

Article

Pairwise Performance Comparison of Docking Scoring Functions: Computational Approach Using InterCriteria Analysis

Maria Angelova , Petko Alov , Ivanka Tsakovska , Dessislava Jereva, Iglia Lessigiarska, Krassimir Atanassov , Ilza Pajeva  and Tania Pencheva * 

Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences, 105 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria; maria.angelova@biomed.bas.bg (M.A.); petko@biophys.bas.bg (P.A.); itsakovska@biomed.bas.bg (I.T.); dessislava.jereva@biomed.bas.bg (D.J.); iglika@biomed.bas.bg (I.L.); k.t.atanassov@gmail.com (K.A.); pajeva@biomed.bas.bg (I.P.)

* Correspondence: tania.pencheva@biomed.bas.bg

Abstract

Scoring functions are key elements in docking protocols as they approximate the binding affinity of a ligand (usually a small bioactive molecule) by calculating its interaction energy with a biomacromolecule (usually a protein). In this study, we present a pairwise comparison of scoring functions applying a multi-criterion decision-making approach based on InterCriteria analysis (ICrA). As criteria, the five scoring functions implemented in MOE (Molecular Operating Environment) software were selected, and their performance on a set of protein–ligand complexes from the PDBbind database was compared. The following docking outputs were used: the best docking score, the lowest root mean square deviation (RMSD) between the predicted poses and the co-crystallized ligand, the RMSD between the best docking score pose and the co-crystallized ligand, and the docking score of the pose with the lowest RMSD to the co-crystallized ligand. The impact of ICrA thresholds on the relations between the scoring functions was investigated. A correlation analysis was also performed and juxtaposed with the ICrA. Our results reveal the lowest RMSD as the best-performing docking output and two scoring functions (Alpha HB and London dG) as having the highest comparability. The proposed approach can be applied to any other scoring functions and protein–ligand complexes of interest.

Keywords: molecular operating environment; scoring functions; molecular docking; InterCriteria analysis



Academic Editor: Chan Kyung Kim

Received: 22 May 2025

Revised: 13 June 2025

Accepted: 25 June 2025

Published: 27 June 2025

Citation: Angelova, M.; Alov, P.; Tsakovska, I.; Jereva, D.; Lessigiarska, I.; Atanassov, K.; Pajeva, I.; Pencheva, T. Pairwise Performance Comparison of Docking Scoring Functions: Computational Approach Using InterCriteria Analysis. *Molecules* **2025**, *30*, 2777. <https://doi.org/10.3390/molecules30132777>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The search for hit compounds with specific biological activity requires considerable effort and resources. The primary efforts concern the synthesis and biological testing of a large number of compounds. This defines the strong interest of the R&D pharmaceutical sector towards computer-aided methods that facilitate rational drug design [1]. Among these methods, molecular docking is particularly important when the structure of a biomacromolecule is known. This method is used to predict the binding affinity of a ligand in the active site (by employing scoring functions to calculate docking scores) based on the ligand docking poses (conformations and positions) in the binding site. Docking is also applied in virtual screening to identify potential active ligands for a specific macromolecular target among thousands to millions of compounds. As the calculations of docking scores can be time-consuming, the selection of appropriate scoring functions could impact

the overall effectiveness of the computational procedure. Over the past decades, a number of studies have conducted comparative assessments of scoring functions implemented in both free and commercial docking software [2–5]. To evaluate the predictive power and examine the strengths and limitations of the scoring functions, several benchmark test sets have been developed [3,6,7]. These datasets are highly diverse, encompassing a wide range of protein families, ligand chemotypes, and binding affinities [8], thus providing good opportunities for the comparison of docking scoring functions. Despite the studies [2–5] assessing and comparing the scoring functions in docking, the question regarding which functions are more appropriate for use compared to others remains open.

In this study, the docking scoring functions implemented in Molecular Operating Environment (MOE, <https://www.chemcomp.com/>, accessed on 22 May 2025) [9], the widely used drug discovery software platform, were selected for comparative analysis. The MOE docking module operates with five scoring functions, namely London dG, ASE, Affinity dG, Alpha HB, and GBVI/WSA dG [8,9]. The first four are empirical, while GBVI/WSA dG is a force-field function [5]. Empirical functions evaluate the binding affinity of protein–ligand complexes based on a set of weights defined by multiple linear regression to experimentally measured affinities. The equation terms describe important contributions to the protein–ligand binding, like hydrogen bonding, ionic and hydrophobic interactions, loss in ligand flexibility, etc. On the other hand, the force-field-based scoring functions use classical force fields to evaluate protein–ligand interactions. In their simplest forms, Lennard-Jones and Coulomb potentials are applied to describe the enthalpy terms.

The comparison of MOE scoring functions in this research is based on the Comparative Assessment of Scoring Functions benchmark subset (CASF-2013) [3] of the PDBbind database [6], which is a comprehensive collection of 195 protein–ligand complexes with available binding affinity data. The CASF-2013 dataset was used for the assessment of scoring functions implemented in free docking software such as AutoDock and Vina [10]. In [11], the CASF-2013 dataset supported the evaluation of 12 scoring functions based on machine learning techniques. The high-quality set of 195 test complexes enabled the comparison and assessment of scoring functions embedded in leading software platforms like Schrodinger, MOE, Discovery Studio, SYBYL, and GOLD [12]. CASF-2013, along with another independent benchmark dataset named Community Structure-Activity Resource (CSAR 2014), was used in an evaluation analysis of an improved version of the scoring function SPecificity and Affinity (SPA) [7]. CASF-2013 was applied to validate 3D convolutional neural network (3D-CNN) scoring functions and end-to-end 3D-CNNs with a mechanism for spatial attention [13,14], as well as for extensive validation of the scoring power of geometric graph learning with extended atom-type features for protein–ligand binding affinity prediction [15]. Thus, CASF-2013 was selected as the dataset in our study.

The InterCriteria analysis (ICrA) [16] was applied for pairwise performance comparison of the MOE scoring functions on CASF-2013. The expectations were for the ICrA to reveal relations between the scoring functions by analyzing a variety of data outputs extracted from molecular docking. ICrA was elaborated as a multi-criterion decision-making approach that detects possible relations between pairs of criteria when multiple objects are considered. Over the past years of research, the approach has already demonstrated its potential when applied to economic, biomedical, industrial, and various other data sets and problem formulations [17]. One of the research objectives was to specify the general framework of the problems that may be solved by the approach, as well as the connections between the new approach and classical multi-criteria decision-making methods such as correlation analyses [<https://intercriteria.net>, accessed on 22 May 2025]. Some case studies of the ICrA applications are freely accessible [<https://intercriteria.net/studies/>, accessed on 22 May 2025], e.g., in relation to the World Economic Forum's global competitiveness

reports, genetic algorithm performance, ant colony optimization performance, and analysis of calorimetric data of blood serum proteome.

Recently, the ICRA approach was applied in the field of computer-aided drug design and computational toxicology in comparative studies of various scoring functions [18,19], as well as in *in silico* studies of biologically active molecules [20]. In contrast to our previous studies, here we apply ICRA on the CASF-2013 benchmark subset of the PDBbind database covering several important docking outputs. The current investigation is a significant extension of our previous studies [18–20] as it is based on a much larger dataset of structurally and functionally different proteins and their ligands, with a detailed investigation of the impact of key ICRA thresholds on the comparison results. A correlation analysis was also performed and juxtaposed with the ICRA results. Our results outline the most and the least similar scoring functions and the most comparable docking outputs. In addition, they demonstrate the applicability of ICRA to reveal new relations between the studied criteria.

2. Results and Discussion

For the CASF-2013 dataset, re-docking of the ligands in the protein–ligand complexes was performed. The following data were extracted from 30 saved poses:

- (i) the best docking score (a lower score suggests better protein–ligand binding)—hereinafter referred to as BestDS;
- (ii) the lowest root mean square deviation (RMSD) between the predicted poses and the ligand in the co-crystallized complex—hereinafter referred to as BestRMSD;
- (iii) the RMSD between the pose with the best docking score and the ligand in the co-crystallized complex—hereinafter referred to as RMSD_BestDS;
- (iv) the docking score of the pose with the lowest RMSD to the ligand in the co-crystallized complex—hereinafter referred to as DS_BestRMSD.

For the sake of completeness, we have also investigated the possible relations to the experimental data available in CASF-2013. Based on the availability, K_d or K_i , expressed as $(-\log K_d)$ or $(-\log K_i)$, respectively, were used.

The collected information was formatted for subsequent application of the ICRA: the ligands from protein–ligand complexes were considered objects of research in terms of the ICRA, while the different scoring functions (represented by BestDS, BestRMSD, RMSD_BestDS, or DS_BestRMSD), along with the binding affinity data, were considered as criteria.

ICRA was applied in two steps. Initially, it was implemented on the data outputs extracted from molecular docking results as described above, with the conditionally defined (default in ICRAData) values $\alpha = 0.75$ and $\beta = 0.25$, based on which the ranges of *consonance* and *dissonance* are defined [21].

Table 1 summarizes the results obtained for the degrees of agreement μ between the five scoring functions of MOE and the binding affinity data of the studied complexes after the ICRA application. The colors in Table 1 reproduce the ones used in ICRAData (green represents *positive consonance*, magenta represents *dissonance*) for the conditionally defined values of $\alpha = 0.75$ and $\beta = 0.25$. The ICRA relations between the scoring functions and experimental data are highlighted in grey at the bottom of the table.

As seen from Table 1, “varicolored” ICRA results were obtained only when BestRMSD was considered. For the rest of the docking outputs, ICRA does not outline any significant relations (other than in *dissonance*) between the five investigated scoring functions in MOE, as well as between the scoring functions and the binding affinity data—all hit the *dissonance* zone only (colored in magenta).

Table 1. Degrees of agreement μ obtained after ICRA application with the conditionally defined ICRAData threshold values of $\alpha = 0.75$ and $\beta = 0.25$.

	BestDS	BestRMSD	RMSD_BestDS	DS_BestRMSD
Affinity dG–Alpha HB	0.60	0.81	0.67	0.59
Affinity dG–ASE	0.62	0.77	0.68	0.57
Affinity dG–GBVI/WSA dG	0.55	0.83	0.67	0.61
Affinity dG–London dG	0.56	0.78	0.63	0.56
Alpha HB–ASE	0.66	0.79	0.64	0.62
Alpha HB–GBVI/WSA dG	0.47	0.76	0.69	0.45
Alpha HB–London dG	0.72	0.84	0.68	0.70
ASE–GBVI/WSA dG	0.44	0.73	0.66	0.36
ASE–London dG	0.62	0.77	0.65	0.60
GBVI/WSA dG–London dG	0.48	0.73	0.64	0.46
Affinity dG–($-\log K_d$) or ($-\log K_i$)	0.45	0.57	0.50	0.53
Alpha HB–($-\log K_d$) or ($-\log K_i$)	0.40	0.53	0.44	0.49
ASE–($-\log K_d$) or ($-\log K_i$)	0.35	0.56	0.37	0.57
GBVI/WSA dG–($-\log K_d$) or ($-\log K_i$)	0.58	0.56	0.60	0.55
London dG–($-\log K_d$) or ($-\log K_i$)	0.45	0.53	0.45	0.48

In the second step, the impact of the variations in α and β values on the investigated relations was explored. The thresholds α and β might be determined algorithmically or chosen by the user. Since these parameters define the thresholds for *positive consonance* / *dissonance* / *negative consonance* between the studied pairs of criteria, we investigated the impact of their values on the relations between the scoring functions. We varied the difference between α and β (keeping their sum at 1.00) starting from 0.5 ($\alpha = 0.75$ and $\beta = 0.25$) and decreasing it gradually to 0.2 ($\alpha = 0.6$ and $\beta = 0.4$) to follow how these values affect the relations between the investigated scoring functions (for clarity, these analyses are reported using ICRA colors in Supplementary Tables S1–S5). The results are summarized in Tables 2 and 3 based on the following two indicators: the number of pairs of scoring functions in *positive consonance* for each docking output (Table 2) and the number of docking outputs in *positive consonance* for each pair of scoring functions (Table 3).

Table 2. Number of scoring function pairs in *positive consonance* for each docking output.

α/β	BestDS	BestRMSD	RMSD_BestDS	DS_BestRMSD
0.75/0.25	0	8	0	0
0.70/0.30	1	10	0	1
0.67/0.33	1	10	5	1
0.65/0.35	2	10	7	1
0.60/0.40	5	10	11	4

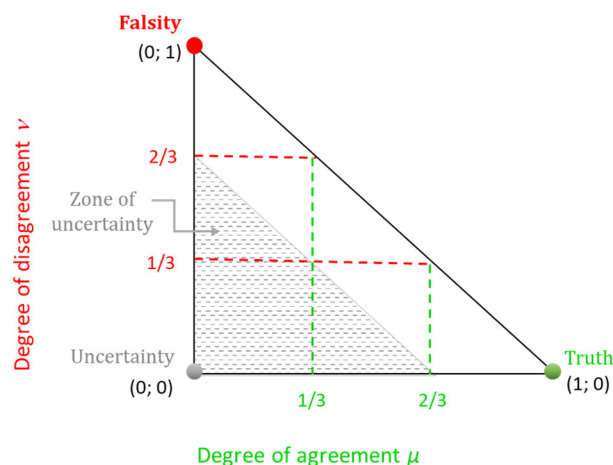
The first indicator, the number of pairs of scoring functions in *positive consonance*, could be considered as a measure of the “sensitivity” of the corresponding docking output to the α and β values. The second one, the number of docking outputs for each pair in *positive consonance*, allows for a comparison of the scoring functions based on the number of the docking outputs—a higher number of docking outputs in *positive consonance* indicates a more similar performance of the pairs of scoring functions studied.

As expected, the decrease in the difference between α and β results in a higher number of pairs in *positive consonance*, thus raising the question regarding the most appropriate α and β values. The values of 0.67/0.33 and 0.65/0.35 show almost identical results, and 0.67/0.33 appears to be a good compromise between “no” and “many” pairs in *positive consonance*, also allowing for a certain tolerance in comparability between the studied criteria.

Table 3. Number of docking outputs in *positive consonance* for each pair of scoring functions.

Pairs of Scoring Functions	α/β				
	0.75/0.25	0.70/0.30	0.67/0.33	0.65/0.35	0.60/0.40
Affinity dG–Alpha HB	1	1	2	2	3
Affinity dG–ASE	1	1	2	2	3
Affinity dG–GBVI/WSA dG	1	1	2	2	3
Affinity dG–London dG	1	1	1	1	2
Alpha HB–ASE	1	1	1	2	4
Alpha HB–GBVI/WSA dG	1	1	2	2	2
Alpha HB–London dG	1	3	4	4	4
ASE–GBVI/WSA dG		1	1	2	2
ASE–London dG	1	1	1	2	4
GBVI/WSA dG–London dG		1	1	1	2
Affinity dG–($-\log K_d$) or ($-\log K_i$)	0	0	0	0	0
Alpha HB–($-\log K_d$) or ($-\log K_i$)	0	0	0	0	0
ASE–($-\log K_d$) or ($-\log K_i$)	0	0	0	0	0
GBVI/WSA dG–($-\log K_d$) or ($-\log K_i$)	0	0	0	0	1
London dG–($-\log K_d$) or ($-\log K_i$)	0	0	0	0	0

In addition, the values $\alpha = 0.67$ and $\beta = 0.33$ allow for the presentation of an alternate scale for *consonance/dissonance*. If the scale of *consonance/dissonance* outlined in Atanassov et al. [21] might be considered as a scale of “quarters”, here a kind of scale of “thirds” is intentionally considered, corresponding to $\alpha = 0.67$ (approximately two-thirds) and $\beta = 0.33$ (approximately one-third). For better understanding, Figure 1 represents the respective interpretative intuitionistic triangle.

**Figure 1.** Interpretative intuitionistic triangle in the case of “thirds”.

In this case, the pair of criteria are said to be in one of the following:

- *positive consonance*, whenever $\mu_{C_k, C_l} \geq 2/3$ and $\nu_{C_k, C_l} < 1/3$;
 - *negative consonance*, whenever $\mu_{C_k, C_l} < 1/3$ and $\nu_{C_k, C_l} \geq 2/3$;
 - *dissonance*, whenever $0 \leq \mu_{C_k, C_l} < 2/3$, $0 \leq \nu_{C_k, C_l} < 2/3$, and $2/3 \leq \mu_{C_k, C_l} + \nu_{C_k, C_l} \leq 1$;
- and
- *uncertainty*, $0 \leq \mu_{C_k, C_l} < 2/3$, $0 \leq \nu_{C_k, C_l} < 2/3$, and $0 \leq \mu_{C_k, C_l} + \nu_{C_k, C_l} < 2/3$.

As seen in Figure 1, the presented scale of “thirds” preserves the symmetry in the interpretative intuitionistic triangle in accordance with the original “quarters”-based symmetry [21]. Thus, the combination $\alpha = 0.67$ and $\beta = 0.33$ appears to be an appropriate selection in this investigation. It should be noted that such a systematic investigation of the

impact of different threshold values for the *consonance* and *dissonance* intervals in the ICrA is conducted for the first time in the current study and may be considered as a methodological contribution to the presented analysis.

According to the pairwise performance comparison of the MOE scoring functions (Table 3), the following results can be outlined:

- (1) an absence of any kind of agreement in all explored values of α and β with the experimental data (except one, between GBVI/WSA dG and $(-\log K_d)$ or $(-\log K_i)$ at $\alpha = 0.60$ and $\beta = 0.40$, rows highlighted in grey in Table 3). This result is in accordance with our previous studies [19]. The lack of any agreement might also be explained by the fact that, even when implementing a variety of scoring terms and becoming more sophisticated, the scoring functions are still a computational approximation mostly aimed at assisting in the prediction of ligand binding poses. This is confirmed by the results of the BestRMSD docking output (Table 2).
- (2) a *positive consonance* between two scoring functions, Alpha HB and London dG: in particular, for 0.67/0.33 threshold values, they are comparable in all four docking outputs (a row in bold in Table 3). The result suggests that these scoring functions might be used interchangeably. At the same time, some pairs show small comparability (Affinity dG–London dG and GBVI/WSA dG–London dG), suggesting that they can complement each other in consensus docking studies.

Figures 2 and 3 demonstrate the results from ICrA implementation for the explored values of thresholds α and β , respectively, for BestDS and RMSD_BestDS. Both figures show the ICrAData screenshots at conditionally defined threshold values of $\alpha = 0.75$ and $\beta = 0.25$ (subplot a), $\alpha = 0.70$ and $\beta = 0.30$ (subplot b), $\alpha = 0.67$ and $\beta = 0.33$ (subplot c), $\alpha = 0.65$ and $\beta = 0.35$ (subplot d), and $\alpha = 0.60$ and $\beta = 0.40$ (subplot e). Subplot (c) represents the full screenshot from the ICrAData software, while the other subplots represent only the results for the degrees of agreement μ and the intuitionistic fuzzy triangle at different values of thresholds.

As seen in Figure 2, applying new values of thresholds leads to the appearance of pairs in *positive consonance*. For $\alpha = 0.70$ and $\beta = 0.30$ (Figure 2b) and $\alpha = 0.67$ and $\beta = 0.33$ (Figure 2c), *positive consonance* appears between the scoring functions Alpha HB and London dG, with no identified significant relations with the conditionally defined threshold values (Figure 2a). Further decreasing the difference between α and β leads to one more pair in *positive consonance*—Alpha HB–ASE for $\alpha = 0.65$ and $\beta = 0.35$ (Figure 2d), and additionally three more pairs in *positive consonance*—namely Affinity dG–Alpha HB, Affinity dG–ASE, and ASE–London dG at $\alpha = 0.60$ and $\beta = 0.40$ (Figure 2e).

As seen in Figure 3 and Table 1, RMSD_BestDS has the highest number of significant relations that appear when the new values of the thresholds are applied. Altogether, five pairs of scoring functions show *positive consonance* at $\alpha = 0.67$ and $\beta = 0.33$, namely, Affinity dG–Alpha HB, Affinity dG–ASE, Affinity dG–GBVI/WSA dG, Alpha HB–GBVI/WSA dG, and Alpha HB–London dG (Figure 3c), in comparison to no significant relations identified at the conditionally defined threshold values (Figure 3a) of $\alpha = 0.70$ and $\beta = 0.30$ (Figure 3b). Further decreasing the difference between α and β leads to two more pairs in *positive consonance*—between ASE–GBVI/WSA dG and ASE–London dG at $\alpha = 0.65$ and $\beta = 0.35$ (Figure 3d), and even to three additional pairs between Affinity dG–London dG, Alpha HB–ASE, and GBVI/WSA dG–London dG at $\alpha = 0.60$ and $\beta = 0.40$ (Figure 3e).

As mentioned above, the most “varicolored” picture from ICrA implementation is when the BestRMSD is considered (the results are presented in Supplementary Tables S1–S5). As seen in Table 1, *positive consonance* was observed for almost all pairs of scoring functions, while the other two pairs of scoring functions, ASE–GBVI/WSA dG and GBVI/WSA dG–London dG, are very close to the range of *positive consonance*. Then, still at the threshold

values $\alpha = 0.70$ and $\beta = 0.30$, all pairs of scoring functions show *positive consonance*. Based on this analysis, one may conclude that, according to the BestRMSD, all scoring functions give quite similar results.

For the DS_BestRMSD output, only one pair of scoring functions, Alpha HB–London dG, falls into the interval of *positive consonance* (Tables 1 and 2; whole information is available in Supplementary Tables S1–S5,) when applying values of thresholds $\alpha = 0.70$ and $\beta = 0.30$. Further decreases in the difference between α and β lead to three more pairs in *positive consonance*—between Affinity dG–GBVI/WSA dG, Alpha HB–ASE, and ASE–London dG, only at $\alpha = 0.60$ and $\beta = 0.40$.

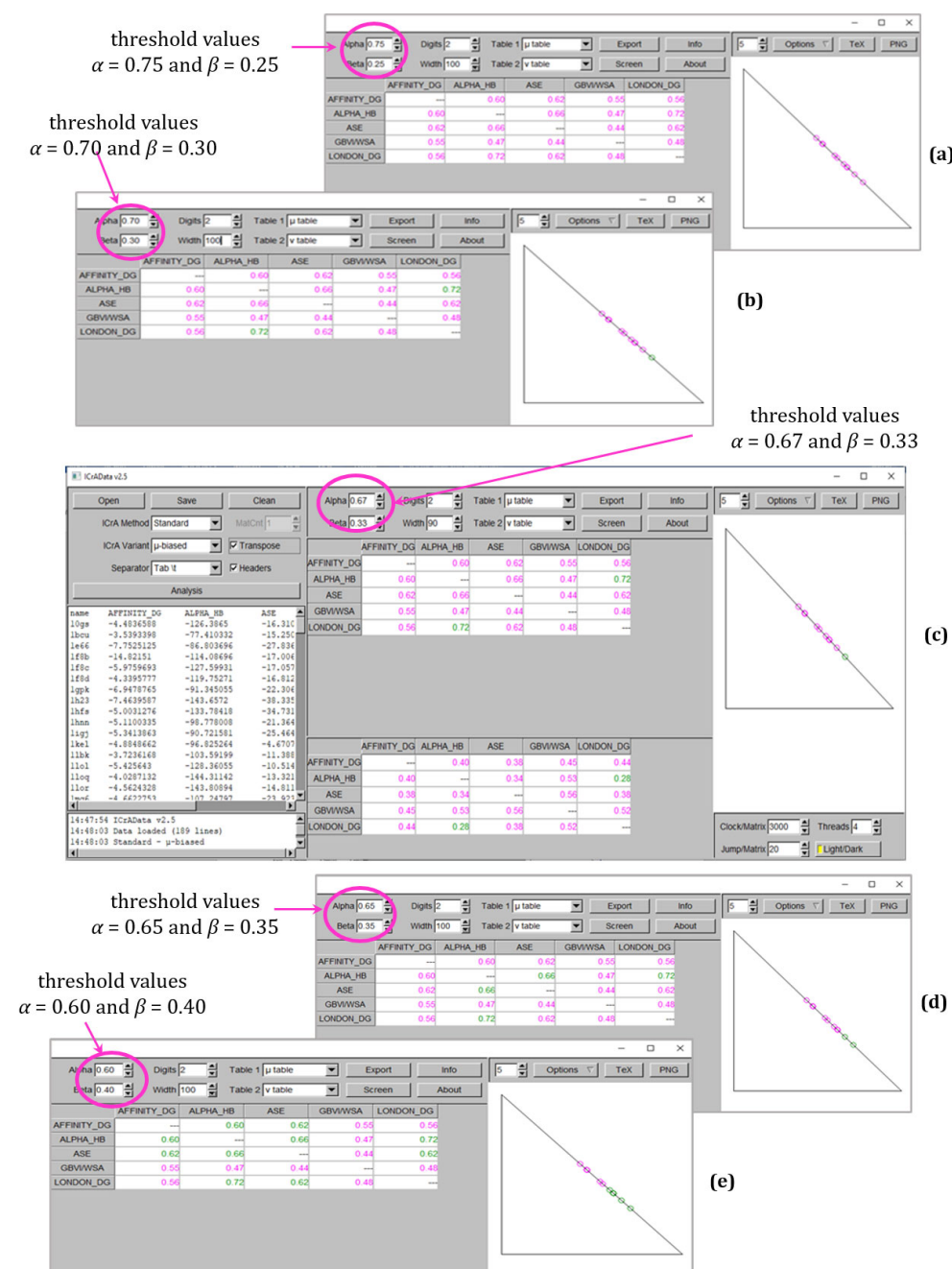


Figure 2. ICrA implementation with different values of α and β thresholds to assess the scoring functions based on BestDS as follows: (a) $\alpha = 0.75$ and $\beta = 0.25$; (b) $\alpha = 0.70$ and $\beta = 0.30$; (c) $\alpha = 0.67$ and $\beta = 0.33$; (d) $\alpha = 0.65$ and $\beta = 0.35$; and (e) $\alpha = 0.60$ and $\beta = 0.40$.

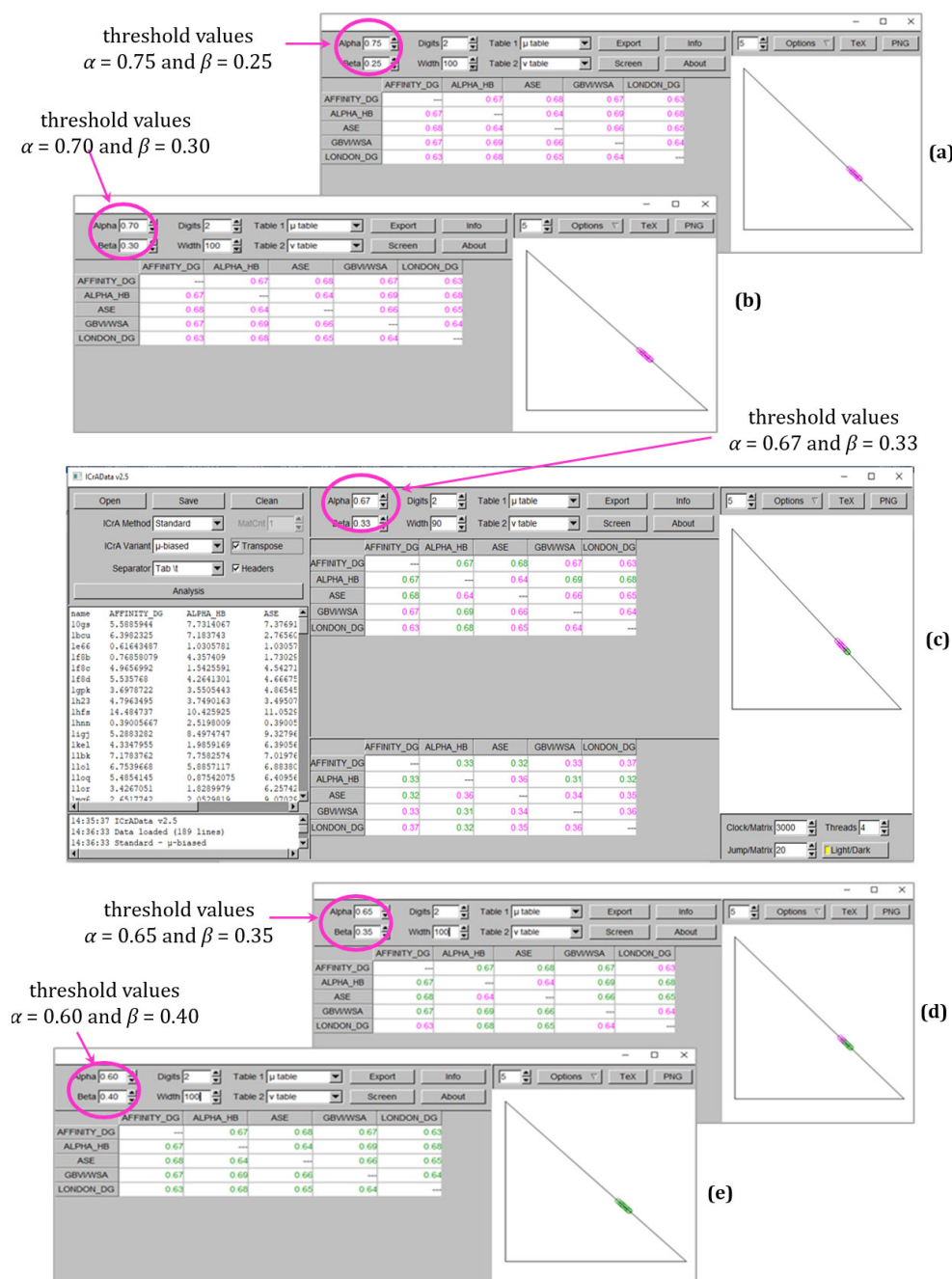


Figure 3. ICRA implementation with different values of α and β thresholds to assess the five scoring functions based on RMSD_BestDS as follows: (a) $\alpha = 0.75$ and $\beta = 0.25$; (b) $\alpha = 0.70$ and $\beta = 0.30$; (c) $\alpha = 0.67$ and $\beta = 0.33$, (d) $\alpha = 0.65$ and $\beta = 0.35$; and (e) $\alpha = 0.60$ and $\beta = 0.40$.

Positive consonance between the scoring functions Alpha HB and London dG was outlined in all investigated docking outputs. The close comparability between London dG and Alpha HB could be related to the fact that both functions include explicit terms related to hydrogen bonding. This type of protein–ligand interaction is among the most common, and it typically has a significant contribution to the binding energy of the protein–ligand complex. In addition, both scoring functions include terms that implicitly account for the conformational changes in the ligands upon binding to achieve the best fit of the ligand in the protein active site. In the case of London dG, this is estimated by the loss of ligand flexibility (calculated from ligand topology only); for Alpha HB, the geometrical fit to the protein active site is estimated by summing up the attractive and repulsive interactions between the protein and ligand atoms. In addition, London dG accounts for metal ligation

and certain entropic effects (the average gain/loss of rotational and translational entropy and desolvation effects). The decision on which scoring function to use could be made based on the preliminary analysis of the protein–ligand interactions in the experimental complex of interest.

For the completeness of the comparison of the five scoring functions, a correlation analysis (CA) has also been performed. Table 4 summarizes the results obtained by the ICRA (the degrees of agreement μ are reported) and CA (Pearson correlation coefficient R) for all docking outputs. The CA shows a higher correlation only for BestRMSD, while for the other docking outputs, the observed correlations are relatively low. In the case of BestRMSD, the highest correlation coefficients coincide with the highest values of ICRA degrees of agreement. In particular, the relation between Alpha HB and London dG is evaluated with the highest values of degree of agreement by ICRA and with the second-best correlation coefficient by CA. Figure 4 illustrates the correlation in terms of CA between Alpha HB and London dG for BestRMSD (Table 4, correlation coefficient 0.79).

Table 4. Results obtained after ICRA and CA applications.

	BestDS		BestRMSD		RMSD_BestDS		DS_BestRMSD	
	ICRA μ	CA R	ICRA μ	CA R	ICRA μ	CA R	ICRA μ	CA R
Affinity dG–Alpha HB	0.60	0.20	0.81	0.74	0.67	0.55	0.59	0.19
Affinity dG–ASE	0.62	0.23	0.77	0.68	0.68	0.53	0.57	0.03
Affinity dG–GBVI/WSA dG	0.55	0.12	0.83	0.81	0.67	0.55	0.61	0.22
Affinity dG–London dG	0.56	0.14	0.78	0.66	0.63	0.35	0.56	0.06
Alpha HB–ASE	0.66	0.54	0.79	0.77	0.64	0.45	0.62	0.42
Alpha HB–GBVI/WSA dG	0.47	−0.10	0.76	0.63	0.69	0.53	0.45	−0.05
Alpha HB–London dG	0.72	0.55	0.84	0.79	0.68	0.52	0.70	0.47
ASE–GBVI/WSA dG	0.44	−0.19	0.73	0.57	0.66	0.48	0.36	−0.16
ASE–London dG	0.62	0.29	0.77	0.68	0.65	0.42	0.60	0.24
GBVI/WSA dG–London dG	0.48	−0.16	0.73	0.57	0.64	0.36	0.46	−0.06
Affinity dG–(−log K_d) or (−log K_i)	0.45	−0.08	0.57	0.21	0.50	0.19	0.53	0.06
Alpha HB–(−log K_d) or (−log K_i)	0.40	−0.29	0.53	0.10	0.44	0.03	0.49	−0.18
ASE–(−log K_d) or (−log K_i)	0.35	−0.42	0.56	0.12	0.37	0.24	0.57	−0.37
GBVI/WSA dG–(−log K_d) or (−log K_i)	0.58	0.13	0.56	0.23	0.60	0.19	0.55	0.09
London dG–(−log K_d) or (−log K_i)	0.45	−0.13	0.53	0.06	0.45	−0.02	0.48	−0.11

The absence of correlation (estimated by the Pearson correlation coefficients) between the docking scores and the experimental data on the binding affinity of the ligands in the complexes is not surprising. This result could not be related only to the heterogeneity of the experimental data (different measures of affinity, different methods and experimental protocols, various protein–ligand complexes) in the dataset used but rather to the fact that molecular docking has not initially been designed for the correlation of docking scores with experimental binding affinities [22]. The absence of such correlations has also been confirmed in our previous studies employing more consistent experimental data on the binding affinities of a homologous series of 88 benzamidine-type ligands toward thrombin, trypsin, and factor Xa [19]. Similar findings have been reported in systematic evaluations of scoring functions by other research groups. Notably, Li et al. [12] assessed a panel of 20 scoring functions, including those implemented in MOE, using the same CASF-2013 dataset. They concluded that the investigated scoring functions generally performed better at predicting binding poses than affinity assessments. These results are in accordance with our findings on the best-performing docking output (BestRMSD) and the absence of consonance with the experimental values; however, in contrast to [12], we confirmed these

results by estimating similarity/dissimilarity between the scoring functions. In this way, our analysis adds a new value to the comparison of the scoring functions.

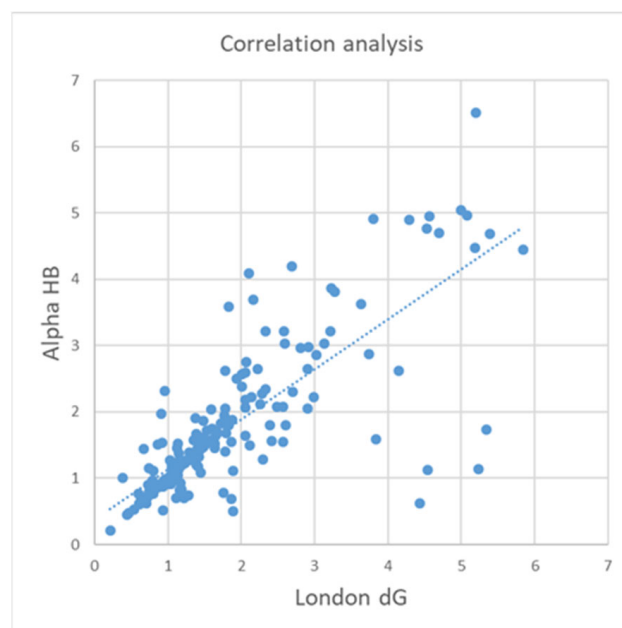


Figure 4. Correlation plot for Alpha HB and London dG based on the BestRMSD.

As seen in Table 4 and Figure 4, ICrA may offer enhanced capabilities over CA by enabling the identification of additional relationships among the evaluated scoring functions. ICrA, as well as CA, reports the degree of coincidence (in terms of ICrA, this is *positive consonance*), and both CA and ICrA report negative correlation (*negative consonance* in terms of ICrA), in which the values for one criterion increase while at the same time the values of the other criterion decrease. Unlike CA, ICrA also allows for classifying the criteria relations in *dissonance*, as well as accounting for the degree of uncertainty, which represents the advantage of ICrA over CA.

In this way, the results from ICrA implementation for the pairwise performance comparison of docking scoring functions would allow the user to decide on the selection of an appropriate scoring function in order to optimize the computational costs. This is relevant, especially when applied to virtual screening tasks. If two or more scoring functions produce similar results, only one of them can be used; instead, the scoring functions that give different results can be combined in consensus docking studies.

3. Materials and Methods

3.1. Dataset

In the current investigation, CASF-2013 was selected as a benchmark dataset widely recognized by researchers dealing with the development and assessment of scoring functions in structure-based studies due to its representativeness, diversity, and quality of complex structures and binding affinity data [3] (Figure 5).

The final CASF-2013 dataset consists of 195 protein–ligand complexes with binding affinity data selected out of 8302 protein–ligand complexes recorded in the PDBbind database (<https://www.pdbbind-plus.org.cn/casf>, v. 2013, accessed on 22 May 2025). The qualified complexes were classified in 65 clusters by 90% similarity in protein sequences. Three representative complexes are chosen from each cluster to control sample redundancy.

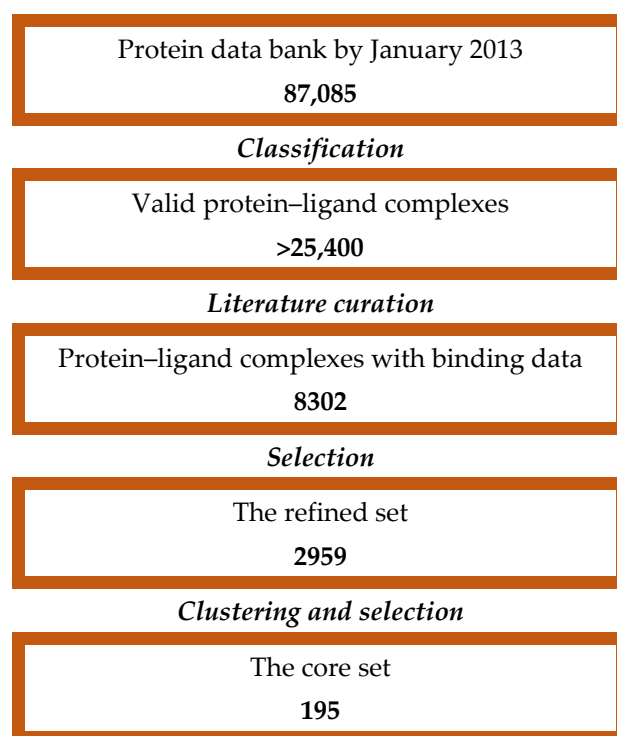


Figure 5. CASF-2013 dataset selection (adapted from [3]).

3.2. Molecular Docking

MOE v. 2022 was used for the molecular docking studies. Prior to re-docking, all 195 protein–ligand complexes were protonated using the Protonate3D tool in MOE. The tool assigns hydrogens to structures following the proton geometry with optimal free energy and the ionization states of titratable protein groups. The rigid protein/flexible ligand option was used. The active site was defined by ligand atoms, and the water molecules were removed from the binding sites of the studied proteins. The placement was obtained by a triangle matcher algorithm.

Re-docking of the ligands in the protein–ligand complexes was performed, applying all available scoring functions in MOE v. 2022, namely London dG, ASE, Affinity dG, Alpha HB, and GBVI/WSA dG, as briefly described below [8,9]:

- ASE is based on the Gaussian approximation and depends on the radii of the atoms and the distance between the ligand atom and receptor atom pairs. ASE is proportional to the sum of the Gaussians over all ligand atom–receptor atom pairs.
- Affinity dG is a linear function that calculates the enthalpy contribution to the binding free energy, including terms based on interactions between H-bond donor and acceptor pairs, ionic interactions, metal ligation, hydrophobic interactions, unfavorable interactions (between hydrophobic and polar atoms,) and favorable interactions (between any two atoms).
- Alpha HB is a linear combination of two terms: (i) the geometric fit of the ligand to the binding site with regard to the attraction and repulsion depending on the distance between the atoms and (ii) H-bonding effects.
- London dG estimates the free binding energy of the ligand, accounting for the average gain or loss of rotational and translational entropy, the loss of flexibility of the ligand, the geometric imperfections of H-bonds and metal ligations compared to the ideal ones, and the desolvation energy of atoms.
- GBVI/WSA dG estimates the free energy of ligand bindings considering the weighted terms for the Coulomb energy, solvation energy, and van der Waals contributions.

Up to 30 poses per ligand were saved for each of the protein–ligand complexes and each of the investigated scoring functions.

3.3. InterCriteria Analysis Approach

The ICRA approach developed by Atanassov et al. in 2014 [16] is based on two mathematical formalisms: index matrices (IMs) [23] and intuitionistic fuzzy sets (IFSs) [24]. The algebraic apparatus of IMs allows for the processing of data arrays of diverse dimensions, while the IFS is a mathematical tool for handling uncertainty. By relying on IMs and IFSs, the ICRA allows for the identification of intercriteria relations in terms of *consonance* or *dissonance* between each pair of criteria, thus differentiating from the classical correlation analysis.

In the concept of IFSs [17], Atanassov builds the ICRA on Zadeh’s theory of fuzzy sets [25], which, on their side, are an extension of the classical notion of set. In classical set theory, an element either belongs or does not belong to the set; therefore, the membership of the element to the set is represented by the values 0 for non-membership or 1 otherwise. The fuzzy set theory introduces a degree of membership μ of the element x to the set, such that $\mu \in [0; 1]$. The theory of IFSs further expands this notion by including the degree of non-membership ν of the element x to the set, such that $\nu \in [0; 1]$.

In mathematical terms, set A is defined as an intuitionistic fuzzy set as follows:

$$A = \{ \langle x, \mu_A(x), \nu_A(x) \rangle \mid x \in X \},$$

where X is the universum, and the two mappings $\mu_A(x), \nu_A(x): A \rightarrow [0, 1]$ are, respectively, the degree of membership and the degree of non-membership of each element $x \in X$, such that $0 \leq \mu_A(x) + \nu_A(x) \leq 1$. The boundary conditions are 0 and 1, both giving the classical set.

The ICRA approach uses a two-dimensional (2D) IM as the input. The 2D IM is represented by a set indexing the rows, a set indexing the columns, and a set of elements corresponding to each pair of row and column index. In the case of ICRA, the IM can be written as $[O, C, e_{o,c}]$, where O and C are, respectively, the set of row and column indices, O stands for objects, C stands for criteria, and $e_{o,c}$ corresponds to the evaluation of each object O against the criterion C . If we denote m as the number of objects and n as the number of criteria, the input IM for the ICRA can be represented as follows:

	C_1	...	C_k	...	C_n
O_1	e_{O_1,C_1}	...	e_{O_1,C_k}	...	e_{O_1,C_n}
...	...	...	...	...	...
O_i	e_{O_i,C_1}	...	e_{O_i,C_k}	...	e_{O_i,C_n}
...	...	...	...	...	...
O_m	e_{O_m,C_1}	...	e_{O_m,C_k}	...	e_{O_m,C_n}

In the next step of the ICRA, relations are formed between every two elements of the matrix; thus, the evaluations of the criteria are compared in pairs for all objects in the matrix (Figure 6).

The relation $R(e_{O_i,C_k}, e_{O_j,C_k})$ has dual relation $\bar{R}$, which is true in the cases when relation R is false and vice versa [16]. Two intuitionistic fuzzy counters, $S_{k,l}^\mu$ and $S_{k,l}^\nu$, are formed and are incremented based on the following rules:

- $S_{k,l}^\mu$ is the number of cases in which the relations $R(e_{O_i,C_k}, e_{O_j,C_k})$ and $R(e_{O_i,C_l}, e_{O_j,C_l})$ (or the relations $\bar{R}(e_{O_i,C_k}, e_{O_j,C_k})$ and $\bar{R}(e_{O_i,C_l}, e_{O_j,C_l})$) are simultaneously satisfied.

- $S_{k,l}^v$ is the number of cases in which the relation $R(e_{O_i,C_k}, e_{O_j,C_k})$ and $\bar{R}(e_{O_i,C_l}, e_{O_j,C_l})$ (or the relations $\bar{R}(e_{O_i,C_k}, e_{O_j,C_k})$ and $R(e_{O_i,C_l}, e_{O_j,C_l})$) are simultaneously satisfied.

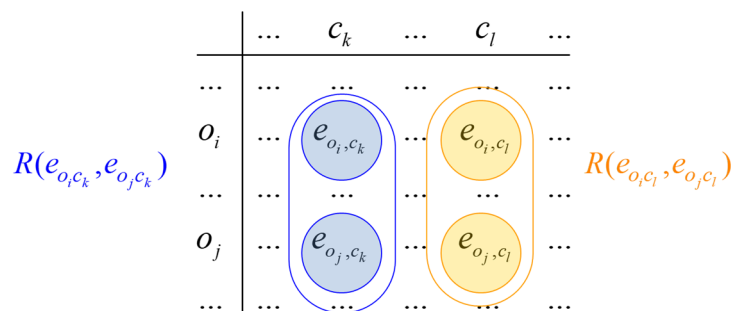


Figure 6. Formation of relations in the ICrA.

The total number of pairwise comparisons between the evaluations of m objects is $m(m-1)/2$; therefore, $0 \leq S_{k,l}^\mu + S_{k,l}^v \leq \frac{m(m-1)}{2}$.

For every k, l ($1 \leq k \leq l \leq m$ and $m \geq 2$), two normalized values are obtained from the above counters:

- $\mu_{C_k,C_l} = 2 \frac{S_{k,l}^\mu}{m(m-1)}$, called the *degree of agreement* in terms of ICrA, and
- $\nu_{C_k,C_l} = 2 \frac{S_{k,l}^v}{m(m-1)}$, called the *degree of disagreement* in terms of ICrA.

The pair $\langle \mu_{C_k,C_l}, \nu_{C_k,C_l} \rangle$ constructed from these two numbers plays the role of intuitionistic fuzzy evaluation of the relation between any two criteria, C_k and C_l . The output IM puts together all these intuitionistic fuzzy pairs as follows:

	C_1	...	C_k	...	C_n
C_1	$\langle 1, 0 \rangle$	...	$\langle \mu_{C_1,C_k}, \nu_{C_1,C_k} \rangle$	...	$\langle \mu_{C_1,C_n}, \nu_{C_1,C_n} \rangle$
...	...	...	...	...	...
C_k	$\langle \mu_{C_k,C_1}, \nu_{C_k,C_1} \rangle$	...	$\langle 1, 0 \rangle$	...	$\langle \mu_{C_k,C_n}, \nu_{C_k,C_n} \rangle$
...	...	...	...	...	...
C_n	$\langle \mu_{C_n,C_1}, \nu_{C_n,C_1} \rangle$	...	$\langle \mu_{C_n,C_k}, \nu_{C_n,C_k} \rangle$	...	$\langle 1, 0 \rangle$

The thresholds for μ_{C_k,C_l} and ν_{C_k,C_l} are determined algorithmically or are indicated by the user. Let $\alpha, \beta \in [0, 1]$ (with $\alpha > \beta$) be the threshold values; then the pair of criteria C_k and C_l are said to be in one of the following:

- *positive consonance*, whenever $\mu_{C_k,C_l} > \alpha$ and $\nu_{C_k,C_l} < \beta$;
- *negative consonance*, whenever $\mu_{C_k,C_l} < \beta$ and $\nu_{C_k,C_l} > \alpha$;
- *dissonance*, otherwise.

Further clarifications on determining *consonance* or *dissonance* can be found in [21].

3.4. Software Implementation of ICrA

A software application that implements the ICrA algorithm called ICrData, v.2.5, was used during the calculations. It is freely available at <https://intercriteria.net/software/> (accessed on 22 May 2025), along with the source code, a README file, and proper examples. The ICrData interface allows for different variants of the ICrA and different algorithms for intercriteria relations calculations to be chosen, the input matrix to be read from a file or pasted from the clipboard, the input matrix to be transposed according to the investigation aims, the index matrices for the degrees of the agreements and disagreements to be visualized according to the user's needs, the threshold values for α and β to be defined by the user (conditionally defined in ICrData as $\alpha = 0.75$ and $\beta = 0.25$, as introduced

in [21]), the graphical interpretation of the results to be visualized by the intuitionistic fuzzy triangle, etc. For better visualization, the tables presenting index matrices for the degrees of the agreements and disagreements use colors for the cells' values: green to indicate *positive consonance*, red for *negative consonance*, and magenta for the cases of *dissonance*.

3.5. Correlation Analysis

The correlations between every two variables investigated in this study were calculated in MS Excel using Pearson's correlation coefficient (Pearson product moment correlation coefficient) [26].

4. Conclusions

Altogether, five scoring functions available in the MOE software package (London dG, Affinity dG, Alpha HB, ASE, and GBVI/WSA dG) were compared for their performance by using four docking outputs (BestDS, BestRMSD, RMSD_BestDS, and DS_BestRMSD) on the benchmark subset CASF-2013 consisting of 195 protein–ligand complexes from the PDBbind database.

The collected information was subjected to the ICRA using the ICRAData software and, additionally, to the CA. The most important outcomes from this investigation might be summarized as follows: (i) to the best of our knowledge, this is the first systematic investigation of the ICRA to assess the pairwise performance comparison of scoring functions available in a molecular software package, considering in parallel a variety of data outputs from molecular docking simulations; (ii) the analysis is also the first systematic investigation of the impact of different threshold values for the intervals of *consonance* and *dissonance* in terms of the ICRA; (iii) the scoring functions Alpha HB and London dG were outlined as a pair appearing in *positive consonance* in all investigated docking outputs, suggesting that these scoring functions might be used interchangeably; (iv) two pairs of scoring functions (Affinity dG–London dG and GBVI/WSA dG–London dG) were mostly in *dissonance* (except for BestRMSD), suggesting that they can complement each other in consensus docking studies; (v) for the docking output BestRMSD, all pairs of scoring functions were in *positive consonance*, thus confirming that the scoring functions studied perform best in reproducing the binding poses of the co-crystallized ligands; (vi) the comparison between the ICRA and CA results illustrates the ability of the ICRA compared to the CA to identify new relations between the investigated criteria.

The proposed approach can be applied to any other scoring functions and software packages as well as any other datasets of protein–ligand complexes of research interest.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/molecules30132777/s1>: Table S1: Results obtained after ICRA application at $\alpha = 0.75$ and $\beta = 0.25$; Table S2: Results obtained after ICRA application at $\alpha = 0.70$ and $\beta = 0.30$; Table S3: Results obtained after ICRA application at $\alpha = 0.67$ and $\beta = 0.33$; Table S4: Results obtained after ICRA application at $\alpha = 0.65$ and $\beta = 0.35$; Table S5: Results obtained after ICRA application at $\alpha = 0.60$ and $\beta = 0.40$.

Author Contributions: Conceptualization, T.P., I.P. and K.A.; methodology, T.P., I.P. and K.A.; software, M.A., P.A., I.T., D.J., I.L. and T.P.; validation, M.A., P.A., I.T., I.L. and T.P.; formal analysis, M.A., P.A., I.T., D.J., I.L., K.A., I.P. and T.P.; investigation, M.A., P.A., I.T., D.J., I.L., K.A., I.P. and T.P.; resources, M.A., P.A., I.T., D.J. and I.L.; data curation, M.A., P.A., I.T., D.J. and I.L.; writing—original draft preparation, M.A., I.P. and T.P.; writing—review and editing, M.A., P.A., I.T., D.J., I.L., K.A., I.P. and T.P.; visualization, M.A., D.J., I.P. and T.P.; supervision, K.A., I.P. and T.P.; project administration, K.A., I.P. and T.P.; funding acquisition, I.T., K.A. and T.P. All authors have read and agreed to the published version of the manuscript.

Funding: K.A. and T.P. gratefully acknowledge the funding by the Bulgarian National Science Fund under the grant KP-06-N72/8 from 14 December 2023, “Intuitionistic Fuzzy Methods for Data Analysis with an Emphasis on the Blood Donation System in Bulgaria”, while M.A., P.A., I.T., D.J., I.L. and T.P. gratefully acknowledge the grant KP-06-COST/3 from 23 May 2023 with regard to COST Action CA21145 “European Network for Diagnosis and Treatment of Antibiotic-resistant Bacterial Infections (EURESTOP)”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article and Supplementary Material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Höltje, H.-D.; Sippl, W.; Rognan, D.; Folkers, G. *Molecular Modeling: Basic Principles and Applications*, 3rd rev. expanded ed.; Wiley-VCH: Weinheim, Germany, 2008.
- Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079–1093. [CrossRef] [PubMed]
- Li, Y.; Liu, Z.H.; Han, L.; Li, J.; Liu, J.; Zhao, Z.X.; Li, C.K.; Wang, R.X. Comparative Assessment of Scoring Functions on an Updated Benchmark: I. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, *54*, 1700–1716. [CrossRef] [PubMed]
- Xu, W.; Lucke, A.J.; Fairlie, D.P. Comparing Sixteen Scoring Functions for Predicting Biological Activities of Ligands for Protein Targets. *J. Mol. Graph. Model.* **2015**, *57*, 76–88. [CrossRef] [PubMed]
- Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, *59*, 895–913. [CrossRef] [PubMed]
- Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977–2980. [CrossRef] [PubMed]
- Yan, Z.; Wang, J. Incorporating specificity into optimization: Evaluation of SPA using CSAR 2014 and CASF 2013 benchmarks. *J. Comput. Aided Mol. Des.* **2016**, *30*, 219–227. [CrossRef] [PubMed]
- Kalinowsky, L.; Weber, J.; Balasubramanian, S.; Baumann, K.; Proschak, E. A Diverse Benchmark Based on 3D Matched Molecular Pairs for Validating Scoring Functions. *ACS Omega* **2018**, *3*, 5704–5714. [CrossRef] [PubMed]
- Molecular Operating Environment. The Chemical Computing Group, v. 2016.08. Available online: <http://www.chemcomp.com> (accessed on 22 May 2025).
- Gaillard, T. Evaluation of AutoDock and AutoDock Vina on the CASF-2013 Benchmark. *J. Chem. Inf. Model.* **2018**, *58*, 1697–1706. [CrossRef] [PubMed]
- Khamis, M.A.; Gomaa, W. Comparative assessment of machine-learning scoring functions on PDBbind 2013. *Eng. Appl. Artif. Intell.* **2015**, *45*, 136–151. [CrossRef]
- Li, Y.; Su, M.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54*, 1717–1736. [CrossRef] [PubMed]
- Wang, Y.; Qiu, Z.; Jiao, Q.; Chen, C.; Meng, Z.; Cui, X. Structure-based Protein-drug Affinity Prediction with Spatial Attention Mechanisms. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021; pp. 92–97.
- Wang, Y.; Wei, Z.; Xi, L. Sfcnn: A novel scoring function based on 3D convolutional neural network for accurate and stable protein-ligand affinity prediction. *BMC Bioinform.* **2022**, *23*, 222. [CrossRef] [PubMed]
- Rana, M.M.; Nguyen, D.D. Geometric graph learning with extended atom-types features for protein-ligand binding affinity prediction. *Comput. Biol. Med.* **2023**, *164*, 107250. [CrossRef] [PubMed]
- Atanassov, K.; Mavrov, D.; Atanassova, V. Intercriteria Decision making: A new approach for multicriteria decision making, based on index matrices and intuitionistic fuzzy sets. *Issues IFSs GNs* **2014**, *11*, 1–8.
- Chorukova, E.; Marinov, P.; Umlenski, I. Survey on Theory and Applications of InterCriteria Analysis Approach. In *Research in Computer Science in the Bulgarian Academy of Sciences. Studies in Computational Intelligence*, 1st ed.; Atanassov, K., Ed.; Springer: Cham, Switzerland, 2021; Volume 934, pp. 453–469.
- Jereva, D.; Angelova, M.; Tsakovska, I.; Alov, P.; Pajeva, I.; Miteva, M.; Pencheva, T. InterCriteria Analysis Approach for Decision-making in Virtual Screening: Comparative Study of Various Scoring Functions. *Lect. Notes Netw. Syst.* **2022**, *374*, 67–78.
- Jereva, D.; Pencheva, T.; Tsakovska, I.; Alov, P.; Pajeva, I. Exploring Applicability of InterCriteria Analysis to Evaluate the Performance of MOE and GOLD Scoring Functions. *Stud. Comput. Intell.* **2021**, *961*, 198–208.

20. Tsakovska, I.; Alov, P.; Ikonov, N.; Atanassova, V.; Vassilev, P.; Roeva, O.; Jereva, D.; Atanassov, K.; Pajeva, I.; Pencheva, T. Intercriteria analysis implementation for exploration of the performance of various docking scoring functions. *Stud. Comput. Intell.* **2021**, *902*, 88–98.
21. Atanassov, K.; Atanassova, V.; Gluhchev, G. Intercriteria analysis: Ideas and problems. *Notes Intuit. Fuzzy Sets* **2015**, *21*, 81–88.
22. Medina-Franco, J.L. Grand Challenges of Computer-Aided Drug Design: The Road Ahead. *Front. Drug Discov.* **2021**, *1*, 728551. [[CrossRef](#)]
23. Atanassov, K. *Index Matrices: Towards an Augmented Matrix Calculus*, *Studies in Computational Intelligence*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2014; Volume 573.
24. Atanassov, K. *Intuitionistic Fuzzy Logics*; Springer: Cham, Switzerland, 2017.
25. Zadeh, L.A. Fuzzy Sets. *Inf. Control.* **1965**, *8*, 338–353. [[CrossRef](#)]
26. Correlation Analysis. Available online: <https://www.questionpro.com/features/correlation-analysis.html> (accessed on 9 January 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.