Supplemental Information File #1

Enhancing ADMET property Models Performance through Combinatorial Fusion Analysis

Nan Jiang,[†,§] Mohammed Quazi,[‡,§] Christina Schweikert,[¶] D. Frank Hsu,[*,†]

Tudor I. Oprea,[*,‡] and Suman Sirimulla[*,‡]

†*Laboratory of Informatics and Data Mining, Department of Computer and Information Science, Fordham University, New York, NY 10023, USA*

‡*Expert Systems, Inc 12760 High Bluff Drive, Suite 370, San Diego, CA 92130, USA*

¶*Division of Computer Science, Mathematics and Science, St. John's University, 8000 Utopia Parkway, Queens, NY 11439, USA*

§*indicates authors contributed equally*

E-mail: hsu@fordham.edu; toprea@expertsystems.edu; ssirimulla@expertsystems.edu

This supplemental information file #1 provides three supplemental sections for the paper "Leveraging Combinatorial Fusion Analysis for Enhanced Prediction Models in Drug Discovery": 1. CFA characteristics; 2. Groups, graphs, and Kemeny rank space; 3. Figure skating judgement: rank vs. score Combination.

Contents

1 CFA Characteristics

In this paper, we use Combinatorial Fusion Analysis (CFA) for combining multiple ML/AI models. Compare to other approaches ([1] and references), we summarize our method w.r.t. the following six distinctive characteristics: (a) model fusion, (b) space duality, (c) method of combination, (d) number of base models, (e) base models, and (f) the Kemeny rank space.

As it was demonstrated in previous sections, CFA is a robust learning and modeling framework for drug discovery in molecular science.

(a) Model Fusion: Instead of optimizing or boosting a single model, CFA combines a group of relatively good but diverse models. Moreover, CFA uses the rank-score function f_A of the scoring system (model) A to characterize A. Then it uses the cognitive diversity (a.k.a. rank-score diversity) $cd(A,B) = d(f_A, f_B)$ between A and B in terms of the difference (or dissimilarity) between f_A and f_B, to measure the diversity between models A and B [2–6].

(b) Space Duality: Most of the prediction models for the drug discovery and the ADMET properties are operating on the Euclidean space. CFA considers each scoring system on the set of n molecules, resulted from each model prediction, as having a score function in the Euclidean space $\mathbb{R}_n$ and a derived rank function in the rank space of B_n, the Bubble sort Cayley graph space without ties, or K_n, the Kemeny rank space when ties are allowed (see Section 2) [5–8]. Method of ranking and real life application have been studied in preference learning and social choice [9–11].

(c) Method of Combination: In the combination of multiple scoring systems (attributes, algorithms, or models), CFA considers score combination in the Euclidean space $\mathbb{R}_n$ and rank combination in the rank space (B_n or K_n). In each of the score or rank combination, three combination types are considered: average combination (AC), weighted combination by performance (WCP), and weighted combination by diversity strength (WCDS). We note that diversity strength of model A is the average of cognitive diversity measurement between A and other models in the model space [2–5, 12].

(d) Number of Base Models: On binding affinity estimation in virtual screening using multiple scoring systems [13] demonstrated that the relatively optimal number of scoring systems to com-

bine is from three to five. For example, four or five scoring systems are used in virtual screening ([2, 14],[15]). In cheminformatics and material science, four ML/AI models are used [16].

(e) Base Models: According to (a) and (d), 4-6 base models are used for model fusion. As to what base models to be used, it may depend on the data representation (molecular encoding scheme), problem task (ADMET properties), and performance evaluation criteria (Spearman's rho, MAE, AUROC or AUPRC). However, the general consensus is that combination of two models (or algorithms) A and B performs better than or equal to the best of A and B only if models A and B are relatively good and they are diverse [7, 12, 16–18].

(f) The Kemeny Rank Space: In Section 2 of supplemental information file #1, we discuss the information structure for a rank function within the CFA framework. Recall that rank function r_A of a scoring system A is obtained from sorting the score values of the score function s_A in descending order and assign rank value in $[1,n]$ to the data items in $D = \{d_1, d_2, \ldots, d_n\}$ accordingly. When the rank function is an 1-1 function, the resulting rank function is considered as a permutation of the positive integers from 1 to n. In this case, the permutation representing the rank function r_A is an element of the symmetric group of order n, S_n, and is a vertex (or node) of the graph B_n (Bubble sort Cayley graph) when the graph distance between vertices (permutations) A and B is the same as Kendall's tau distance between A and B. Other metrics on the permutation space are available [19, 20]. In the case that the rank function r_A is not one-one, two or more data items in D have the same score values in s_A (see also Section 2.1 in the paper). CFA considers such a rank function r_A as a vertex of the Kemeny rank space where each vertex A is defined as a score matrix a_{ij} and the distance between two vertices A and B is calculated by the rank correlation $\tau_x(A,B)$ (see also Equation (2) in Section 2). The Kemeny rank space K_n has been used to calculate Kemeny consensus (the median) of aggregating a set of given rank functions (for example [21, 22]).

2 Groups, Graphs, and Kemeny Rank Space

For every scoring system A on the data set $D = \{d_1, d_2, \ldots, d_n\}$ with n items (objects, subjects, etc.), rank function r_A is a function which maps each of the n items to a positive integer less than or equal to n. We first assume that the mapping is an one-to-one function (i.e.,ranks $r_A(d_i)$ are all different considering $r_A(D) = [r_A(d_1), r_A(d_2), \ldots, r_A(d_n)]$ as a permutation of the set $N = \{1, 2, \ldots, n\} = [1, n]$ and $\bigcup_{i=1}^{n} r_A(d_i) = [1, n]$). The set of all permutations is a group, denoted by the symmetric group S_n, with the composition $\alpha \circ \beta$ as a binary operation between permutation α and β. The set of rank values $r_A(D)$ is considered as a permutation of the set $[1, n]$ and an element of the group S_n.

Let T_n be the set of all $n-1$ adjacent transpositions in S_n. The Cayley graph $Cay(S_n, T_n)$ is a graph with vertex set $V = S_n$ and edge set $E = \{(\alpha, \alpha \circ t) | \alpha \text{ in } S_n \text{ and } t \in T_n\}$. In the graph, any two points (permutations or rankings) A_1 and A_2 are connected by a path with distance equal to the Kendall's tau distance between A_1 and A_2 in the Rank Correlation Analysis (RCA) [2, 4, 9, 23]. Moreover, since the distance from A_1 to A_2 is equal to the number of adjacent interchanges (swaps), using bubble sort, needed to take A_1 to A_2, this graph is also called Bubble Sort Cayley graph (denoted by B_n). The graph B_n is $(n-1)$ regular, bipartite, and $(n-1)$ connected. It was also shown that B_n consists of $(n-1)$ mutually independent Hamiltonian Cycle [24]. Even though RCA in statistics and CFA in combinatorics and computing are all dealing with ranks (or permutations) in S_n, they have quite different properties and emphasis. RCA treats the set S_n as a population sample space and aims to calculate the static correlation and sufficient P-value so that hypothesis testing can be validated. On the other hand, CFA considers the set S_n as a rank space and aims to produce a process using rank-score diversity (or cognitive diversity) so that a better combined rank function or scoring system can be found [2–7].

In the CFA framework, if the score function s_A of a scoring system A is not an one-to-one function, the number of score values $\bigcup_{i=1}^{n} s_A(d_i)$ is strictly less than the number of data set items n. In such case, one has to resolve the "tie ranking" situation. One immediate solution is to use the mean of these ranks which has ties. However, this will become ineffective if the number of ties

is large. Since the Kendall's tau correlation (or distance) can not handle ties, Kemeny and Snell [25, 26] proposed a distance metric d_K which can handle ties. However, this metric is not practical as it is calculated as sums of absolute values. Emond and Mason reexamined the concept of the half-flip in [26] and defined a new rank correlation τ_x [27]. They showed that the half-flip metric, the Kemeny-Snell metric, and the τ_x rank correlation coefficient are equivalent. More specifically, suppose a weak order A, a rank order with possible ties, of n data items $D = \{d_1, d_2, \ldots, d_n\}$ is defined as an $n \times n$ score matrix a_{ij} as follows:

$$a_{ij} = \begin{cases} 1 & \text{if } d_i \text{ is ranked ahead of or tie with } d_j, \\ -1 & \text{if } d_i \text{ is ranked behind } d_j, \\ 0 & \text{if } i = j. \end{cases} \tag{1}$$

Rank correlation τ_x between two weak orders A and B is taken as inner product of their score matrices:

$$\tau_x(A,B) = \left(\frac{1}{n(n-1)}\right)\left(\sum_{i=1}^{n} a_{ij} b_{ij}\right) \tag{2}$$

Therefore the relationship between the rank correlation $\tau_x(A,B)$ and the rank distance $d_K(A,B)$ in the Kemeny rank space K_n has the relationship:

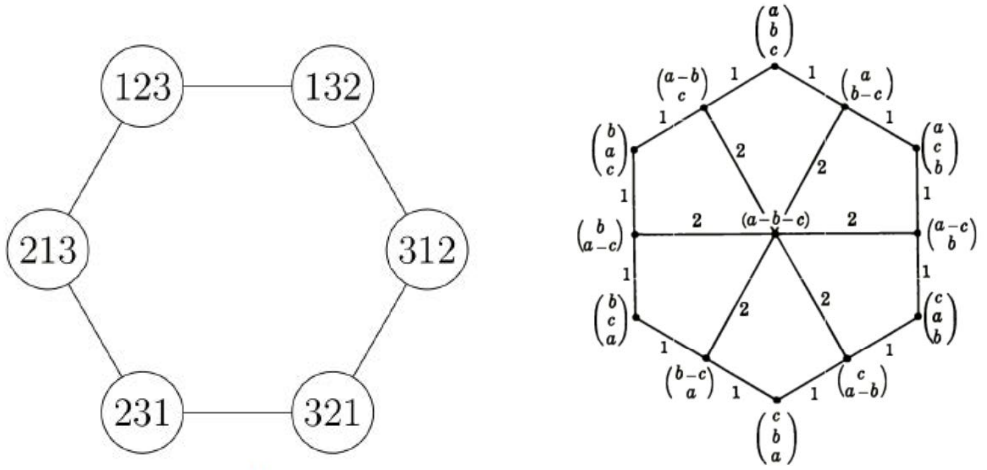
$$\tau_x(A,B) = 1 - \left(\frac{1}{n(n-1)}\right)(d_K(A,B)) \tag{3}$$

which leads to an efficient calculation of $d_K(A,B)$ as:

$$d_K(A,B) = n(n-1) - \left(1 - \frac{\tau_x(A,B)}{2}\right) \tag{4}$$

We note that the vertex set of the Kemeny rank space $V(K_n)$ is a superset of the Bubble sort Cayley graph vertex set $V(B_n)$. For $n = 3$, B_3 and K_3 are shown in the following [7, 8, 26].

Figure 1: Bubble Sort Cayley Graph (B_3) and Kemeny Rank Space (K_3) *



(a) B_3 (b) K_3

Ref. Kemeny John, G.; Snell, J. L. Preference Rankings. An Axiomatic Approach. in Mathematical Models in the Social Sciences, 1962.

3 Figure Skating Judgement: Rank vs. Score Combination

In a figure skating competition, three judges $J_1,\ J_2,$ and J_3 give the eight skaters $\{d_1, d_2, \ldots, d_8\}$ the following scores (Table 1(a)). The goal is to reach a final ranking J^*. Tables 1(a) and 1(b) use score combination and rank combination respectively. Table 2(a) consists of the three score function $s_{J_1},\ s_{J_2},$ and s_{J_3} with scores normalized to the interval $[0, 1]$. Table 2(b) lists the three rank-score functions $f_{J_1},\ f_{J_2},$ and f_{J_3}. Figure 2 consists of rank-score function graphs for $f_{J_1},\ f_{J_2},$ and f_{J_3} on the same coordinate system.

In this example, judge J_3 manipulates the score toward skater d_8. CFA is able to: (a) characterize each judge's scoring/ranking behavior using the rank-score function (Table 2(b)), (b) detect the variation or bias between three judges using the rank-score diversity between two rank-score functions (Figure 2) [28], and (c) mitigate the bias using rank combination instead of score combination (Tables 1(a) and 1(b)).

In addition to average combination, CFA is able to use weighted combination with either performance (WCP) or diversity strength (WCDS) as weight to each system (see Section 2.1.2 in the paper). Figure 3 illustrates these six types of combinations (score combination vs. rank combina-

tion) as well as AC (average combination), WCP (weighted combination by performance strength), and WCDS (weighted combination by diversity strength).

Table 1(a): Final ranking J^* by Score Combination

	J_1	J_2	J_3	SC	J^*
d_1	8.6	7.2	9.3	25.1	5
d_2	7.0	8.8	9.8	25.6	4
d_3	8.4	8.5	9.5	26.4	3
d_4	6.5	7.4	9.1	23.0	8
d_5	9.4	8.5	9.4	27.3	1
d_6	9.2	6.3	10.0	26.5	2
d_7	7.5	7.6	9.0	24.1	7
d_8	10.0	10.0	5.0	25.0	6

Table 1(b): Final ranking J^* by Rank Combination

	J_1	J_2	J_3	RC	J^*
d_1	4	7	5	16	6
d_2	7	2	2	11	3
d_3	5	3.5	3	11.5	4
d_4	8	6	6	20	8
d_5	2	3.5	4	9.5	1
d_6	3	8	1	12	5
d_7	6	5	7	18	7
d_8	1	1	8	10	2

Table 2(a): Normalized Score Function with Score Values in [0,1]
(Note: Linear normalization is used.)

	s_{J_1}	s_{J_2}	s_{J_3}
d_1	0.6	0.24	0.86
d_2	0.14	0.68	0.96
d_3	0.54	0.59	0.9
d_4	0	0.3	0.82
d_5	0.83	0.59	0.88
d_6	0.77	0	1
d_7	0.29	0.35	0.8
d_8	1	1	0

Table 2(b): Rank-score Function f_{J_1},f_{J_2},f_{J_3}

	f_{J_1}	f_{J_2}	f_{J_3}
d_1	1	1	1
d_2	0.83	0.68	0.96
d_3	0.77	0.59	0.9
d_4	0.6	0.59	0.88
d_5	0.54	0.35	0.86
d_6	0.29	0.3	0.82
d_7	0.14	0.24	0.8
d_8	0	0	0

Figure 2: Rank-score Function Graph for f_{J_1}, f_{J_2}, f_{J_3}

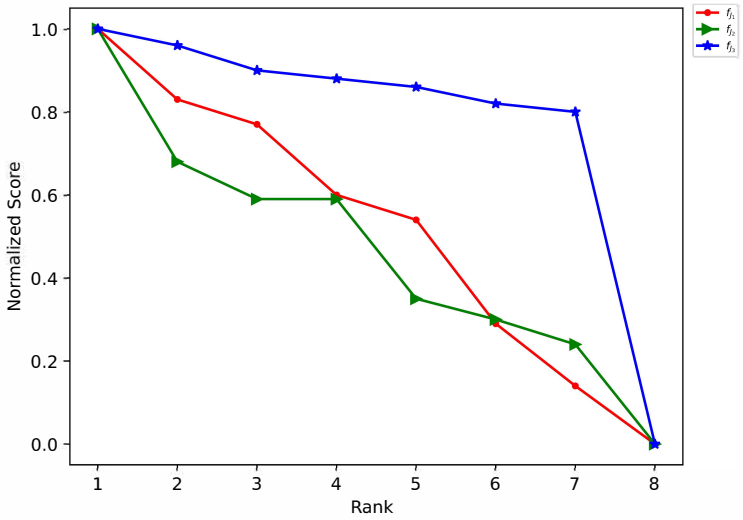
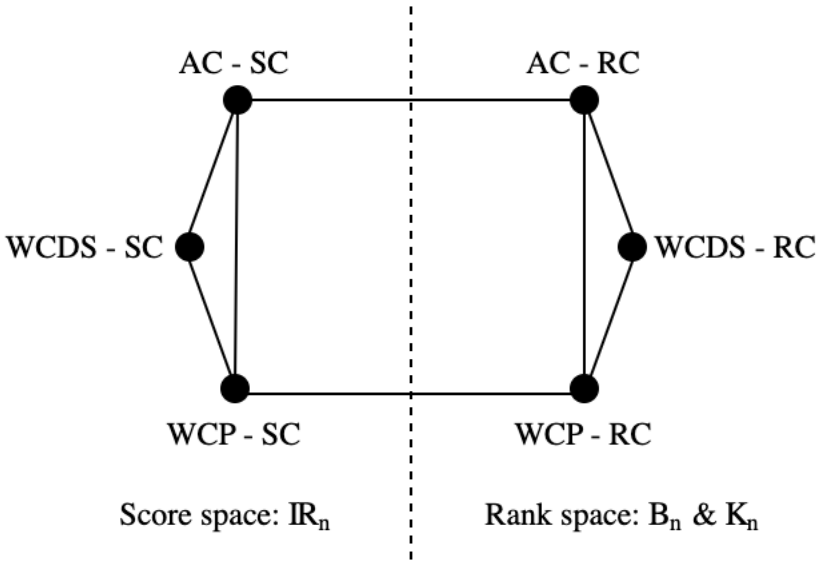


Figure 3: CFA Architecture in The Dual Space of $\mathbb{R}_n$ (Euclidean Space) and B_n & K_n (Bubble Sort Cayley Graph Space & Kemeny Rank Space) *



Ref. Kemeny John, G.; Snell, J. L. Preference Rankings. An Axiomatic Approach. in Mathematical Models in the Social Sciences, 1962.

References

(1) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548* **2021**,

(2) Hsu, D. F.; Taksa, I. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval* **2005**, *8*, 449–480.

(3) Hsu, D. F.; Chung, Y.-S.; Kristal, B. S. Combinatorial fusion analysis: methods and practices of combining multiple scoring systems. *Advanced Data Mining Technologies in Bioinformatics* **2006**, 32–62.

(4) Hsu, D.; Shapiro, J.; Taksa, I. Methods of data fusion in information retreival: Rank vs. Score combination. *DIMACS Technical Report* **2002**, *58*, 662–667.

(5) Hsu, D. F.; Kristal, B. S.; Hao, Y.; Schweikert, C. Cognitive diversity: A measurement of dissimilarity between multiple scoring systems. *Journal of Interconnection Networks* **2019**, *19*, 1940001.

(6) Hsu, D. F.; Kristal, B. S.; Schweikert, C. Rank-score characteristics (RSC) function and cognitive diversity. Brain Informatics: International Conference, BI 2010, Toronto, ON, Canada, August 28-30, 2010. Proceedings. 2010; pp 42–54.

(7) Hurley, L.; Kristal, B. S.; Sirimulla, S.; Schweikert, C.; Hsu, D. F. Multi-Layer Combinatorial Fusion Using Cognitive Diversity. *IEEE Access* **2020**, *9*, 3919–3935.

(8) Zhong, X.; Hurley, L.; Sirimulla, S.; Schweikert, C.; Hsu, D. Combining multiple ranking systems on the generalized permutation rank space. Proc. IEEE 5th Int. Conf. Big Data Intell. Comput.(DATACOM). 2019; pp 123–129.

(9) Marden, J. I. *Analyzing and Modeling Rank Data*; CRC Press, 1996.

(10) Fürnkranz, J.; Hüllermeier, E. *Preference Learning*; Springer, 2010; pp 65–82.

(11) Fligner, M. A.; Verducci, J. S. *Probability Models and Statistical Analyses for Ranking Data*; Springer, 1993; Vol. 80.

(12) Zhang, Z.; Schweikert, C.; Shimojo, S.; Hsu, D. Improving prediction quality of face image preference using combinatorial fusion algorithm. *Brain Informatics* **2023**,

(13) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 1422–1426.

(14) Salim, N.; Holliday, J.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *Journal of chemical information and computer sciences* **2003**, *43*, 435–442.

(15) Ginn, C. M.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. Virtual Screening: An Alternative or Complement to High Throughput Screening? Proceedings of the Workshop 'New Approaches in Drug Design and Discovery', special topic 'Virtual Screening', Schlob Rauischholzhausen, Germany, March 15–18, 1999. 2002; pp 1–16.

(16) Tang, Y.; Li, Z.; Nellikkal, M. A. N.; Eramian, H.; Chan, E. M.; Norquist, A. J.; Hsu, D. F.; Schrier, J. Improving data and prediction quality of high-throughput perovskite synthesis with model fusion. *Journal of Chemical Information and Modeling* **2021**, *61*, 1593–1602.

(17) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *Journal of Chemical Information and Modeling* **2005**, *45*, 1134–1146.

(18) Sniatynski, M. J.; Shepherd, J. A.; Ernst, T.; Wilkens, L. R.; Hsu, D. F.; Kristal, B. S. Ranks

underlie outcome of combining classifiers: Quantitative roles for diversity and accuracy. *Patterns* **2022**, *3*.

(19) Schiavinotto, T.; Stützle, T. A review of metrics on permutations for search landscape analysis. *Computers & Operations Research* **2007**, *34*, 3143–3153.

(20) Diaconis, P. Group representations in probability and statistics. *Lecture Notes - Monograph Series* **1988**, *11*, i–192.

(21) Jiao, Y.; Korba, A.; Sibony, E. Controlling the distance to a kemeny consensus without computing it. International Conference on Machine Learning. 2016; pp 2971–2980.

(22) Dwork, C.; Kumar, R.; Naor, M.; Sivakumar, D. Rank aggregation methods for the web. Proceedings of The 10th International Conference on World Wide Web. 2001; pp 613–622.

(23) Grammatikakis, M. D.; Hsu, D. F.; Kraetzl, M. *Parallel system interconnections and communications*; CRC press, 2000.

(24) Ho, T.-Y.; Lin, C.-K.; Tan, J. J.; Hsu, D. F.; Hsu, L.-H. On the extremal number of edges in hamiltonian connected graphs. *Applied Mathematics Letters* **2010**, *23*, 26–29.

(25) Kemeny, J. G. Mathematics without Numbers. *Daedalus* **1959**, *88*, 577–591.

(26) Kemeny John, G.; Snell, J. L. Preference Rankings. An Axiomatic Approach. W: Mathematical Models in the Social Sciences. 1962.

(27) Emond, E. J.; Mason, D. W. A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis* **2002**, *11*, 17–28.

(28) Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; Hall, P., et al. Towards A Standard for Identifying and Managing Bias in Artificial Intelligence. *NIST special publication* **2022**, *1270*.