CHEM452/CHBE413
Chemical Data Science and Engineering
Quiz 2
Released: Tuesday, November 18th, 2025 – 5:00pm
Due Date: Thursday, November 20th, 2025 – 5:00pm

Note: *This is a 48-hour take-home exam. There is no collaboration allowed on this exam. You may use your previous homeworks, lecture notes, slide sets, provided Jupyter notebook examples, and the course textbooks.*

1. **A Machine Learning Model for Radical Copolymerization Prediction.** (datasets: *Ptype_dataset_quiz_training.csv*, *Ptype_dataset_quiz_test.csv*. These datasets can be found under Files/Datasets/ on the Canvas webpage.)

The microstructure of copolymers, which encompasses both block length statistics and variations in chain composition, plays a crucial role in determining physical and chemical properties. In a binary copolymer, in which the polymer chain consists of two different monomers, researchers have classified copolymers into different types by described by the monomer sequence microstructures as shown in **Figure 1**.

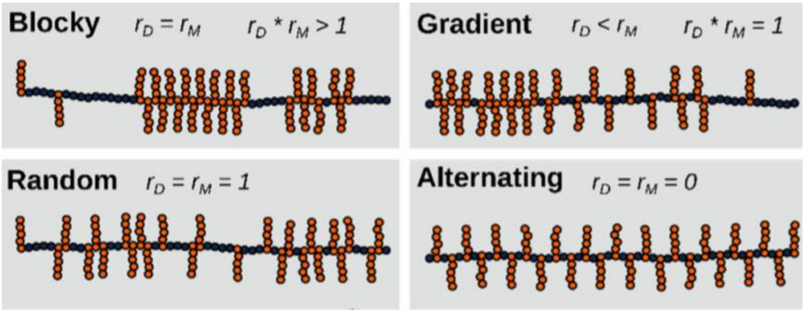

Figure 1. The accompanying figure illustrates the four example types of copolymerization—blocky, gradient, random, and alternating—distinguished by their respective arrangements of the two different monomer units along the polymer chain. In the figure, the two monomer types are represented by orange sidechains and black single-beads. r_D and r_M are reactivity ratios between the two monomers, which is irrelevant to our quiz. (source: K. C. McCleary-Petersen, et al. *Macromolecules*, **2025,** *58 (1),* 18-31.)

In this quiz, the task is to use features of the two monomer molecules (monomer pair) to train a classification model predicting polymerization type between: **{0:'alternating', 1:'random', 2:'gradient', 3:'pseudo-alternating', 4:'blocky'}.**

Contained in the attached dataset files are the following columns, with each row corresponding to an experimental record extracted from literature from different sources with monomer information and polymerization type label.

Feature	Feature Description
Monomer_1	Name of monomer 1
Monomer_2	Name of monomer 2
Monomer_1_smi	SMILES of monomer 1
Monomer_2_smi	SMILES of monomer 2
mol1_xxx, mol2_xxx	Descriptors calculated for monomer 1 and monomer 2 (Detailed descriptor explanation see the Appendix Table)
label	The reported copolymerization type

In answering these questions, you are expected to employ **all best practices** discussed in the course including feature scaling, cross-validation, and hyperparameter tuning.

(a) **(15 points)** Pre-process and load the training dataset, adhering to all best practices covered in this course. After completion, present the .info() output for the imported DataFrame. Additionally, identify and list the 10 most frequently occurring monomers in the training data (note that each row contains two monomers) and use RDKit to draw these monomers. **Note: Retain all duplicated rows, as they represent distinct experimental entries sourced from various literature**.

(b) **(10 points)** Intuitively, copolymerization type should be related to the frontier molecular orbital (FMO) energy (i.e. ["mol1_HOMO", "mol2_HOMO", "mol1_LUMO", "mol2_LUMO"]) of the two monomers. Visualize the distribution of each of these FMO energy descriptors using a histogram and illustrate the pair-wise feature correlations with a heatmap.

(c) **(10 points)** Analyze the distribution (proportion) of each copolymerization label in both the training and test datasets. Evaluate the consistency of this distribution between the two datasets and determine whether the classification labels are balanced.

(d) **(15 points)** As you may have noticed, identical monomer pairs are sometimes classified under different copolymerization schemes across various literature sources (different rows in the dataset). What potential experimental reasons could explain this discrepancy?

This inconsistency introduces significant noise into our dataset, establishing a natural prediction limitation for any model attempting to predict the correct copolymerization scheme. If we employ a "voting" mechanism to assign the true label within the dataset, we can quantify this internal noise by determining the upper limit of classification performance, as illustrated in the example below. **Note: the confusion matrix shall be normalized, which means each row shall add up to unity, which also applies for 1e-1g**

***For example**: Consider three data entries for the (ethylene, styrene) monomer pair, labeled as (blocky, blocky, alternating). Since "blocky" is the majority vote, it is considered the true polymerization type for this pair. The three data entries are then evaluated as predictions: (True, True, False). Consequently, the ceiling accuracy for this monomer pair is calculated as 0.67.*

(e) **(30 points)** Train a Random Forest (RF) classification model on the training dataset, optimizing for classification accuracy, and employ 5-fold cross-validation to identify the optimal set of hyperparameters. Then, the best-performing model will be tested on the dedicated test set, reporting the final accuracy and F1 score, and generating and displaying the confusion matrix. What are the top-10 important features given by RF model’s native feature importance analysis?

Conduct PCA analysis and visualize the training dataset on PC1 and PC2 with the top-10 important features, color the datapoints with copolymerization labels.

Use SHAP analysis on the top of trained model and make a beeswarm plot showing the top-10 important features contributing to each classification label. What are the top-10 important features given by SHAP analysis? Are they consistent with scikit-learn RF model native method's result?

(f) **(20 points)** Finally, use classification accuracy as the metric to train a fully connected neural network (FCNN) with two hidden layers (128 and 64 neurons, ReLU activation) on the training dataset. Plot the learning curve during the training process. After training, test the model on test set, and report accuracy, F1 score and plot the confusion matrix. Compared with RF model in (e), is the FCNN achieving a better result?

(g) **(optional, bonus 10 points)** Considering the inherent chemical symmetry in copolymerization, swapping the identities of Monomer 1 and Monomer 2 should not alter the fundamental copolymerization scheme. Does your existing FCNN or RF model inherently possess this invariance property? If yes, rationalize your judgement. If no, propose possible way of solving this problem.

Appendix Table: Descriptors Explanation

Descriptor Notation	Descriptions
AtomNum	Number of atoms in molecule
Weight	Molecular weight
Volume	Molecular volume
Density	Molecular density defined as Weight / Volume
Farthest Distance	Farthest distance between atoms
Mol_Radius	Molecular radius
Mol_Size_Short	Shortest molecular dimension length size
Mol_Size_2	Medium molecular dimension length size
Mol_Size_L	Longest molecular dimension length size
Length_Ratio	$\frac{\text{Mol\_Size\_L}}{(\text{Mol\_Size\_L}+\text{Mol\_Size\_2}+\text{Mol\_Size\_Short})}$
Len_Div_Diameter	$\frac{\text{Mol\_Size\_L}}{2\cdot\text{Mol\_Radius}}$
MPP	Molecular planarity parameter
SDP	Span of deviation from plane
HOMO	Highest occupied molecular orbital energy
HOMO_1	HOMO-1 energy
LUMO	Lowest unoccupied molecular orbital energy
LUMO_Add1	LUMO+1 energy
ODI_HOMO_1	Orbital delocalization index (ODI) of HOMO-1
ODI_HOMO	ODI of HOMO
ODI_LUMO	ODI of LUMO
ODI_LUMO_Add1	ODI of LUMO+1
ODI_Mean	Mean of ODI of FMOs
ODI_Std	Standard deviation of ODI of FMOs