# Supporting Information

# GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics

*Christina de Bruyn Kops[1], Martin Šícho[2], Angelica Mazzolari[3], Johannes Kirchmair[1,4]**

[1] Center for Bioinformatics (ZBH), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany

[2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, 166 28 Prague 6, Czech Republic

[3] Facoltà di Scienze del Farmaco, Dipartimento di Scienze Farmaceutiche "Pietro Pratesi", Università degli Studi di Milano, I-20133 Milan, Italy

[4] Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

* Correspondence:  johannes.kirchmair@univie.ac.at; Tel.: +43 1-4277-55104

# TABLES

**Table S1.** Enzymes Excluded From Consideration When Extracting Relevant Metabolites From DrugBank.

| **Enzyme** |
|---|
| Cocaine esterase |
| Thymidine phosphorylase |
| Serum albumin |
| Ribulose-phosphate 3-epimerase |
| UDP-galactose 4-epimerase |
| cGMP-specific 3'5'-cyclic phosphodiesterase |
| Dihydropyrimidinase-related protein 2 |
| Aromatic-L-amino-acid decarboxylase |
| Elongation of very long chain fatty acids protein 4 |
| Elongation of very long chain fatty acids protein 5 |
| Hemoglobin subunit beta |
| Hemoglobin subunit alpha |
| Selenocysteine lyase |
| Lysosomal protective protein |
| Enoyl-CoA hydratase mitochondrial |
| NADPH--cytochrome P450 reductase |
| Cytochrome b |

**Table S2.** Descriptors Used for Principal Component Analysis.

| **Name** | **Description (1)** |
|---|---|
| a_acc | Hydrogen bond acceptor atom count |
| a_acid | Acidic atom count |
| a_aro | Aromatic atom count |
| a_base | Basic atom count |
| a_don | Hydrogen bond donor atom count |
| a_heavy | Heavy atom count |
| a_hyd | Hydrophobic atom count |
| a_nB | Boron atom count |
| a_nBr | Bromine atom count |
| a_nC | Carbon atom count |
| a_nCl | Chlorine atom count |
| a_nF | Fluorine atom count |
| a_nH | Hydrogen atom count |
| a_nI | Iodine atom count |
| a_nN | Nitrogen atom count |
| a_nO | Oxygen atom count |
| a_nP | Phosphorus atom count |
| a_nS | Sulfur atom count |
| b_ar | Number of aromatic bonds |
| b_count | Number of bonds |
| b_double | Number of double bonds |
| b_rotN | Number of rotatable bonds |
| b_rotR | Fraction of rotatable bonds[a] |
| b_single | Number of single bonds |
| b_triple | Number of triple bonds |
| chiral | Number of chiral centers |

| | |
|---|---|
| FCharge | Total charge of the molecule |
| logP(o/w) | Log of the octanol/water partition coefficient |
| logS | Log of the aqueous solubility (mol/L) |
| mr | Molecular refractivity |
| PC+ | Total positive partial charge |
| PC- | Total negative partial charge |
| rings | Number of rings |
| TPSA | Polar surface area (Å$^2$) |
| vdw_area | Area of van der Waals surface (Å$^2$) |
| vdw_vol | van der Waals volume (Å$^3$) |
| vsa_acc | Approximation of the sum of VDW[b] surface areas (Å$^2$) of pure hydrogen bond acceptors[c] |
| vsa_acid | Approximation of the sum of VDW surface areas of acidic atoms (Å$^2$) |
| vsa_base | Approximation of the sum of VDW surface areas of basic atoms (Å$^2$) |
| vsa_don | Approximation of the sum of VDW surface areas of pure hydrogen bond donors[d] |
| vsa_hyd | Approximation of the sum of VDW surface areas of hydrophobic atoms (Å$^2$) |
| vsa_other | Approximation of the sum of VDW surface areas (Å$^2$) of atoms typed as "other" |
| vsa_pol | Approximation of the sum of VDW surface areas (Å$^2$) of polar atoms |
| Weight | Molecular weight |

[a] b_rotN divided by the number of bonds between heavy atoms

[b] VDW = van der Waals

[c] Not counting acidic atoms and atoms that are both hydrogen bond donors and acceptors

[d] Not counting basic atoms and atoms that are both hydrogen bond donors and acceptors

**Table S3.** Number of Molecules Used to Train the FAME 3 Reaction Type-Specific SoM Prediction Models.

| Reaction class | Number of molecules | ClassID(s) from MetaQSAR |
|---|---|---|
| Glucuronidations & glycosylations | 440 + 153 = 593 | 14, 15 |
| GSH & RSH[a] conjugations | 243 | 17 |
| Sulfonations | 148 | 16 |
| Methylations | 94 | 20 |
| Acetylations & acylations | 83 | 18 |

[a] RSH = protein thiol

**Table S4.** Ranking Performance of Phase 2 Metabolite Prediction using the Reaction Rules from SyGMa and Various Formulas for Combining the Predicted SoM Probabilities[a] with SyGMa's Reaction Probabilities.

| Score equation[b] | AUC of rank-based ROC curve |
|---|---|
| S x R | 0.85 |
| (S + R) / 2 | 0.82 |
| (2S + R) / 3 | 0.81 |
| (3S + R) / 4 | 0.80 |
| (5S + R) / 6 | 0.80 |
| (10S + R) / 11 | 0.80 |
| (S + 2R) / 3 | 0.82 |
| (S + 3R) / 4 | 0.82 |
| (S + 5R) / 6 | 0.82 |
| (S + 10R) / 11 | 0.83 |

[a] The SoM probabilities were predicted with FAME 3 model P2

[b] S = SoM probability, R = reaction probability

# FIGURES

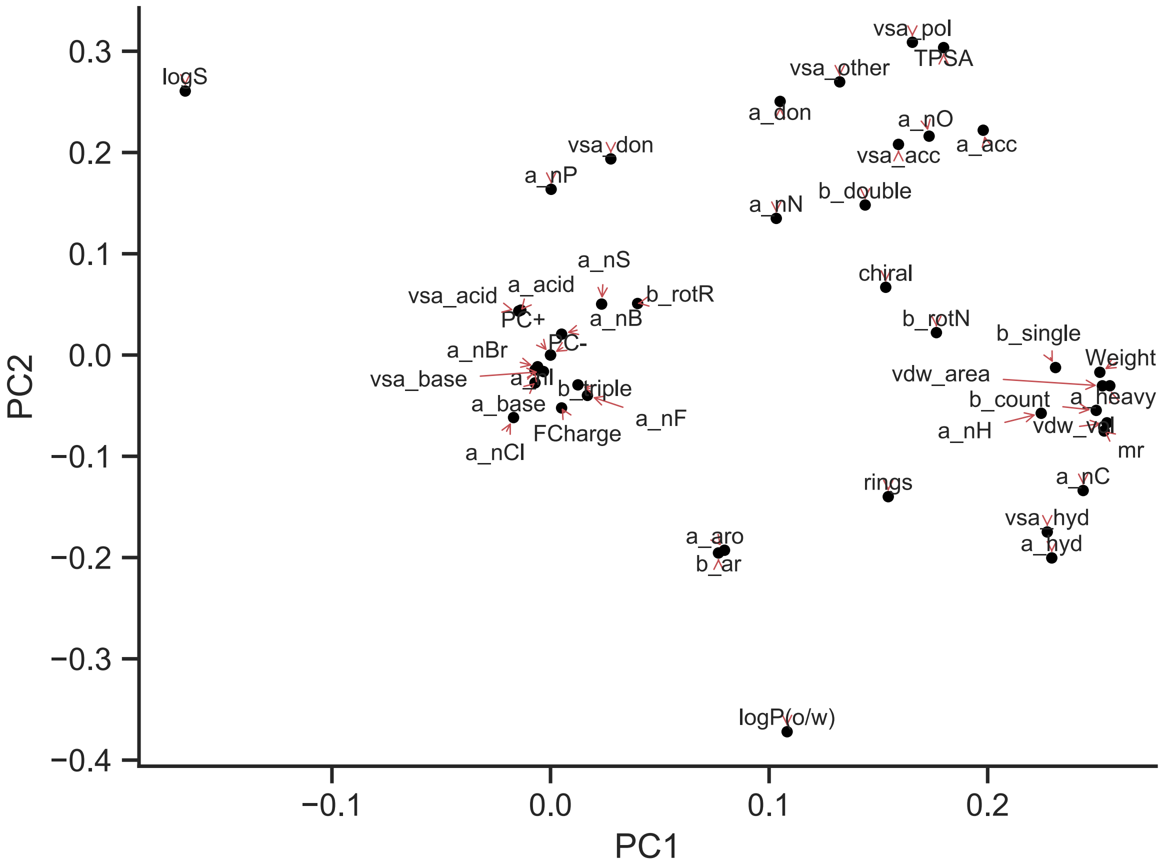

**Figure S1.** PCA loading plot for the PCA plot shown in Figure 1D. The PCA compares parent molecules from DrugBank and MetXBioDB using 44 physicochemical descriptors (Table S2). The percentage of the total variance explained by each of the first two principal components is 35.81% for PC1 and 10.69% for PC2.

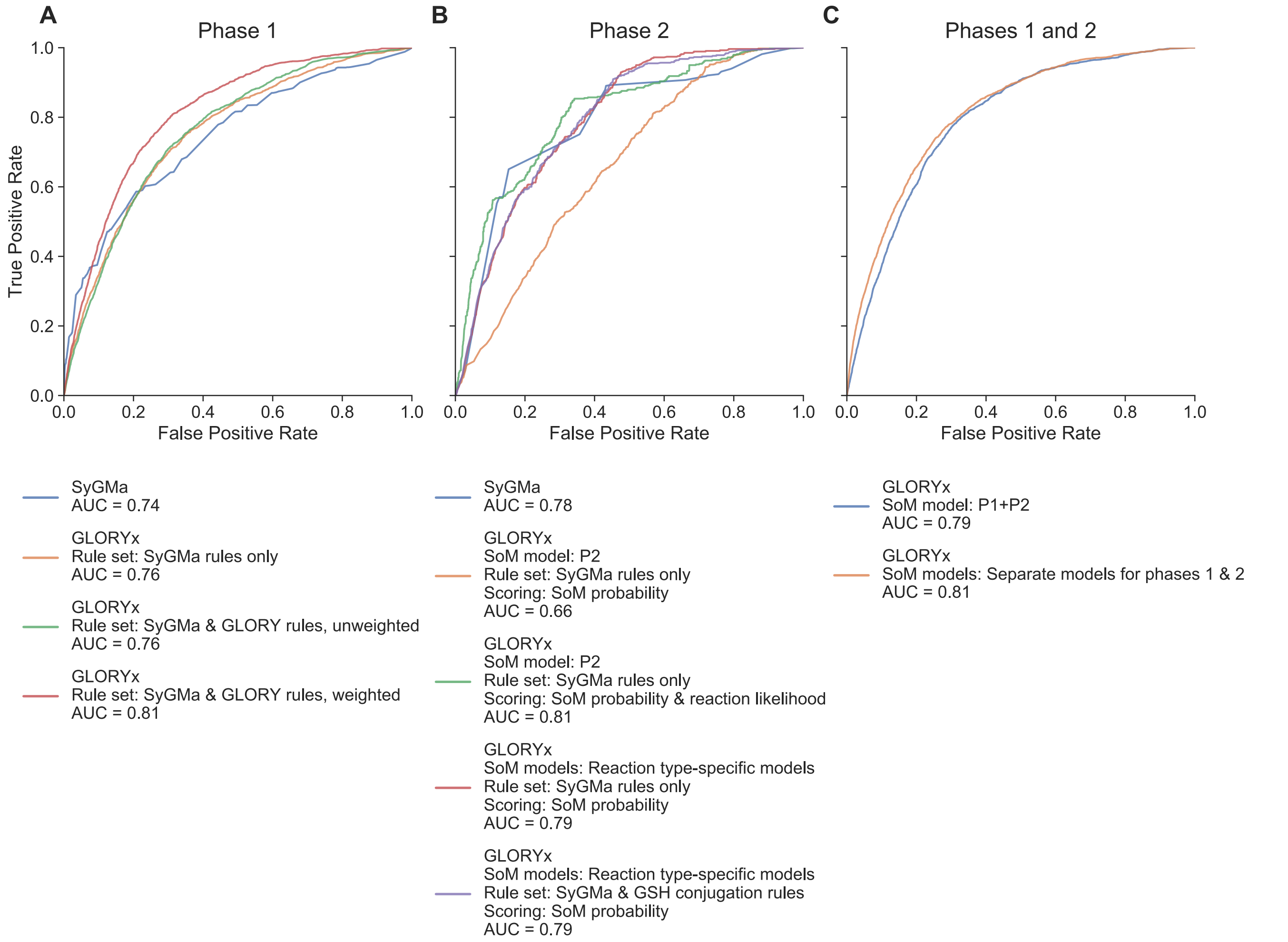

**Figure S2.** Score-based ROC curves for the evaluation of metabolite prediction performance on the reference dataset. (A) Comparison of GLORYx, which scores its predicted metabolites based on predicted SoM probability, to SyGMa, which uses reaction probability-based scoring, for phase 1 metabolite prediction. Weighted rules refer to the weighting of the SoM probability-based score based on whether the reaction type is designated common or uncommon. (B) Comparison of the ranking performance of GLORYx with different scoring approaches and rule sets, as well as a direct comparison to SyGMa's performance, for phase 2 metabolite prediction. The scoring approach that is based on both SoM probability and reaction probability is achieved by a simple multiplication of the two components. (C) Comparison of the ranking performance of GLORYx for combined prediction of metabolites for phases 1 and 2 metabolism, using different SoM prediction approaches to score the predicted metabolites. The predicted metabolites are scored based on predicted SoM probability. The rule set in both cases is the same and is made up of the final phase 1 rule set (SyGMa and GLORY rules) and final phase 2 rule set (SyGMa and GSH conjugation rules). Note that the score-based ROC curves for SyGMa should be viewed cautiously because SyGMa's scoring approach was only intended to compare scores among predicted metabolites of the same parent molecule (i.e. a rank-based comparison).

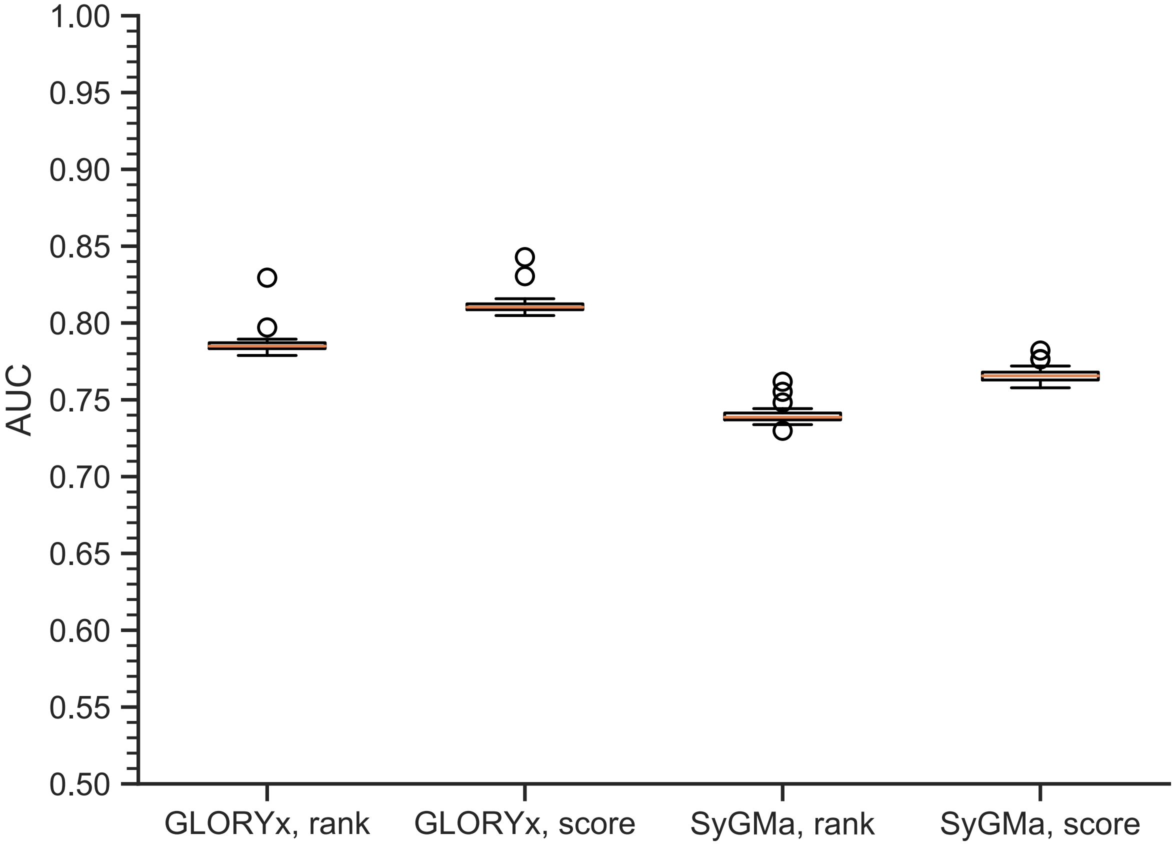

**Figure S3.** Variability in the ranking performance of SyGMa and GLORYx on the test set based on the rank and the score of the predicted metabolites. The data points were calculated by systematically removing one parent molecule from the test set at a time and calculating the AUC from the remaining predictions. There are therefore 37 AUC data points for each combination of tool and AUC type, corresponding to the size of the test set.

# REFERENCES

(1) Chemical Computing Group ULC. *MOE User Guide, MOE 2018.01*. Chemical Computing Group ULC: Montreal, Canada, 2018.