Supporting Information

Synthesis of Challenging Cyclic Tetrapeptides by Machine Learning Assisted High-throughput Continuous Flow Technology

Chaoyi Li,†[a] Jiaping Yu,†[a] Wanchen Li,[a] Jingyuan Liao,[a] Junrong Huang,[a] Jiaying Liu,[a] Wei Zhao,[a] Yinghe Zhang,*[a] Yuxiang Zhu,*[a] and Hengzhi You*[a]

[a] School of Science, Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, China.

† These authors contributed equally to this work.

*Corresponding authors E-mail: Hengzhi You (youhengzhi@hit.edu.cn); Yuxiang Zhu (zhuyuxiang@hit.edu.cn); Yinghe Zhang (zhangyinghe@hit.edu.cn)

Table of Content

Supporting Information

1 General Information

All commercially available reagents including the substrates were used as received. Dry dichloromethane was purchased from Energy-Chemical. Phenylglyoxal Hydrate, and other reagents without specified were purchased from Adamas. Column chromatography purifications were performed by flash chromatography using Merck silica gel 60. The reversed-phase medium pressure liquid chromatography (RP-MPLC) performed on Santai Science Inc. SepaBean® machine T with SW-5222-040-SP C18 26 × 185 mm column. The semi-preparative high-performance liquid chromatography (SP-HPLC) performed on Agilent 1260 with Nouryon Kromasil® C18 10 × 250 mm column. ^{1}H NMR, and ^{13}C NMR spectra were recorded using Q.One Instruments Quantum-I 400M spectrometer. ^{1}H NMR and ^{13}C NMR chemical shifts were reported in parts per million (ppm) downfield from tetramethylsilane. Coupling constants (J) are reported in Hertz (Hz). The residual solvent peak was used as an internal reference: ^{1}H NMR (chloroform δ 7.26) and ^{13}C NMR (chloroform δ 77.16). The following abbreviations were used to explain the multiplicities: s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet, br = broad. High Resolution Mass Spectrometer (HRMS) were obtained on Waters Xevo G2-XS QTof. Ultra Performance Liquid Chromatography Mass spectrometry (UPLC-MS) spectra were acquired on an Agilent Technologies 1290 Infinity LC equipped with an Agilent Technologies 6270 Quadrupole mass spectrometer. The infrared (IR) spectra were acquired on a Thermo Scientific™ Nicolet™ iS50 FTIR. Melting points of compounds were test on The SGWX-4 Melting Point Apparatus.

2. Preparation of the cyclization precursors

All tetrapeptides were synthesized through solid-phase peptide synthesis (SPPS) method.[1]

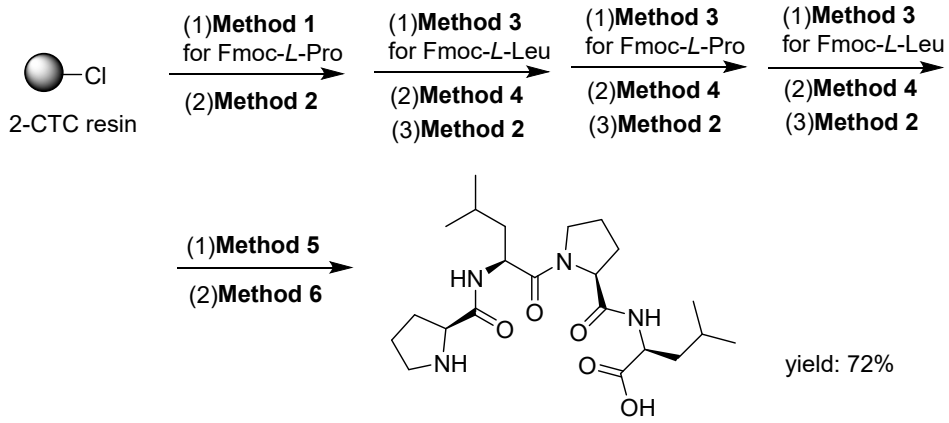


Figure S1. Synthesis procedure for Pro-Leu-Pro-Leu

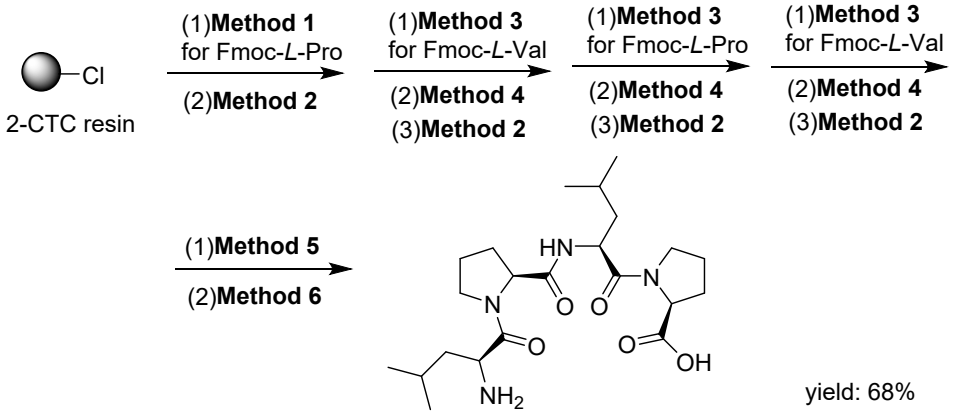


Figure S2. Synthesis procedure for Leu-Pro-Leu-Pro

Method 1 (General Method for Fmoc-AA-OH Attachment on CTC Resin):

Swelling the 2-chloro-trimethylphenol resin (CTC resin, 0.44 g, 0.5 mmol) resin in anhydrous dichloromethane for 15 min, then the solvent was removed under reduced pressure. Fmoc-AA-OH (2 eq, 0.100 M) and N,N-diisopropylethylamine (DIPEA) (5 eq, 0.250 M) was dissolved in 5 mL of anhydrous DCM (10.0 mL/g resin), which was then mixed with the resin and shaken at room temperature for 2 h. After 2 hours, methanol (0.1 mL g^{-1} resin) was added to the reaction mixture and the resulting mixtures was shaken for

15 min. to cap the vacant CTC resin sites. The solvent was then removed under reduced pressure, and the resulting resin was washed sequentially with DMF (1 min × 3), DCM (1 min × 3), and MeOH (1 min × 3). Nitrogen flush was employed to remove the solvents, ensuring the dryness of the resin.

Method 2 (General Method for Fmoc Deprotection):

A 20% vol. solution of 4-methylpiperidine in DMF (5 mL) was mixed with the resin and was shaken for 10 min., which was washed once with DMF. Next, 20% vol. 4-methylpiperidine solution (in DMF) was added and was shaken to achieve the deprotection of Fmoc. The resulting resin was washed three times with DMF, and nitrogen flush was employed to remove the solvents, ensuring the dryness of the resin.

Method 3 (General Method for Solid Phase Synthesis Using Oxyma-B/DIC):

Fmoc-AA-OH (2 eq., 0.100 M), Oxyma-B (0.2 eq.), and DIC (4 eq.) were dissolved in 5 mL of DMF, it was then mixed with the resin and shook for 2 h. Upon completion, the resin was washed sequentially with DMF (1 min × 3) and DCM (1 min × 3) to ensure removal of unreacted amino acids.

Method 4 (General Method for Capping Free Amino Groups):

Fmoc-deprotected peptide resin was treated with a pyridine solution containing 10% acetic anhydride (5 mL) and shook at room temperature for 5 min. to cap any exposed amino groups that might not have fully reacted in the previous step, which would prevent the side reactions in the subsequent coupling step. The resin was then filtered and washed with DMF (1 min × 3).

Method 5 (General Method for Resin Cleavage Mediated by HFIP):

After completing the Fmoc deprotection, the resulting resin was added 5 mL of a 20% vol. HFIP (hexafluoroisopropanol) in DCM solution and shook for 40 min to cleave the peptide sequence. Nitrogen flush was employed to remove the solvents, and the resin was subsequently washed several times with DCM to

obtain the filtrate.

Method 6 (General Method for Linear Peptide Purification):

The filtrate was concentrated under reduced pressure, before adding an appropriate amount of cold ether. Once the white solids precipitate was formed, centrifugation was performed to obtain the product as a white solid.

3 General procedure for the cyclization reaction under batch conditions

To a solution of linear precursor (0.1mM) in MeCN (100 mL) was added a solution of HATU (0.3 mM, 4 mL) and DIPEA (0.6 mM) in MeCN. The reaction mixture was heated to room temperature (r.t.) and stirred for 4 h. LC-MS analysis was performed to confirm the completion of the starting materials (Figure S3). Upon completion, the reaction mixture was concentrated in *vacuum* and purified by reverse-phase separation (using methanol and water as the mobile phase). The collected fraction was then concentrated in *vacuum* to afford the product as a white solid. HPLC spectrum was used to analysis the ratio of the products and epimers (MeOH:H_2O = 90%:10%, Signal=215nm, Figure S4). These peaks represent epimer (RT = 13.054 min.), product (RT = 13.363 min.), unknown intermediate (RT = 14.594 min.), dimer (RT = 15.389 min.), respectively from left to right.

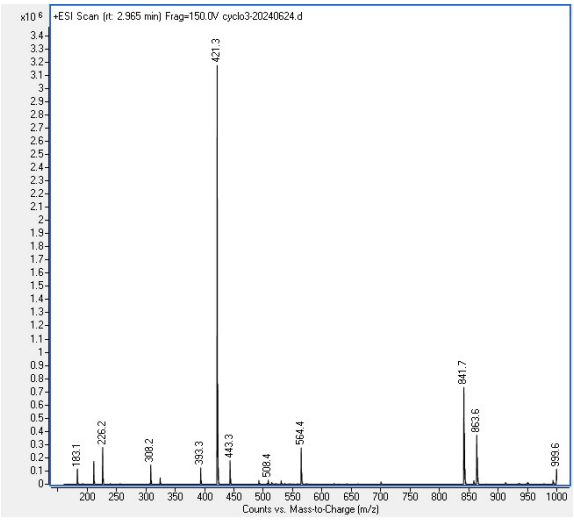
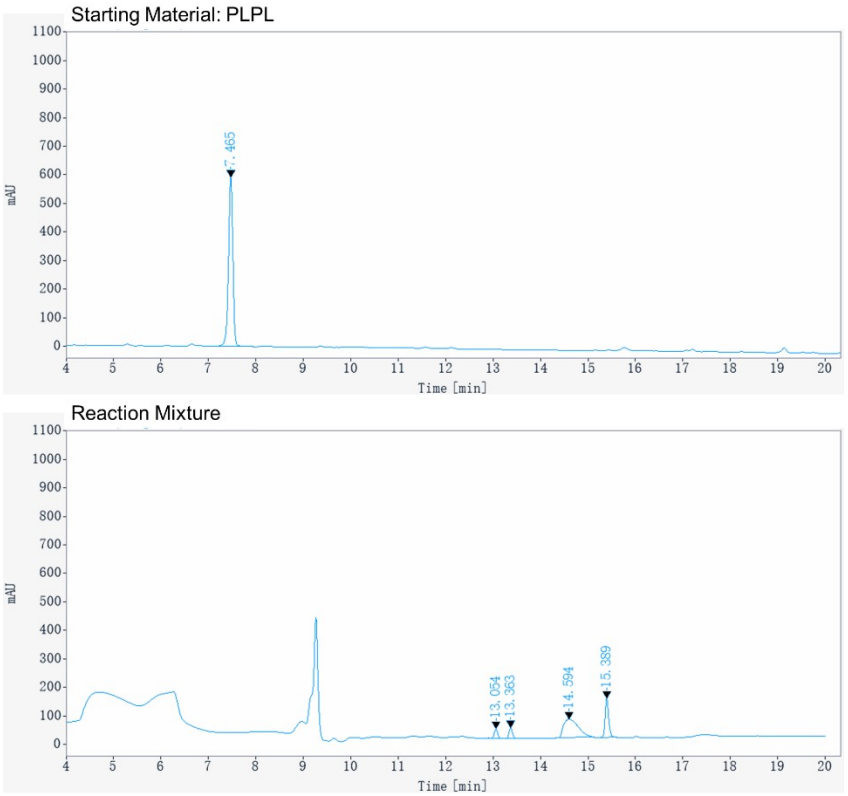


Figure S3. The LC-MS spectrum for analyses of the crude product under batch condition after 4 h at 100 ℃.



RT [min]	Width [min]	Area	Height	Area%
13.054	0.0794	163.2149	34.2545	6.4804
13.363	0.0860	197.4114	38.2565	7.8381
14.594	0.3690	1443.1584	65.1847	57.2998
15.389	0.0865	714.8265	137.7744	28.3818

Figure S4. The HPLC spectrum for the crude product under batch condition after 4 h at 100 °C (t_R = 13.36 min, 10% to 90% MeOH over 20min).

4 General procedure for the cyclization reaction under flow conditions

For the flow reactor: Polypropylene (PP) T-shape mixers, PP fittings, PP unions, and Teflon® tubes (inner diameter: 0.8 mm) were purchased from Nanjing Runze Fluid Control Equipment Co., LTD. Solutions were introduced to a micro-flow system with syringe pumps (TYD01, purchased from Lead Fluid Technology Co., Ltd.) equipped PP syringes (10 mL).

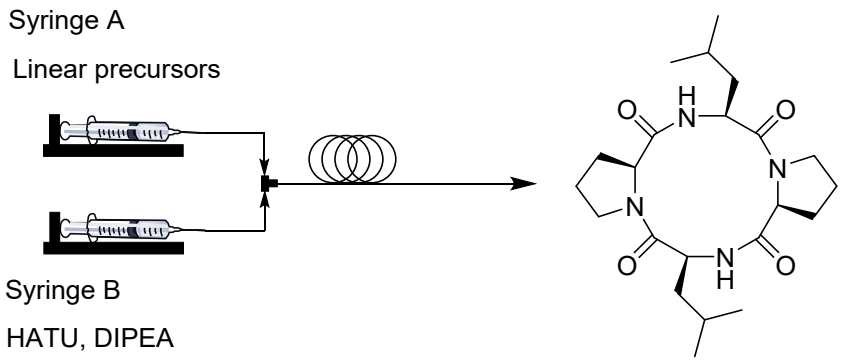


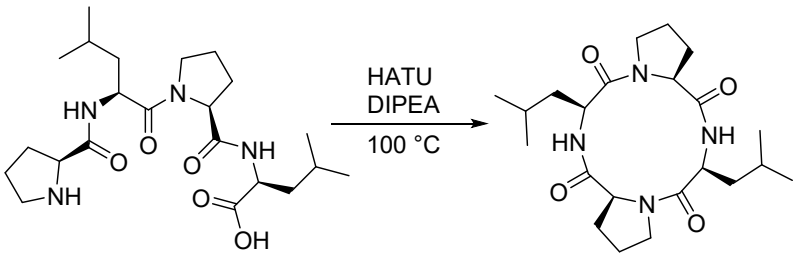
Figure S5. The overview of *cyclo*-(Pro-Leu)$_2$ synthesis under continuous flow condition

Syringe A: a solution of linear tetrapeptide (Pro-Leu-Pro-Leu) (0.01 mmol) in 10 mL of MeCN. Syringe B: a solution of HATU (11.4 mg, 3 eq.) and DIPEA (10.5 µL, 6 eq.) in 10 mL of MeCN. The reaction time (residence time) was controlled by adjusting the injection rate of the syringe pump (Figure S5). The reaction coil was heated in an oil bath at 50°C. With a residence time of 2 minutes, LC-MS analysis showed that the product was formed, but starting material was still detected. At a residence time of 5 minutes, the conversion of starting material reached 90%.

4.1 Investigation into the influence of coupling reagent and base loading on the synthesis of *cyclo*-(Pro-Leu)$_2$.

Typically, this cyclization requires an excess of coupling reagents and bases to achieve satisfactory results, as this approach accelerates the desired cyclization and suppresses epimerization.[2,3] Optimal performance was achieved with 3 equivalents of coupling reagent and 6 equivalents of base (Table 2, entry 7). However, adding more coupling reagents was unbeneficial, which is attributed to the increased dimer formation at higher coupling reagent concentration.

Table S1. Investigation of the effect of coupling reagent and base loading on the synthesis of *cyclo*-(Pro-Leu)$_2$



Entry	HATU (eq.)	DIPEA (eq.)	Yield[a] (%)
1	1.0	2.0	5.2
2	1.5	2.0	9.6
3	2.0	2.0	13.2
4	2.5	2.0	13.5
5	2.0	3.0	15.7
6	2.0	4.0	21.2
7	**3.0**	**6.0**	**30.1**
8	3.5	6.0	27.6
9	4.0	6.0	27.3

[a] Yields were determined by HPLC-UV analysis.

5 General screening procedure using HTE continuous-flow platform

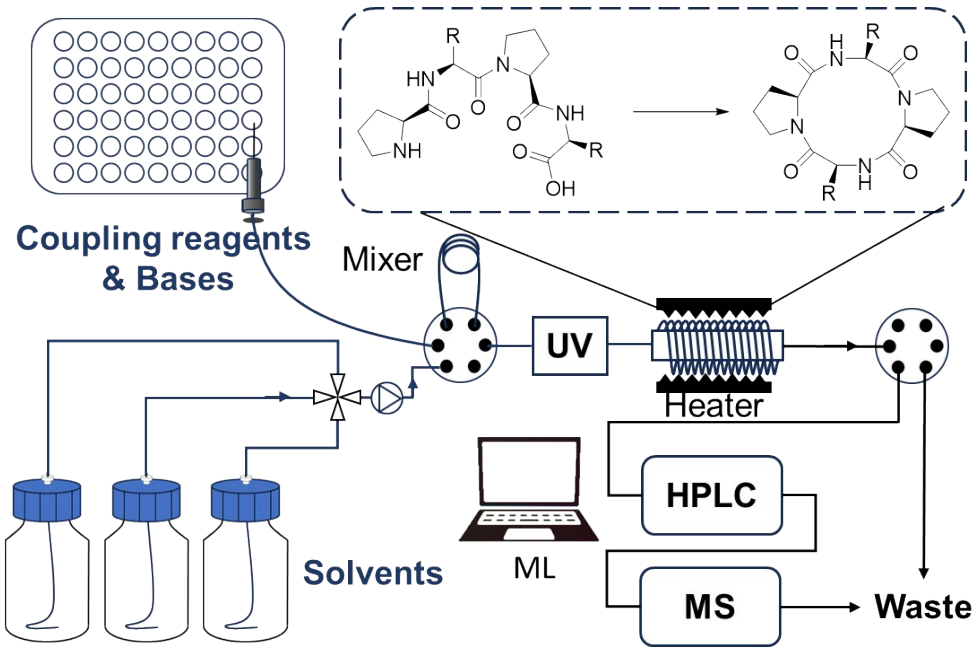


Figure S6. Overview of the operating principles of the HTE continuous-flow platform

(1) Firstly, the solvent channel is set through computer software (a 12-port valve connecting different solvent bottles). Once the reaction solvent is chosen, the pump provides the driving force. We use a series of reciprocating piston pumps to drive the solvent flow in the tubing (the pump can operate under both high and low pressure; high pressure ensures high precision and low delay). The solvent is first degassed using a degassing module to avoid bubbles that could affect the flow rate, ensuring it reaches the set precision.

(2) Next, set the reaction sample sequence and the automatic sampler start for sampling. The sampling needle then sequentially takes samples from the designated sample bottles. Before and after each reaction sampling, the automatic sampler draws a specified volume of isolation solvent to separate the reaction components at both ends from the carrier solvent, reducing

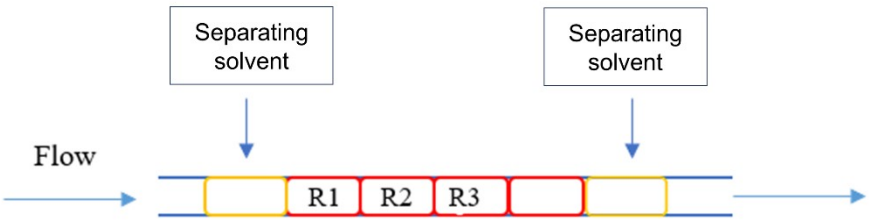


Figure S7. Overview of the mechanism of each reaction in flow

diffusion, as shown in Figure S7.

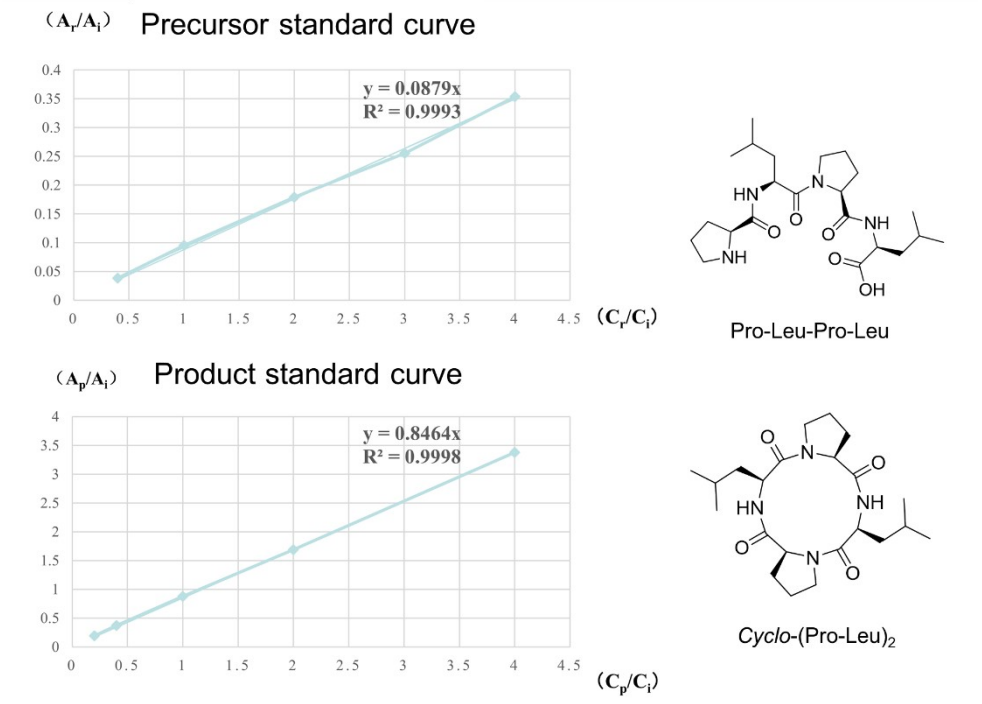
(3) The various reaction components are injected into a quantification loop, and then a six-port valve directs the reactants from the quantification loop into the solvent system. The solvent pushes the reactants to the UV detector, producing a signal intensity response, and then flows into the reactor for the chemical reaction (the reaction temperature can be set via the software). The reaction residence time is determined by the length of the tubing on the heating block and the flow rate.

(4) After the reaction is complete, the reaction liquid in the tubing is directed through the six-port valve's quantification loop (5 μL) into high-performance liquid chromatography (HPLC) for separation and quantitative analysis. The remaining reaction liquid is discarded as waste. A portion of the eluent (1 μL) is then diverted from the HPLC's backend to the mass spectrometer for qualitative analysis.

(5) Finally, based on the large amount of experimental data generated from rapid and efficient reactions, artificial intelligence is used to screen process routes, and automated devices validate and conduct online analysis of the selected processes to quickly identify the optimal reaction conditions.

In this study, the HTE conversion rate and yield calculations were determined using both internal standard and standard curve method. The anisole was selected as the internal standard to determine the concentration of starting materials. These values were determined based on the ratio of the peak area of the product to that of the internal standard. Standard solutions of 1 mM, 2 mM, 5 mM, 10 mM, and 20 mM were prepared, with 10 μL of each sample

Figure S8. The standard curve of precursor and product



injected. Standard curves for both the reactant and the product were generated based on peak area (using the same HPLC elution gradient as in reaction analysis). The standard curve for the reactant is shown in Figure S8.

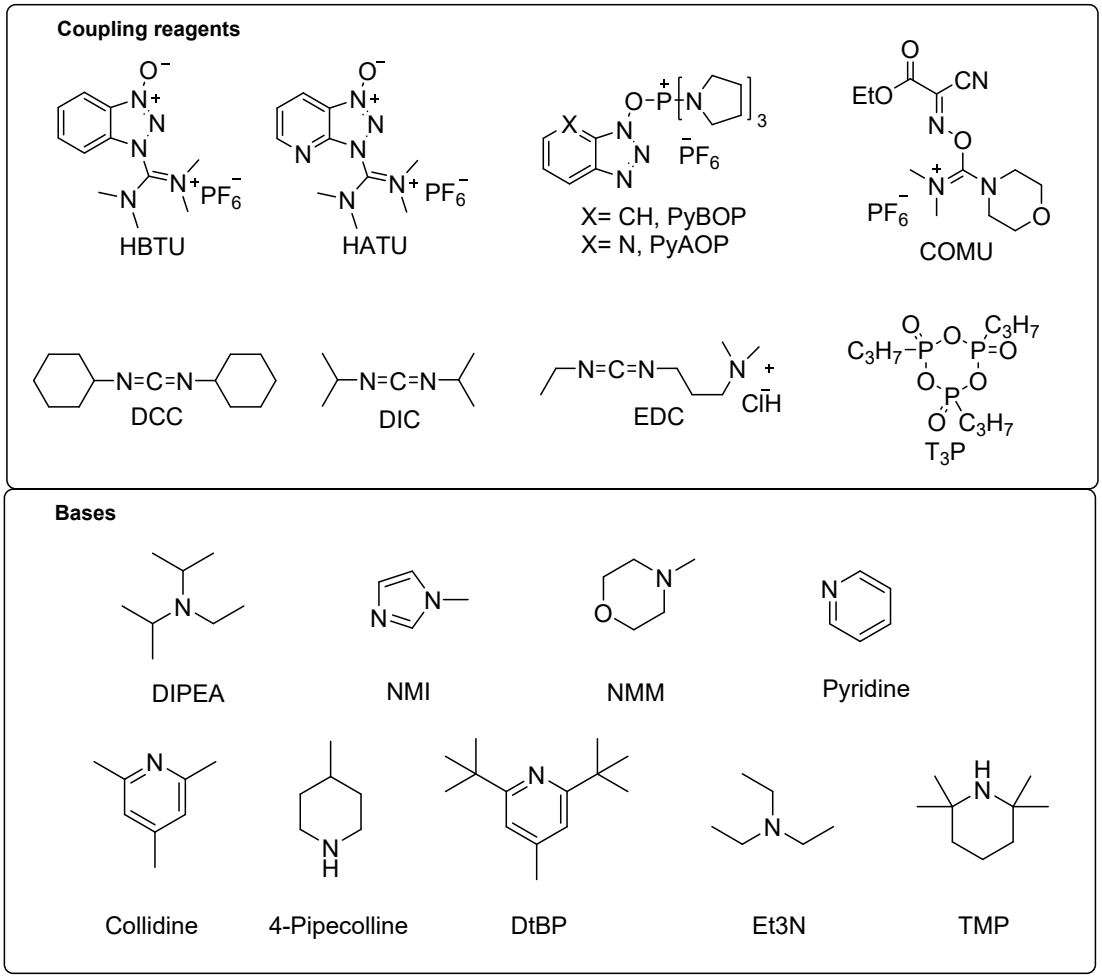


Figure S9. Structures of coupling reagents and bases screened.

17.6 mg of linear tetrapeptide (Leu-Pro-Leu-Pro) was dissolved in acetonitrile to prepare a 20 mM solution in a 2 mL sample vial. Anisole was used as the internal standard (5 mM). A coupling reagent solution (60 mM) and a base solution (120 mM) was prepared. All chose coupling reagents and bases are showed in Figure S9 The reactor temperature was set to 100°Cand the solvent pump flow rate was set to 1.0 mL/min. The reaction sequence was programmed to sequentially draw 10 μL of the linear tetrapeptide solution, 10 μL of the coupling reagent, and 10 μL of the base solution for each reaction. The reaction components were then injected into the tubing via a quantification loop for flow reaction. After the reaction, the six-port valve at the back end directed the reaction solution into the HPLC-MS system for automated data collection.

The effect of coupling reagents and bases in the HTE screening reactions

The table S2 shows the HTE screening results for the reaction yields and the product-to-epimer ratio (P:P*) under different coupling reagent and base combinations in DMF. The highest yield occurs with PyBOP-DIPEA (35.1%), followed by PyAOP-DIPEA (32.0%) and HATU-DIPEA (30.1%).

Table S2. HTE screening results in DMF with different coupling reagents and bases.

Reaction conditions	Yields(%)[a]	P:P*[b]
HATU-DIPEA	30.1	82:18
HBTU-DIPEA	28.7	76:24
TCFH-DIPEA	23.6	72:28
PyAOP-DIPEA	32.0	80:20
PyBOP-DIPEA	35.1	75:25
COMU-DIPEA	21.6	70:30
DCC-DIPEA	10.2	76:24
DIC-DIPEA	8.3	56:44
EDC-DIPEA	trace	*
T3P-DIPEA	7.2	55:45
HATU-NMI	25.5	88：12
HATU-NMM	24.6	83：17
HATU-Collidine	19.4	73：27
HATU-Pyridine	21.5	86：14
HATU-TMP	10.6	70：30
HATU-Et3N	16.8	68：32
HATU-DtBP	6.8	*
HATU-4Pipecoline	13.2	85：15

* The epimer of the product

[a] Yields were determined by HPLC-UV analysis.

[b] The ratio of the peak area

6 NMR study on the cyclization site

The choice of cyclization sites for peptides has a significant impact on the synthesized peptide molecules. Different cyclization sites can lead to varying stereo configurations, which affect the spatial structure and activity of the

peptide, as well as its stability.[4] In this study, the cyclic tetrapeptide molecules have a cyclic symmetrical structure, thus requiring exploration of two different cyclization sites. Cyclization experiments were conducted using two different linear tetrapeptide precursors, H-Pro-Leu-Pro-Leu-OH and H-Leu-Pro-Leu-Pro-OH, under the same conditions. After reverse-phase column separation and LC-MS analysis, preparative liquid chromatography was used to isolate the isomers. The analysis revealed that the cyclization reaction of linear precursor 2 was more effective, while linear precursor 1 was more prone to side reactions, resulting in the detection of isomer formation.

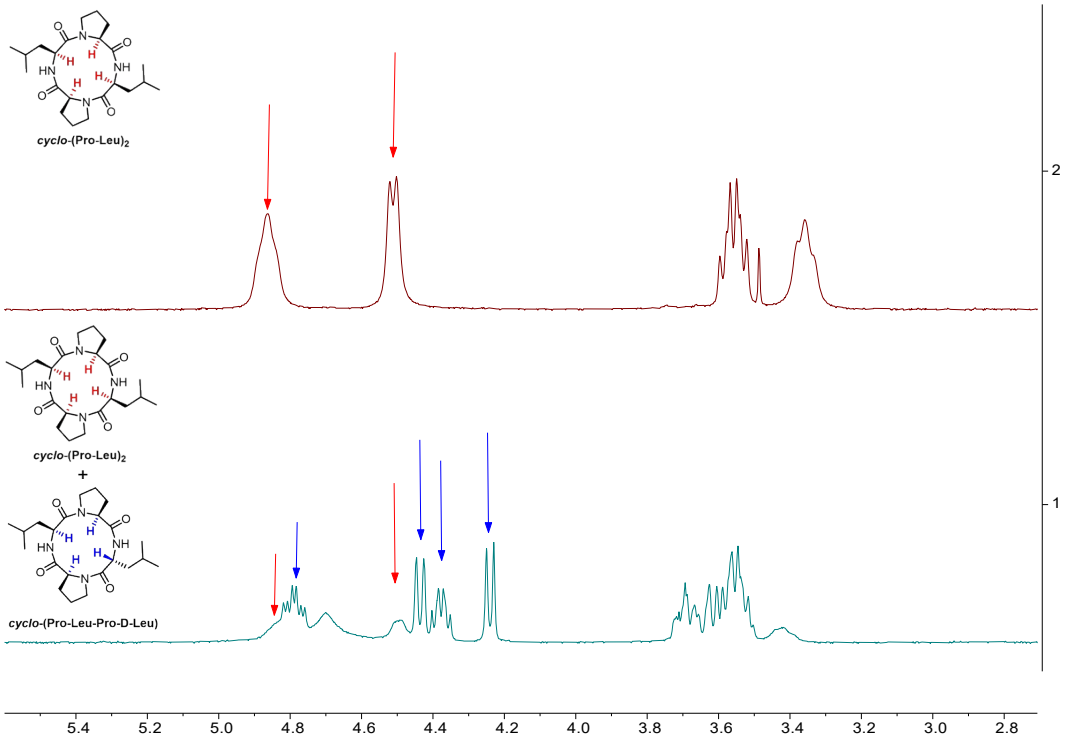


Figure S10. Comparison of amine hydrogen NMR shifts before and after isolated with preparation HPLC. Top side represents the product has been isolated. The bottom side is the mixtures (contain product and epimer) before isolated.

NMR comparison of the solated product *cyclo*-(Pro-Leu)$_2$ and crude mixtures suggested the formation of an isomer (Figure S10). To determine the structure of this isomer, we synthesized H-Pro-Leu-Pro-D-Leu-OH and H-Leu-Pro-Leu-Pro-OH and conducted the cyclization reaction. After NMR analysis,

this main side product was determined to be *cyclo*-(Pro-Leu-Pro-D-Leu), a epimer of the main product *cyclo*-(Pro-Leu) (Figure S11).

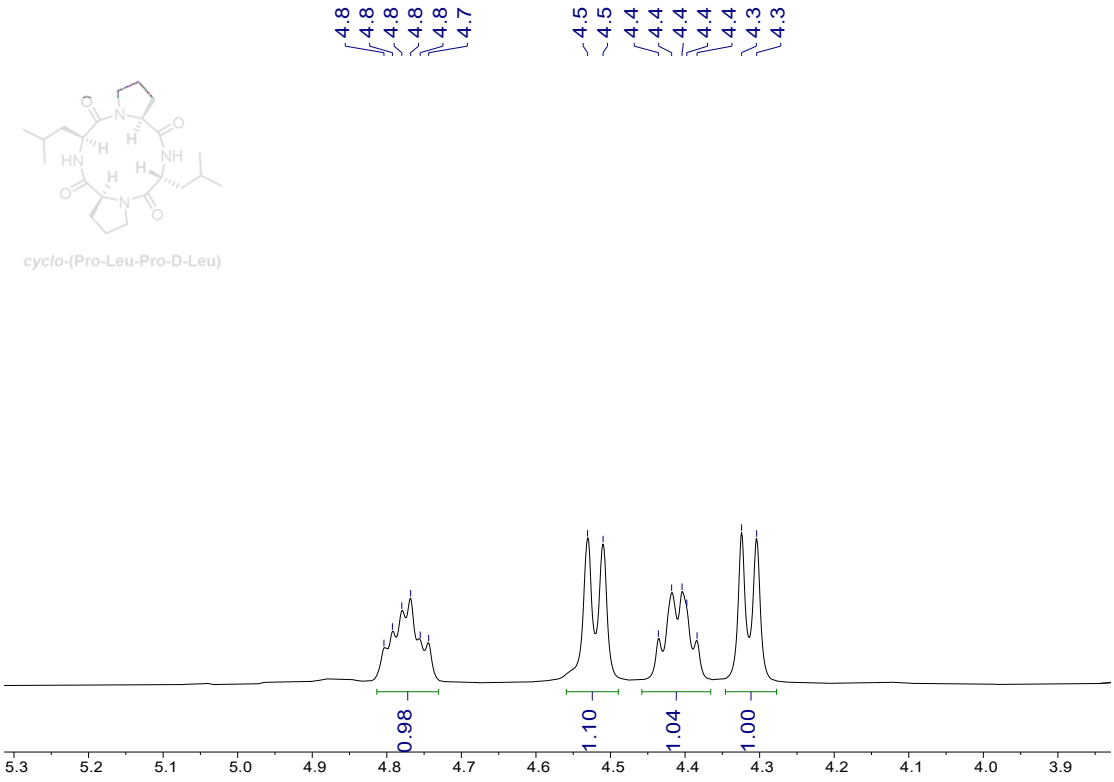


Figure S11. The amine hydrogen NMR shifts of *cyclo*-(Pro-Leu-Pro-D-Leu)

7 Application of Machine learning to predict yields

All scripts and computation are based on Python and Sci-kit Learn package.[5-7] With the database of 270 cyclization HTE reactions, we wonder if we could leverage machine learning and proper chemical descriptors to predict reaction yields. As the screening conditions is mainly about coupling reagents, bases, and solvents, we consider to use chemical informatic descriptors to transform the structure to suitable format. The circular morgan molecular fingerprints was calculated with RDKit packages, setting radius 2 and 256 dimensions.

To optimize the basic hyperparameters of random forest model, we test several different numbers of decision trees and the max depth of each tree (Table S3). Although more decision trees lead to better prediction performance, considering the size of dataset and we select 100 as the investigated

parameter. For the depth of each decision tree, we noticed when it over 8, the R^2 is increased while the RMSE is also increased. Thus, to avoid probable overfitting, we decided depth of 8.

Table S3 Investigation of basic hyperparameter of random forest model.

Model	R^{2*}	RMSE*
RF (n_estimators=20)	0.8361	3.65
RF ((n_estimators=50)	0.8414	3.60
RF ((n_estimators=100)	**0.8426**	**3.59**
RF ((n_estimators=250)	0.8439	3.57
RF ((n_estimators=500)	0.8449	3.56
RF (max_depth=3)	0.7760	4.29
RF (max_depth =4)	0.8266	3.77
RF (max_depth =5)	0.8311	3.72
RF (max_depth =8)	**0.8445**	**3.56**
RF (max_depth =10)	0.8425	3.59

* Calculated the average results over 10 times 30/70 random split training.

For out of sample test, we propose two kinds of data splitting strategies. One is splitting dataset with different coupling reagents, and the other on is splitting with different bases. For coupling reagents splitting strategy, each out of sample test set will choose 3 different coupling reagent and the other 7 reagents as training set. For bases splitting strategy, each out of sample test will choose 3 bases as out of sample test set and the other 6 bases as training set. The details are shows in Table S4.

Table S4. The details of two out-of-sample training strategies.

Strategy	Out of sample task	Compounds in test set	Num. of Training set	Num. of Test set
Coupling reagent	Coupling-1	HATU+HBTU+TCFH	189	81
	Coupling-2	PyAOP+PyBOP+COMU	189	81
	Coupling-3	DCC+DIC+EDC	189	81
	Base-1	DIPEA+Et3N+TMP	180	90

	Base-2	Pyridine+Collidine+DtBP	180	90
	Base-3	NMM+NMI+4-Pipecoline	180	90

Feature selection and the importance investigation

The MACCS fingerprint (Molecular Access System Computational Chemistry Software) is a popular method for encoding molecular structures into a binary vector format. Essentially, it is a simple method of representing molecule structures, such as its functional groups, rings, bonds, and other structural characteristics, in a way that a computer can process efficiently. We applied the MACCS fingerprint consists of a predefined set of 166 bits, where each bit represents the presence or absence of a particular structural feature. The binary encoding of these features allows for comparison between different molecules and analysis the substructure features contributions to the ML models.

In this study, we aim to investigate the structural features of various components involved in chemical reactions, including precursors, solvents, coupling reagents, and bases, and their potential impact on reaction outcomes. To achieve this, we first compute MACCS fingerprints for each of the four components, which encode the molecular structures of these entities into binary vectors representing their key chemical features. By generating these fingerprints, we can analyze the structural similarities and differences between the components based on predefined substructural features. Next, to explore the relationship between these components, we perform a Spearman correlation analysis on the computed fingerprints. This non-parametric test helps to identify any monotonic correlations between the structural features of the components, providing insights into how these structural elements may interact or influence each other in the context of the reaction (Figure S12). The color scale on the right indicates the strength of the correlation between the different components and fingerprints. Positive correlations are shown in red, indicating a strong positive relationship between the features, while blue shows

negative correlations. We kept only those molecular fingerprints with Spearman correlations between 0.9 and -0.9 that effectively reducing the complexity of the dataset while retaining the most relevant features. This would help improve computational efficiency without sacrificing the essential information.

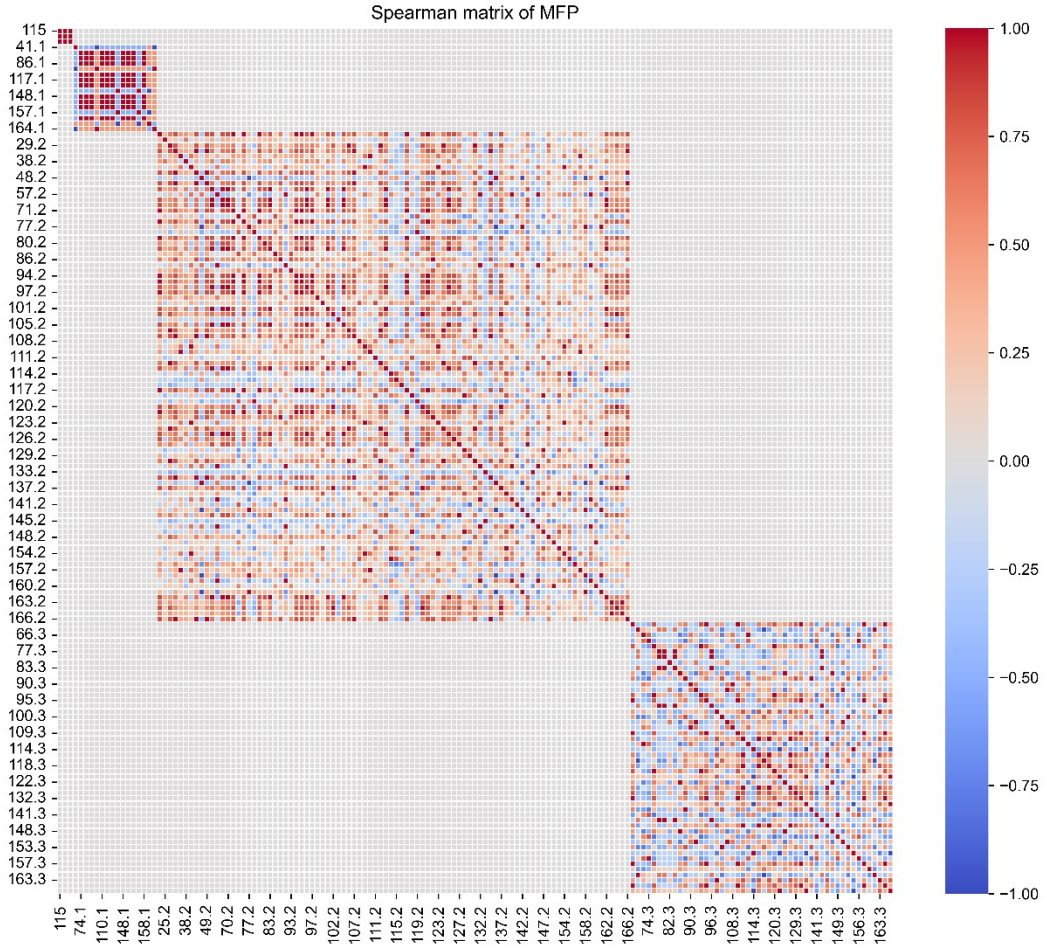


Figure S12. Spearman matrix of MFP features. Each label on the axes represents a bit position in the MACCS fingerprint, and the number after the decimal indicates the reaction components: none: Precursors 1: Solvents 2: Coupling reagents 3: Bases

Figure S13 shows the top 15 most important features for the random forest model in predicting yield. The x-axis represents the importance of each feature, and the y-axis lists the corresponding MACCS fingerprint positions. 166.2, 162.2 and 121.2 has the highest importance related to the coupling reagent, which represent the ring structure and amido and heterocyclic groups for predicting yield. 148.3 and 138.3 indicate the contribution of the amido groups and heterocycle in bases. For solvents, 164.1 and 158.1 represents the C-O

bond and the oxygen atoms. Overall, these most important features seem to be indicating the significant roles in determining reaction yield. The higher the feature importance, the stronger the component's influence on the model's predictions.

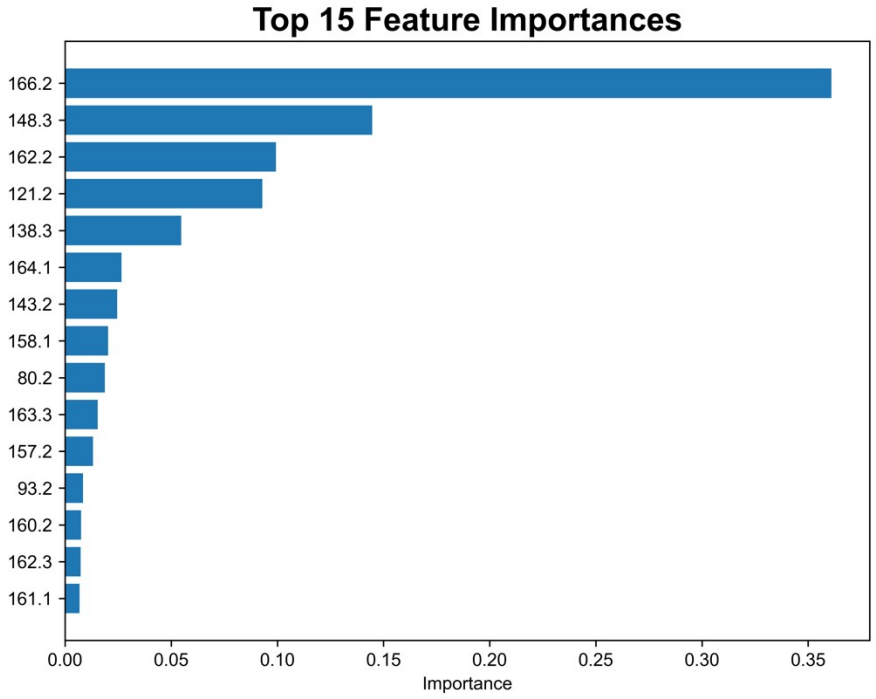


Figure S13. The top 15 most important features for the RF model.

Molecular Fingerprint calculation:

```
import pandas as pd

from rdkit import Chem

from rdkit.Chem import AllChem

from peptide_mols.Pep_lib_construct import get_structure, pep_seq_transform

from tqdm import    tqdm

import time


# load smiles from files
```

```
pep_lib = Chem.SmilesMolSupplier('peptide_mols/peptide_smiles.smi', delimiter='\t')

pep_smi = [Chem.MolToSmiles(mol) for mol in pep_lib] # get smiles of peptides in silico library

pep_seq = [mol.GetProp('sequence') for mol in pep_lib] # get sequence of peptides in silico library


fps = [AllChem.GetMorganFingerprintAsBitVect(x,2,256) for x in pep_lib]

clusters = ClusterFps(fps,cutoff=0.3)

for i in range(len(clusters)):
    df = pd.DataFrame()
    seq = []
    smi = []
    for j in clusters[i]:
        seq.append(pep_seq[j])
        smi.append(pep_smi[i])
    data_temp = list(zip(clusters[i],seq, smi))
    columns = ['Cluster_{}'.format(i), 'sequence','SMILES']
    data_temp_df = pd.DataFrame(data_temp, columns=columns)

data_temp_df.to_csv('peptide_mols/morgan_cluster_0.3/cluster{}.csv'.format(i))
```

Train Machin leanring model

```
import os
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import matplotlib.patches as mpatches
import random

from ModelFits import fit_models
from ModelFits import fit_models_train_val
from ModelFits import plot_models
from ModelFits import run_fold_pipline
# Import relevant scikit-learn modules
# Used ML methods: MLR, PLS, ANN, SVR, RF, ET, Bag

from sklearn.model_selection import train_test_split
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error, r2_score

from sklearn.cross_decomposition import PLSRegression # n_components=9
from sklearn.linear_model import LinearRegression # Default
```

```
# from sklearn.neighbors import KNeighborsRegressor
from sklearn.neural_network import MLPRegressor #  n_estimators = 300
from sklearn.svm import LinearSVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVR #linearSVR 和 SVR #  Default; C, epsilon and gamma optimised per solvent
from sklearn.ensemble import RandomForestRegressor # n_trees = 500
from sklearn.ensemble import ExtraTreesRegressor # n_trees = 500
from sklearn.ensemble import BaggingRegressor # n_trees = 500

df_PLPL = pd.DataFrame(df_ori[270:])
df_fp_PLPL = pd.DataFrame(df_fp_all[270:])
df_dft_PLPL = pd.DataFrame(df_dft[270:])

df_all_desc = {
        # 'onehot':df_oh,
        'fp':df_fp_PLPL          }

yields = (df_PLPL['yield'])

models = [LinearRegression(),
        PLSRegression(n_components=9),
        MLPRegressor(hidden_layer_sizes=300,max_iter=1000),
```

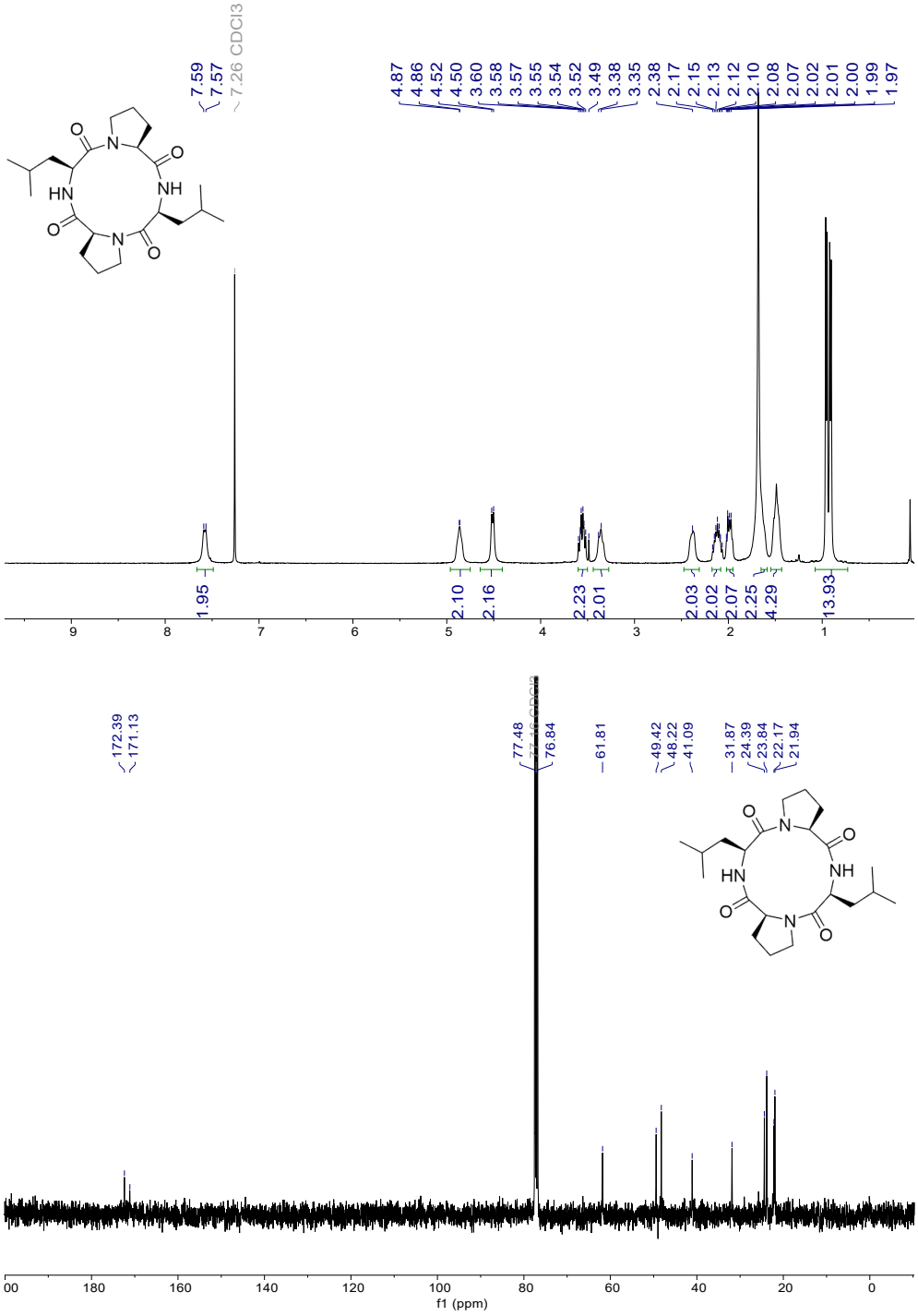
```
        KNeighborsRegressor(n_neighbors=7), # use k = 7 as in papers
        LinearSVR(),
      #  SVR(),
        RandomForestRegressor(n_estimators=500, random_state=42),
      #  ExtraTreesRegressor(n_estimators=500),
      #  BaggingRegressor(n_estimators=500)
        ]
keys = list(df_all_desc.keys()) # all descriptors will run


for i in range(len(keys)):
    r2 = []
    rmse = []
    df_repr = df_all_desc[keys[i]]
    print(f'The running descriptor: {keys[i]} ')
    for seed in range(70,100):
        random.seed(seed)


        A = list(range(0,270,1)) # index
        train_num = [] #index number
        test_num = []
        train_num = random.sample(A, int(270*0.7))
        for j in A:
```

```
            if j not in train_num:
                test_num.append(j)

        train_set = []
        test_set = []
        train_yield = []
        test_yield = []

        train_set = df_repr.iloc[train_num]
        test_set = df_repr.iloc[test_num]

        train_yield = yields.iloc[train_num].values
        test_yield = yields.iloc[test_num].values

        x = pd.DataFrame(train_set)
        y = np.array(train_yield)

        preds, r2_values, rmse_values =
fit_models(x,y,test_set,test_yield,models)
        r2.append(r2_values)
        rmse.append(rmse_values)
        print(f'******************The {seed} time run******************')
    df_r2 = pd.DataFrame(r2,columns= models)
```

```
    df_rmse = pd.DataFrame(rmse,columns=models)

    df = pd.concat([df_r2,df_rmse], axis=1, names=['R2','RMSE'])

    df.to_csv(f'results/HTE_Cyclopep/Diff_descriptors/Descriptor-{keys[i]}-
240918.csv') # change exact folder to save trained results

    print(f'The {keys[i]} run complete')
```

NMR-spectra

Cyclo-(Pro-Leu)$_2$



^{1}H NMR (400 MHz, Chloroform-*d*) δ 7.58 (d, *J* = 9.6 Hz, 2H), 4.96 – 4.75 (m, 2H), 4.51 (d, *J* = 8.2 Hz, 2H), 3.62 – 3.48 (m, 2H), 3.37 (d, *J* = 11.5 Hz, 2H), 2.38 (s, 2H), 2.19 – 1.93 (m, 4H),1.64 – 1.60 (m, 2H), 1.48 (q, *J* = 8.7, 6.3 Hz,

4H), 0.93 (dd, *J* = 17.1, 6.3 Hz, 14H).

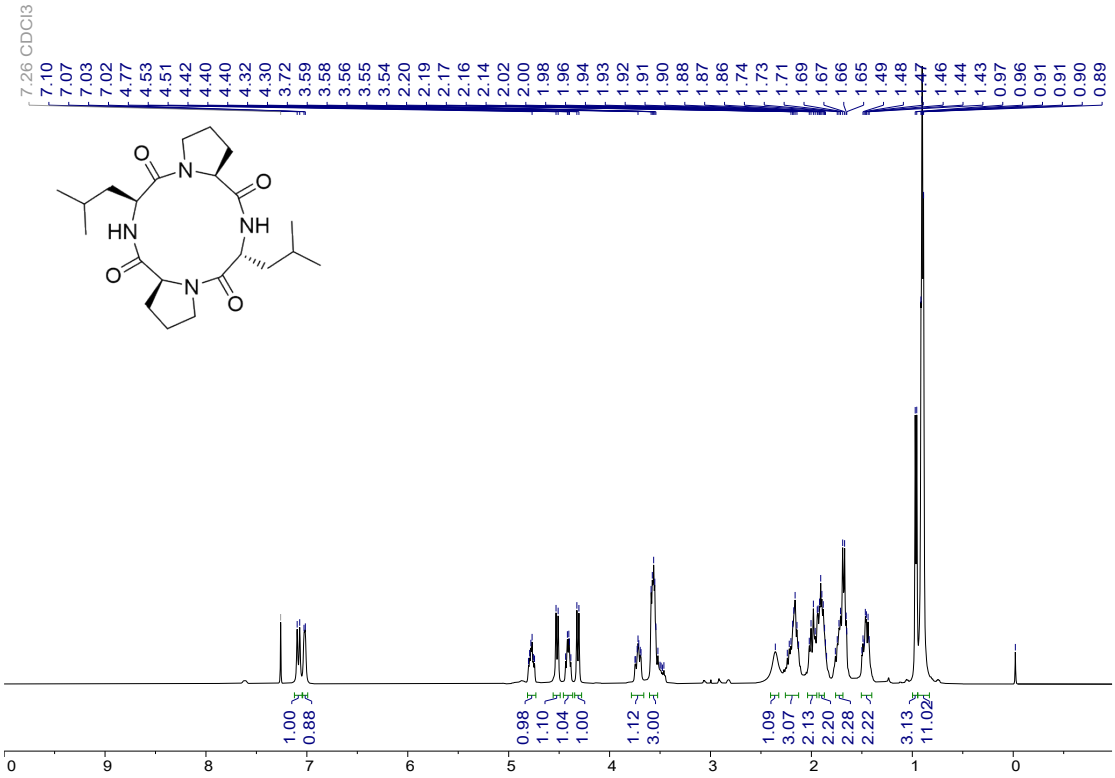
^{13}C NMR (101 MHz, Chloroform-*d*) δ 172.4, 171.1, 61.8, 49.4, 48.2, 41.1, 31.9, 24.4, 23.8, 22.2, 21.9.

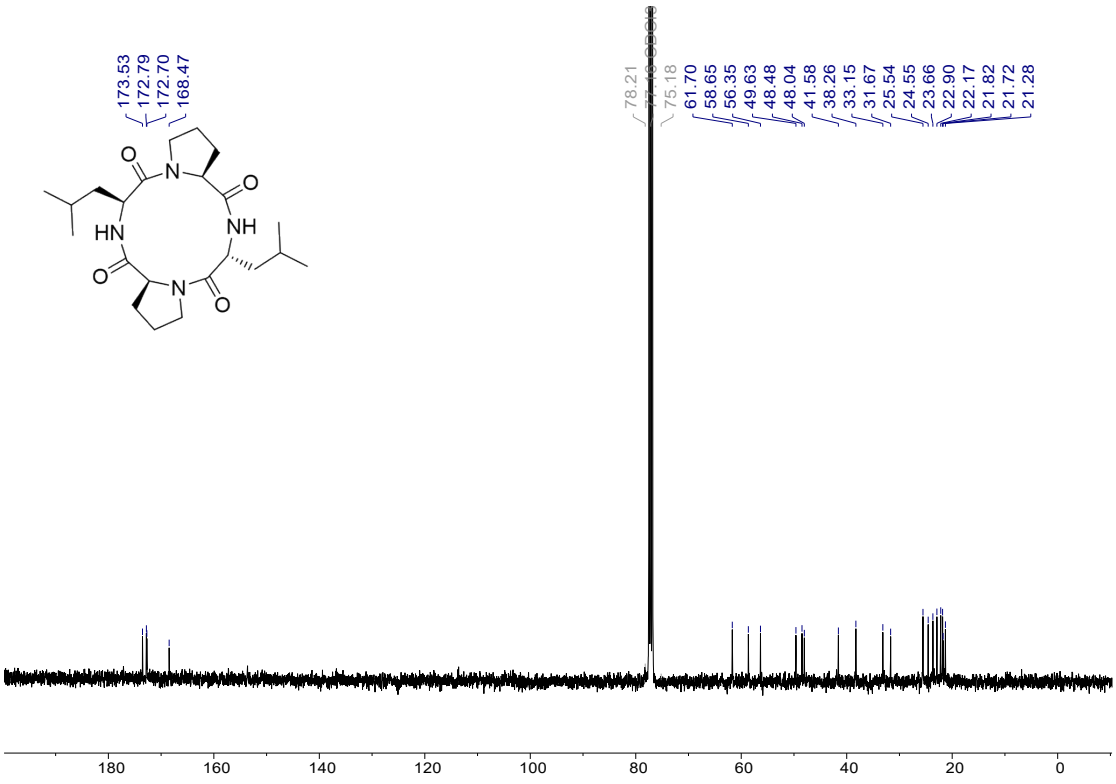
HRMS (ESI): *m/z* $[M+Na]^+$ calcd. for $C_{22}H_{37}N_4O_4$: 443.2809; found: 443.2823.

IR (KBr, cm^{-1}): 3549, 3473, 3414, 3324, 2957, 2929, 1669, 1651, 1636, 1532, 1421, 1400, 1204, 1185, 1134, 599, 480.

Melting Point: 156-158 °C

Cyclo-(Pro-Leu-Pro-D-Leu)





^{1}H NMR (400 MHz, Chloroform-*d*) δ 7.08 (d, *J* = 10.1 Hz, 1H), 7.02 (d, *J* = 5.3 Hz, 1H), 4.77 (td, *J* = 9.6, 4.5 Hz, 1H), 4.52 (d, *J* = 8.3 Hz, 1H), 4.46 – 4.37 (m, 1H), 4.31 (d, *J* = 8.1 Hz, 1H), 3.72 (ddd, *J* = 11.9, 8.7, 3.1 Hz, 1H), 4.61 – 4.49 (m, 3H),, 2.40 – 2.30 (m, 1H), 2.25 – 2.10 (m, 3H),, 2.03 – 1.94 (m, 2H), 1.93 – 1.84 (m, 2H), 1.76 – 1.69 (m, 2H),, 1.51 – 1.39 (m, 2H), 0.96 (d, *J* = 6.1 Hz, 3H), 0.90 (dd, *J* = 6.3, 4.1 Hz, 11H).

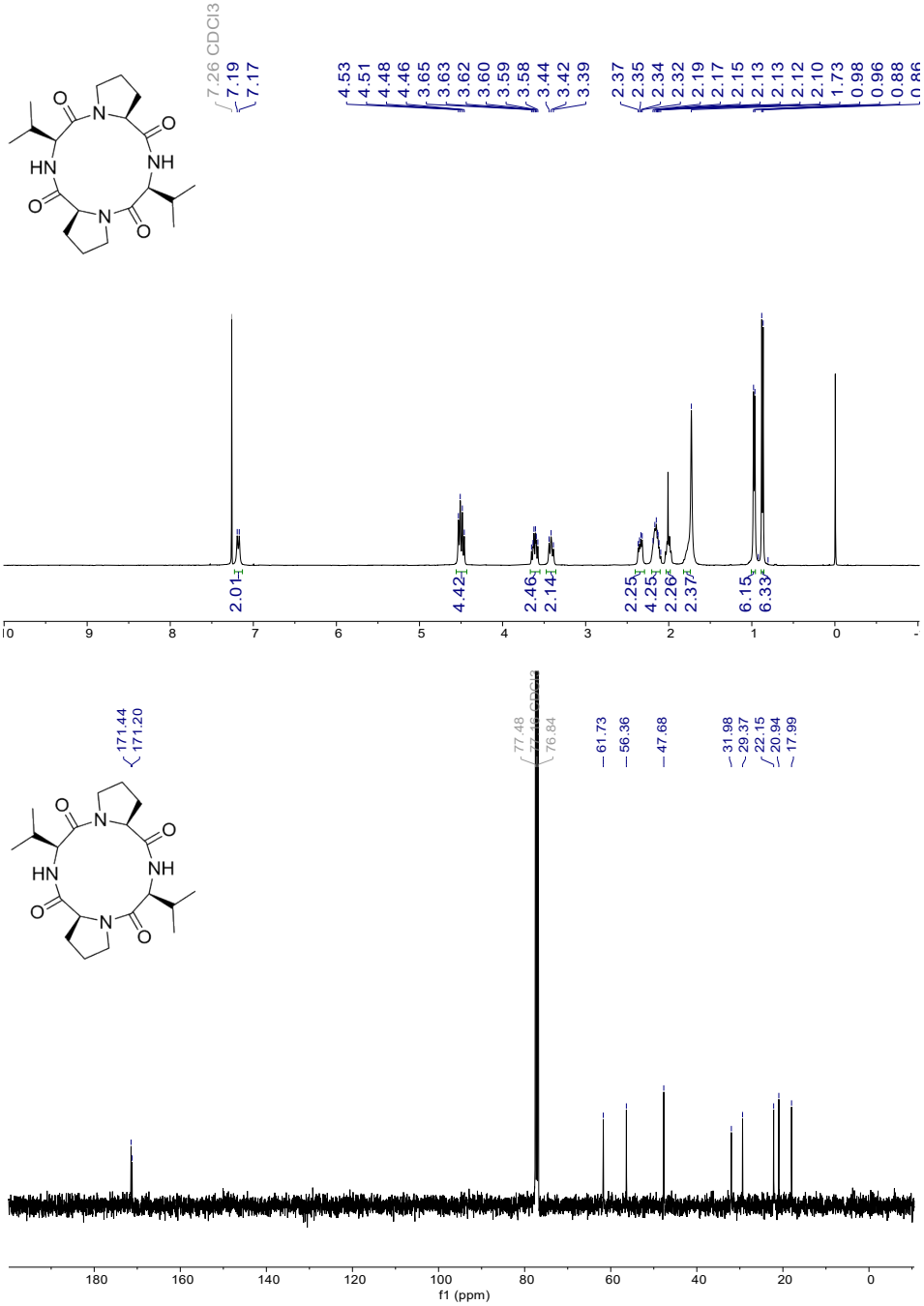
^{13}C NMR (101 MHz, Chloroform-*d*) δ 173.5, 172.8, 172.7, 168.5, 61.7, 58.7, 56.4, 49.6, 48.5, 48.0, 41.6, 38.3, 33.2, 31.7, 25.5, 24.6, 23.7, 22.9, 22.2, 21.8, 21.7, 21.3.

HRMS (ESI): *m/z* $[M+Na]^+$ calcd. for $C_{22}H_{37}N_4O_4$: 443.2809; found: 443.2817.

IR (KBr, cm^{-1}): 3472, 3414, 3324, 3238, 2929, 2870, 1652, 1636, 1617, 1531, 1452, 1421, 1204, 1185, 1134, 599, 480.

Melting Point: 156-158 °C

Cyclo-(Pro-Val)$_2$



^{1}H NMR (400 MHz, Chloroform-*d*) δ 7.18 (d, *J* = 9.9 Hz, 2H), 4.56 – 4.43 (m, 4H), 3.61 (td, *J* = 11.5, 7.4 Hz, 2H), 3.47 – 3.36 (m, 2H), 2.34 (dd, *J* = 12.8, 6.4 Hz, 2H), 2.13 (dt, *J* = 12.9, 7.6 Hz, 4H),2.03 – 1.97 (m, 2H), 1.81– 1.73 (m, 2H), 0.97 (d, *J* = 6.5 Hz, 6H), 0.87 (d, *J* = 6.8 Hz, 6H).

^{13}C NMR (101 MHz, Chloroform-*d*) δ 171.4, 171.2, 61.7, 56.4, 47.7, 32.0, 29.4, 22.2, 20.9, 18.0.

HRMS (ESI): *m/z* $[M+Na]^+$ calcd. for $C_{20}H_{33}N_4O_4$: 415.2496; found: 415.2485.

IR (KBr, cm^{-1}): 3447, 3434, 2962, 2930, 2847, 1668, 1652, 1649, 1640, 1637, 1534, 1417, 1384, 1203, 1136, 627, 612.

Melting point: 178-180 °C

Notes and references

1 J. Liao, X. Jia, F. Wu, J. Huang, G. Shen, H. You and F. Chen, Rapid mild macrocyclization of depsipeptides under continuous flow: total syntheses of five cyclodepsipeptides, *Org. Chem. Front.*, 2022, **9**, 6640-6645.

2 R. Wills, V. Adebomi, C. Spancake, R. D. Cohen and M. Raj, Synthesis of L-cyclic tetrapeptides by backbone amide activation CyClick strategy, *Tetrahedron*, 2022, **126**, 133071.

3 G. J. Saunders, S. A. Spring, E. Jayawant, I. Wilkening, S. Roesner, G. J. Clarkson, A. M. Dixon, R. Notman and M. Shipman, Synthesis and Functionalization of Azetidine-Containing Small Macrocyclic Peptides, *Chemistry–a European Journal*, 2024, **30**, e202400308.

4 J. Han, H. Wang, R. Zhang, H. Dai, B. Chen, T. Wang, J. Sun, W. Wang, F. Song, E. Li, Z. Lyu and H. Liu, Cyclic Tetrapeptides with Synergistic Antifungal Activity from the Fungus Aspergillus westerdijkiae Using LC-MS/MS-Based Molecular Networking, *Antibiotics*, 2022, **11**, 166.

5 A. Swami and R. Jain, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2013, **12**, 2825-2830.

6 S. Wang, J. Witek, G. A. Landrum and S. Riniker, Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences, *J. Chem Inf. Model.*, 2020, **60**, 2044-2058.

7 G. Landrum., RDKit: Open-Source Cheminformatics Software.2016.