Discriminating physiological from non-physiological interfaces in structures of protein complexes: a community-wide study

by

Hugo Schweke[1], Qifang Xu[2], Gerardo Tauriello[3], Lorenzo Pantolini[3] , Torsten Schwede[3], Frédéric Cazals[4], Alix Lhéritier[5], Juan Fernandez-Recio[6], Luis Angel Rodríguez-Lumbreras[6], Ora Schueler-Furman[7], Julia K. Varga[7], Brian Jiménez-García[8,9], Manon F. Réau[8], Alexandre M.J.J.Bonvin[8], Castrense Savojardo[10], Pier-Luigi Martelli[10], Rita Casadio[10], Jérôme Tubiana[11], Haim J. Wolfson[11], Romina Oliva[12], Didier Barradas-Bautista[13], Tiziana Ricciardelli[14], Luigi Cavallo[14], Česlovas Venclovas[15], Kliment Olechnovič[15], Raphael Guerois[16], Jessica Andreani[16], Juliette Martin[17], Xiao Wang[18], Genki Terashi[18], Daipayan Sarkar[18], Charles Christoffer[19], Tunde Aderinwale[19], Jacob Verburgt[18], Daisuke Kihara[18,19], Anthony Marchand[20], Bruno E. Correia[20], Rui Duan[21], Liming Qiu[21], Xianjin Xu[21], Shuang Zhang[21], Xiaoqin Zou[21], Sucharita Dey[22], Roland L. Dunbrack[2], Emmanuel D. Levy[1*], Shoshana J. Wodak[23*]

1- Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot 7610001, Israel
2- Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, PA, 19111 USA
3- Biozentrum, University of Basel & SIB Swiss Institute of Bioinformatics, Spitalstrasse 41, 4056 Basel, Switzerland
4- Centre Inria d'Université Côte d'Azur, F-06902Sophia-Antipolis, FRANCE
5- Amadeus SAS, F-06902 Sophia-Antipolis, France
6- Instituto de Ciencias de la Vid y del Vino (ICVV), CSIC-UR-Gobierno de La Rioja, E-26004 Logroño, Spain
7- Department of Microbiology and Molecular Genetics, The Institute for Medical Research Israel-Canada, Hebrew University-Hadassah Medical School, Jerusalem 91120, Israel
8- Computational Structural Biology Group, Department of Chemistry, Bijvoet Centre, Faculty of Science, Utrecht University, Utrecht 3584 CH, The Netherlands
9- Zymvol Biomodeling SL, Barcelona, Spain
10- Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, via San Giacomo 9/2, 40126 Bologna, Italy
11-Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
12- Department of Sciences and Technologies, University of Naples "Parthenope", I-80143, Naples, Italy
13- Kaust Visualization Lab, Core lab Division, King Abdullah University of Science and Technology (KAUST), 23955-6900, Thuwal, Saudi Arabia
14- Physical Sciences and Engineering Division, Kaust Catalysis Center, King Abdullah University of Science and Technology (KAUST), 23955-6900, Thuwal, Saudi Arabia
15- Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania
16- Institute for Integrative Biology of the Cell (I2BC), Commissariat à l'Energie Atomique, CNRS, Université Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, 91190, France
17- Univ Lyon, Université Claude Bernard Lyon 1, CNRS, UMR 5086 MMSB, F-69367, Lyon, France
18- Department of Biological Sciences, Purdue University, West Lafayette, IN 79075
19- Department of Computer Sciences, Purdue University, West Lafayette, IN 79075
20 Laboratory of protein design and immunoengineering, Ecole polytechnique fédérale de Lausanne (EPFL), 1015 - Lausanne, Switzerland
21- Department of Physics and Astronomy, Department of Biochemistry, Dalton Cardiovascular Research Center, Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211
22- Department of Bioscience and Bioengineering, Indian Institute of Technology Jodhpur, Karwar-342037, Rajasthan, India
23- VIB-VUB Center for Structural Biology, Pleinlaan 2, 1050, Brussels Belgium.

*Corresponding Authors:
Emmanuel Levy (emmanuel.levy@weizmann.ac.il)
Shoshana J. Wodak (shoshana.wodak@gmail.com)

Supplementary Tables

Table S1: Master Table of the benchmark dataset of physiological and non-physiological homodimers used in this study

Link:
https://drive.google.com/file/d/1TyOK_a1_hUVmYyT73OVHp_GiYo0iT549/view?usp=share_link
or github: https://github.com/vibbits/Elixir-3DBioInfo-Benchmark-Protein-Interfaces

File: benchmark_annotations.csv

Table recapitulating the main features of the complexes of the dataset. The columns are organized as follow:

- *ID*: pdb identifier of the structure, followed by "_X" for the complexes from the QSalign dataset, with X denoting the number of the biological assembly.
- *InterfaceID*: The unique integer number for an interface in the crystal. Specific to complexes derived from ProtCID [1].
- *AuthChain1*: The author chain identifier of the first chain of the dimer. Specific to complexes derived from ProtCID [1].
- *AuthChain2*: The author chain identifier of the second chain of the dimer. Specific to complexes derived from ProtCID [1].
- *SymmetryOp1*: The symmetry operator used to rotate and translate the asymmetric chain to generate the first chain of a dimer. Crystallographic symmetry operators are used to build the crystal from an asymmetric unit, defined in PDB/mmCIF/XML files. Interfaces are identified from the crystal. Specific to complexes derived from ProtCID [1].
- *SymmetryOp2*: The symmetry operator used to rotate and translate the asymmetric chain to generate the second chain of a dimer. Specific to complexes derived from ProtCID [1].
- *physio*: True if the dimer is a physiological contact, False otherwise.
- *contacts*: The number of inter-residue contacts between the two subunits of the dimer.
- *gene*: the UniProt identifier of the protein.
- *superfamily*: The CATH [2,3] superfamily annotation of the protein.
- *pfam*: The Pfam domain annotation [4,5] of the protein.
- *bsa*: the buried surface area of the dimer
- *bsa_polar*: The polar surface area of the dimer
- *bsa_apolar*: The apolar surface area of the dimer
- *frac_polar*: The polar fraction of the buried surface area of the dimer
- *frac_apolar*: The apolar fraction of the buried surface area of the dimer

Table S2: Benchmark dataset of homodimers annotated with their original assignments and classification results by the 3 classification procedures, and the results of re-evaluation of the original assignments.

Link:https://docs.google.com/spreadsheets/d/1iSc_zxCsdk-X_hQ2GixKdvRlRQ-dcQpt/edit#gid=538228730

or github: https://github.com/vibbits/Elixir-3DBioInfo-Benchmark-Protein-Interfaces
File:***classification_entries_allmethods.xlsx***
Table showing the classification of each entry of the benchmark by the Consensus Score, the Random Forest and AlphaFold2 methods. The columns are organized as follow:

Sheet 1: *classification_entries_allmethods*

- *pdb.id***:** pdb identifier of the structure, followed by "_X" for the complexes from the QSalign dataset, with X denoting the number of the biological assembly.
- *physio***:** 1 if the dimer is a physiological contact, 0 otherwise.
- *consensus.score*: classification according to the consensus score. TRUE if correctly classified, FALSE otherwise.
- *random.forest*: classification according to the random forest score. TRUE if correctly classified, FALSE otherwise.
- *AlphaFold2*: classification according to the AlphaFold score. TRUE if correctly classified, FALSE otherwise.
- *misclassified*: number of methods misclassifying the entry.

Sheet 2: *misclassified_entries_analysis*
The subset of entries misclassified by 2 or more methods. Columns 1-6 contain the same information as in Sheet 1, Column 7 and 8 are organized as follow:

- *reassignment*: lists misclassified entries that are likely misassigned ('1' red), and misclassified entries that are likely not misassigned ('0' blue).
- *protCID*: Comments supporting the re-assignments on the basis of ProtCID or ProtCAD.

Table S3: Summary of the changes in the benchmark set after re-evaluation of the 150 entries misclassified by 2 or more methods (see section 3.5 of main text).

	entry reassigned	**entry deleted**	**total number of entries**
physiological	4	2	840
non-physiological	12	4	831

Table S4: Updated master Table of the benchmark dataset of physiological and non-physiological homodimers following re-evaluation of the 150 entries misclassified by 2 or more methods (not used in this study).
The re-evaluation was performed using updated versions of the ProtCID and ProtCAD resources and resulted in reassignment or deletion of some entries (see summary Table S3 and section 3.5 of the Main text for details). This updated version of the benchmark dataset should be used for any future work.
Link:
https://drive.google.com/file/d/1rUEKym0raq4zLLSWhRVCRDkGRmLdBWR2/view?usp=share_link

or github: https://github.com/vibbits/Elixir-3DBioInfo-Benchmark-Protein-Interfaces

File: benchmark_annotated_updated_30042023.csv

Table recapitulating the main features of the complexes of the dataset. The columns are organized as follow:

- *ID*: pdb identifier of the structure, followed by "_X" for the complexes from the QSalign dataset, with X denoting the number of the biological assembly.
- *InterfaceID*: The unique integer number for an interface in the crystal. Specific to complexes derived from ProtCID [1].
- *AuthChain1*: The author chain identifier of the first chain of the dimer. Specific to complexes derived from ProtCID [1].
- *AuthChain2*: The author chain identifier of the second chain of the dimer. Specific to complexes derived from ProtCID [1].
- *SymmetryOp1*: The symmetry operator used to rotate and translate the asymmetric chain to generate the first chain of a dimer. Crystallographic symmetry operators are used to build the crystal from an asymmetric unit, defined in PDB/mmCIF/XML files. Interfaces are identified from the crystal. Specific to complexes derived from ProtCID [1].
- *SymmetryOp2*: The symmetry operator used to rotate and translate the asymmetric chain to generate the second chain of a dimer. Specific to complexes derived from ProtCID [1].
- *physio*: True if the dimer is a physiological contact, False otherwise.
- *contacts*: The number of inter-residue contacts between the two subunits of the dimer.
- *gene*: the UniProt identifier of the protein.
- *superfamily*: The CATH [2,3] superfamily annotation of the protein.
- *pfam*: The Pfam domain annotation [4,5] of the protein.
- *bsa*: the buried surface area of the dimer
- *bsa_polar*: The polar surface area of the dimer
- *bsa_apolar*: The apolar surface area of the dimer
- *frac_polar*: The polar fraction of the buried surface area of the dimer
- *frac_apolar*: The apolar fraction of the buried surface area of the dimer
- *comment*: comments after manual curation. Can be "physio after curation", "non-physio after curation", "ambiguous", "deleted" or blank.

Supplementary methods

Identifying non-Physiological dimers from the ProtCID database

To identify non-physiological dimers from the ProtCID database [1], if there are at least 3 CFs (crystal forms) for a protein (a UniProt), any interface in only one CF is defined as a non-physiological dimer. The procedure is described as following:

1. For each UniProt with #CFs ≥ 3
   For each interface in each crystal form
   If an interface is symmetric (isologous) and not in ProtCID clusters
   add to {non-physiological dimers}
2. Add in close homologous of UniProt with sequence identity ≥ 80%
   For each group with #CFs ≥ 3
   For each interface in each crystal form

If an interface is symmetric and not in ProtCID clusters
add to {non-physiological dimers}

3. Filter
   Remove EM and NMR structures
   Remove dimers with resolution > 3.5 Å
   Remove dimers with ≥ 3 atomic clashes (Chimera: cutoff 0.6 Å, h-bonds 0.4 Å)
4. Match surface area distribution of physiological dimers, choosing 700 nonphy dimers
   Calculate percent of {physiological dimers} in each bin every 100 Å²
   Take that percent * 700 from each bin of {non-physiological dimers}, sorted by #CFs and sequence identity, distinct from {non-phys-QSalign dimers}.

Supplementary Figures

Figure S1

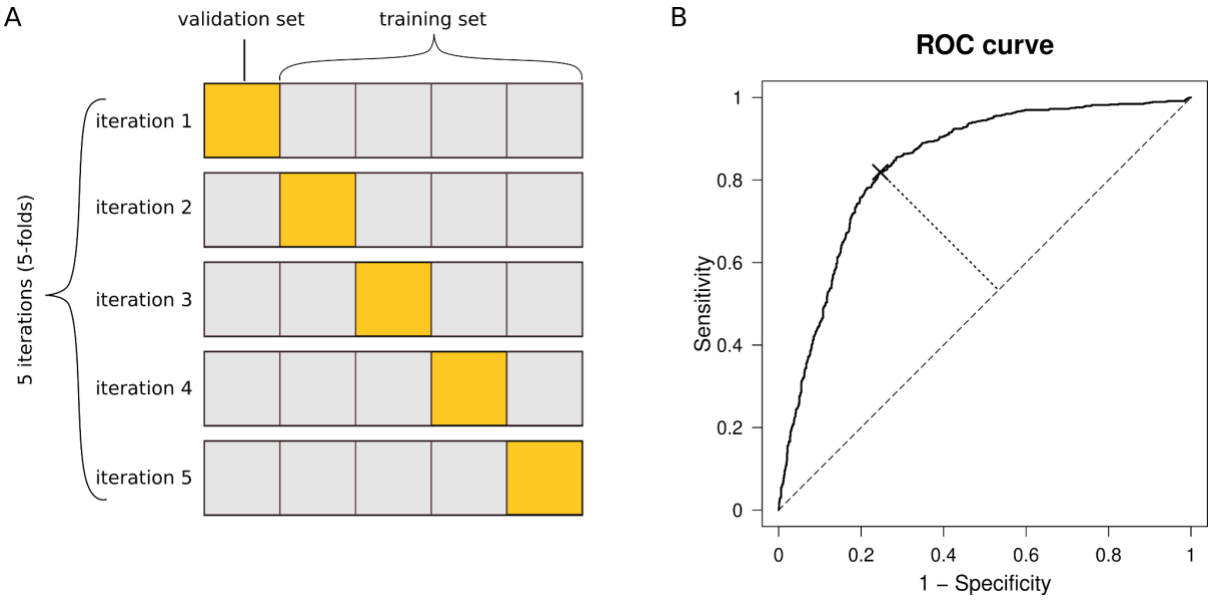

Figure S1: ROC and score threshold generated for individual scores computed by each group for dimers of the benchmark dataset. (a) Five-fold cross validation approach used to compute the Receiver Operator Characteristic (ROC) curve for individual scores computed for the benchmark dimers by each of the 13 groups. (b) Score threshold defined for classification purposes as the point on the ROC most distant from the diagonal.

Figure S2

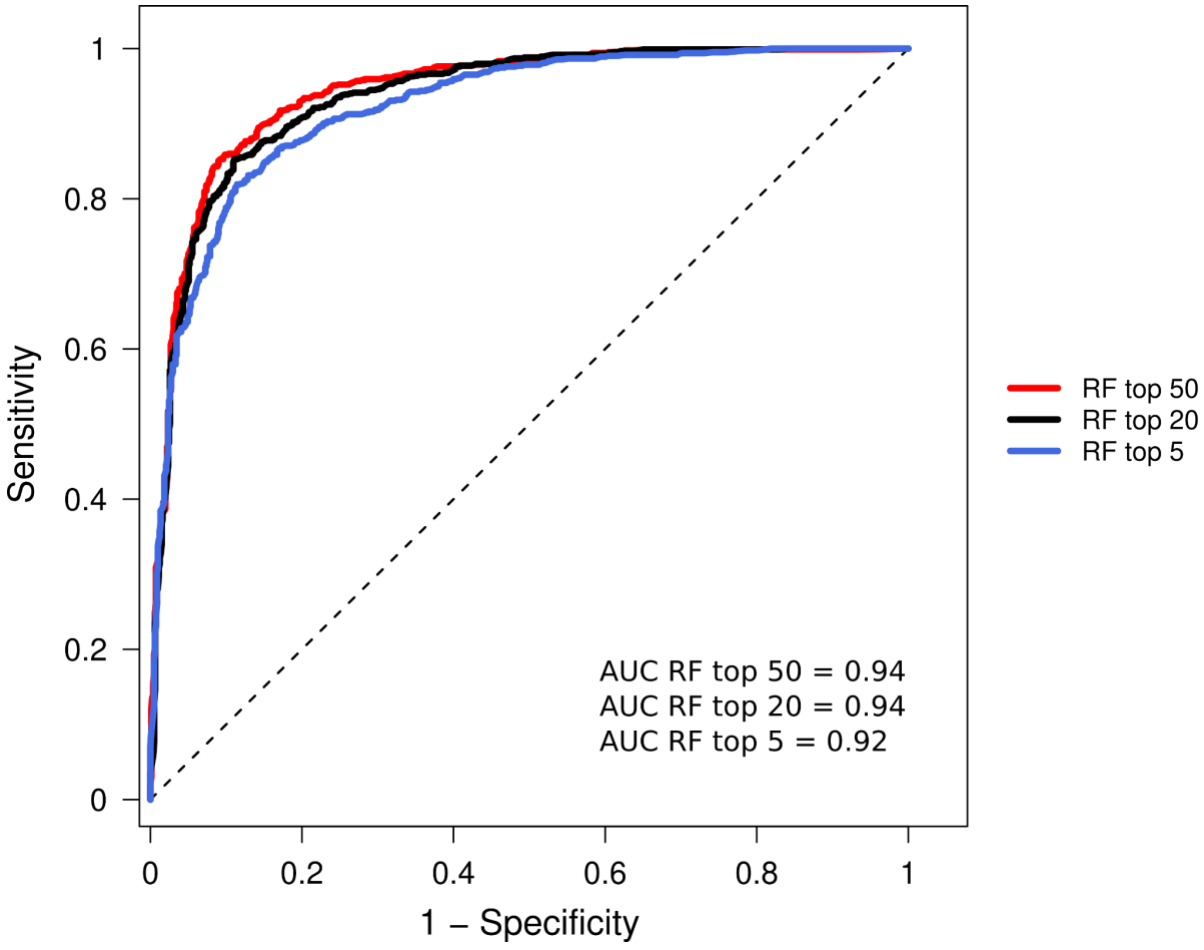

Figure S2: Random Forest (RF) classifier performance for the top 50, 20 and top 5 raw scores. ROC curves and their AUC, for the 50, 20 and 5 top raw scores with the highest impact on the performance of the RF classifier.

Figure S3

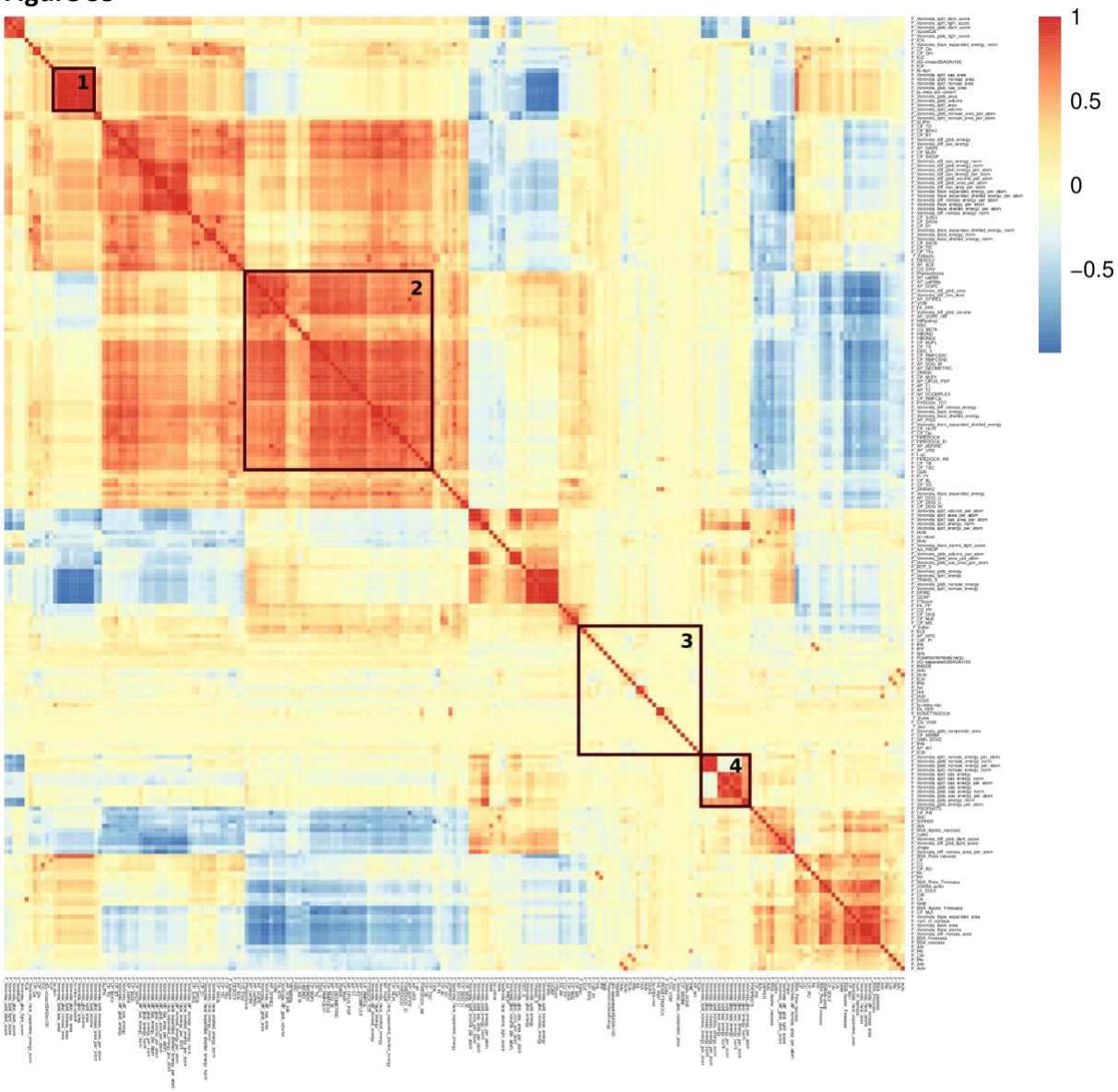

Figure S3: Heatmap of the matrix of the pairwise (Pearson) correlations between the 221 raw scores used in this study clustered along columns and rows. Examples of patterns displayed by the correlation between groups of scores (labeled Group 1-4) are highlighted. Group 1 includes scores from the Venclovas group. These scores evaluate Voronoi tessellation-derived interatomic contact areas, solvent-accessible areas and volumes for the dimer (global) and separate subunits (split). Scores in this group are positively correlated to each other (red square rear the diagonal) but negatively correlated to scores evaluating energetic contributions computed by the same group (dark blue rectangle upper middle region). Group 2 includes a large set of scores from different groups, evaluating energy terms of the dimer interface. These scores tend to be positively correlated to each other and negatively correlated to scores from the Venclovas team that evaluate contributions from geometric features (volumes and area) of the binding interface. Such negative correlations of volume and area features of interfaces, with energetic features are expected, since better packing and larger interface areas are associated with lower (more favorable) binding energies. The scores of group 3 stand out by being weakly correlated to one another, as well as to other scores. These scores include surface area-based scores as well as specific energetic terms (electrostatic, VdW), some of which are evaluated for the dimer as a whole and interactions with the

solvent. The last example of group 4 includes Voronoi tessellation-based scores from the Venclovas group, evaluating mainly interatomic contact area-based energetic properties of the dimer, and of the independent subunits. This group of well correlated scores, are weakly correlated to most other scores.

Figure S4

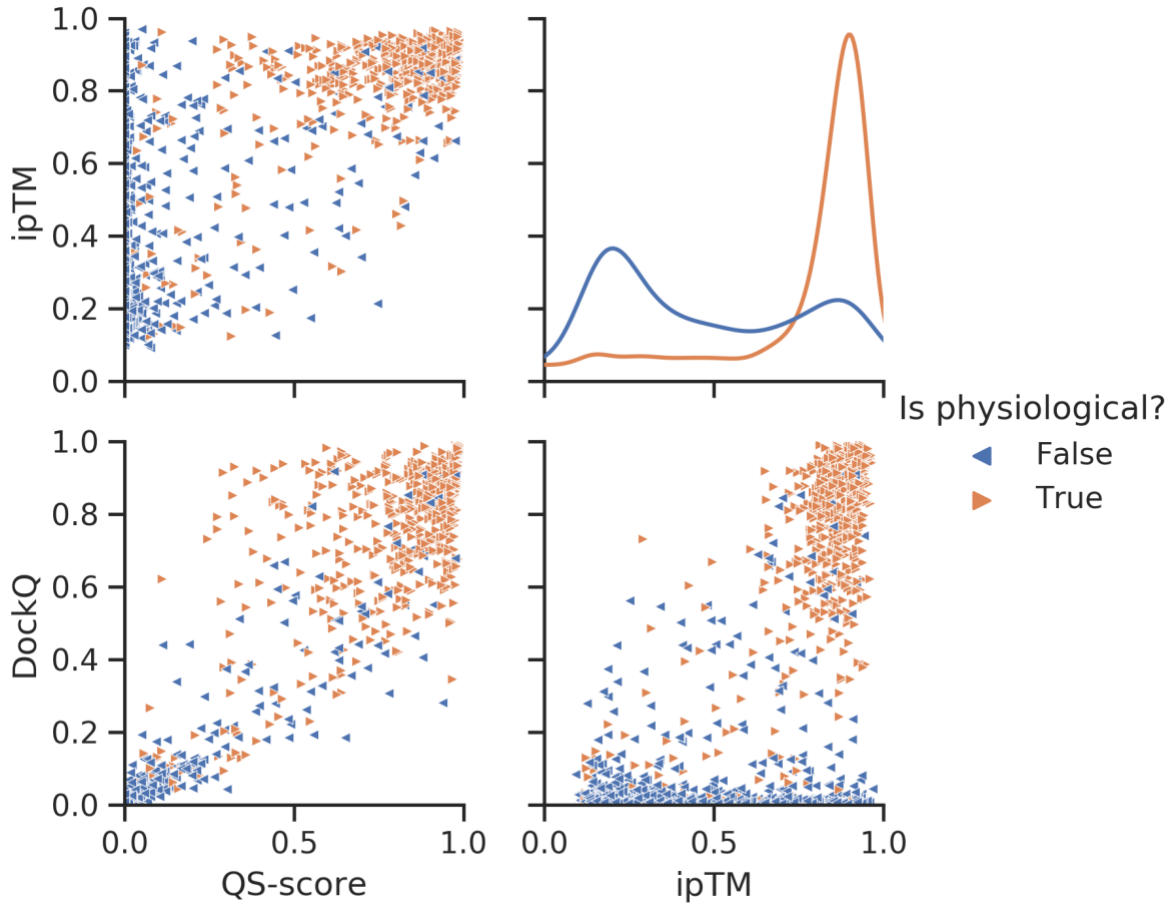

Figure S4: Distributions of similarity metrics for models predicted by AlphaFold-multimer (unrelaxed) for the 1671 dimers of the benchmark dataset which could be successfully evaluated. For each target the model with highest ipTM was selected (among the five given by AlphaFold). The upper right plot depicts the distribution of ipTM values while the scatter plots illustrate the pairwise relationships between the 3 scores. The targets are split (by colors and markers) according to whether the target dimer is physiological or non-physiological.

Figure S5:

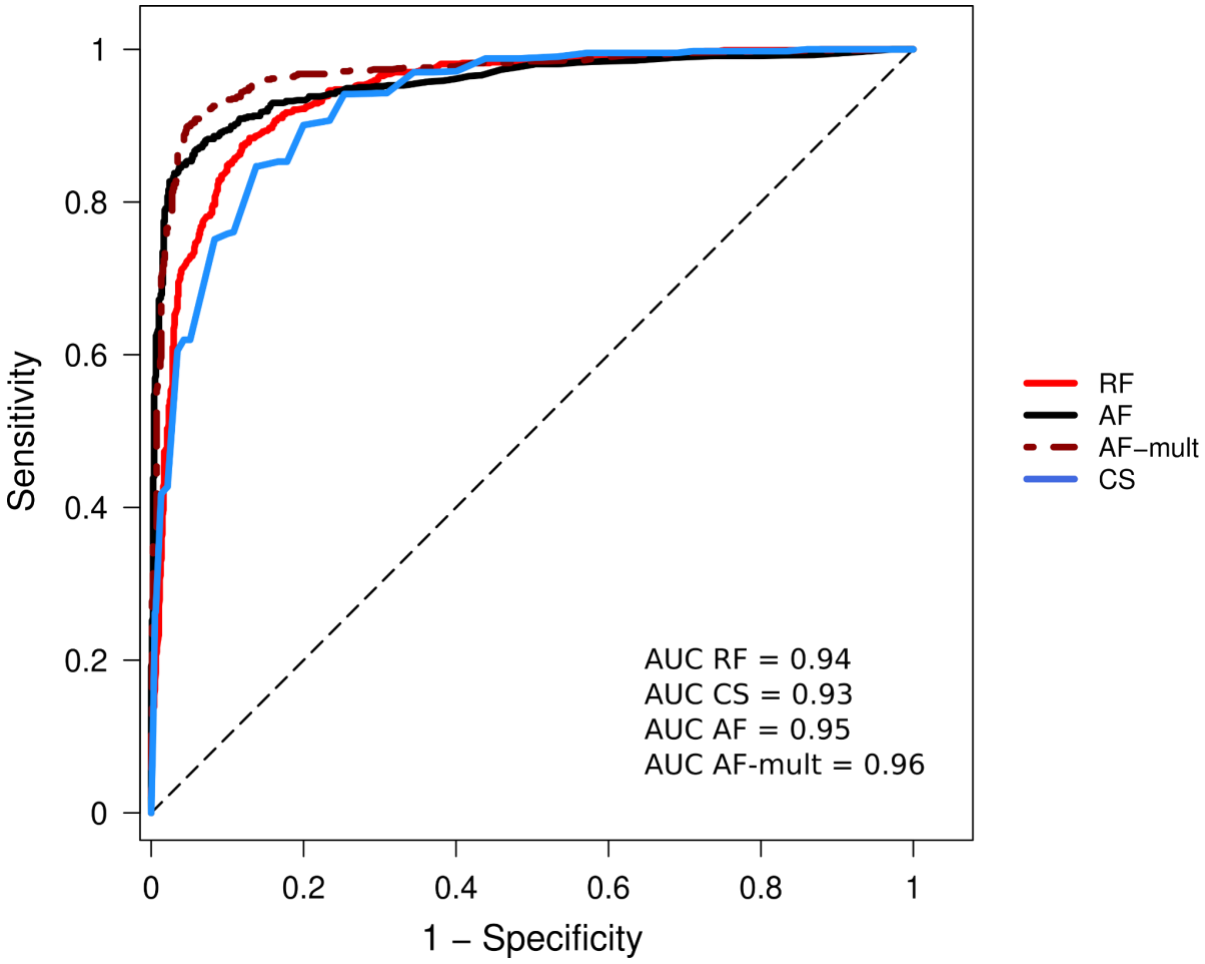

Figure S5: ROCs computed for the Consensus score (CS), Random Forest (RF) classifier and the similarity scores of models predicted by AlphaFold2 (AF) and AlphaFold-multimer (AF-mult), for the homodimers of the benchmark dataset.

Collated Methods of individual groups

https://docs.google.com/document/d/1DAuEGi6qOUTVhVmPPm4FynomsKBbWq-G/edit

References

[1] Xu, Q., Dunbrack, R.L., Jr, ProtCID: a data resource for structural information on protein interactions. *Nat. Commun.* 2020, 11, 711.

[2] Lewis, T.E., Sillitoe, I., Dawson, N., Lam, S.D., et al., Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res.* 2018, 46, D435–D439.

[3] Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., et al., CATH: increased structural coverage of functional space. *Nucleic Acids Res.* 2021, 49, D266–D273.

[4] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., et al., Pfam: The protein families database

in 2021. *Nucleic Acids Res.* 2021, 49, D412–D419.
[5] Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., et al., InterPro in 2022. *Nucleic Acids Res.* 2023, 51, D418–D427.