# Combinatorial Fusion Analysis

**D. Frank Hsu**, Fordham University
**Bruce S. Kristal**, Tufts University
**Christina Schweikert**, St. John's University

*Combinatorial fusion analysis is a paradigm that seeks to provide methods and workflows for combining multiple scoring systems in computational learning and modeling, informatics, and intelligent systems.*

Multiple scoring systems (MSSs) have been used widely in a variety of different settings including multiple regression, multiple classifier systems, multiple expert systems, multiple neural networks, ensemble methods, machine learning (ML) and artificial intelligence (AI), multiple criterion decision making (MCDM), multimodal biometric systems, preference and deep learning, data and information fusion, and multiple large language models (LLMs).[8,13] In the past decades, combining MSSs has achieved numerous successes across many domain applications, but these settings, algorithms, and approaches vary widely, and better understanding of MSS remains an area of active investigation.

*Combinatorial fusion analysis* (CFA), proposed by Hsu, Chung, and Kristal,[8] provides methods and workflows for combining MSSs in the process of converting raw data to actionable knowledge (through informatics) and the design and development of efficient and effective intelligent systems. CFA characterizes each scoring system with a *score function*, a *rank function*, and a function that relates normalized scores to ranks, termed a *rank-score characteristic* (*RSC*) *function*.[8,10] CFA uses the RSC function to measure the diversity between two scoring systems so as to help specify system selection and algorithm combinations. A unique and critical feature of this diversity metric is that it is independent of the associations between observations and variables. As such, CFA has its antecedents in the works of Hsu, Shapiro, and Taksa[11,12] and Yang et al.[21] The differences and complementarity among scores and ranks, defined by the RSC function, were further explored and operationalized by Hsu, Kristal, and Schweikert in 2010[10], where, by analogy with biological systems, we termed it *cognitive diversity* (CD; note that it has also been referred to as *rank-score diversity*).

Using the RSC function and CD, CFA offers a robust combination algorithm for both score combination in Euclidean space and rank combination in Kemeny space.[8,10,12,14]

*Digital Object Identifier 10.1109/MC.2024.3406058*
*Date of current version: 27 August 2024*

*Multilayer combinatorial fusion* (*MCF*), as an extension of CFA, has recently been established to perform reinforcement learning and deep learning.[13,18,22] CFA and MCF were developed and shown to be useful in a variety of domains. These include information retrieval, virtual screening and drug discovery, protein structure prediction, chromatin immunoprecipitation sequencing (ChIP-Seq) peak detection, target tracking and robotics, microarray gene expression analysis and motif detection, information and cybersecurity, data and information fusion, visual informatics, cognitive neuroscience, decision making, equity ranking, and portfolio management.[8,12,13,18]

This article provides an introduction to critical aspects of CFA including the general combination algorithm and global working space, beginning with the characterization of and diversity among the scoring systems through the lens of the RSC function and CD. Second, the working space of learning, modeling, and decision making for combining MSSs in informatics and intelligent systems is established through a sequence of mathematical constructs: group, graph, and geometry. Third, various combination algorithms are considered to facilitate robust and efficient computational learning and modeling. We note five specific advantages of the CFA framework relative to other existing tools. Then, we illustrate the advantage of the CFA framework using a recent example in drug discovery and contrast it with diversity of ranks and accuracy (DIRAC), another system-level fusion method we have developed.[19,20] Finally, we summarize the article and offers a few remarks for future consideration.

## RSC FUNCTION AND COGNITIVE DIVERSITY

MSSs exist at both the data/attribute/feature level and the information/algorithm/model level. Due to space constraints, we focus this article on the latter.

As proposed,[11,12] scoring system $A$ on the dataset $D = \{d_1, d_2, \ldots, d_n\}$ consists of a score function $s_A$ and a derived rank function $r_A$. Sorting the score value in $s_A$: $D \to R$, the set of real numbers, in descending order using the data items in $D$ as the key, a rank function $r_A$: $D \to N$, where $N = \{1, 2, \ldots, n\}$, is obtained. The RSC function $f_A$: $N \to R$ is defined as $f_A(i) = s_A(r_A{}^{-1}(i)) = (s_A \circ r_A{}^{-1})(i)$. For scoring systems $A$ and $B$, CD between $A$ and $B$, CD($A$, $B$), is computed as the difference between $f_A$ and $f_B$[8,9,10]: CD($A$, $B$) $=$ $d(f_A, f_B) = \sqrt{\sum_{i=1}^{n}((f_A(i) - f_B(i))^2/(n-1)}$.

CFA's unique paradigm for combining MSS using RSC and CD has the following advantages.

**Advantage 1.** The scoring system $A$ so defined on the dataset $D = \{d_1, d_2, \ldots, d_n\}$ with score function $s_A$ and rank function $r_A$ is analogous to the variable $x_A$ on the data points in $D$ with score values $s_A(d_i)$ and rank values $r_A(d_i)$.[13]

**Advantage 2.** The RSC function $f_A$ of the scoring system $A$ in informatics is analogous to, but different from, the cumulative distribution function $F_A(x)$ in statistics.[8,10,13] However, $f_A$ is rank-centric, while $F_A(x)$ is score-centric.[9,10,13]

**Advantage 3.** CD is analogous to, but different from, 1) statistical correlations such as Pearson's $r$, Spearman's rho, and Kendall's tau and 2) computational information diversity such as the Q statistic, KW variance, and double fault.[9,10,13] However, CD is data, distribution, and domain independent since it is measured between two RSC functions, which are rank-centric.

## GLOBAL WORKING SPACE: FROM EUCLIDEAN SPACE TO THE KEMENY SPACE

A scoring system $A$ on the dataset $D = \{d_1, d_2, \ldots, d_n\}$ consists of a score function $s_A$: $D \to R$ and a derived rank function $r_A$: $D \to N$. The score values of the score function $s_A(d_i)$ are in Euclidean space. The rank values $r_A(d_i)$ constitute a permutation of the set of natural numbers $N = \{1, 2, \ldots, n\}$ if the score function $s_A$ is a 1–1 function. That means the rank values $r_A(d_i)$ are all different in $N$. The set of all permutations is a group, *the symmetric group* $S_n$ *of order* $n$, where the composition $\pi_A \circ \pi_B$ is a binary operation between permutations $\pi_A$ and $\pi_B$.[2]

Let $T_n$ be the set of all $n - 1$ adjacent transpositions in $S_n$. The Cayley graph Cay($S_n$, $T_n$) is defined to be the graph with vertex set $S_n$ and edge set $= \{(\alpha, \alpha \circ t) \mid \alpha \text{ in } S_n \text{ and } t \in T_n\}$ and denoted by $B_n$ with Kendall's tau correlation distance as the distance between two points. The graph $B_n$ = Cay($S_n$, $T_n$) is also called the *bubble sort Cayley graph* since any two vertices $A_1$ and $A_2$ are connected by a path of distance equal to the number of adjacent interchanges (swaps) using bubble sort. It has several combinatorial properties including being ($n$ – 1)- connected, bipartite, and ($n$ – 1)- regular and consisting of ($n$ – 1) mutually independent Hamiltonian cycles.[13,14]

If the score function $s_A$ of the scoring system $A$ is not a one-to-one function, the number of score values $\cup_{i=1}^{n} S_A(d_i)$ is strictly less than $n$. In this case, "tie rankings" would occur. Since Kendall's tau correlation distance does not apply to tie rankings, Kemeny and Snell[15] proposed a distance metric $d_k$ that can handle ties. Although $d_k$ satisfies all the axioms of a metric, it is not practical, as it is the sum of absolute values. Emond and Mason defined a new rank correlation $T_x$ between two weak order (rank order with ties) $A$ and $B$ values as the inner product of their score matrices.[3] This leads to an efficient calculation of $d_k(A, B)$ as a function of $T_x(A, B)$.

The vertex set $V(K_n)$ of *the Kemeny rank space* $K_n$ includes more vertices than the vertex set of the bubble sort Cayley graph $V(B_n)$, which is the set $S_n$ of all permutations of $N = \{1, 2, \ldots, n\}$. The Kemeny space $K_n$ has been used widely in information retrieval, MCDM, ML, and computational social science.[1] The cardinality $k_n = |V(K_n)|$ of $V(K_n)$ was calculated as $\sum_{i=1}^{n} b_i * S(n, b)$, where $S(n, b)$ is the Stirling number of the second kind.[5] This number $k_n$ was also studied in terms of preferential arrangements on $n$ distinct decisions or objects, allowing indifference, and obtained as a recursive formula: $k_n = 1 + \sum_{i=1}^{n-1} \binom{n}{i} * k_{n-i}$.[6]

It is not surprising that in a field that seeks to address such a broad spectrum of challenges, CFA is one of many approaches with different strengths and weaknesses.

This number $k_n$ is approximately equal to $(0.5)n!(1.4)^{n+1}$, where $n$ is the number of data items in $D$, which is much larger than $n! = \#(V(B_n))$.[1,5]

The flexibility of extending ranking with or without ties from $B_n$ to $K_n$ gives rise to the issue of complexity. This is one of the reasons why most of the results related to the Kemeny metric focus on optimization in finding the median ranking.[1] Here we would like to point out the advantage of using both the score function in the Euclidean space and the rank function in the Kemeny space, in particular from the information theory perspective.

**Advantage 4.** Let $A$ be a scoring system on the dataset items $D = \{d_1, d_2, \ldots, d_n\}$. Converting the score function $s_A(d_i)$ to a rank function $r_A(d_i)$ may seem to lose information about the data items. However, the information entropy of the working space $B_n$ (or $K_n$) has $\log_2(n!)$ bits of information, which is bigger than $\log_2 k^n = n \log_2 k$ with $k$ score functions, where $n$ is the number of data items in $D$. The rank function $r_A(d_i)$ on $B_n$ (or $K_n$) carries more information across all $n$ data items than that by a score function $s_A(d_i)$ on Euclidean space w.r.t. each of the $n$ data items.

## GENERAL COMBINATION ALGORITHM

Given $m$ scoring systems $A_1, A_2, \ldots, A_m$ on the dataset items $D = \{d_1, d_2, \ldots, d_n\}$, there are $2^m - 1 - m$ possible combinations. Each of these combinations may be a score or a rank combination (SC or RC), as well as one of the three types of combinations: an average combination (AC), a weighted combination by performance (WCP), and a weighted combination by diversity strength (WCDS) where the diversity strength of a scoring system $A_{j*}$, $ds(A_{j*})$ is the average of the CD between $A_{j*}$ and all other values of $A_j$, where $j \in [1, m]$ but $j \neq j*$. As such, we are able to generate $3(2^m - 1 - m)$ new scoring systems by score combination in the Euclidean space and the same number of scoring systems by rank combination in the Kemeny space. Here we illustrate the advantages w.r.t. the combinatorial fusion algorithm and the dual working space.

**Advantage 5.** Since the rank values in the rank space do not follow a distribution pattern and CD between two scoring systems is domain independent, weighted rank or score combination by diversity strength provides a useful potentially nonlinear combination of scoring systems.[14]

## RECENT DRUG DISCOVERY EXAMPLE

Successful drug approval requires optimizing and predicting the five core pharmacokinetic properties: absorption, distribution, metabolism, excretion, and toxicity (ADMET). Existing computational models and methods in informatics, however, often lack generalization and robustness. Recent work by Jiang et al.[14] uses CFA to deploy an ML/AI system (CFA4DD), which enhances ADMET model performance. The CFA4DD model utilizes the 22 ADMET benchmark datasets on Therapeutics Data Commons (TDC) and outperforms many traditional and individual state-of-the-art models. It uses five algorithms as base models with three encoding schemes.

CFA4DD employs three encoding techniques to generate molecular features for the representation of the compounds in these 22 datasets: Morgan circular fingerprints, RDKit 2D molecular descriptors, and MCFP, an encoding scheme using fingerprinting. It uses five ML/AI models, A, B, C, D, and E (extended gradient boosted decision trees, random forest, support vector machines with linear kernel, AdaBoost, and convolutional neural networks). Following the guidelines of TDC, CFA4DD uses either score or rank combination[14] each with the three types of combinations noted above (AC, WCP, and WCDS).

CFA4DD achieved very high rankings in many of the 22 datasets. These datasets range from data size as small as 475 to data sizes as big as 13,130. It is worth noting that due to the CD among the five base models, CFA4DD achieved very good results in datasets E1 and E2, which use Spearman's rho to evaluate the performance of the modeling results on the rank space using WCDS.

## CFA VERSUS DIRAC

It is not surprising that in a field that seeks to address such a broad spectrum of challenges, CFA is one of many approaches with different strengths and weaknesses. CFA is not, for example, as deterministic as multiple regression and has fewer tuning options compared with ensemble classifiers, but it is more flexible than the former and requires fewer data and is less likely to overfit than the latter. In our own work, Sniatynski et al.[19,20] established the DIRAC framework using a combination of empirical and theoretical arguments; we showed that DIRAC accurately predicts the utility of score fusions of binary classifiers, and DIRAC has 100% accuracy in predicting the outcomes of fusions within rank-based binary classifiers or ranking systems. Moreover, the DIRAC framework is a precise geometric representation of *diversity* and *accuracy* as angle-based distances within rank-based combinatorial structures (permutahedra), which is analogous to the bubble sort Cayley graph $B_n$. As such, the certainty of the DIRAC framework provides a general working solution that is domain and distribution independent in critical real-world applications such as biomarker development, personalized health, and clinical trial enrollments. Notably, however, in contrast to CFA, DIRAC cannot directly leverage score–rank interrelationships, and DIRAC relies on the specific observation–variable linkages, while CFA does not consider this information.

In summary, CFA provides a global working space and general combination algorithms for combining MSS in computational learning and modeling, informatics, and intelligent systems. The working space consists of the Euclidean space $R_n$ and the Kemeny space $K_n$; the latter is an extension of the bubble sort Cayley graph space $B_n$ that is useful for working algorithms of learning, modeling, and decision making in domain and distribution-independent applications. Advantage 4 further suggests that learning and modeling algorithms on the working space $B_n$ can be more robust and efficient due to information gain from $n\log_2 k$ bits to $\log_2(n!)$ bits, where $k$ is the number of scoring systems and $n$ is the number of dataset items. Although the Kemeny space $K_n$ contains a working space with more than $\log_2(n!)$ information bits, most of the work done on $K_n$ focuses on rank aggregation.[1] Recent results using MCF on the working space $K_n$ in deep learning beyond the neural network model have demonstrated the viability of designing intelligent general ML/AI systems on the working space $K_n$.[13,18,22]

Here we offer some perspectives on the future use of CFA in informatics and intelligent systems.

**Remark A.** AI-driven drug design and discovery, especially for bioactive small molecules, has been a major/growing area of research.[7,16] CFA4DD was competitive in the TDC ADMET benchmark leaderboard and provides an example of generative AI that is different from and much more energy efficient than the LLM approach.[14,17]

Hsu, Shapiro, and Taksa[11,12] showed that if the diversity between two scoring systems exhibits a certain degree of difference, the rank combination RC($A$, $B$) is better than the score combination SC($A$, $B$) of scoring systems $A$ and $B$. Given this, we can say that two highly diverse scoring systems are expected to fuse more beneficially by ranks than by scores. Coupled with evidence that we can, in some cases, further improve performance by weighting the models, we can then say that it is likely that weighted rank models may be optimal for systems that begin with individual models displaying high CD.[13,14,19,20]

> Given this, we can say that two highly diverse scoring systems are expected to fuse more beneficially by ranks than by scores.

**Remark B.** MCF, used in preference detection[18] and combining multiple ranking systems[22] on the Kemeny space $K_n$, provides examples of a general algorithm for deep learning on the working space $K_n$ that is different from and much more energy efficient than the neural network model (see also Fürnkranz and Hüllermeier[4]).

**Remark C.** CFA uses a global working space that includes both a Euclidean space for score combination and a Kemeny space for rank combination. The RSC function $f_A$ provides a crucial link between the Euclidean space and the Kemeny space for a scoring system $A$. Converting a score function $s_A(d_i)$ to a rank function $r_A(d_i)$ is similar to normalization of the scoring system $A$ in the rank space $K_n$ so that two scoring systems $A$ and $B$ can be compared and the distance between these two points in the Kemeny space can be calculated regardless of its application domain or their distribution.[1,2,13,14]

**Remark D.** As stated in Advantage 3 and described in Remark C, CD($A$, $B$) is data independent and distribution free.

**Remark E.** As stated in Advantage 3 and described in Remark C, CD is data and domain independent. This is useful for unsupervised learning and can help identify and manage a variety of biases in AI including statistical, computational, human, and systemic biases.[17]

## ACKNOWLEDGMENT

All authors, in alphabetical order, contributed to this article equally. The study was supported in part by the USDA Agricultural Research Service under Cooperative Agreement 58-8050-9-004. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the USDA.

## REFERENCES

1. S. Akbari and A. R. Escobedo, “Beyond Kemeny rank aggregation: A parameterizable-penalty framework for robust ranking aggregation with ties,” *Omega*, vol. 119, Sep. 2023, Art. no. 102893, doi: 10.1016/j.omega.2023.102893.
2. P. Diaconis, *Group Representations in Probability and Statistics* (Lecture

Notes-Monograph Series). Beachwood, OH, USA: Institute of Mathematical Statistics, 1988.

3. E. J. Emond and D. W. Mason, “A new rank correlation coefficient with application to the consensus ranking problem,” *J. Multi-Criteria Decis. Anal.*, vol. 11, no. 1, pp. 17–28, 2002, doi: 10.1002/mcda.313.
4. J. Fürnkranz and E. Hüllermeier, Eds., *Preference Learning*. New York, NY, USA: Springer-Verlag, 2010, pp. 1–466.
5. I. J. Good, “The number of orderings of n candidates when ties are permitted,” *Fibonacci Quart.*, vol. 13, no. 1, pp. 11–18, 1975.
6. O. A. Gross, “Preferential arrangements,” *Amer. Math. Monthly*, vol. 69, no. 1, pp. 4–8, 1962, doi: 10.2307/2312725.
7. C. Hasselgren and T. I. Oprea, “Artificial intelligence for drug discovery: Are we there yet?” *Annu. Rev. Pharmacol. Toxicol.*, vol. 64, no. 1, pp. 527–550, 2024, doi: 10.1146/annurev-pharmtox-040323-040828.
8. D. F. Hsu, Y. S. Chung, and B. S. Kristal, “Combinatorial fusion analysis: Methods and practice of combining multiple scoring systems,” in *Advanced Data Mining Technologies in Bioinformatics*, Hsu H. H. Ed., Calgary, AB, Canada: Idea Group, 2006, pp. 1157–1181.
9. D. F. Hsu, B. S. Kristal, Y. Hao, and C. Schweikert, “Cognitive diversity: A measurement of dissimilarity between multiple scoring systems,” *J. Interconnection Netw.*, vol. 19, no. 01, 2019, Art. no. 1940001, doi: 10.1142/S0219265919400012.
10. D. F. Hsu, B. S. Kristal, and C. Schweikert, “Rank-score characteristics (RSC) function and cognitive diversity,” in *Brain Informatics*, Y. Yao, R. Sun, T. Poggio, J. Liu, N. Zhong, and J. Huang, Eds., Berlin, Germany: Springer-Verlag, 2010, pp. 42–54.
11. D. F. Hsu, J. Shapiro, and I. Taksa, “Methods of data fusion in information retrieval: Rank vs. score combination,” Piscataway, NJ, USA, DIMACS Tech. Rep. 2002-58, 2002.
12. D. F. Hsu and I. Taksa, “Comparing rank and score combination methods for data fusion in information retrieval,” *Inf. Retrieval*, vol. 8, no. 3, pp. 449–480, 2005, doi: 10.1007/s10791-005-6994-4.
13. L. Hurley, B. S. Kristal, S. Sirimulla, C. Schweikert, and D. F. Hsu, “Multilayer combinatorial fusion using cognitive diversity,” *IEEE Access*, vol. 9, pp. 3919–3935, 2021, doi: 10.1109/ACCESS.2020.3047057.
14. N. Jiang, M. Quazi, C. Schweikert, D. F. Hsu, T. Oprea, and S. Sirimulla, “Enhancing ADMET property models performance through combinatorial fusion analysis,” 2023. [Online]. Available: https://doi: 10.26434/chemrxiv-2023-dh70x
15. J. G. Kemeny and J. L. Snell, “Preference rankings: An axiomatic approach,” in *Mathematical Models in the Social Sciences*. Cambridge, MA, USA: Ginn Company; Blaisdell Publishing Company, 1962, pp. 9–23.
16. D. Merk, L. Friedrich, F. Grisoni, and G. Schneider, “De Novo design of bioactive small molecules by artificial intelligence,” *Mol. Inform.*, vol. 37, nos. 1–2, 2018, Art. no. 1700153, doi: 10.1002/minf.201700153.
17. R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, “Towards a standard for identifying and managing bias in artificial intelligence,” National Institute of Standards and Technology, Gaithersburg, MD USA, Special Publication (NIST SP) 1270, 2022.
18. C. Schweikert, L. Gobin, S. Xie, S. Shimojo, and D. F. Hsu, “Preference prediction based on eye movement using multi-layer combinatorial fusion,” in *Brain Informatics*, vol. 11309, S. Wang, Ed., Cham, Switzerland: Springer-Verlag, 2018, pp. 282–293.
19. M. J. Sniatynski, J. A. Shepherd, T. Ernst, L. R. Wilkens, D. F. Hsu, and B. S. Kristal, “Ranks underlie outcome of combining classifiers: Quantitative roles for diversity and accuracy,” *Patterns*, vol. 3, no. 2, 2022, Art. no. 100415, doi: 10.1016/j.patter.2021.100415.
20. M. J. Sniatynski, J. A. Shepherd, L. R. Wilkens, D. F. Hsu, and B. S. Kristal, “The DIRAC framework: Geometric structure underlies roles of diversity and accuracy in combining classifiers,” *Patterns*, vol. 5, no. 3, 2024, Art. no. 100924, doi: 10.1016/j.patter.2024.100924.
21. J.-M. Yang, Y.-F. Chen, T.-W. Shen, B. S. Kristal, and D. F. Hsu, “Consensus scoring criteria for improving enrichment in virtual screening,” *J. Chem. Inf. Model*, vol. 45, no. 4, pp. 1134–1146, 2005, doi: 10.1021/ci050034w.
22. X. Zhong, L. Hurley, S. Sirimulla, C. Schweikert, and D. F. Hsu, “Combining multiple ranking systems on the generalized permutation rank space,” in *Proc. IEEE 5th Int. Conf. Big Data Intell. Comput. (DATACOM)*, 2019, pp. 123–129, doi: 10.1109/DataCom.2019.00027.

**D. FRANK HSU** is the Clavius Distinguished Professor of Science and a professor of computer and information science at Fordham University, New York, NY 10023 USA. Contact him at hsu@fordham.edu.

**BRUCE S. KRISTAL** is a senior scientist at the Jean Mayer U.S. Department of Agriculture Human Nutrition Research Center on Aging, Tufts University, Boston, MA 02111 USA. Contact him at bruce.kristal@tufts.edu.

**CHRISTINA SCHWEIKERT** is an associate professor of computer science and the program director for the M.S. in data science program at St. John’s University, Queens, NY 11439 USA. Contact her at schweikc@stjohns.edu.