlogP_VSA6', 'SlogP_VSA7', 'SlogP_VSA8', 'SlogP_VSA9', 'SMR_VSA1', 'SMR_VSA2', 'SMR_VSA3', 'SMR_VSA4', 'SMR_VSA5', 'SMR_VSA6', 'SMR_VSA7', 'TPSA', 'VAdjMa', 'VDistEq', 'vsa_don', 'weinerPath']

(5) Caco2-A (92):

['apol', 'ast_fraglike', 'ast_violation', 'ast_violation_ext', 'a_acc', 'a_aro', 'a_don', 'a_donacc', 'a_hyd', 'a_nCl', 'a_nF', 'a_nN', 'a_nO', 'a_nS', 'balabanJ', 'BCUT_SLOGP_2', 'b_1rotN', 'b_1rotR', 'b_double', 'b_max1len', 'b_rotR', 'chiral', 'chiral_u', 'diameter', 'GCUT_PEOE_3', 'h_logD', 'h_logP', 'h_logS', 'h_log_dbo', 'h_log_pbo', 'h_pavgQ', 'h_pKa', 'h_pKb', 'h_pstates', 'h_pstrain', 'Kier3', 'lip_don', 'lip_druglike', 'lip_violation', 'logP(o/w)', 'logS', 'opr_brigid', 'opr_leadlike', 'opr_nring', 'opr_violation', 'PEOE_PC+', 'PEOE_PC-', '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-6', 'PEOE_VSA_FNEG', 'PEOE_VSA_FPOS', 'PEOE_VSA_NEG', 'petitjeanSC', 'radius', 'reactive', 'rings', 'rsynth', 'SlogP', 'SlogP_VSA0', 'SlogP_VSA1', 'SlogP_VSA3', 'SlogP_VSA4', 'SlogP_VSA5', 'SlogP_VSA6', 'SlogP_VSA7', 'SlogP_VSA8', 'SlogP_VSA9', 'SMR_VSA1', 'SMR_VSA2', 'SMR_VSA3', 'SMR_VSA4', 'SMR_VSA5', 'SMR_VSA6', 'SMR_VSA7', 'TPSA', 'VAdjEq', 'VAdjMa', 'VDistEq', 'vsa_acc', 'vsa_don', 'weinerPath']

(3) How to form your own prediction pipeline:

The details of constructing your own workflow are shown in **Figure 1**, **Figure 2** and **Figure 3**.

Explanation of workflow:

1) The local model file is read by the **Model Reader** node above, while the below reads the model-Normalizer.zip 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 RRCK-C.csv, example for Caco2-C.csv and example for PAMPA-C.csv). If you just want to predict molecules without experimental values or labels, you should ignore the 'Permeability' and disconnect the evaluation nodes for example 'Numeric Scorer'.

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

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

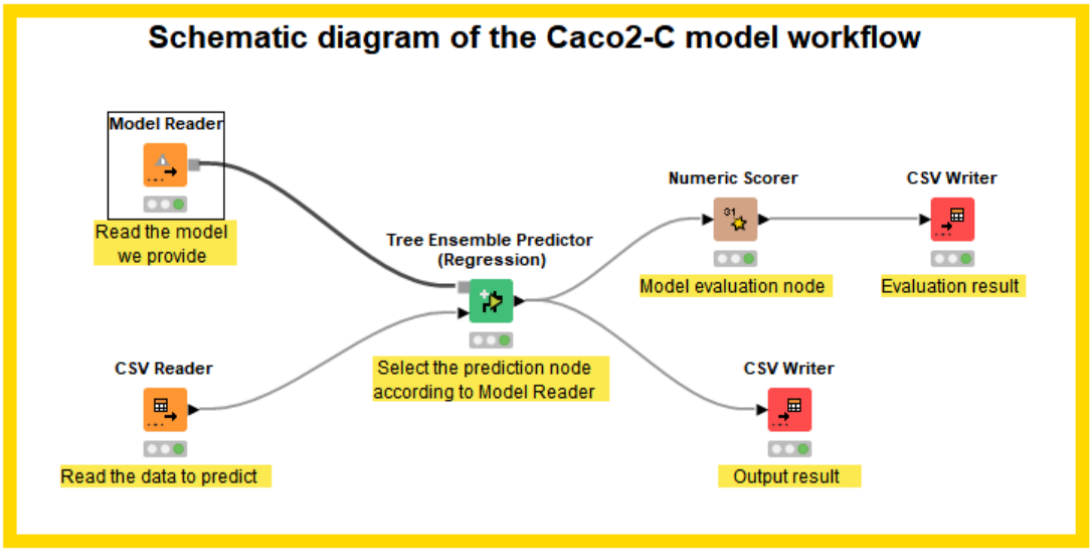

Figure 1. KNIME usage example of Caco2-C model.

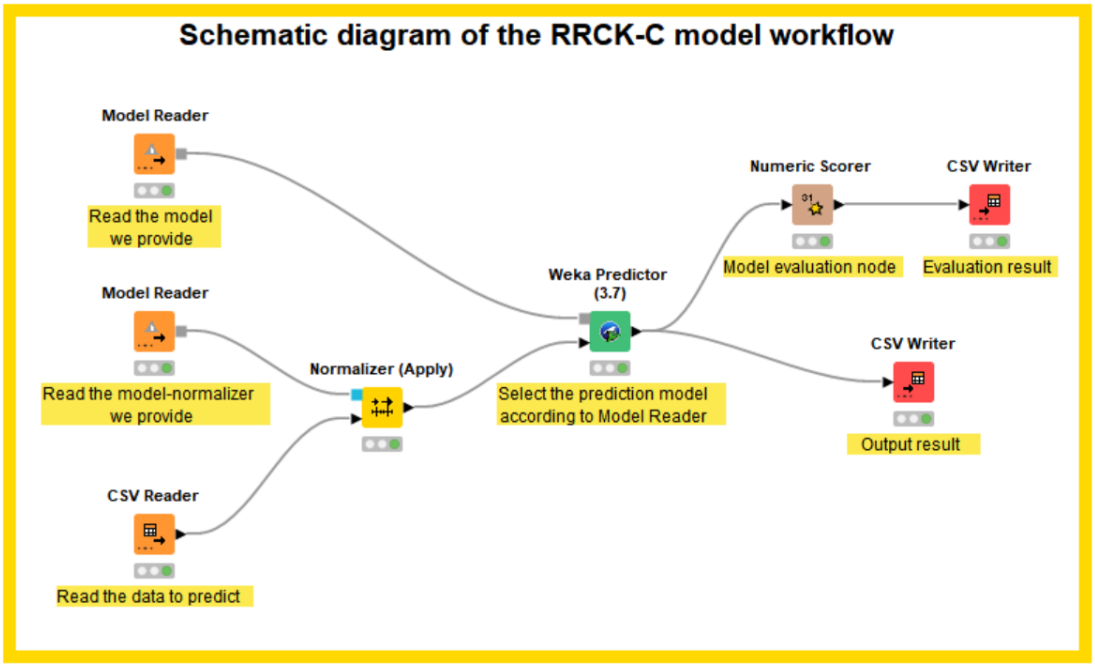

Figure 2. KNIME usage example of RRCK-C regression model.

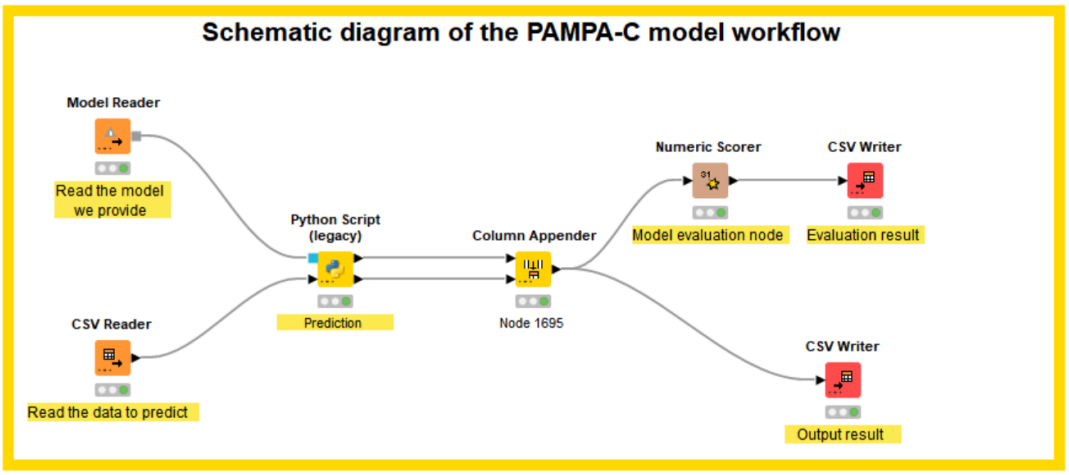

Figure 3. KNIME usage example of PAMPA-C regression model.