SCIENCE CHINA
Information Sciences

• **Supplementary File** •

SBSM-Pro: Support Bio-sequence Machine for Proteins

Yizheng WANG[1,2], Yixiao ZHAI[1,2], Yijie DING[2*] & Quan ZOU[1,2*]

[1]*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu* 611731*, China;*
[2]*Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou* 324003*, China*

Appendix A Datasets and generating sequence kernels

In this study, we used 10 protein identification datasets and plotted the sequence lengths of each dataset as box plots, as shown in Figure A1.

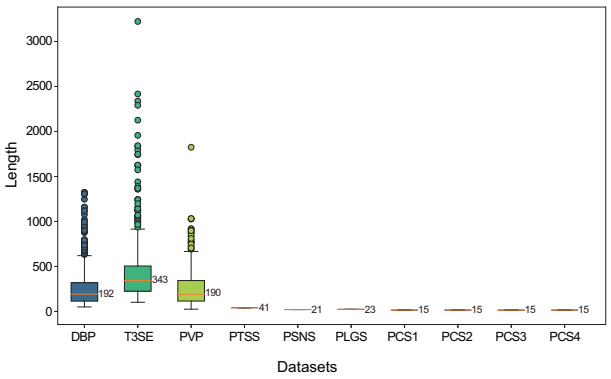

Figure A1 Boxplots of protein sequence lengths for the different datasets.

Spectral clustering, a graph theory-based clustering methodology, utilizes spectral information (i.e., eigenvectors) for data segmentation. Renowned for its robustness and adaptability, this data partitioning technique has garnered widespread attention in recent years.

Consider a set comprising p distinct data points, which are clustered into k_c clusters through spectral clustering. We first construct a similarity matrix $\mathbf{S} \in \mathbf{R^{P\times P}}$. The most prevalent similarity measure implemented is the Gaussian kernel of the Euclidean distance. Hence, the elements of matrix $\mathbf{S}$ can be computed using the following equation:

$$S_{ij} = \exp\left(-\gamma\|x_i - x_j\|^2\right), \tag{A1}$$

where x_i and x_j are the data points. γ is the coefficient of the kernel function, which effectively quantifies the decay rate of the similarity and determines how rapidly the similarity between data points diminishes as their distance increases.

Degree matrix $\mathbf{D}$ is defined as a diagonal matrix that satisfies $\mathbf{D} \in \mathbf{R^{P\times P}}$, and its elements can be calculated as

$$[D]_{i,j} = \sum_{j=1}^{p} [S]_{i,j}\,. \tag{A2}$$

Then, the graph Laplacian matrix $\mathbf{L} \in \mathbf{R^{P\times P}}$ is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{S}. \tag{A3}$$

Next, we proceed with the eigendecomposition of the Laplacian matrix. This decomposes the Laplacian matrix into a set of eigenvalues and their corresponding eigenvectors, thereby offering a more tractable framework for our subsequent analysis. Given that the Laplacian matrix is a real symmetric matrix, it is pertinent to note that all its eigenvalues are real numbers.

Subsequently, we select the k_c smallest eigenvalues and form a matrix $\mathbf{U} \in \mathbf{R^{P\times k_c}}$ with corresponding eigenvectors as columns. The matrix $\mathbf{U}$ is row-normalized to obtain the matrix $\mathbf{T} \in \mathbf{R^{P\times k_c}}$. We can conceptualize each row in the matrix $\mathbf{T}$ as an individual data point and then apply the K-means algorithm for clustering to derive the results.

The Calinski–Harabasz index (CHI), also known as the variance ratio criterion, is a commonly utilized metric for evaluating the outcomes of cluster analysis. It quantifies both the compactness within clusters and the separation between clusters. A higher value of the CHI suggests superior clustering performance.

* Corresponding author (email: wuxi_dyj@163.com, zouquan@nclab.net)

Our dataset comprises p elements, which have been clustered into k_c clusters through spectral clustering. The evaluation metric CHI for this particular clustering outcome can be calculated utilizing the following equation:

$$CHI = \frac{trace\,(B_k)}{trace\,(W_k)} \times \frac{p - k_c}{k_c - 1}, \tag{A4}$$

where B_k and W_k are the between-group dispersion matrix and within-cluster dispersion matrix, respectively.

We define the center of E as C_E. For a particular cluster q, its center is represented as c_q. The set of all data points contained within cluster q is defined as C_q, with n_q representing the number of elements in the set C_q. Subsequently, B_k and W_k can be calculated using the following equations:

$$B_k = \sum_{q=1}^{k} n_q\,(c_q - c_E)(c_q - c_E)^T, \tag{A5}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T. \tag{A6}$$

Proteins are composed of amino acids, fundamental organic compounds in biological processes. Each amino acid molecule consists of an amino group, a carboxyl group, a hydrogen atom, and a side chain. This particular structure of amino acids gives rise to various physicochemical properties. We collated data on the physicochemical properties of amino acids from previous studies, which are frequently employed in bioinformatics research, including alpha-carbon positions (ACP) [1], hydrophobicity (H) [2,3], secondary structure (SS) [4,5], non-bonded energy (NBE) [6], membrane regions (MR) [7], polarity and bulkiness (PB) [8], chemical structure (CS) [9], mean polarities (MP) [10], and side-chain (SC) [11,12]. These properties were numerically represented and retained for the purpose of generating dictionaries for grouping via spectral clustering. Regrettably, the data are not fully complete, necessitating further processing to ensure their usability and integrity. Biological factors can often result in unusable or incomplete data. For example, the simplicity of the side chains in alanine and glycine, composed of a methyl group and a hydrogen atom, respectively, may result in a less pronounced impact during detailed side chain analysis compared to more complex amino acids. This often results in missing data, manifesting as not applicable (NA) in the numerical values for the physicochemical properties of these amino acids. Directly assigning a specific value, such as zero, to missing data could result in a loss of accuracy and interpretability. Thus, we opted to eliminate data entries for amino acids' physicochemical properties containing "NA." The processed data is shown in Table A1. For reference, the removed data entries can be found in Table A2.

Table A1 Summary of processed physicochemical properties of amino acids.

	ACP	H1	SS1	NBE	SS2	MR	PB	CS	MP	H2
Amino acid	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
Ala (A)	1.6	87	0.8	-0.491	16	9.36	9.9	0.33	-0.06	-0.26
Arg (R)	0.9	81	0.96	-0.554	-70	0.27	4.6	-0.176	-0.84	0.08
Asn (N)	0.7	70	1.1	-0.382	-74	2.31	5.4	-0.233	-0.48	-0.46
Asp (D)	2.6	71	1.6	-0.356	-78	0.94	2.8	-0.371	-0.8	-1.3
Cys (C)	1.2	104	0	-0.67	168	2.56	2.8	0.074	1.36	0.83
Gln (Q)	0.8	66	1.6	-0.405	-73	1.14	9	-0.254	-0.73	-0.83
Glu (E)	2	72	0.4	-0.371	-106	0.94	3.2	-0.409	-0.77	-0.73
Gly (G)	0.9	90	2	-0.534	-13	6.17	5.6	0.37	-0.41	-0.4
His (H)	0.7	90	0.96	-0.54	50	0.47	8.2	-0.078	0.49	-0.18
Ile (I)	0.7	105	0.85	-0.762	151	13.73	17.1	0.149	1.31	1.1
Leu (L)	0.3	104	0.8	-0.65	145	16.64	17.6	0.129	1.21	1.52
Lys (K)	1	65	0.94	-0.3	-141	0.58	3.5	-0.075	-1.18	-1.01
Met (M)	1	100	0.39	-0.659	124	3.93	14.9	-0.092	1.27	1.09
Phe (F)	0.9	108	1.2	-0.729	189	10.99	18.8	-0.011	1.27	1.09
Pro (P)	0.5	78	2.1	-0.463	-20	1.96	14.8	0.37	0	-0.62
Ser (S)	0.8	83	1.3	-0.455	-70	5.58	6.9	0.022	-0.5	-0.55
Thr (T)	0.7	83	0.6	-0.515	-38	4.68	9.5	0.136	-0.27	-0.71
Trp (W)	1.7	94	0	-0.839	145	2.2	17.1	-0.011	0.88	-0.13
Tyr (Y)	0.4	83	1.8	-0.656	53	3.13	15	-0.138	0.33	0.69
Val (V)	0.6	94	0.8	-0.728	123	12.43	14.3	0.245	1.09	1.15

The SW algorithm is a widely used sequence alignment method in bioinformatics for identifying optimal local alignments between two sequences. Using this method, the similarity between proteins can be calculated.

To perform sequence alignment between two protein sequences, denoted as Sx_i and Sy_i, and compute their Smith-Waterman (SW) scores, we employ the SW algorithm. The core of this algorithm can be formulated using a scoring matrix A, where each element $A_{i,j}$ represents the best score for aligning the prefixes of the two sequences up to positions i and j. The equation for calculating the elements of the scoring matrix A is as follows:

$$[A]_{i,j} = \max \begin{cases} [A]_{i,j-1} - g,\ if\ j > 0\ and\ i \geqslant 0 \\ [A]_{i-1,j} - g,\ if\ j \geqslant 0\ and\ i > 0 \\ [A]_{i-1,j-1} + p\,(i,j)\,,\ if\ j > 0\ and\ i > 0, \end{cases} \tag{A7}$$

Table A2 Summary of processed physicochemical properties of amino acids.

Amino acid	SC1 P_{11}	SC2 P_{12}
Ala (A)	0.54	NA
Arg (R)	-0.16	0.62
Asn (N)	0.38	0.76
Asp (D)	0.65	0.66
Cys (C)	-1.13	0.83
Gln (Q)	0.05	0.59
Glu (E)	0.38	0.73
Gly (G)	NA	NA
His (H)	-0.59	0.92
Ile (I)	-2	0.88
Leu (L)	-1.08	0.89
Lys (K)	0.48	0.77
Met (M)	-0.97	0.77
Phe (F)	-1.51	0.92
Pro (P)	-0.22	0.94
Ser (S)	0.65	0.58
Thr (T)	0.27	0.73
Trp (W)	-1.61	0.86
Tyr (Y)	-1.13	0.93
Val (V)	-0.75	0.88

where $A_{i,j-1}$ denotes a gap at position j of sequence Sy, $A_{i-1,j}$ denotes a gap at position i of sequence Sx, and $A_{i-1,j-1}$ indicates an alignment without gaps at positions i and j. The p_{ij} is a function that allocates scores based on matches or mismatches at positions i and j. It is defined as follows:

$$p\left(i,j\right)=\begin{cases} m_1,\ if\ Sx_i=Sy_j \\ m_2,\ if\ Sx_i\neq Sy_j, \end{cases} \tag{A8}$$

where m_1 and m_2 represent the scores for a match and a mismatch between elements at positions i and j, respectively.

After computing the scoring matrix, a traceback can be performed. In contrast to the Needleman–Wunsch algorithm used for global alignment, which backtracks from the bottom right corner of the scoring matrix to the bottom left corner, the SW algorithm initiates the traceback from the highest value within the scoring matrix and stops when it reaches a score of zero, thereby identifying the optimal local alignment. However, our primary goal in incorporating the SW algorithm is to obtain the SW score. Therefore, we do not need to perform the traceback process. Instead, we simply choose the maximum value from the scoring matrix as the SW score.

The schematic diagram of the SW algorithm is shown in Figure A2. In the example diagram, a gap is introduced at the fifth position of protein sequence S_x. Starting from the second amino acid of both proteins, a local alignment region comprising seven amino acids emerges, resulting in a final SW score of 4.

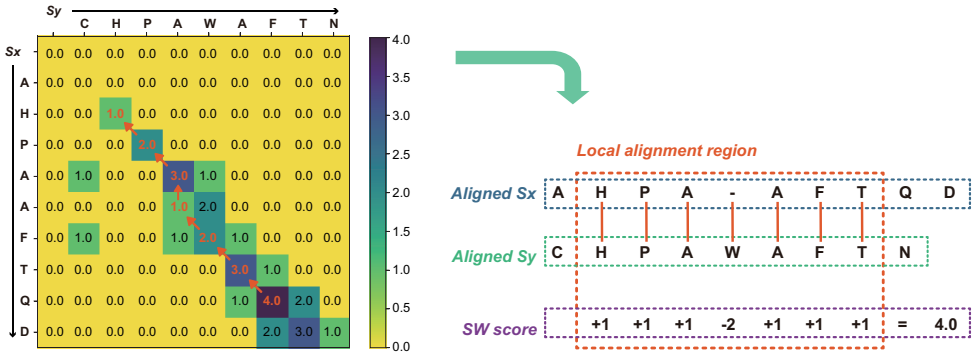

Figure A2 Schematic of the SW algorithm.

Appendix B Introduction of Hilbert–Schmidt Independence Criterion

We define $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\}^T\in\mathbf{R}^{N\times d}$ as the original feature of d dimensions of samples, and $\mathbf{Y}\in\mathbf{R}^{N\times 1}$ is the label of these samples. We can derive a series of observations from the probability distribution $\Pr_{xy}$, defined as

$$Z\equiv\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_N,y_N)\}\subseteq\mathbf{X}\times\mathbf{Y} \tag{B1}$$

HSIC calculates the cross-covariance operator on the domain $\mathbf{X} \times \mathbf{Y}$ to determine the independence between $\mathbf{X}$ and $\mathbf{Y}$. The feature set $\mathbf{X}$ and label set $\mathbf{Y}$ can be mapped to $\mathbf{F}$ and $\mathbf{G}$ by the mapping $\phi : \mathbf{X} \to \mathbf{F}$ and $\psi : \mathbf{Y} \to \mathbf{G}$. Then, we defined their expectations as $\mu_{\mathbf{x}}$ and μ_y, respectively. The kernel function is as follows:

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left\langle \phi\left(\mathbf{x}_i\right), \phi\left(\mathbf{x}_j\right) \right\rangle \tag{B2}$$

Similarly, the kernel function of $\mathbf{Y}$ is defined as

$$l\left(y_i, y_j\right) = \left\langle \psi\left(y_i\right), \psi\left(y_j\right) \right\rangle \tag{B3}$$

The following equation can be used to determine the cross-covariance operator $C_{\mathbf{x}y}$:

$$C_{\mathbf{x}y} = E_{\mathbf{x},y}\left[\phi\left(\mathbf{x}\right) \otimes \psi\left(y\right)\right] - \mu_{\mathbf{x}}\mu_y \tag{B4}$$

where $E_{\mathbf{x},y}$ denotes the common expectation of $\mathbf{x}$ and y. Then, we can write the HSIC operator is:

$$HSIC\left(\mathbf{F}, \mathbf{G}, \Pr_{xy}\right) = \left\|C_{\mathbf{x}y}\right\|_{HS}^2 \tag{B5}$$

Then, we define as the $\mathbf{I}$ identity matrix, and it satisfies $\mathbf{I} \in \mathbf{R}^{N \times N}$. By defining $\mathbf{e} = [1, 1, \cdots, 1]^T \in \mathbf{R}^{1 \times N}$, we can obtain

$$\mathbf{H} \equiv \mathbf{I} - \frac{\mathbf{e}\mathbf{e}^T}{N} \tag{B6}$$

Note that $\mathbf{H}$ is the centering matrix, and it satisfies $\mathbf{H} \in \mathbf{R}^{N \times N}$. Then, we can make an empirical estimate of $\mathbf{Z}$ set as

$$\begin{aligned} HSIC\left(\mathbf{F}, \mathbf{G}, \mathbf{Z}\right) &= \tfrac{1}{N^2} tr(\mathbf{KU}) - \tfrac{2}{N^3}\mathbf{e}^T\mathbf{KUe} + \tfrac{1}{N^4}\mathbf{e}^T\mathbf{Kee}^T\mathbf{Ue} \\ = \tfrac{1}{N^2}\left[tr\left(\mathbf{KU}\right) - \tfrac{1}{N} tr\left(\mathbf{Kee}^T\right) - \tfrac{1}{N} tr\left(\mathbf{Uee}^T\right) + \tfrac{1}{N^2} tr(\mathbf{Uee}^T\mathbf{Kee}^T)\right] \\ = \tfrac{1}{N^2} tr\left[\mathbf{K}\left(\mathbf{I} - \tfrac{1}{N}\mathbf{e}\mathbf{e}^T\right)\mathbf{U}\left(\mathbf{I} - \tfrac{1}{N}\mathbf{e}\mathbf{e}^T\right)\right] \\ = \tfrac{1}{N^2} tr(\mathbf{KHUH}) \triangleq HSIC(\mathbf{K}, \mathbf{U}) \end{aligned} \tag{B7}$$

where $\mathbf{K}, \mathbf{U} \in \mathbf{R}^{N \times N}$ are kernel matrices, $k\left(\mathbf{x}_i, \mathbf{x}_j\right)$ and $l\left(y_i, y_j\right)$.

Appendix C Dictionaries for grouping

Through the PSD process, spectral clustering results for 10 different physicochemical properties of amino acids were obtained, resulting in 10 dictionaries used for grouping. In each dictionary, the correspondence between amino acids and group numbers is shown in Tables C1 to C10. Using LS distance and SW scoring, two amino acid similarity measurement methods, we evaluate the effectiveness of different dictionaries. The detailed results are shown in Table C11 and C12.

Table C1 Dictionaries for grouping D_1.

Dictionary	Group	Amino Acid
D_1	G_1	A, W
	G_2	D
	G_3	R, Q, G, F, S
	G_4	E
	G_5	L, Y
	G_6	N, H, I, P, T, V
	G_7	C, K, M

Table C2 Dictionaries for grouping D_2.

Dictionary	Group	Amino Acid
D_2	G_1	R, S, T, Y
	G_2	C, I, L, M, F
	G_3	Q, K
	G_4	P
	G_5	W, V
	G_6	N, D, E
	G_7	A, G, H

Table C3 Dictionaries for grouping D_3.

Dictionary	Group	Amino Acid
D_3	G_1	D, Q, Y
	G_2	A, I, L, T, V
	G_3	N, F, S
	G_4	G, P
	G_5	C, W
	G_6	E, M
	G_7	R, H, K

Table C4 Grouping for amino acid D_4.

Dictionary	Group	Amino Acid
D_4	G_1	R, G, H, T
	G_2	A, P, S
	G_3	W
	G_4	D, K
	G_5	C, L, M, Y
	G_6	I, F, V
	G_7	N, Q, E

References

1 Richardson JS, Richardson DC. Amino Acid Preferences for Specific Locations at the Ends of α Helices. Science, 1988, 240(4859): 1648-1652.
2 Meirovitch H, Rackovsky S, Scheraga HA. Empirical Studies of Hydrophobicity. 1. Effect of Protein Size on the Hydrophobic Behavior of Amino Acids. In: Proceedings of Macromolecules, 1980. 1398-1405.
3 Geisow MJ, Roberts RDB. Amino acid preferences for secondary structure vary with protein class. International Journal of Biological Macromolecules, 1980, 2(6): 387-389.
4 Oobatake M, Ooi T. An analysis of non-bonded energy of proteins. Journal of Theoretical Biology, 1977, 67(3): 567-584.
5 Biou V, Gibrat JF, Levin JM, Robson B, Garnier J. Secondary structure prediction: combination of three different methods. Protein Engineering, Design and Selection, 1988, 2(3): 185-191.
6 Nakashima H, Nishikawa K. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS Letters, 1992, 303(2): 141-146.
7 Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. Journal of Theoretical Biology, 1968, 21(2): 170-201.
8 Sneath PHA. Relations between chemical structure and biological activity in peptides. Journal of Theoretical Biology, 1966, 12(2): 157-195.
9 Radzicka A, Wolfenden R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry, 1988, 27(5): 1664-1670.
10 Cornette JL, Cease KB, Margalit H, et al. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. Journal of Molecular Biology, 1987, 195(3): 659-685.
11 Guy HR. Amino acid side-chain partition energies and distribution of residues in soluble proteins. Biophysical Journal, 1985, 47(1): 61-70.
12 Yang JM, Tsai CH, Hwang MJ, et al. GEM: A Gaussian evolutionary method for predicting protein side-chain conformations. Protein Science, 2002, 11(8): 1897-1907.

Table C5 Grouping for amino acid D_5.

Dictionary	Group	Amino Acid
D_5	G_1	H, Y
	G_2	A, G, P, T
	G_3	K
	G_4	I, L, M, W, V
	G_5	R, N, D, Q, E, S
	G_6	C, F

Table C6 Grouping for amino acid D_6.

Dictionary	Group	Amino Acid
D_6	G_1	G, S, T
	G_2	N, C, M, P, W, Y
	G_3	L
	G_4	I, V
	G_5	A, F
	G_6	R, D, Q, E, H, K

Table C7 Grouping for amino acid D_7.

Dictionary	Group	Amino Acid
D_7	G_1	M, P, Y, V
	G_2	F
	G_3	H, S
	G_4	R, N, G
	G_5	D, C, E, K
	G_6	A, Q, T
	G_7	I, L, W

Table C8 Grouping for amino acid D_8.

Dictionary	Group	Amino Acid
D_8	G_1	A, G, P
	G_2	C, I, L, T
	G_3	D, E
	G_4	R, H, K, M, Y
	G_5	F, S, W
	G_6	V
	G_7	N, Q

Table C9 Grouping for amino acid D_9.

Dictionary	Group	Amino Acid
D_9	G_1	C, I, L, M, F
	G_2	K
	G_3	N, G, S, T
	G_4	W, V
	G_5	H, Y
	G_6	R, D, Q, E
	G_7	A, P

Table C10 Grouping for amino acid D_{10}.

Dictionary	Group	Amino Acid
D_{10}	G_1	L
	G_2	A, N, G, P, S
	G_3	D
	G_4	I, M, F, V
	G_5	C, Y
	G_6	Q, E, K, T
	G_7	R, H, W

Table C11 Comparison of the different dictionaries for grouping by the LS distance.

	DBP	T3SE	PVP	PTSS	PSNS	PLGS	PCS1	PCS2	PCS3	PCS4
not	0.7419	0.7368	0.8085	0.7438	0.7378	0.8219	0.8667	0.8627	0.8616	0.8476
d1	0.7742	0.7368	0.8191	0.8063	0.7378	0.8259	0.8667	0.8638	0.8616	0.8581
d2	0.7957	0.7763	0.7979	0.7688	0.7683	0.8219	0.8667	0.8627	0.8616	0.8617
d3	0.8064	0.7632	0.8085	0.8375	0.7439	0.8259	0.8667	0.8649	0.8616	0.8593
d4	0.7957	0.7368	0.7766	0.7625	0.7378	0.8259	0.8671	0.8627	0.8616	0.8581
d5	0.7796	0.8026	0.8085	0.8063	0.7622	0.8219	0.8667	0.8649	0.8626	0.8593
d6	0.8333	0.7237	0.7872	0.7000	0.7378	0.8259	0.8667	0.8627	0.8616	0.8581
d7	0.7796	0.7895	0.8085	0.7188	0.7378	0.8219	0.8667	0.8627	0.8616	0.8581
d8	0.8172	0.7632	0.8191	0.8063	0.7378	0.8259	0.8667	0.8627	0.8626	0.8581
d9	0.8011	0.7763	0.7979	0.7625	0.7378	0.8219	0.8688	0.8660	0.8636	0.8593
d10	0.8226	0.7237	0.7872	0.8063	0.7378	0.8259	0.8667	0.8627	0.8616	0.8581

Table C12 Comparison of the different dictionaries for grouping by the SW score.

	DBP	T3SE	PVP	PTSS	PSNS	PLGS	PCS1	PCS2	PCS3	PCS4
not	0.7957	0.7237	0.7234	0.7875	0.7012	0.8219	0.8524	0.8627	0.8616	0.8464
d1	0.8548	0.7632	0.8191	0.8000	0.7378	0.8259	0.8667	0.8627	0.8636	0.8581
d2	0.7957	0.8158	0.7872	0.8063	0.7622	0.8259	0.8519	0.8638	0.8626	0.8581
d3	0.8441	0.7895	0.8191	0.8375	0.7439	0.8300	0.8667	0.8627	0.8616	0.8581
d4	0.8817	0.7895	0.7766	0.7625	0.7378	0.8219	0.8670	0.8627	0.8616	0.8593
d5	0.8656	0.7763	0.7766	0.8063	0.7378	0.8219	0.8670	0.8627	0.8616	0.8581
d6	0.8701	0.8026	0.7766	0.7313	0.7439	0.8219	0.8670	0.8649	0.8544	0.8593
d7	0.8763	0.8026	0.7872	0.7750	0.7378	0.8300	0.8582	0.8627	0.8616	0.8581
d8	0.8656	0.7763	0.8191	0.8063	0.7378	0.8219	0.8670	0.8638	0.8616	0.8581
d9	0.8602	0.7632	0.7979	0.7875	0.7500	0.8219	0.8667	0.8681	0.8616	0.8581
d10	0.8387	0.7105	0.8191	0.7813	0.7439	0.8219	0.8670	0.8649	0.8616	0.8581