

Recent developments in the eHiTS ligand docking and scoring software



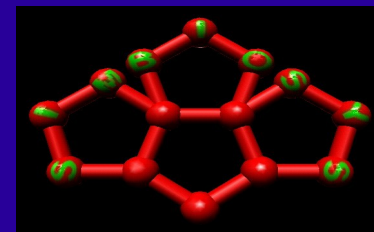
Zsolt Zsoldos
SimBioSys Inc.



<http://www.simbiosys.ca/>

Contents:

- Hardware acceleration technology on the Cell/BE platform
- Interacting surface point based statistical scoring function
- Automated score tuning mechanism for protein families
- Protonation state optimization on the fly
- Inclusion of ligand based activity estimation into the score
- Pose accuracy benchmark results on the Astex set
- Virtual screening performance on the DUD set

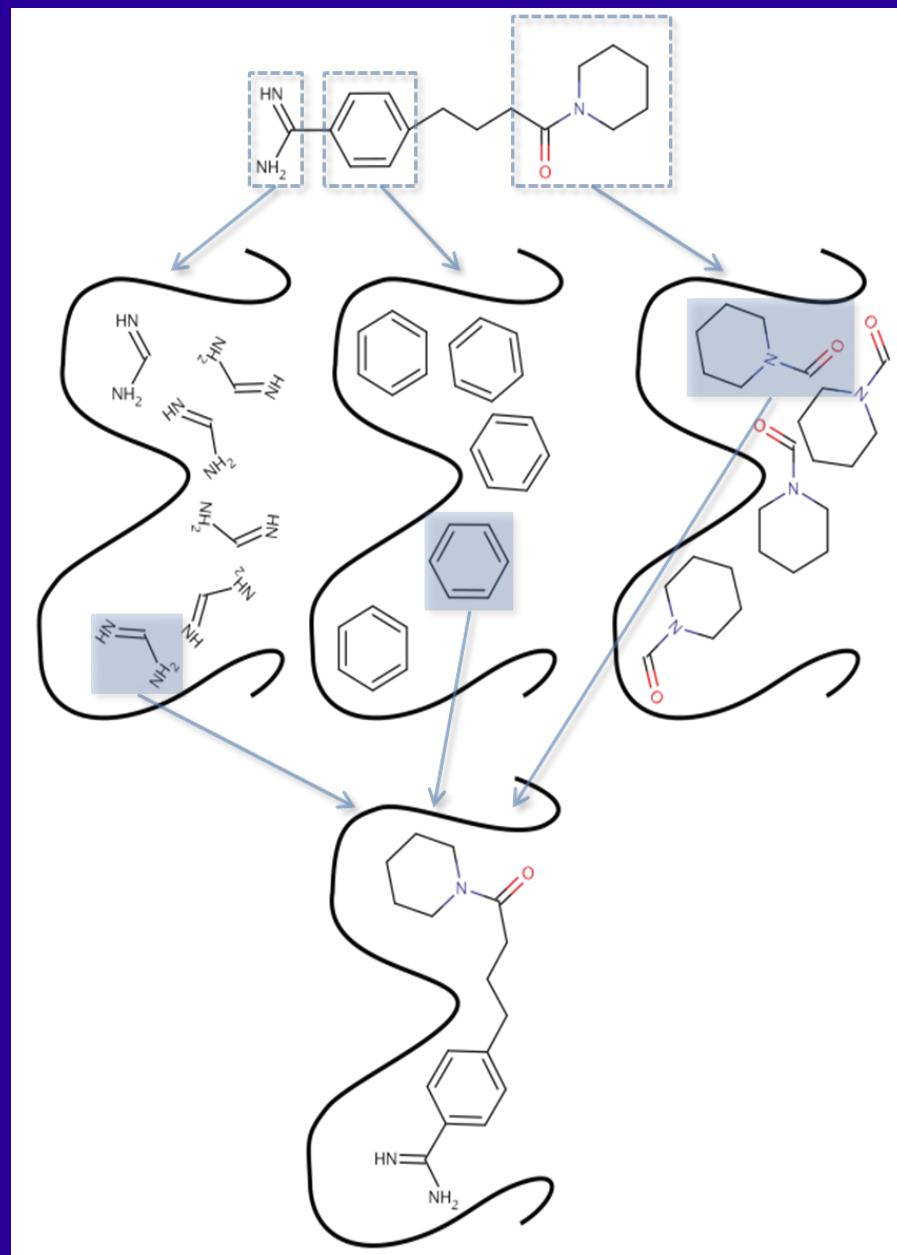


2. Overview of eHiTS

- Ligand is divided into rigid fragments, flexible chains
- All rigid fragments are docked **independently** (many poses)
- Pose matching (clique detection)
- Flexible chain fitting (continuous)
- Local energy minimization

J.MGM (26) #1, July 2007, pp 198-212

doi: 10.1007/s10822-007-9164-5



3. Speed-up factors on the Cell/BE architecture vs Intel CPU

Code component (task)	Speed-up factor	
	PS3	QS21
eHiTS Scoring (with rotomer opt.)	53x	125x
Rigid Fragment Docking	31x	76x
Pose Matching	7x	18x
Conformation Minimization	34x	45x
Final Optimization	12x	33x
Total (complete flexible docking)	5-56x	8-117x

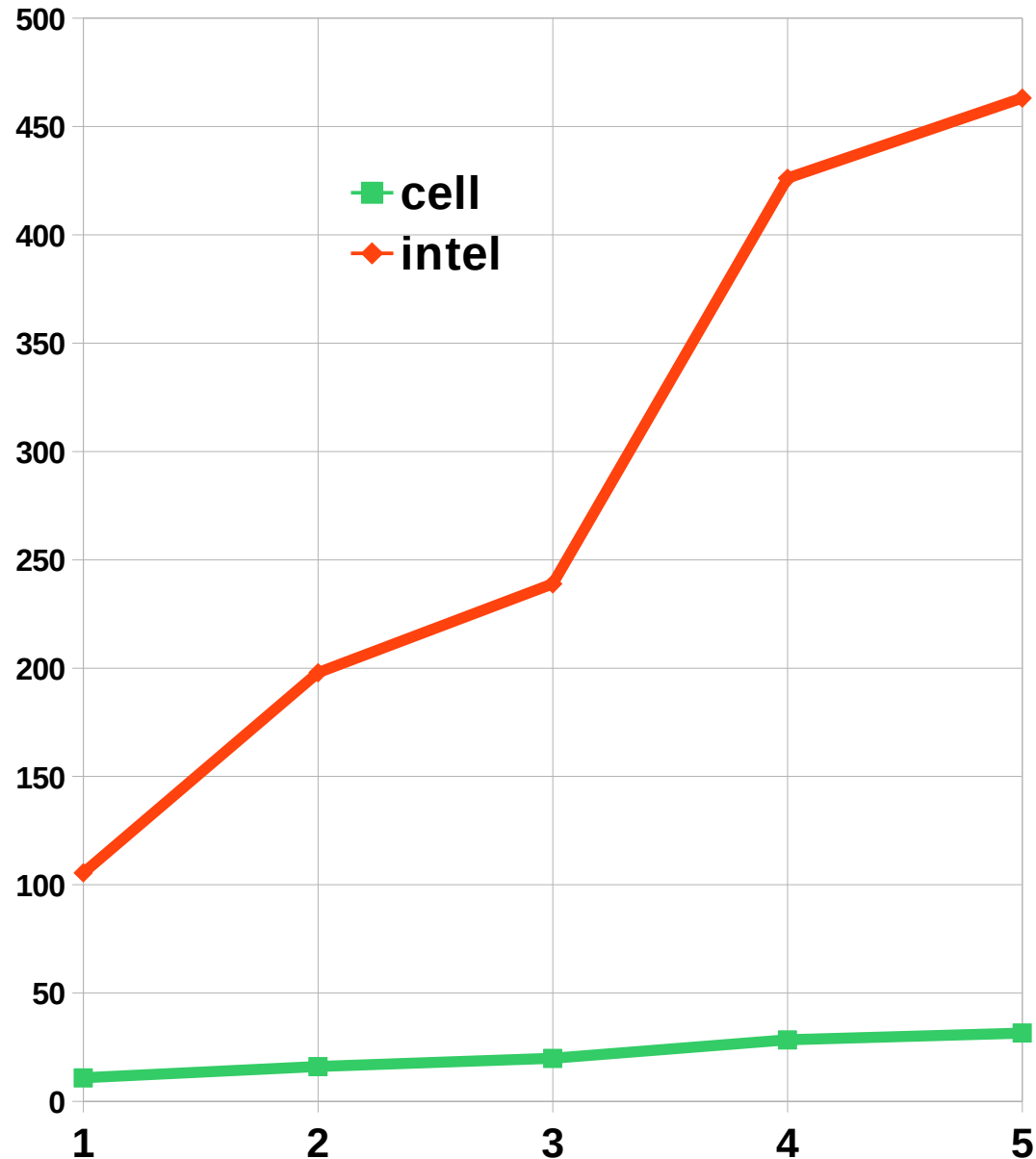


4. Speed-up dependence on accuracy parameter

Graph shows the run time in seconds for accuracy levels 1-5

Values are averaged over test cases with various complexity

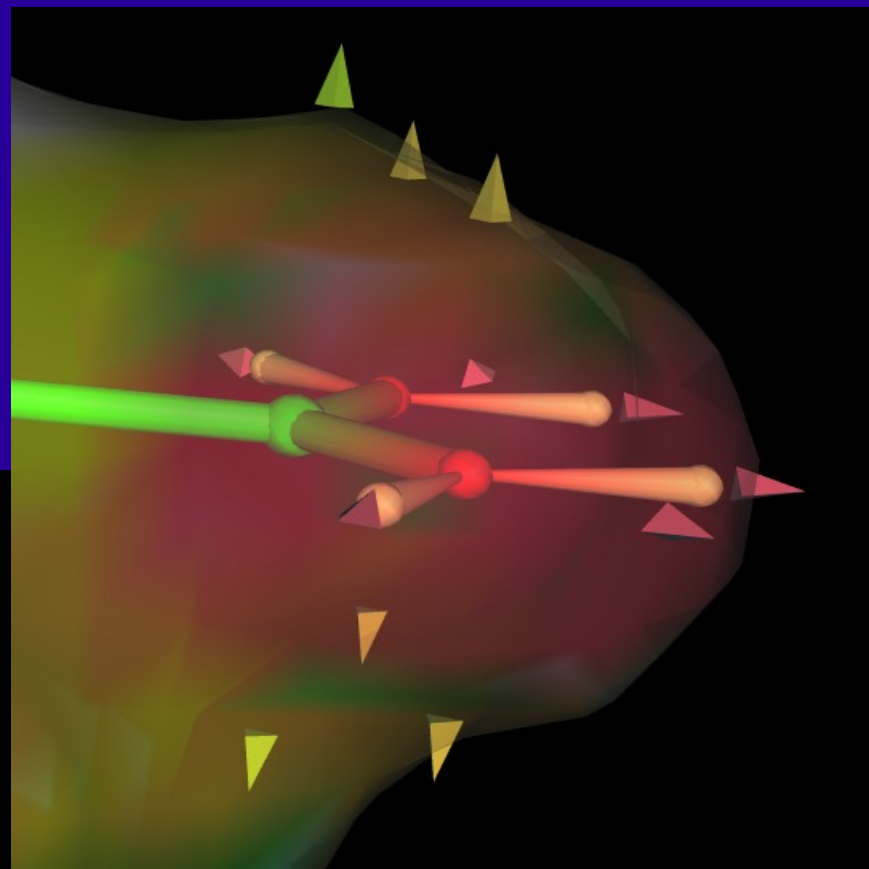
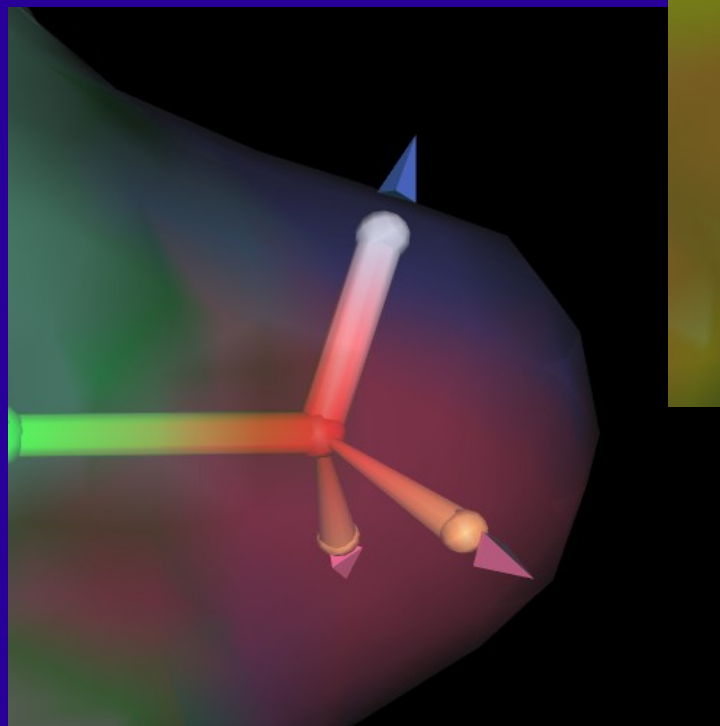
The Cell/BE speed-up increases with the accuracy



5. Interaction Surface Points (ISP)

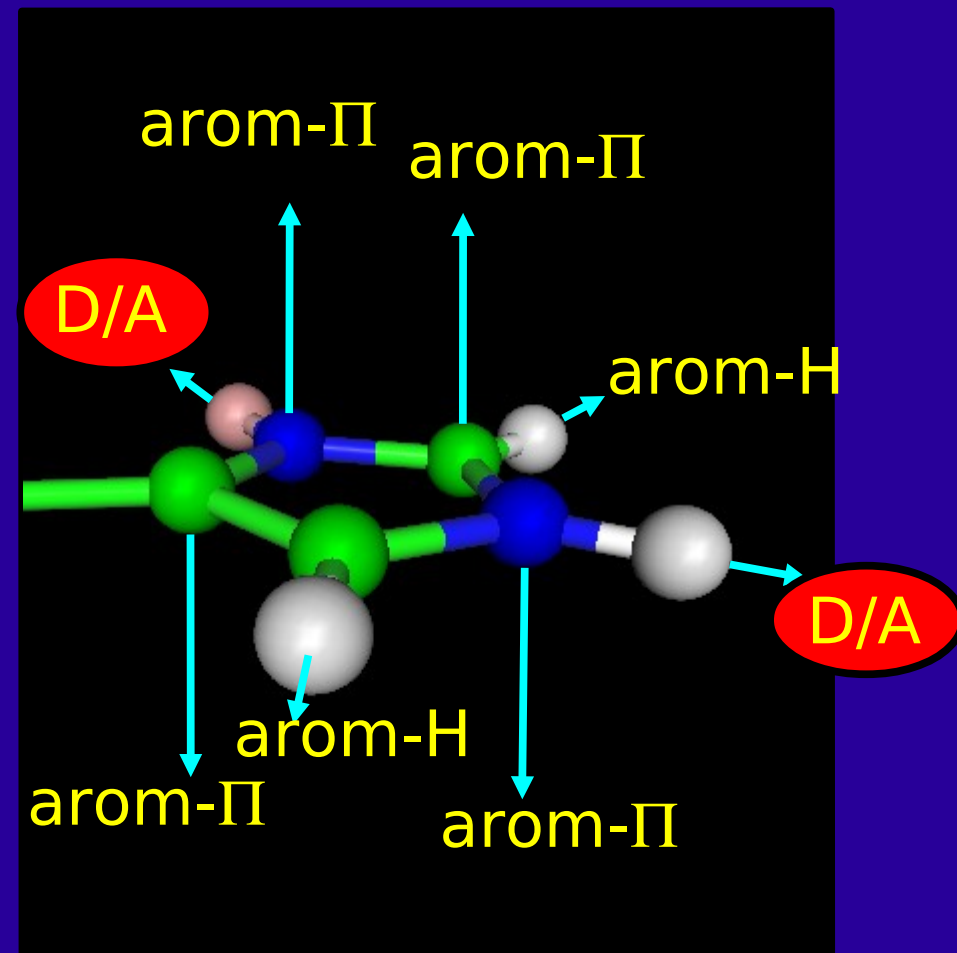
- eHiTS places directional surface points in specific locations on the surface of molecules to represent various interaction capabilities:

- H atoms,
- lone electron pairs,
- π electrons

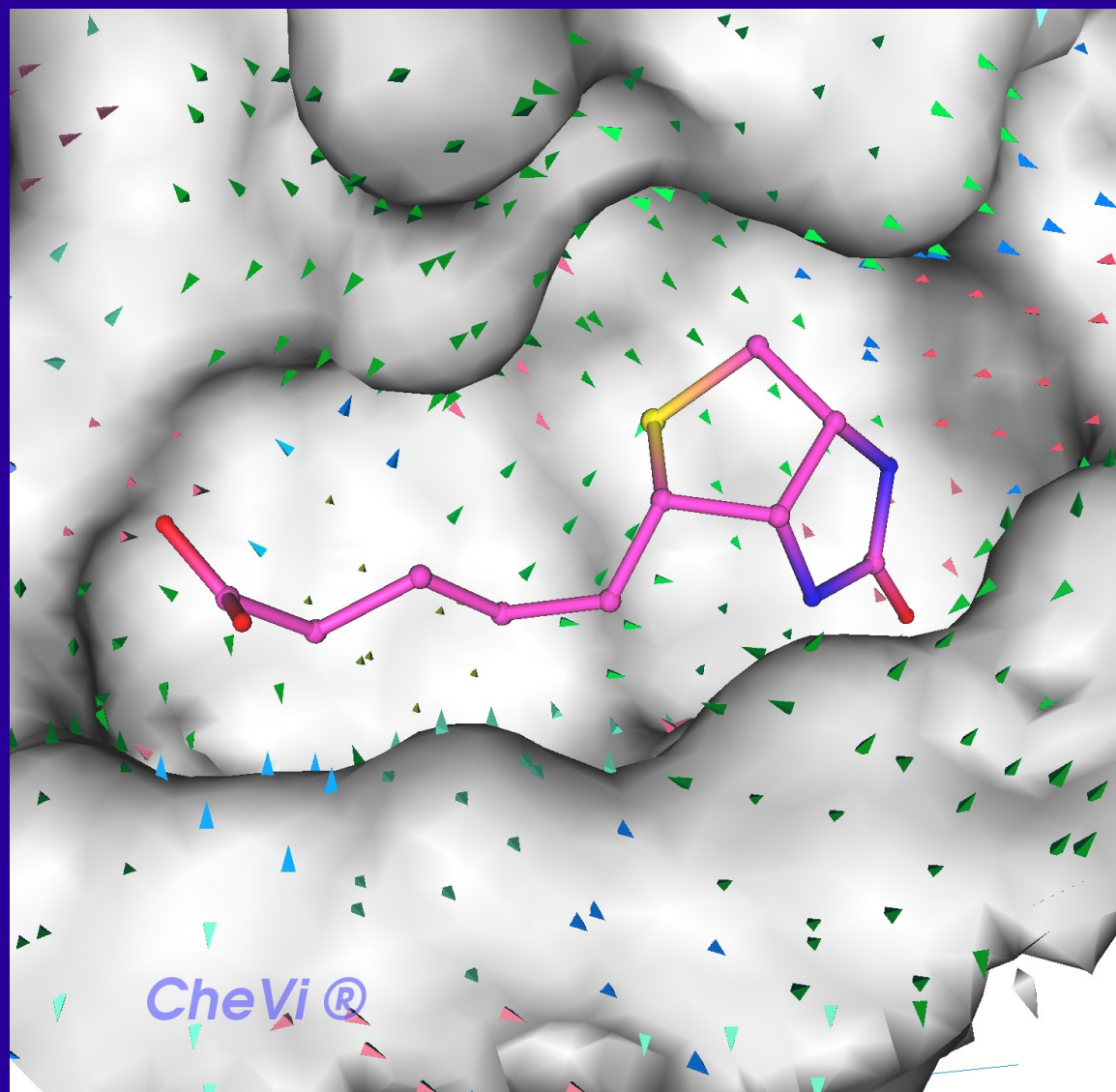
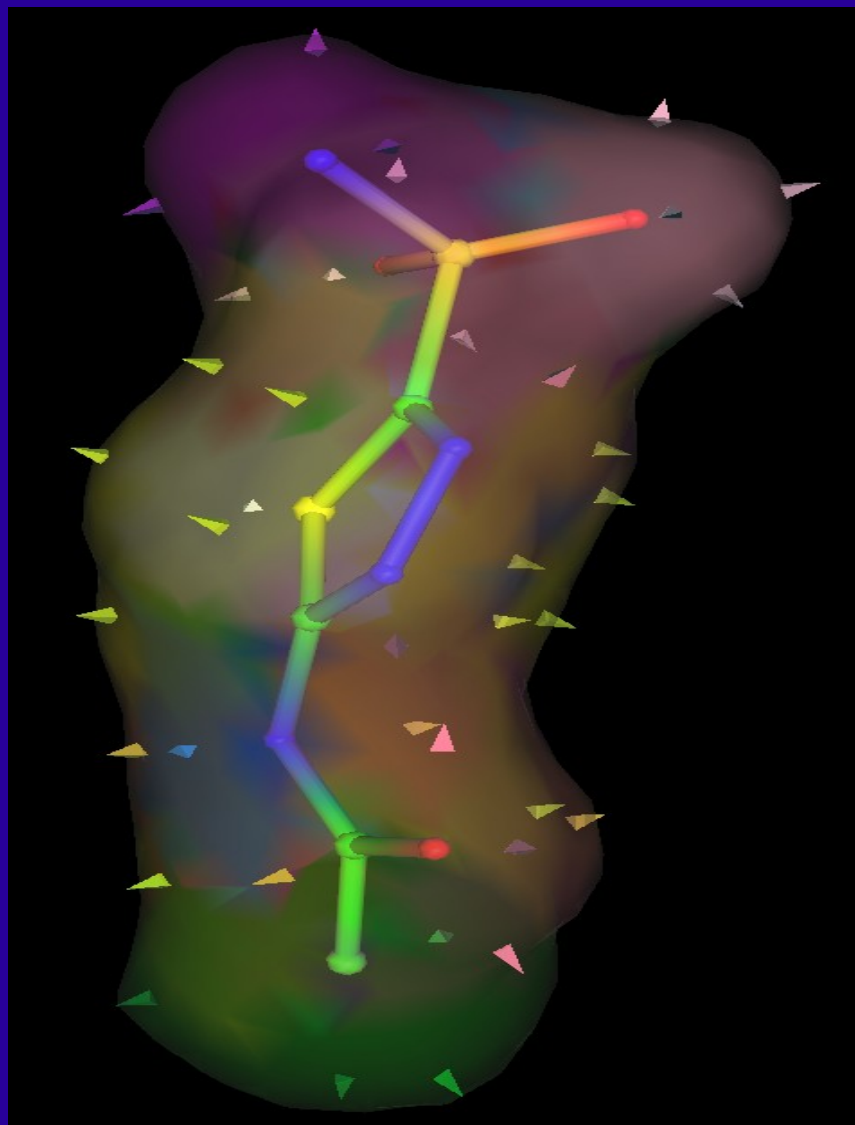


6. Interaction surface point (ISP) types

- **H-bond Donors:**
 - charged, amine, strong, weak, rotatable
- **H-bond Acceptors:**
 - charged, acid, strong, weak, rotatable
- **Ambivalent H donor/acceptor**
- **Aromatic Pi-stacking:**
 - carbon, polar, resonance, edge-H, arom- π
- **Hydrophobic:**
 - strong / weak lipophil, neutral
- **Metal ions**
- **Misc (Sulfur, Halogens)**

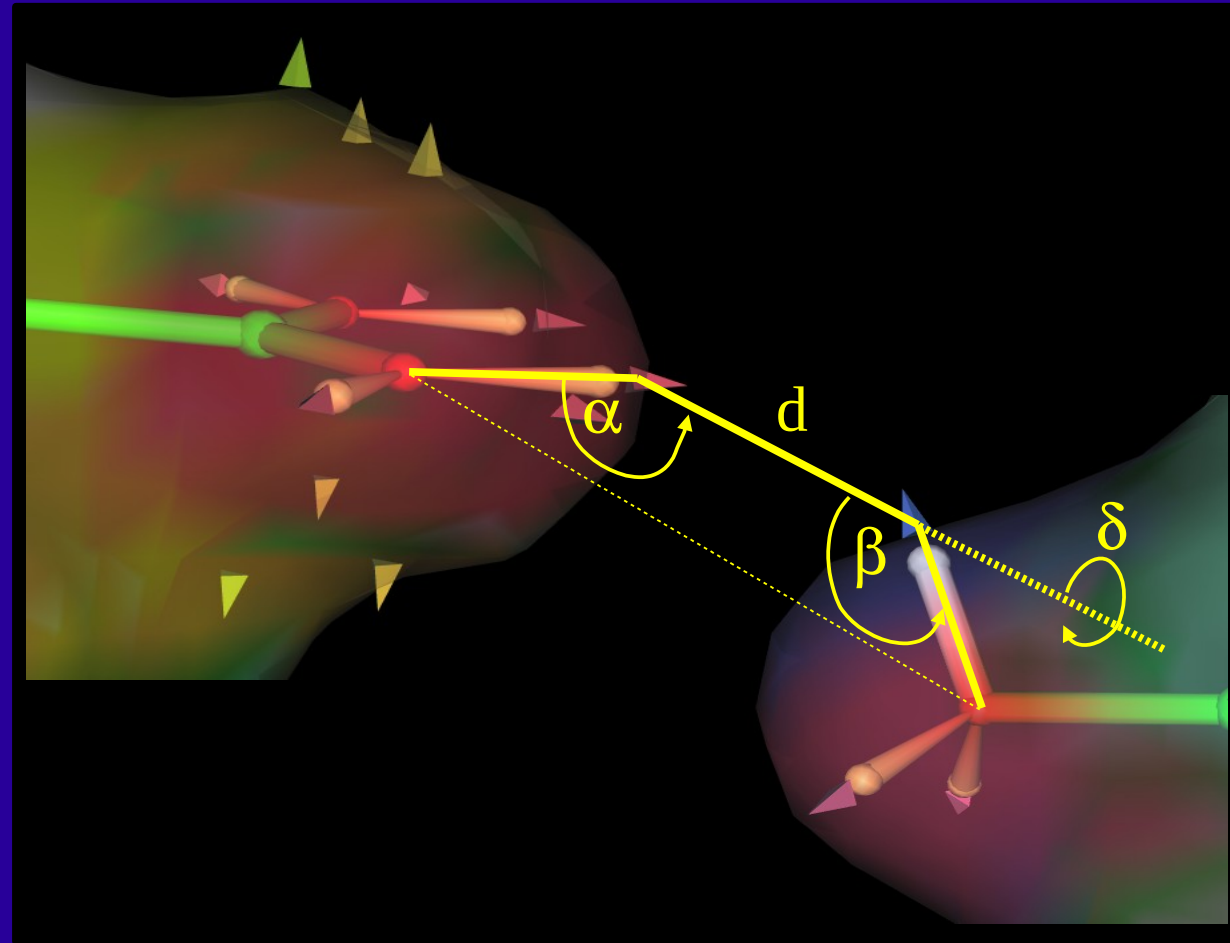


7. ISP set examples



8. Interaction Geometry scoring

- Interactions can not be described by distance (d) alone, the angles between ISP directions and interaction directions (α, β) as well as the torsions (δ) between them must be considered:



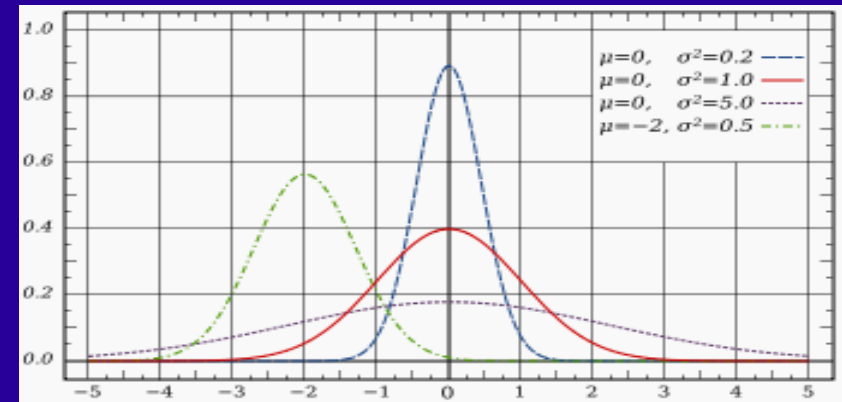
9. PDB file filtering and curation

1. Protein-ligand complexes from the PDB, xray resolution 2.5Å or better: ~21,000
2. The PDB-report created by the WHAT_CHECK software was used for filtering:
Errors in protein structures. R.W.W. Hooft, G. Vriend, C. Sander, E.E. Abola, Nature (1996) 381, 272-272
 - Major modeling errors
 - High bond length or bond angle deviations
 - Ramachandran Z-score very low
 - chi-1/chi-2 angle correlation Z-score very low
 - Abnormal packing environment or Z-score
 - Backbone conformation Z-score very low
 - Side chain planarity problems
 - C/N-terminal problems
 - Unusual residues or torsional angles
 - Connections to aromatic rings out of plane
 - Abnormal packing for sequential residues
 - Low packing Z-score for some residues
5. HIS, ASN, GLN side chain flips are detected (H-bonding) and corrected
6. Duplicate, unexpected atoms and water clusters without H-bonding are omitted
7. The Uppsala Electron-Density Server was used to detect and filter local errors
GJ Kleywegt, MR Harris, JY Zou, TC Taylor, A Wählby & TA Jones (2004), Acta Cryst. D60, 2240-2249
3. Structures with major errors or too many residue errors are omitted: ~12,000 left
4. Residues with significant errors (RSCC<0.85, RSR>0.2, OWAB>40) are omitted

10. Statistical data collection

- ~12000 high resolution (<2.5Å) crystal structures – millions of inter.
- Probability of atom being at distance d (Gaussian distribution):

$$p(d) = \left(\frac{B}{4\pi}\right)^{-3/2} \int_0^\pi \int_0^{2\pi} \exp\left(\frac{-4\pi^2 r_{\alpha\beta}^2}{B}\right) d^2 \sin(\alpha) d\alpha d\beta$$



- Probability of distance d to occur between two heavy atoms:

$$P(d) = \left(\frac{4\pi}{B_0 + B_1}\right)^{\frac{3}{2}} d^2 \int_0^\pi \int_0^{2\pi} \exp\left(\frac{-4\pi^2}{B_0 + B_1} \|P_0 - P_1 + P_s\|^2\right) \sin \alpha d\alpha d\beta$$

- Similar formulae for angle and torsional components
- 4D data collection using fine numerical integral sampling

11. Additional scoring terms

- De-solvation: continuous model, ISP type dependent
- Steric clash penalty: distance-square from Connolly surface
- Pocket depth: signed distance of atoms from convex hull
- Protein family data based coverage (ISP type pairs)
- Ligand strain energy (torsional probability + vdw LJ 6-12)
- Ligand intra-molecular interaction score (ISP pair ~ receptor)
- Ligand entropy loss (frozen rotatable bonds)

12. Protein “family” clustering

~12,000 PDB Complexes are clustered automatically into ~500 protein sets.

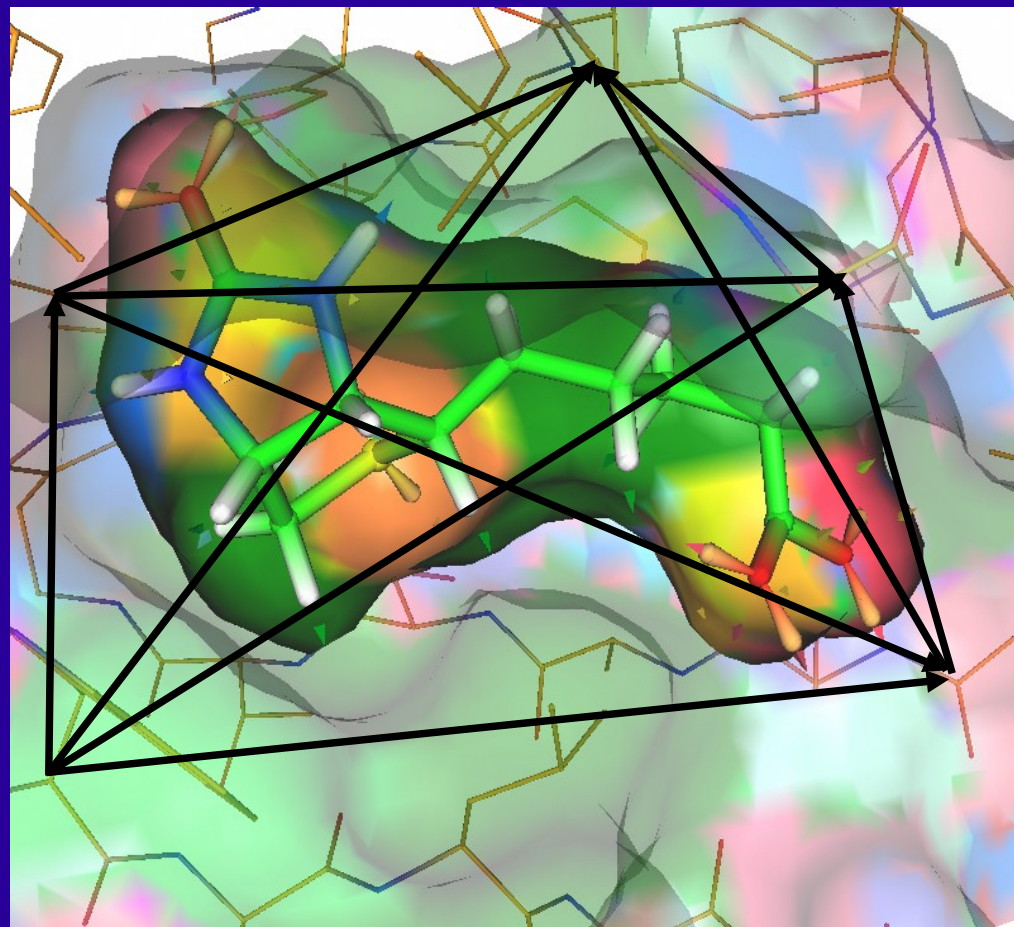
Geometric clustering is based on binding site residue C α distance matrix.

- distance tolerance (default 3Å)
- matching subset size minimum (5)
- minimum set-size (5 entries)

Correspondence to biological activity family is not exact, e.g. Kinase DFG-in DFG-out is separate, but thrombin and trypsin in same set.

Under represented sets and singletons are treated as a fall-back general set

The same matching criteria is used to find the “family” of the target protein in the preprocessing step of a docking run



13. Protein “family” based weight tuning

Docking is performed for all members of a “family” (training PDB set) to generate 300+ poses using default scoring weight parameters

All scoring term values are recorded for each pose along with the RMSD from the x-ray pose

The interaction score-matrix is divided into 5 categories (metal, H-bond, hydrophobic, pi-pi and other): we use 1 weight parameter per category

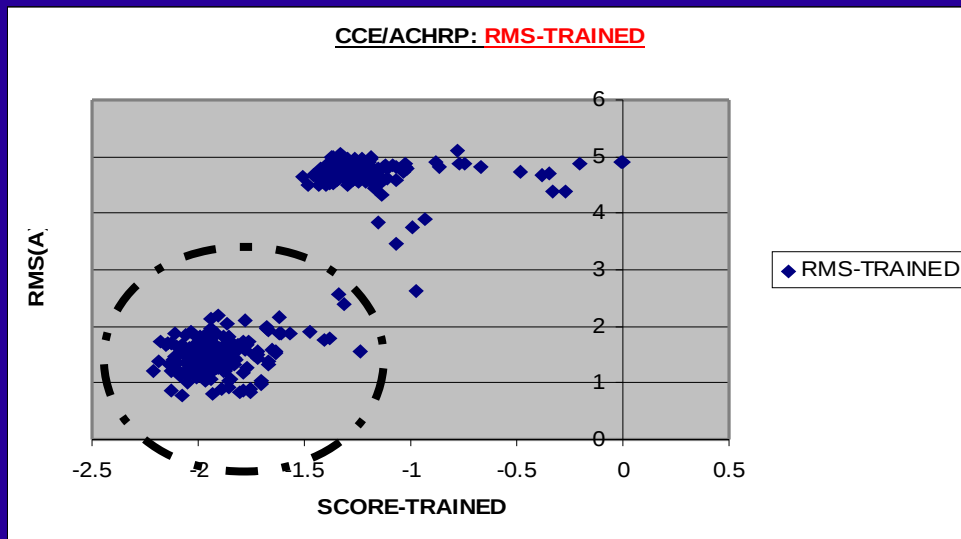
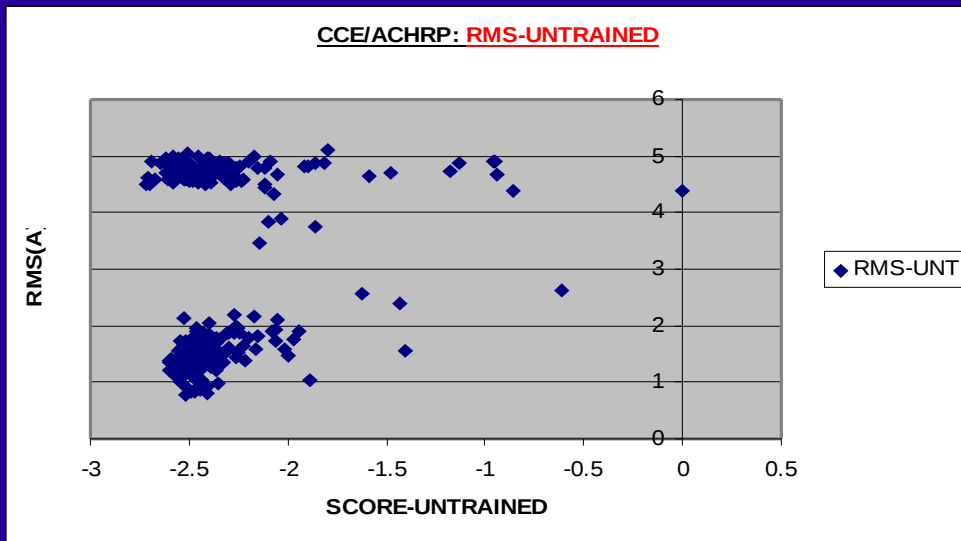
Along with the additional terms, we have 20 weight parameters to tune

The goal function of the weight tuning optimization process includes:

- RMSD of the top-rank pose from each of the complexes in the set
- rank position of the closest pose to the x-ray among all poses
- score difference between the closest pose and the top-rank pose

Tuning does not influence the generated set of poses (rank-order only)

14. Effect of rank tuning

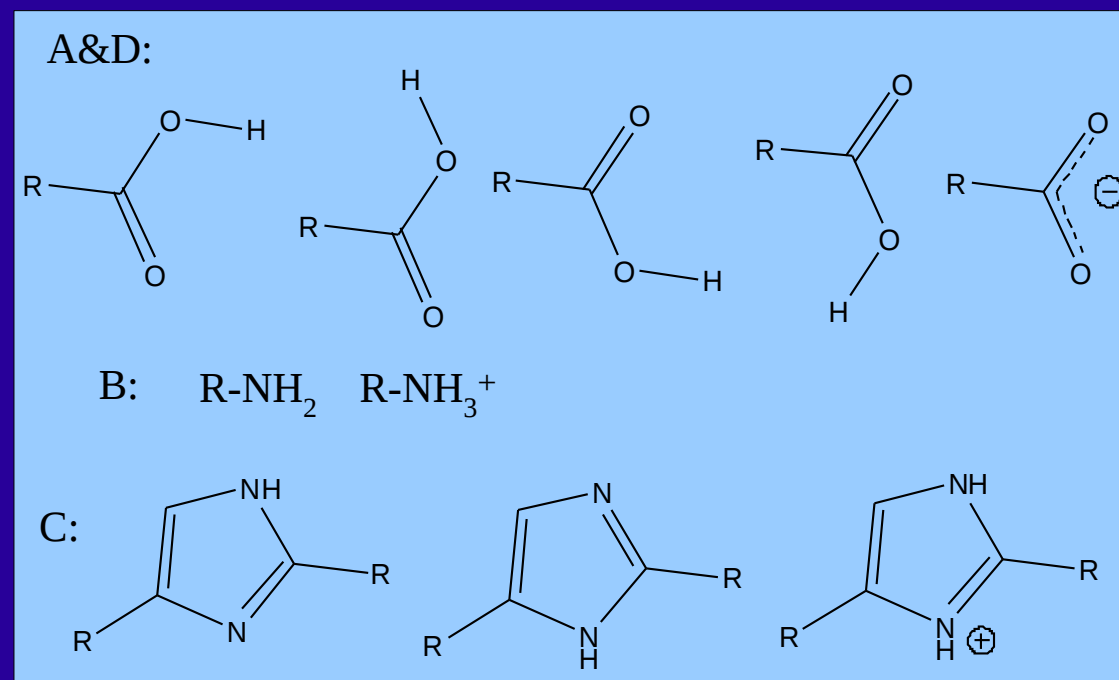
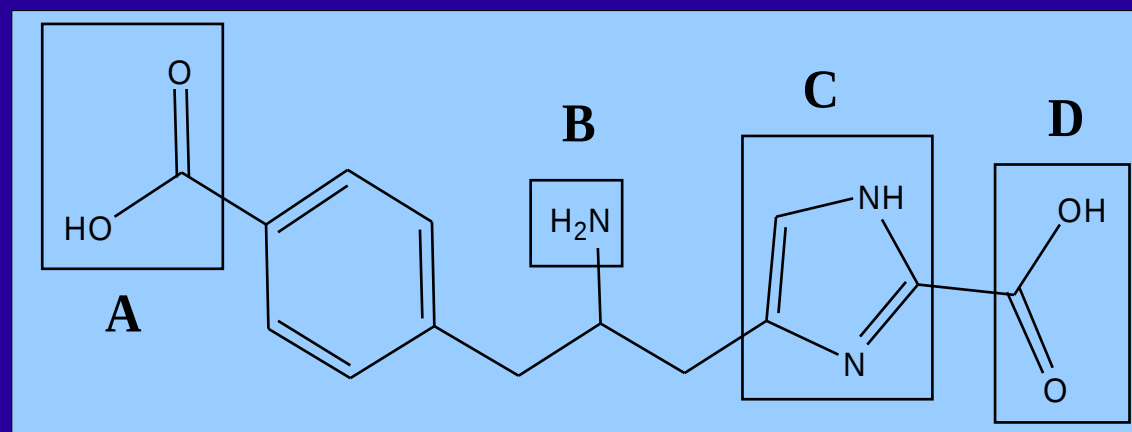


Untrained Scoring: Note that while there are many low RMS solutions in the good score regime there are also high RMS solutions with same score range

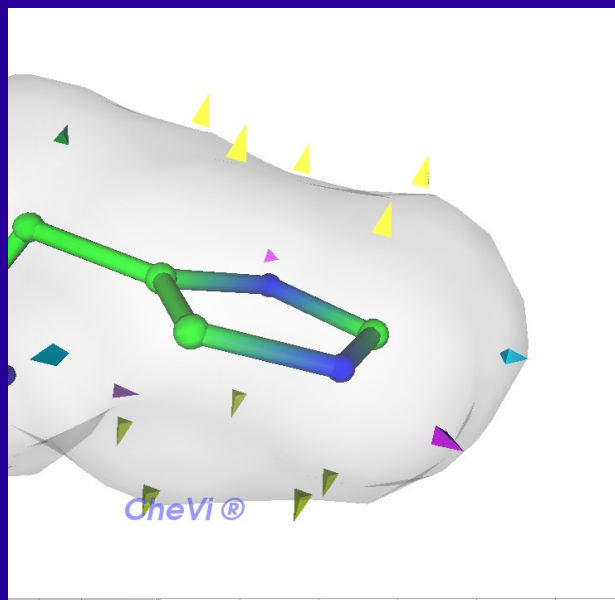
- **Trained Scoring:** There is dramatic score-separation of the 'correct' pose RMS-regime (circled) at low scores from poor scoring –high RMS results

15. Universal protonation handling

- Generic form using alternative flags (H/Lp)
- Scoring picks better one for each atom
- Example:
 - 150 states enumerated
 - 11 independent H/Lp



16. Ligand Surface Descriptor



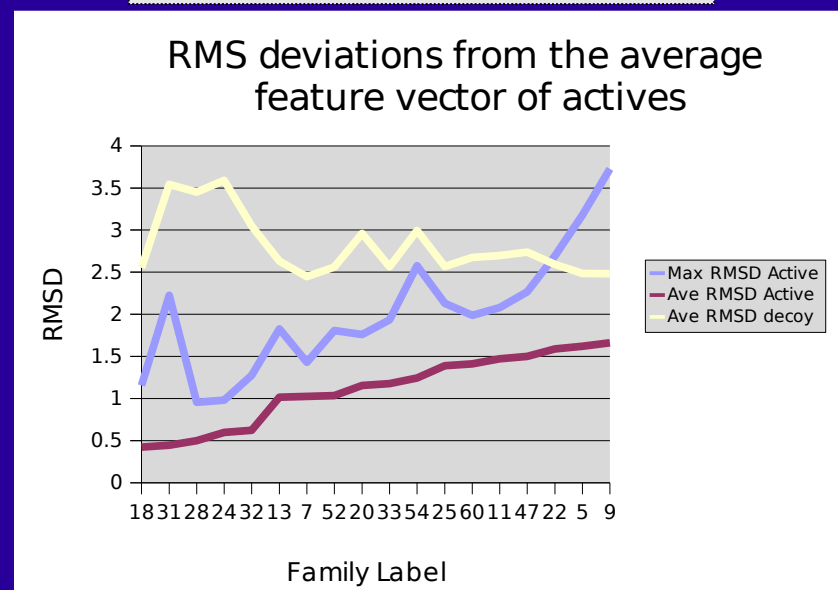
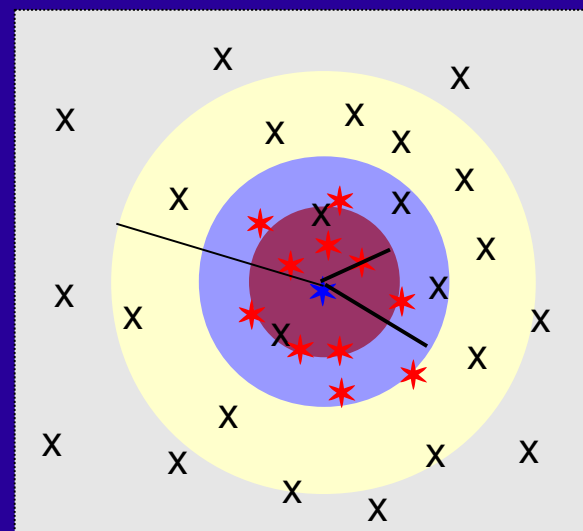
The interaction surface points (ISP), indicated by arrow heads are shown on the histidine ring. The colours correspond to the associated chemical properties.

It is counted how many of each ISP type (ISPT) occurs on the ligand. The feature count vector is the QSAR descriptor.

This descriptor is based on the assumption that ligands with similar feature vectors have similar binding activity.

17. Diversity of Actives and decoys

- For each set of actives, the average feature vectors was calculated (blue star)
- The RMSD from this feature vector was calculated for each active and decoy. The plot below shows the average RMSD for the actives and the decoys, as well as the MAX RMSD for the actives
- For most cases even the max RMSD of the actives is less than the average RMSD of the decoys



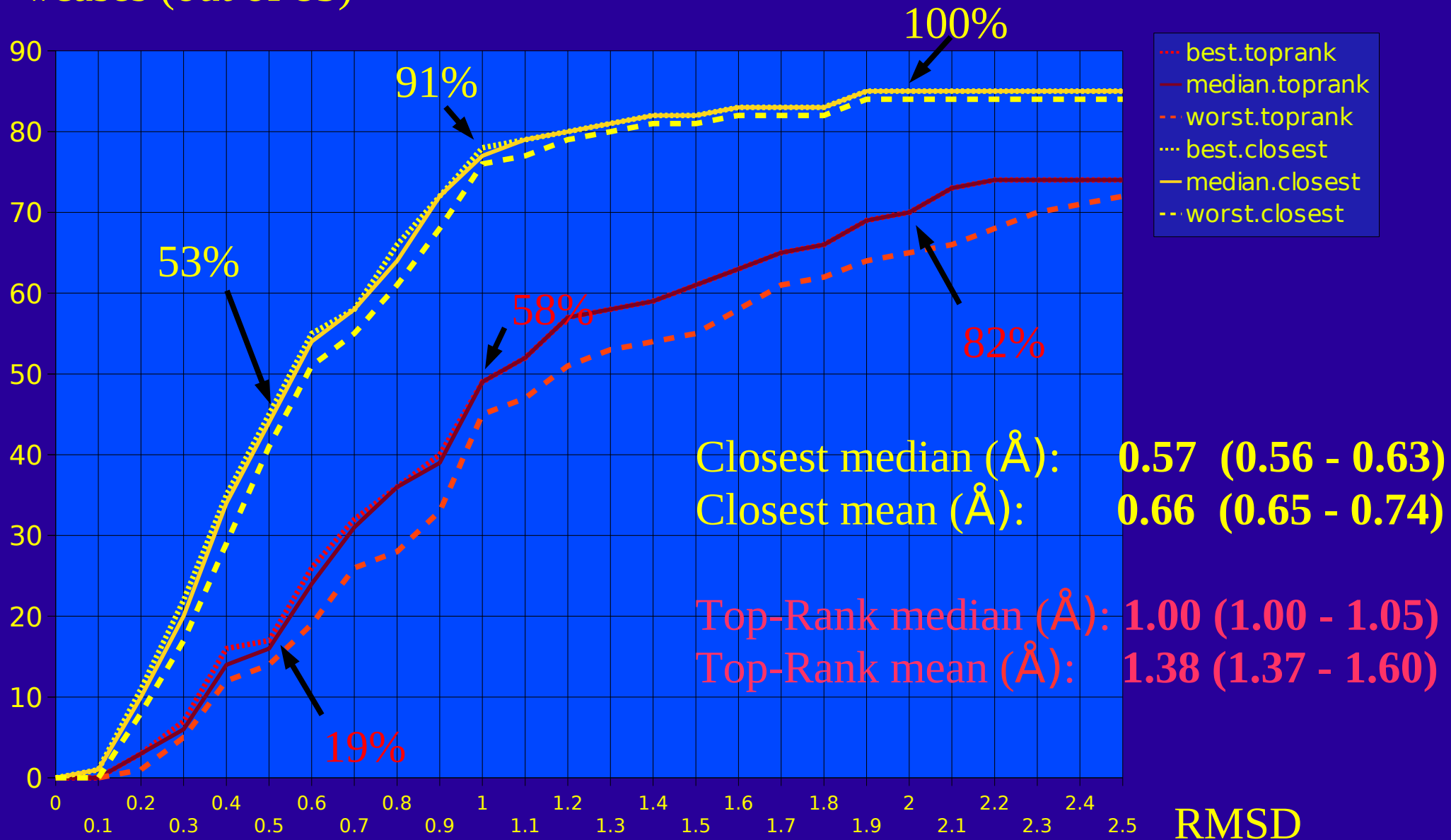
18. Astex benchmark test runs

- Receptor files used as prepared by the symposium organizers *_prot_con.pdb with a small format error correction:
ATOM → HETATM for non-standard residues, e.g. cofactors
- Ligands docked from the given *_start.sdf files without modification
- Multiple docking runs were performed for each pocket identified by the xray positions given in the files *_lig.sdf (-clip and -xray options)
- The 2009.1 (Nov'09) release of eHiTS was used on the Cell/BE platform – the Cell version provides about 5-10% better accuracy than the Intel version due to implementation differences
- Standard, default parameter set and default accuracy (3) was used
- The out-of-box scoring weight sets were used from the release

Overall: no special tweaking of any kind, software used as given in release and data used as given, results reproducible by any user

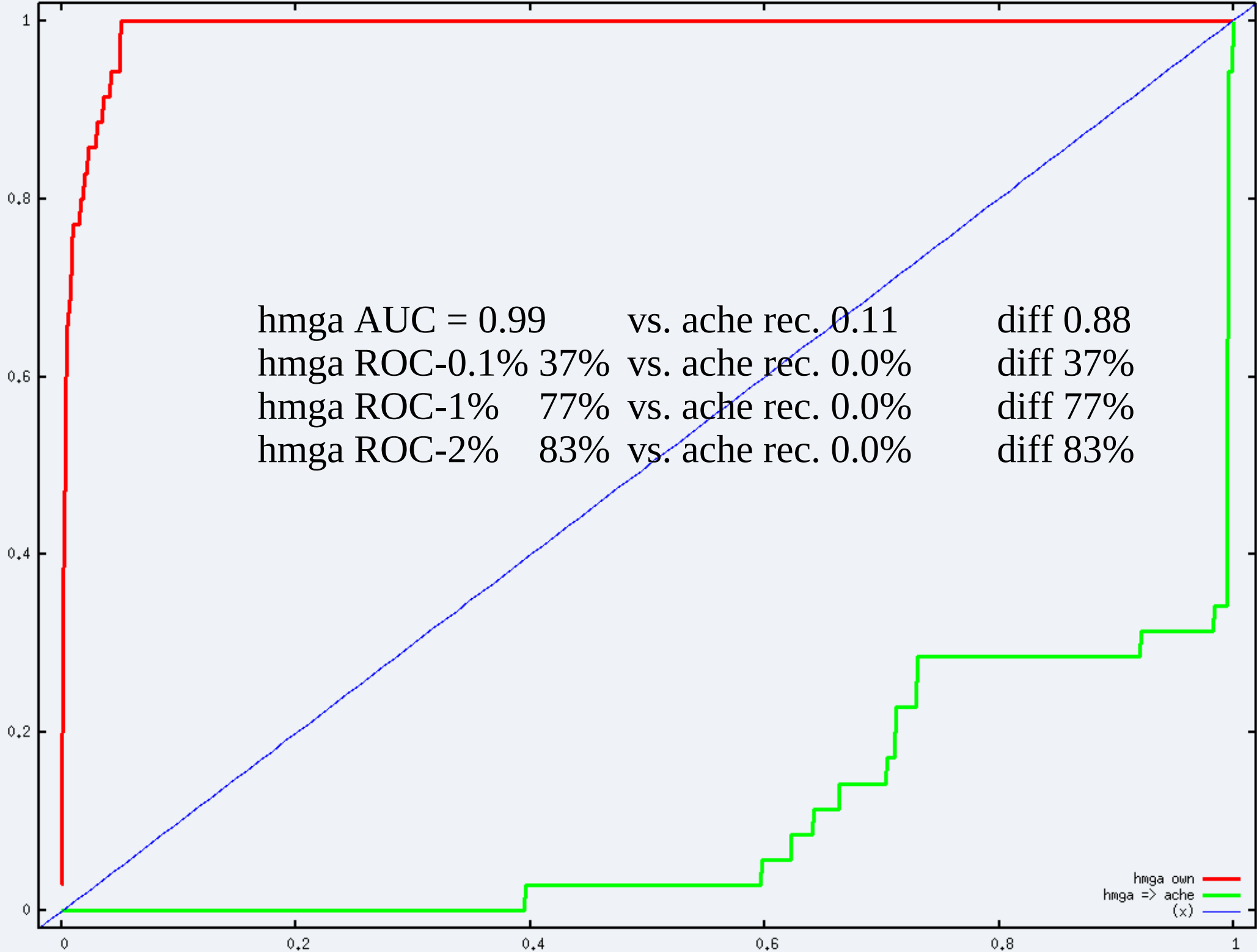
19. Pose accuracy on the Astex 85 set

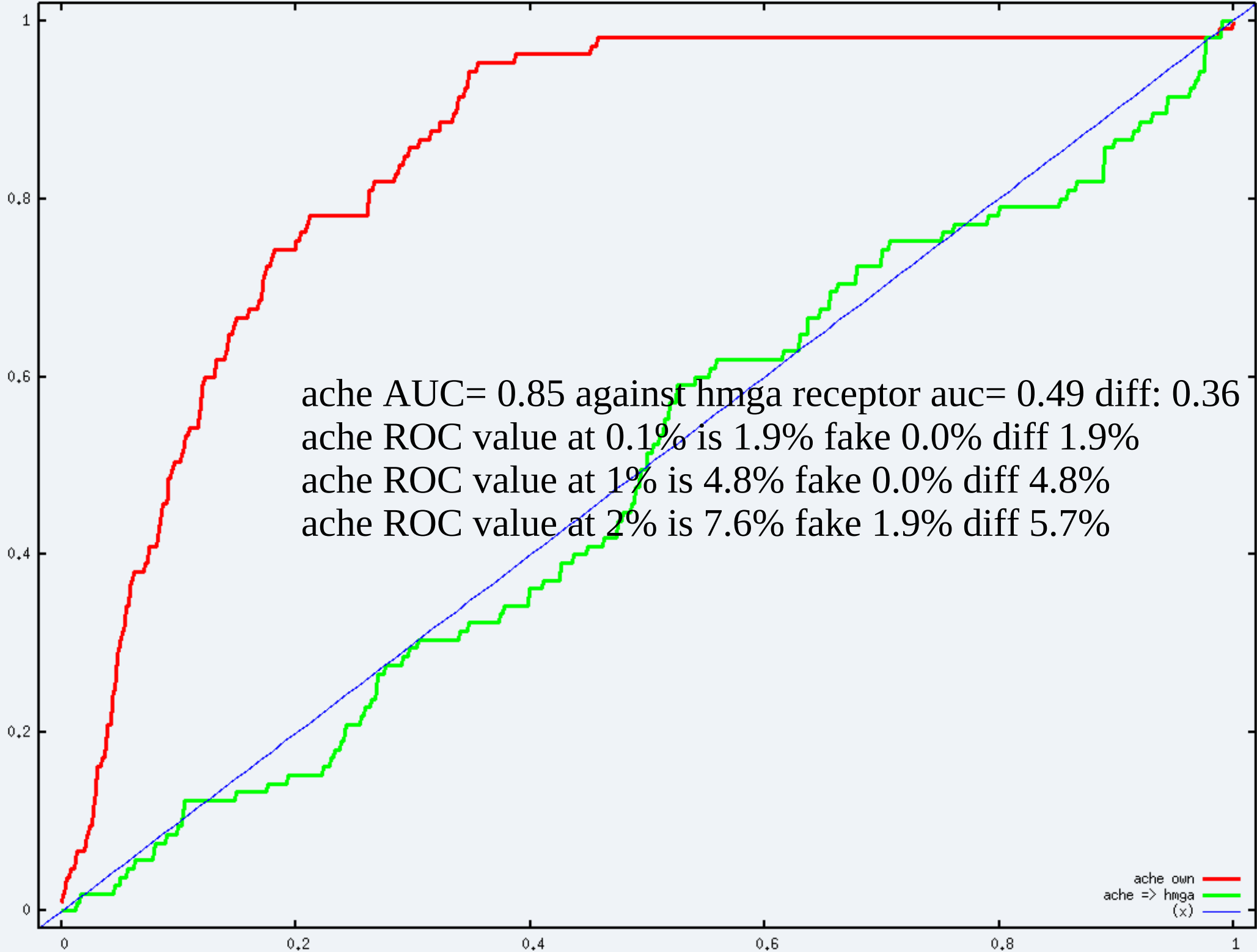
#cases (out of 85)

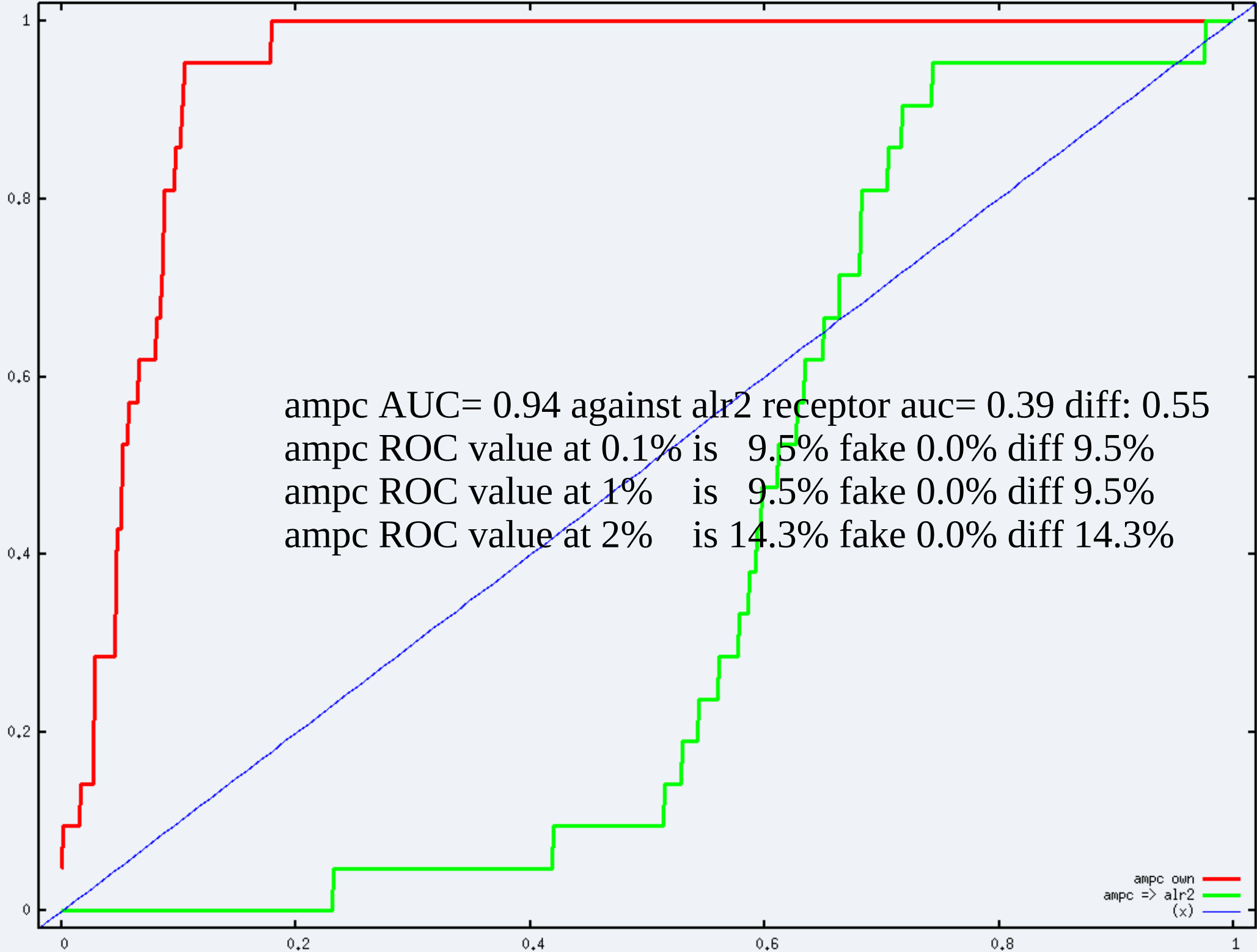


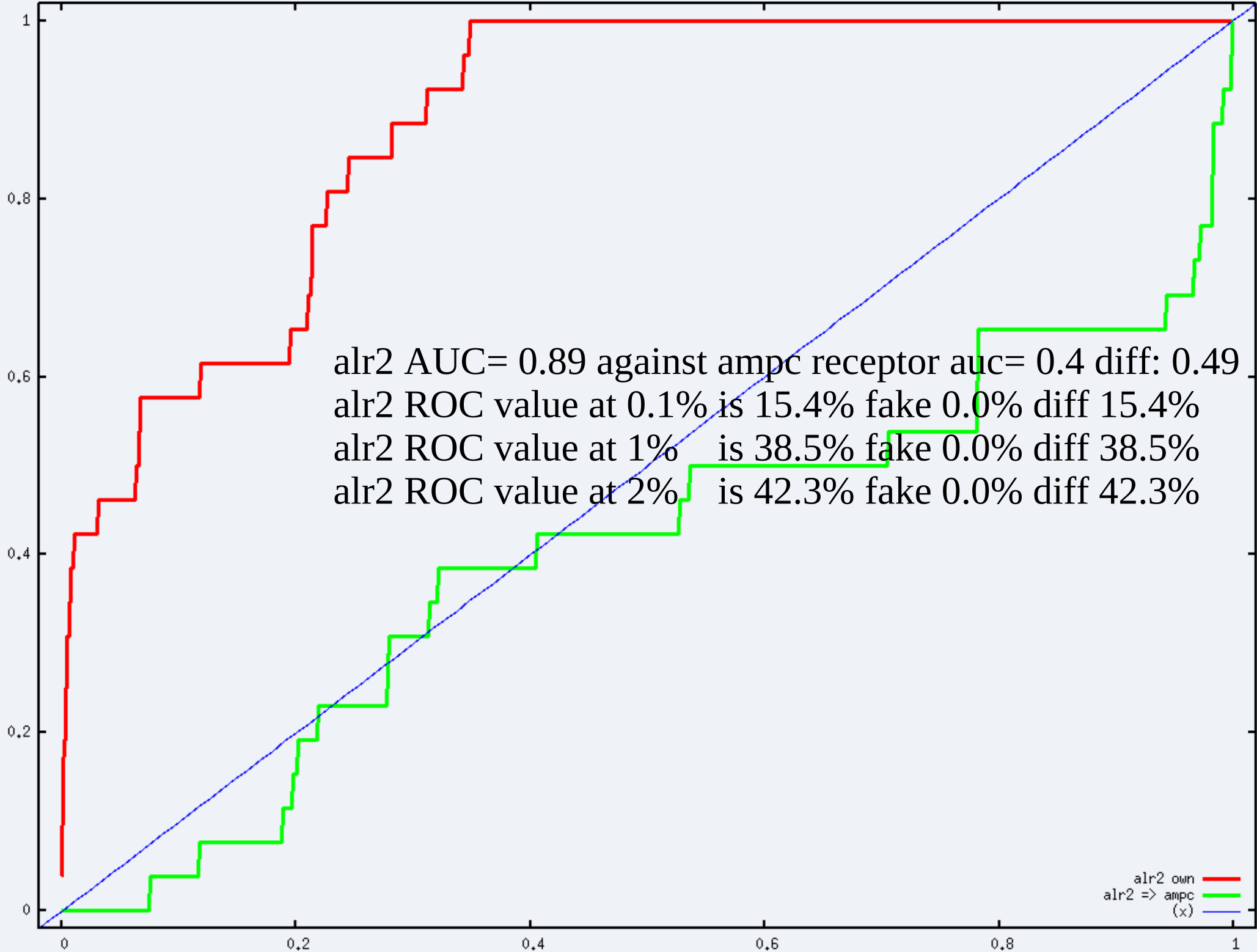
20. Virtual Screening on the DUD set

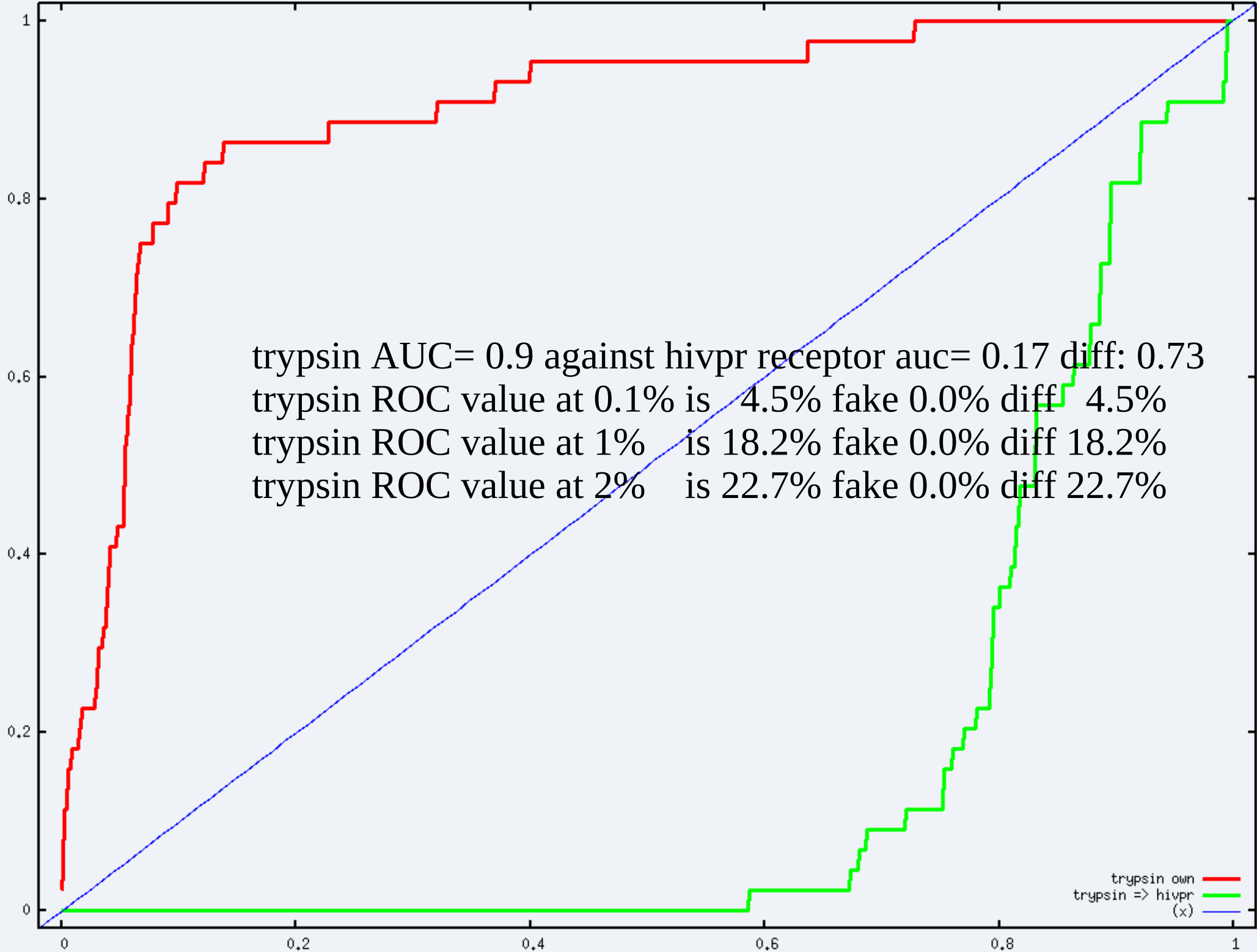
- Receptor files used as prepared by the symposium organizers: *_rec_min.mol2
- Active and decoy Ligands docked from the given sdf files
- The 2009.1 (Nov'09) release of eHiTS was used on the Cell/BE platform
- Standard, default parameter set and default accuracy (3) was used
- The out-of-box scoring weight sets were used from the release
- Cross—target screening “null-hypotesis” test was performed
- *Note: a real null-hypotesis requires the fake target to be very different from the real ones. This is not the case for some of the suggested pairs, e.g. factor Xa vs thrombin*

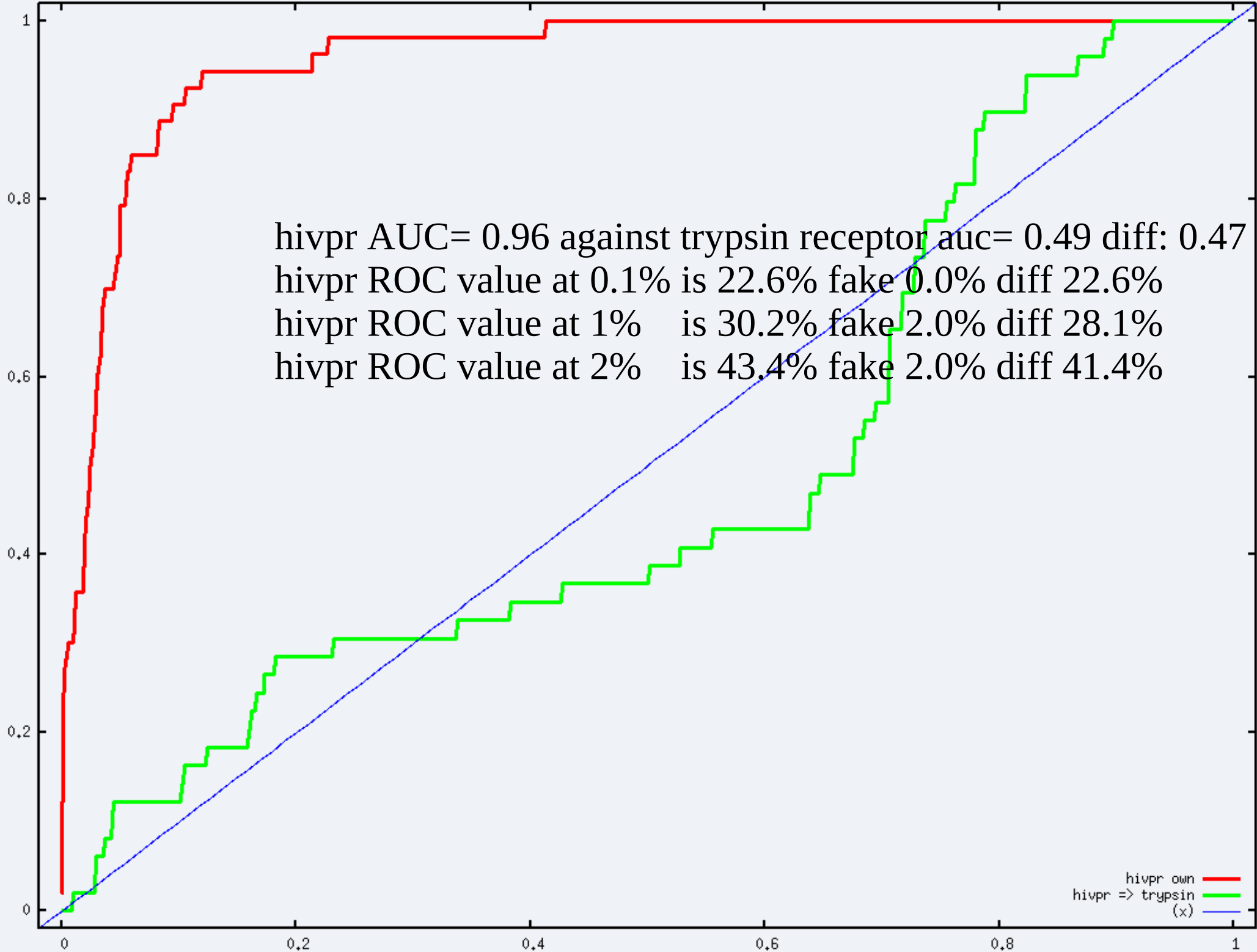


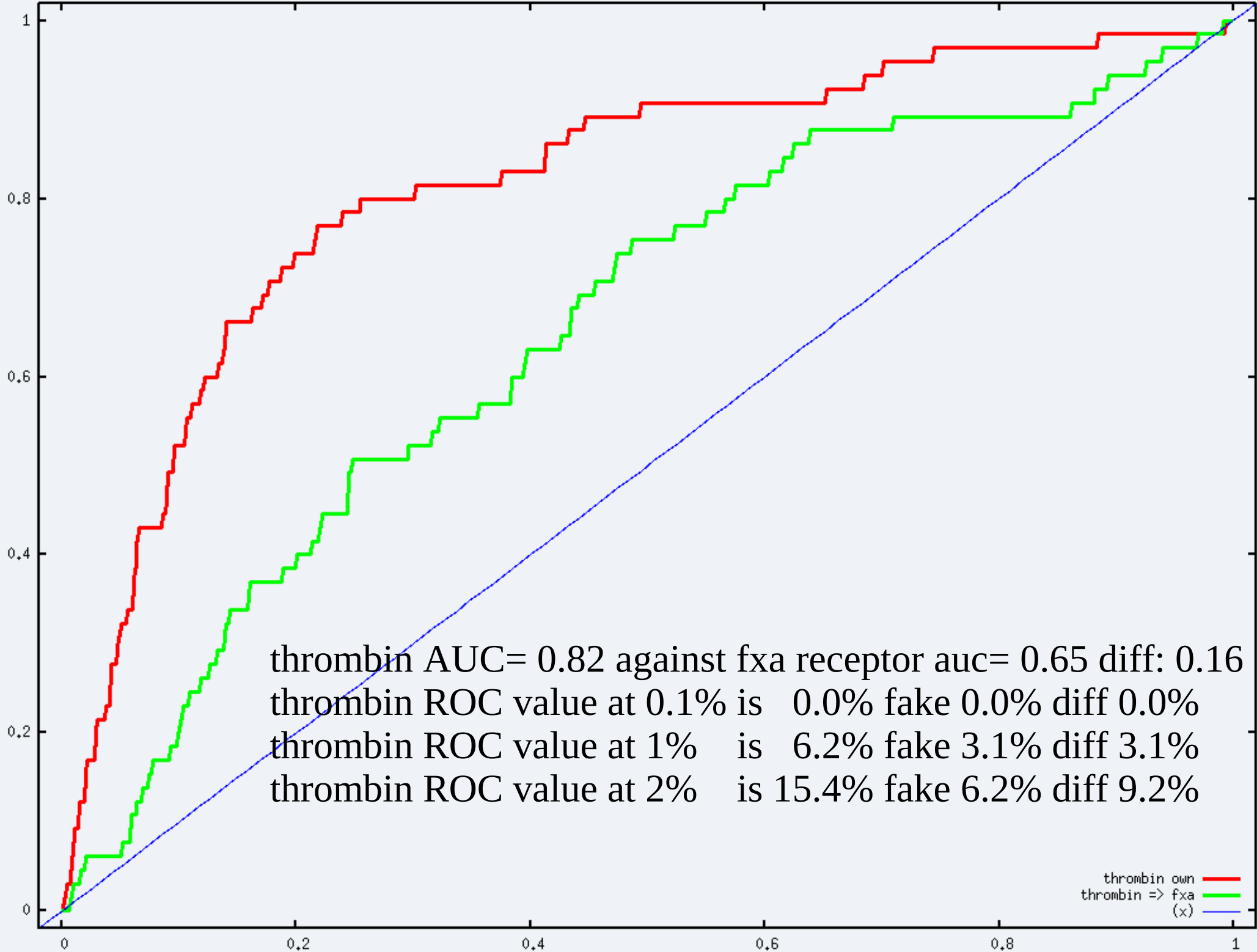


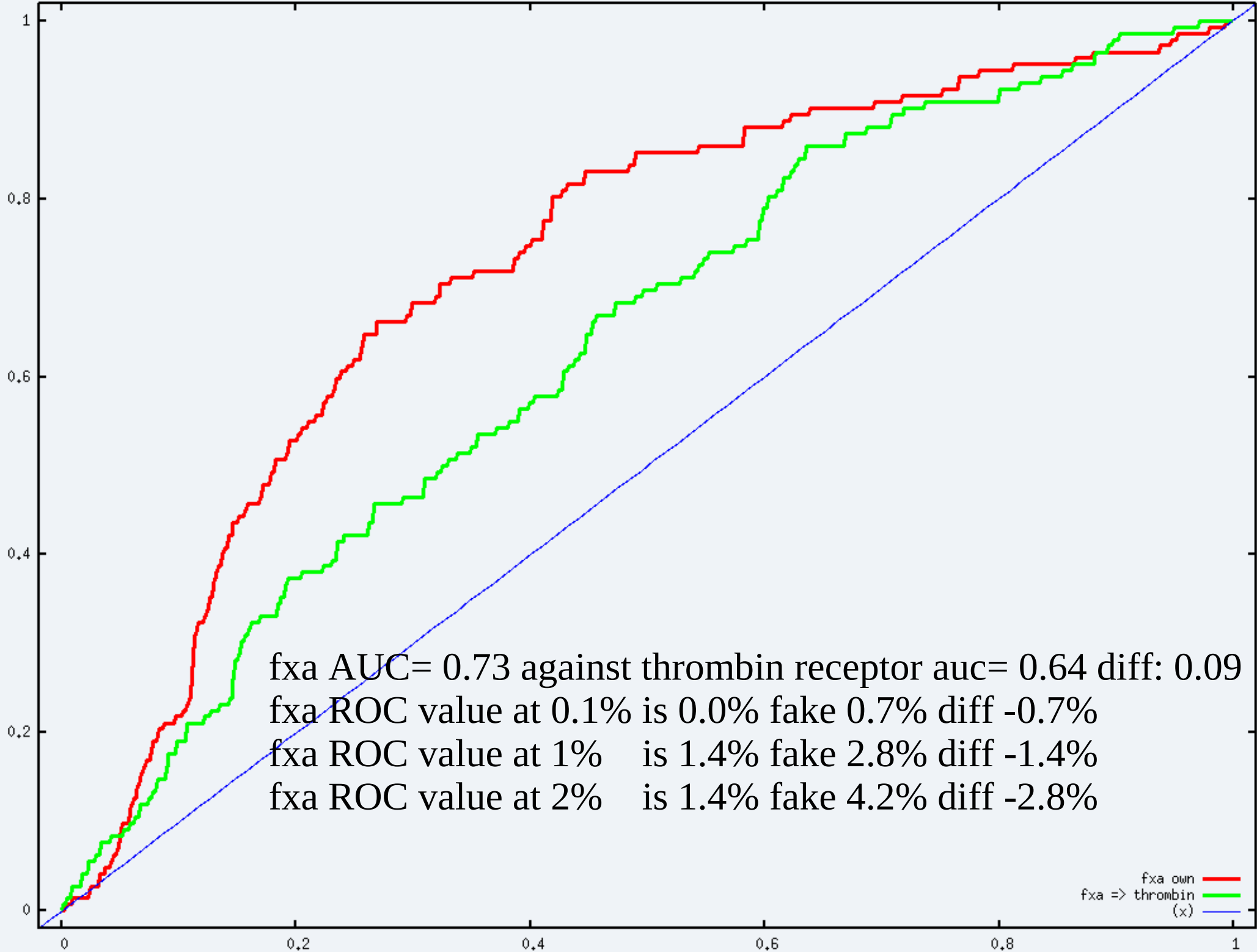


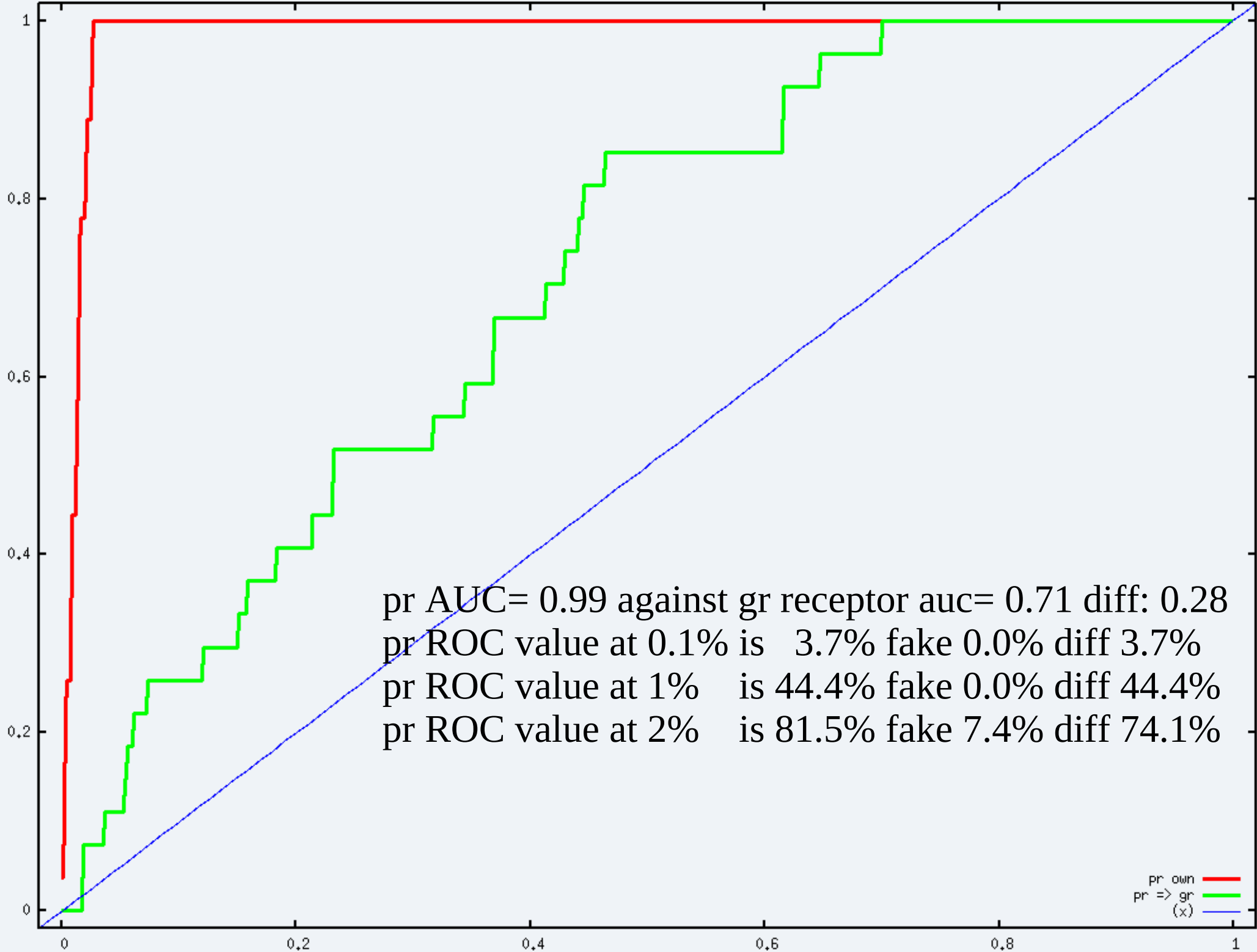






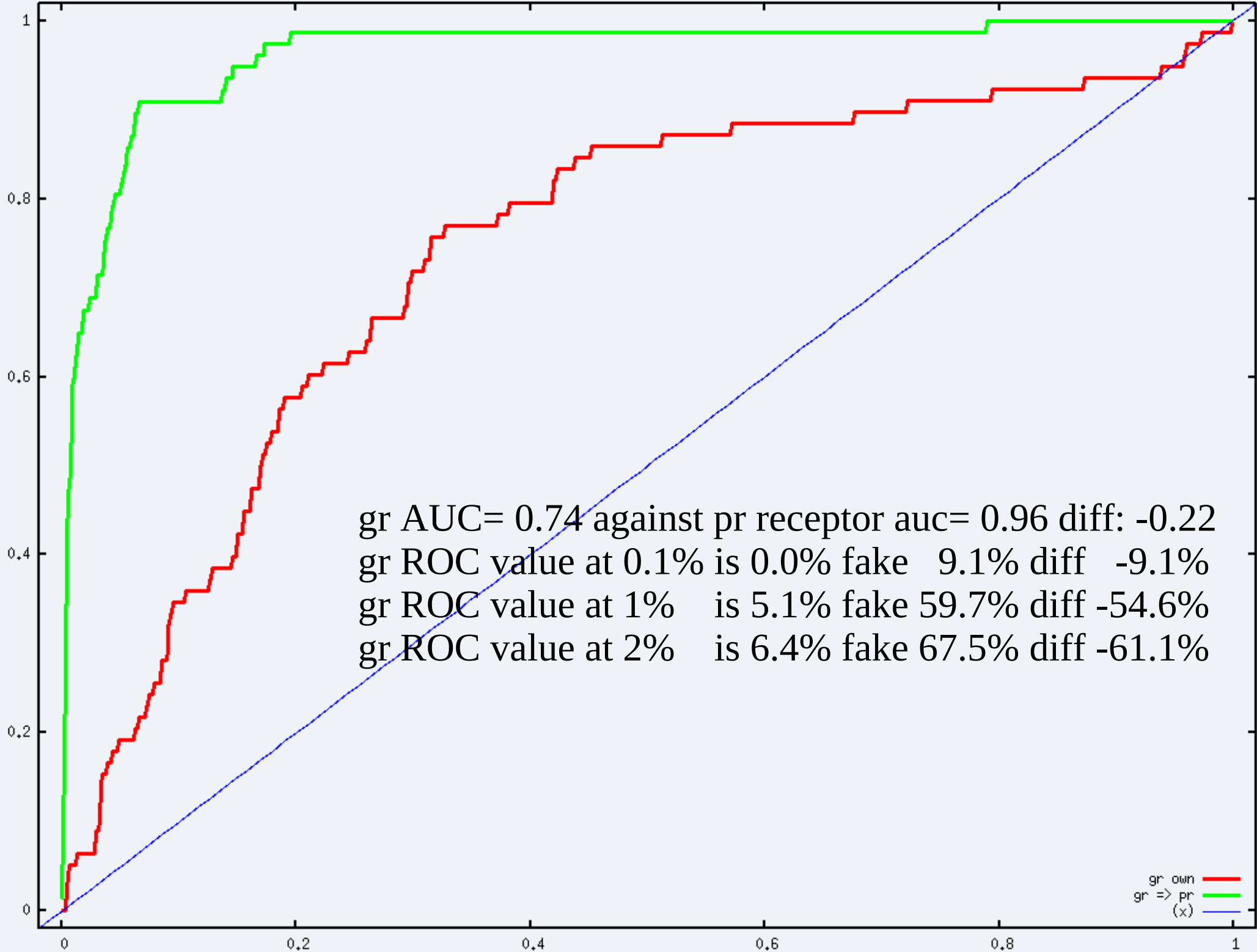




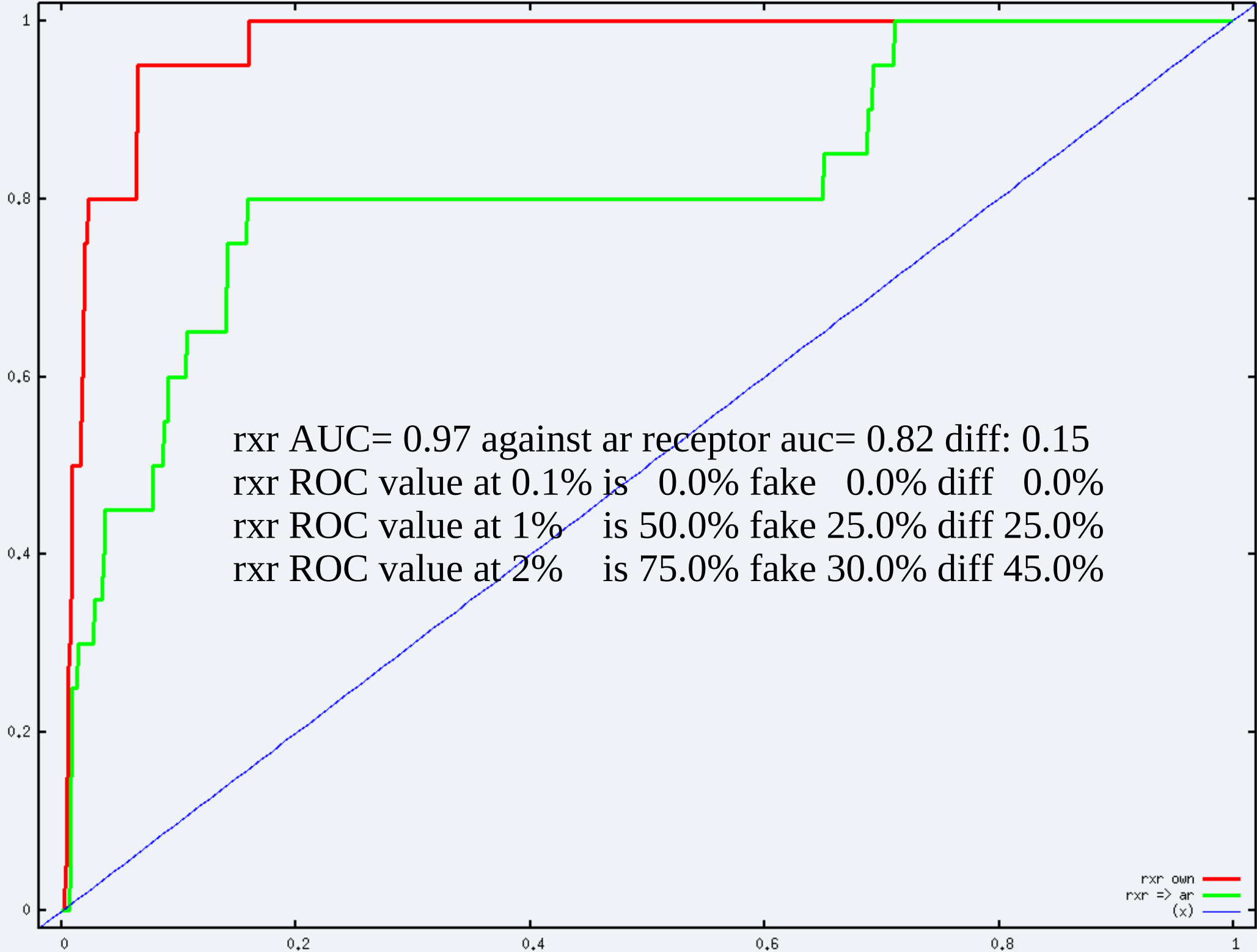


pr AUC= 0.99 against gr receptor auc= 0.71 diff: 0.28
pr ROC value at 0.1% is 3.7% fake 0.0% diff 3.7%
pr ROC value at 1% is 44.4% fake 0.0% diff 44.4%
pr ROC value at 2% is 81.5% fake 7.4% diff 74.1%

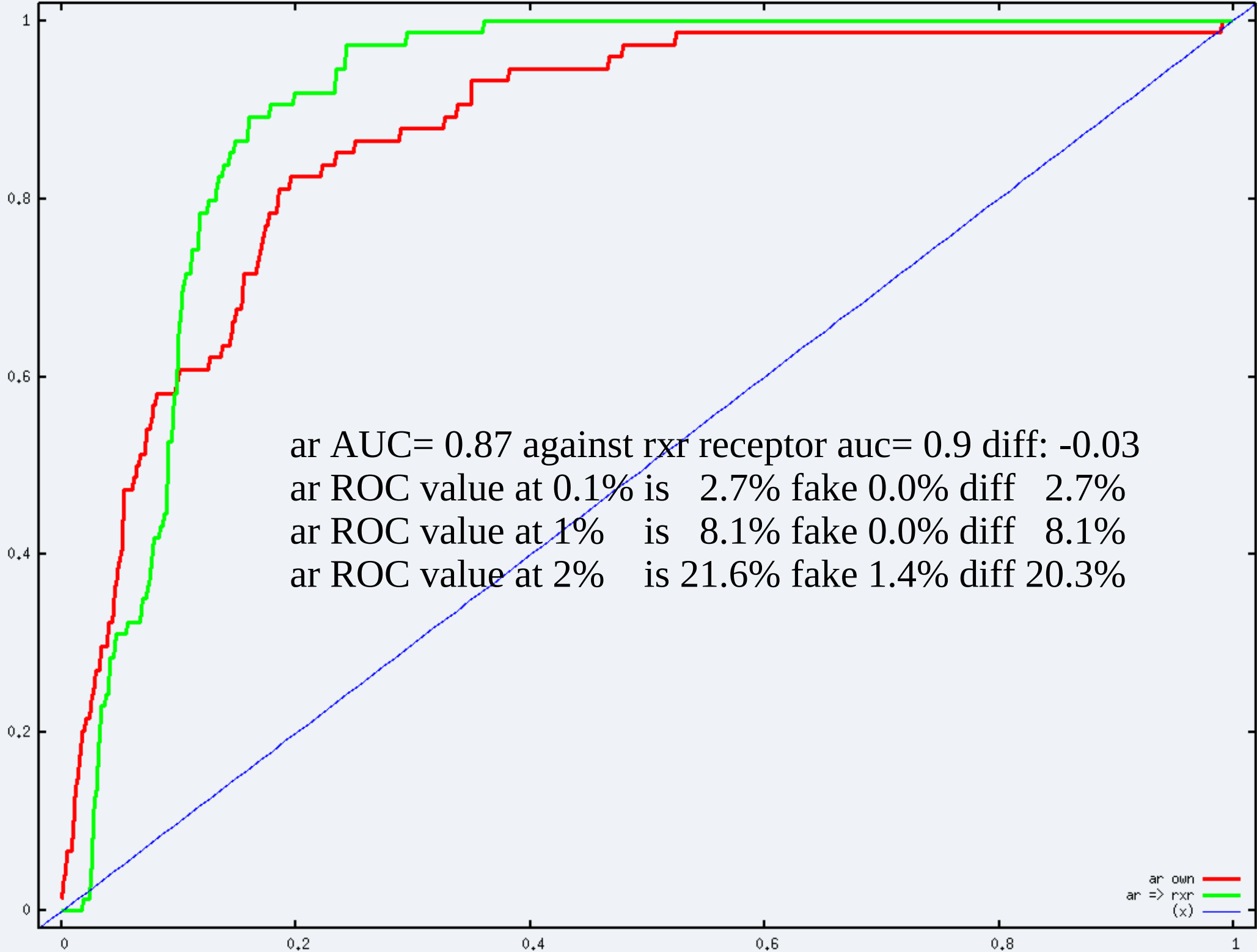
pr own (red)
pr => gr (green)
(x) (blue)

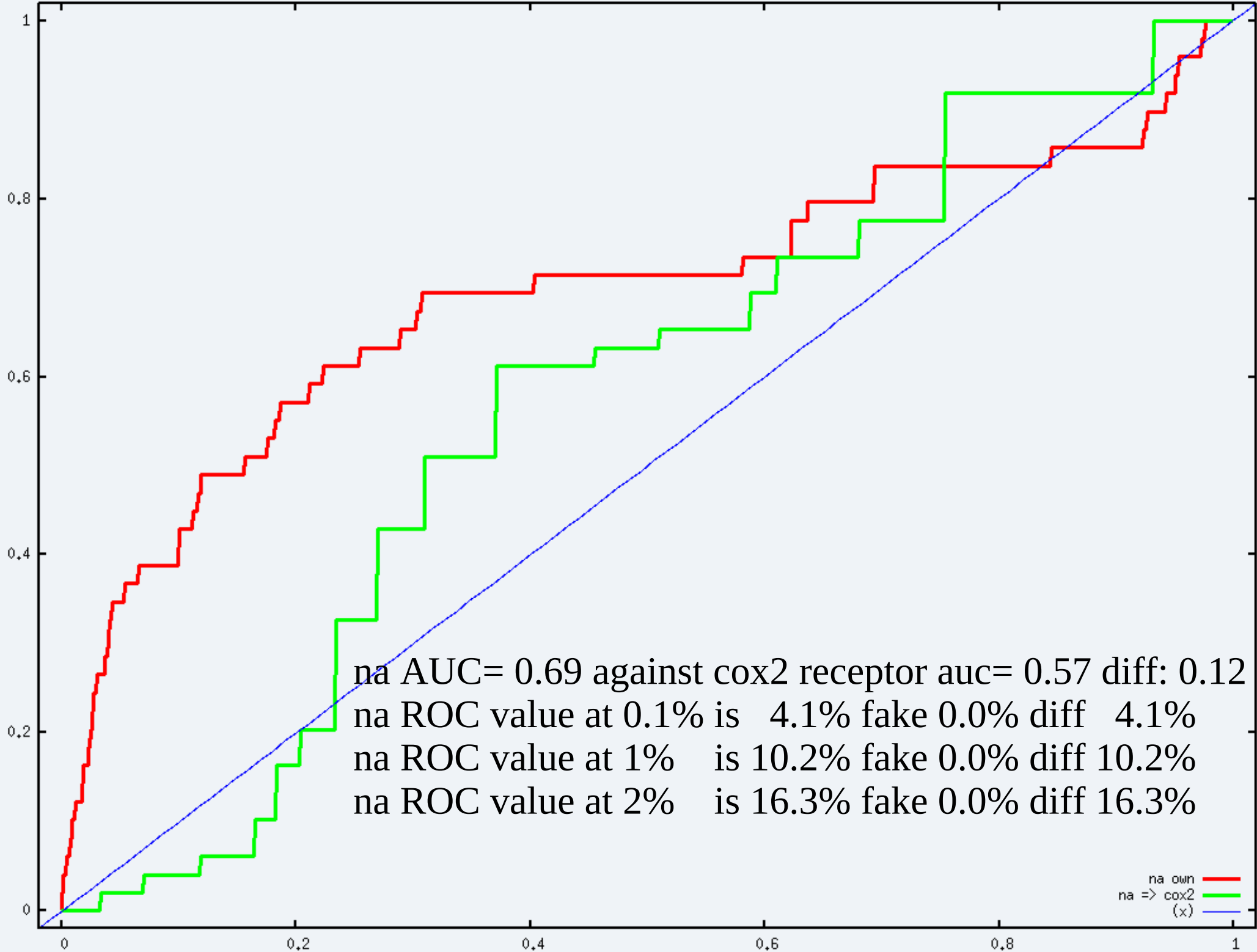


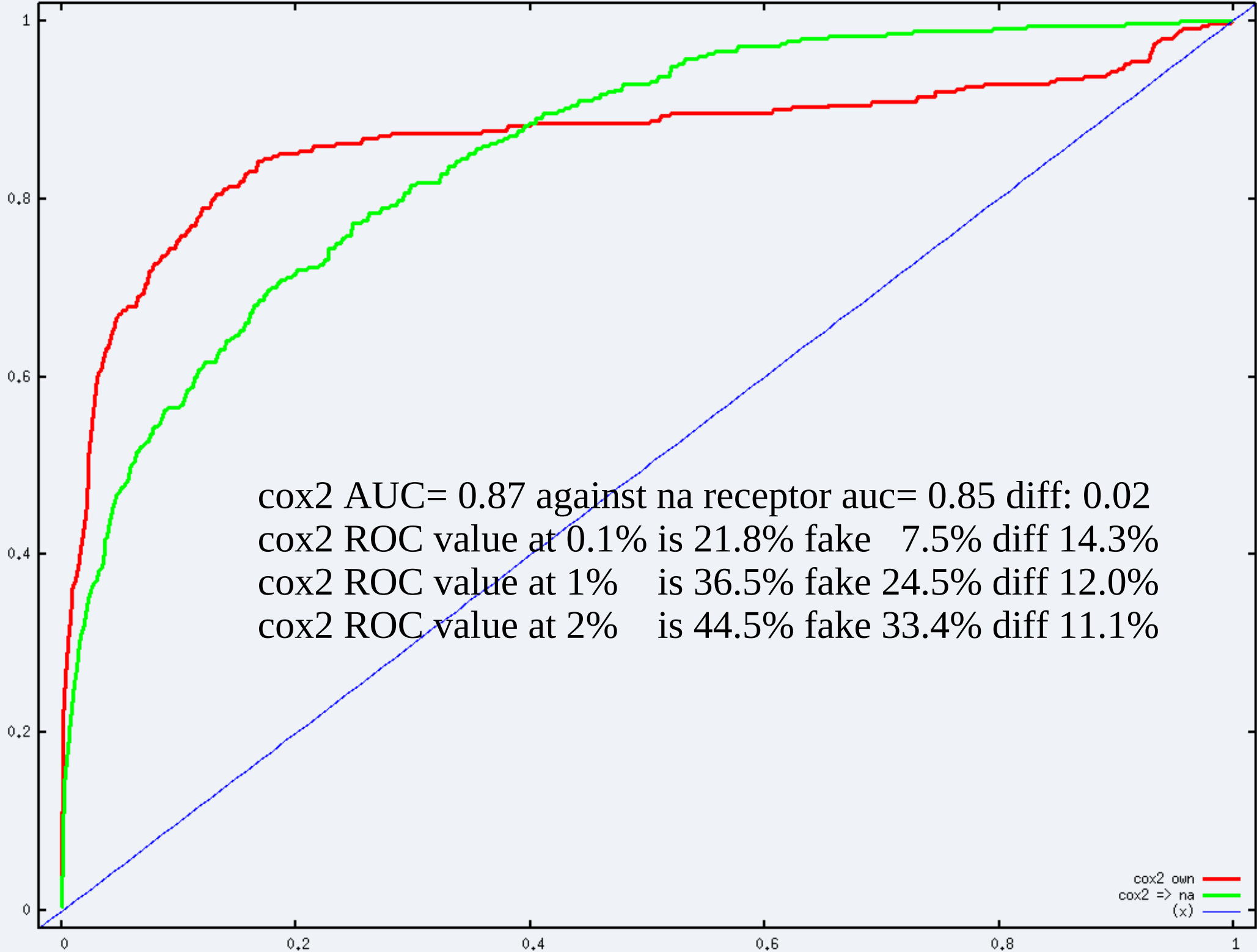
rxr AUC= 0.97 against ar receptor auc= 0.82 diff: 0.15
rxr ROC value at 0.1% is 0.0% fake 0.0% diff 0.0%
rxr ROC value at 1% is 50.0% fake 25.0% diff 25.0%
rxr ROC value at 2% is 75.0% fake 30.0% diff 45.0%



rxr own (red line)
rxr => ar (green line)
(x) (blue line)





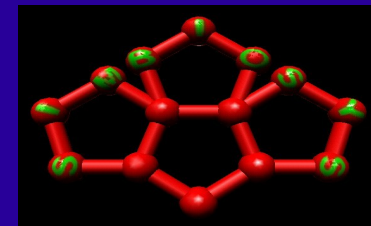


35. Virtual Screening result summary

	Real Target Median	Real Target Average	Standard Deviation of mean	Median Diff. vs. Fake Target	Average Diff. vs. Fake Target	Standard Deviation of diff.	Diff. range
ROC AUC	0.88	0.85	0.11	0.28	0.29	0.29	-0.22 0.89
ROC at 0.1%	4.5%	8.8%	10.9%	4.1%	7.5%	11.2%	-9.1% 37.1%
ROC at 1%	16.2%	24.3%	21.6%	14.2%	15.9%	28.0%	-54.6% 77.1%
ROC at 2%	22.1%	33.9%	27.3%	18.3%	22.9%	33.7%	-61.1% 82.9%



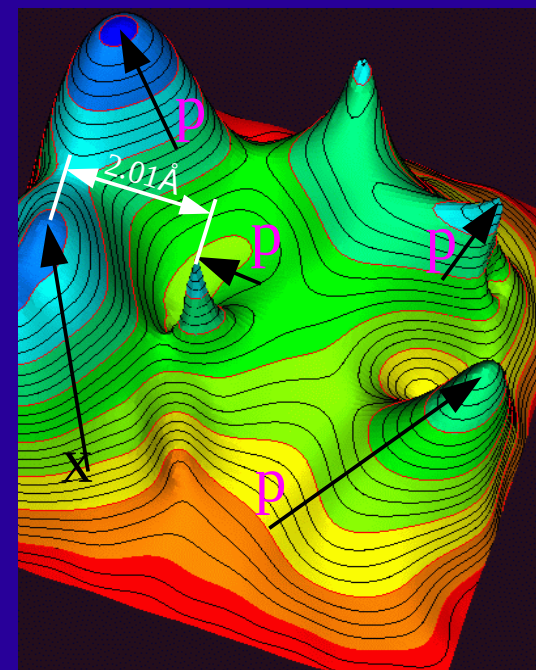
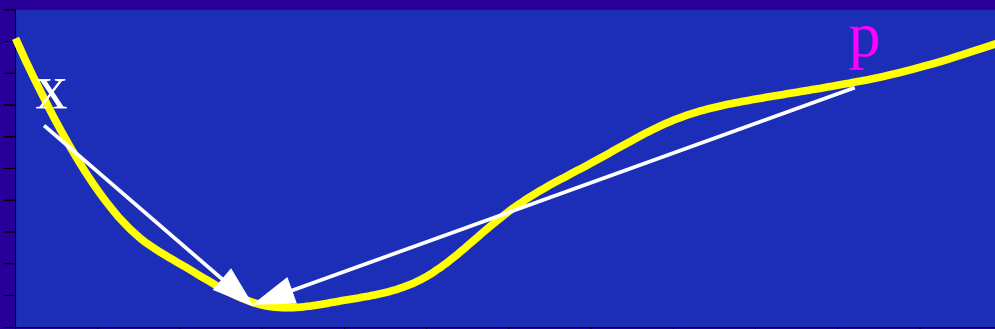
36. Summary



- Exhaustive, high accuracy - **don't miss a potential drug!**
- Very fast – **electronic High Throughput Screening** on the Cell B.E. “supercomputer in a chip”
- Surface point based statistical scoring with auto-tuning system
- Samples all feasible protonation states on-the-fly
- Good virtual screening performance via combination of interaction scoring and ligand based surface similarity metric
- For more information, free evaluation see our web site:
<http://www.simbiosys.com/>

37. Warning about RMSD result comparisons

Some vendors publish RMSD results measured against pre-optimised ligand instead of the original X-ray structure



Such RMSD is a property of the scoring function shape (distance of two local minima) rather than a measure of docking accuracy. This number should be **ZERO!**