**Supplementary Information for**

**Title** Cross-Species Molecular Docking Method to Support Predictions of Species Susceptibility to Chemical Effects

**Authors** Peter G. Schumann a, Daniel Chang b, Sally Mayasich c, d, Sara Vliet d, Terry Brown e, Carlie A. LaLone \*d

a Oak Ridge Institute for Science and Education, Duluth, Minnesota, USA

b U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, Chemical Characterization and Exposure Division, Research Triangle Park, North Carolina, USA

c Aquatic Sciences Center, University of Wisconsin‐Madison, Madison, Wisconsin, USA

d U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division, Duluth, Minnesota, USA

e U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, Scientific Computing and Data Curation Division, Duluth, Minnesota, USA

\***Corresponding Author:** lalone.carlie@epa.gov

Any use of trade, firm, or product, names is for descriptive purposes only and does not imply endorsement by the authors or the U.S. Government.

**This PDF file includes:**

Supplementary text

Supplementary Figures S1 to S4

Supplementary Table S1

**Supplementary Information Text**

**Advanced search criteria for generating the human androgen receptor ensemble**

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) was accessed on 8/21/23 to generate an ensemble set comprised of 47 human AR structures: 1e3g, 1t5z, 1t63, 1t65, 1xow, 1xq3, 2am9, 2ama, 2amb, 2ao6, 2ax9, 2axa, 2hvc, 2pio, 2pip, 2piq, 2pir, 2pit, 2piu, 2piv, 2pix, 2pkl, 2pnu, 2q7i, 2q7j, 2yhd, 2ylo, 2ylp, 2ylq, 3b5r, 3b65, 3b66, 3b67, 3b68, 3l3x, 3l3z, 3rlj, 3v49, 3v4a, 4hlw, 4k7a, 4ql8, 5cj6, 5jjm, 5v8q, 5vo4, 8e1a.

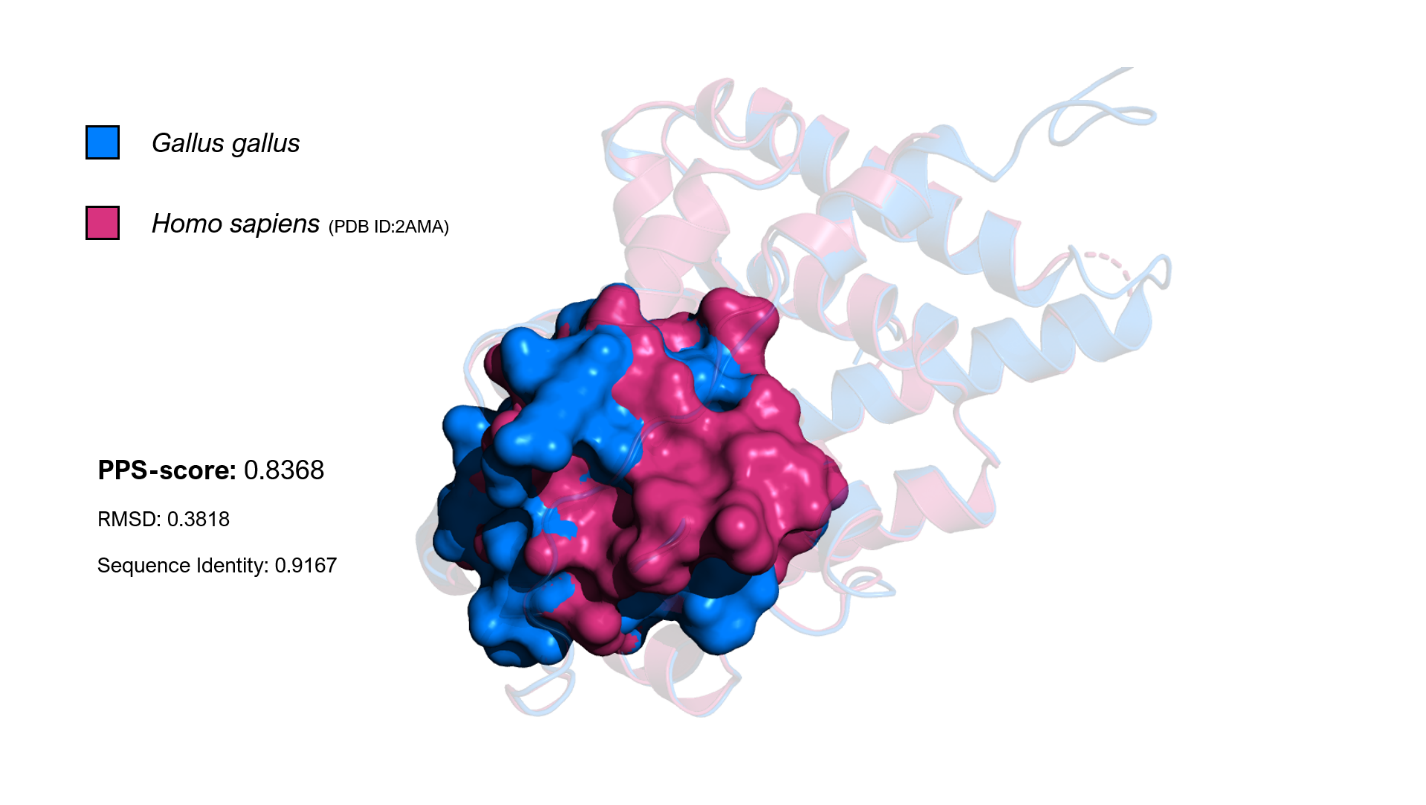
The advanced search criteria were as follows:

**QUERY:** (Gene Name = "AR" **AND** Entry Polymer Types = "Protein (only)" **AND** Refinement Resolution **<=** 2.5 **AND** Scientific Name of the Source Organism = "Homo sapiens" **AND** Polymer Entity Mutation Count **=** 0 **AND** Structure Title **NOT** **HAS EXACT PHRASE** "mutant") **AND** (Full Text = "androgen receptor" **AND** Full Text = "AR") **AND** Structure Similarity **WHERE** (Entry ID = "2AMA" **AND** Chain ID = "A" **AND** Shape Match = "Strict")

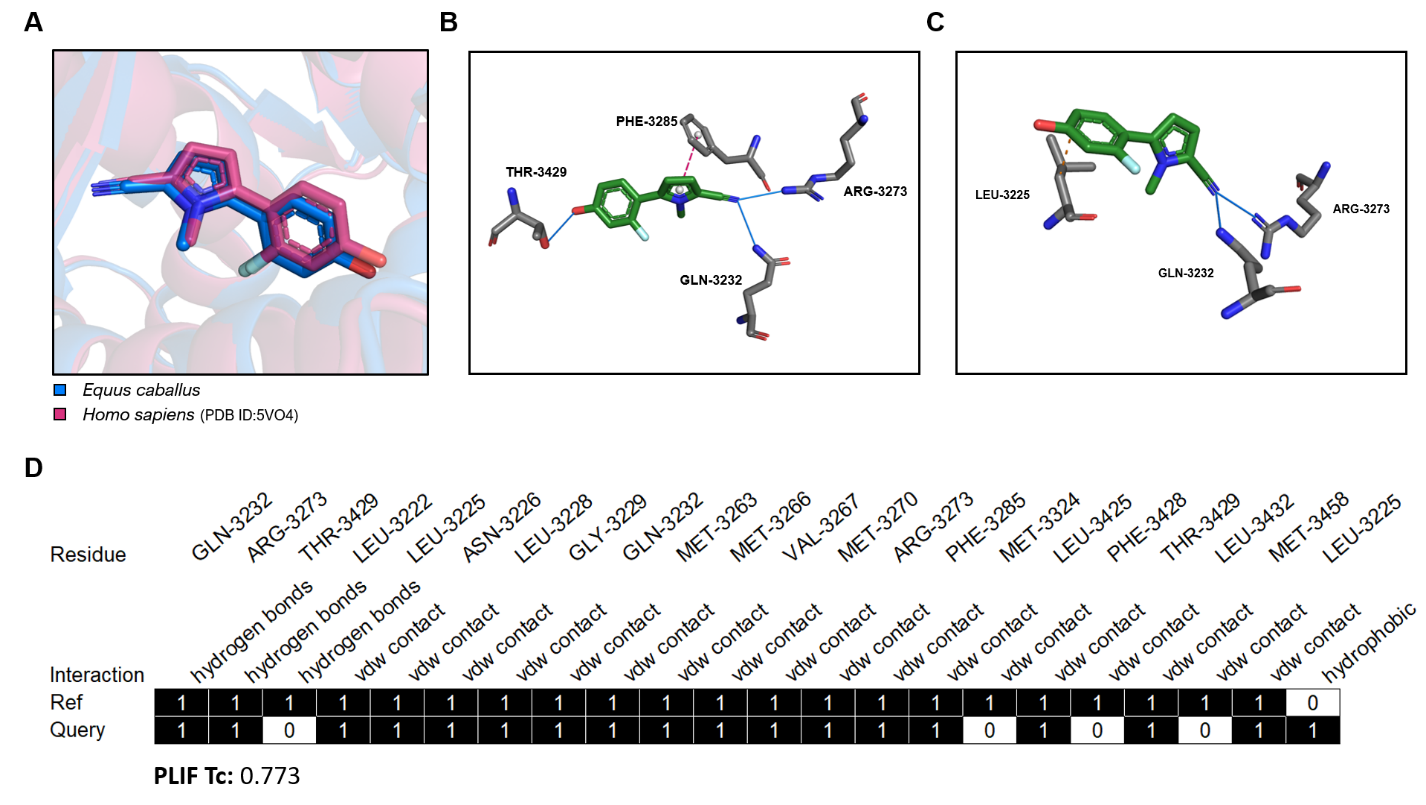
Chart, histogram

Description automatically generated

**Figure S1.** Histograms of the ligand RMSD values calculated for each of the binding modes generated from the docking simulations of the 268 species structures for (**A**) DHT and (**B**) FHPMPC.



**Figure S2.** An example of determining the binding pocket shape similarity between the reference structure (*Homo sapiens*; PDB ID:2AMA) and a query structure (*Gallus gallus*) using the PPS-align algorithm ([www.zhanggroup.org/PPS-align/](http://www.zhanggroup.org/PPS-align/)). In this example, the PPS-score was 0.8368 with a pocket RMSD of 0.3818 and a sequence identity of 91.67%.



**Figure S3.** An example of calculating the protein-ligand interaction fingerprint (PLIF) similarity using the Tanimoto coefficient (Tc). (**A**) The position of FHPMPC after the simulated binding using the *Equus caballus* AR structure is shown (blue) in relation to the position of FHPMPC according to the reference structure (*Homo sapiens*; PDB ID:5VO4; magenta). The major intermolecular interactions determined using the Protein-Ligand Interaction Profiler (PLIP; <https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index>) tool are shown for the reference structure (**B**) and the *E*. *caballus* query structure (**C**). After determining the Van der Waals (vdw) contacts, all the intermolecular interactions for each protein-ligand model are converted to bit vectors (**D**) where “1” represents the presence of an interaction and “0” represents the absence of that interaction. The similarity between these bit vectors is calculated using the Tc. In this example, the Tc = 0.773.

Chart, line chart

Description automatically generated

**Figure S4.** Elbow plots for determining the ideal value of *k* for the *k*-means clustering analyses are shown for the ensemble-docking results for DHT (**A**) and FHPMPC (**B**). The within-clusters sum of squared error (SSE) is plotted against various values of *k*. The red dots represent the point at which the SSE values begin to level-out (*k* = 3 for DHT and FHPMPC), which could be considered the ideal value for *k*. Silhouette plots were created for the same purpose for (**C**) DHT and (**D**) FHPMPC. In this case, a silhouette score, which is a measure of how distinct clusters are from one another, is plotted against various values of *k*. Again, the red dots represent the ideal value of *k* (i.e., the number of clusters that results in the lowest degree of cluster overlap), which was *k* = 2 for DHT and *k* = 4 for FHPMPC. Lastly, plots for determining the ideal value of *k* for the *k*-nearest neighbors (kNN) classification models are shown for (**E**) DHT and (**F**) FHPMPC. The 5-fold cross-validation accuracy is plotted against various values of *k*. The value of *k* that resulted in the highest measure of accuracy in the cross-validation test was selected for the final kNN model and represented by a red dot in the figures (*k* = 36 for DHT and *k* = 2 for FHPMPC).

**Table 1** Summary of species susceptibility calls based on three data sources for 5α-dihydrotestosterone (DHT): in vitro binding/activation assays, Level 4 evaluations of the Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool, and the cross-species molecular docking method. Note: \* = A different androgen receptor (AR) isoform was tested for this species. \*\* = The predicted AR structures did not pass SeqAPASS Level 4 quality filters for these species.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **seqapass susceptibility call** | NA\*\* | NA\*\* | Yes | Yes | Yes | Yes | Yes | Yes | Yes | NA\*\* | NA\*\* | Yes | Yes | Yes |
| **docking susceptibility call** | Yes | Yes | Yes | No | No | No | No | Yes | Yes | NA\* | NA\* | Yes | No | No |
| **in vitro active?** | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes | Yes |
| **variability type** | range | - | - | - | stderror | stderror | stderror | range | - | - | - | - | stdev | stdev |
| **variability** | 0.56-4.7 | - | - | - | 1.55 | 0.04 | 2.22 | 0.065-0.74 | - | - | - | - | 6.1 | 1.5 |
| **value** | 1.6 | 521 | 100 | 1 | 1.7 | 0.1 | 40 | 0.2 | 274 | 0.6 | 5 | - | 14 | 4.3 |
| **units** | nM | % of postive control | nM | nM | nM | nM | max fold induction | nM | % of postive control | nM | nM | - | nM | nM |
| **metric** | EC50 | Relative potency | Fold induction | EC50 | Ki | EC50 | Max fold induction | EC50 | Relative potency | IC50 | EC50 | - | EC50 | EC50 |
| **species** | *A. mississippiensis* | *A. mississippiensis* | *A. japonica* | *D. rerio* | *D. rerio* | *D. rerio* | *D. rerio* | *G. gallus* | *G. gallus* | *O. mykiss* | *O. mykiss* | *P. nelsoni* | *X. laevis* | *X. laevis* |