Kernel Learning for Virtual Screening: Discovery of a New PPARγ Agonist

Matthias Rupp,[1,2] Timon Schroeter,[3,4] Ramona Steri,[5] Ewgenij Proschak,[1,5] Katja Hansen,[3] Heiko Zettl,[5] Oliver Rau,[5] Manfred Schubert-Zsilavecz,[5] Klaus-Robert Müller,[3] Gisbert Schneider[1,6]

[1] Beilstein-Endowed Chair for Cheminformatics, Institute for Organic Chemistry and Chemical Biology, LiFF/ZAFES, Johann Wolfgang Goethe-University, Siesmayerstr. 70, 60323 Frankfurt, Germany. matthias.rupp@bio.uni-frankfurt.de [2] now at Institute for Bioinformatics and Systems Biology, Helmholtz Center Munich, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany [3] Chair for Machine Learning, Technical University of Berlin, Franklinstr. 28/29, 10587 Berlin, Germany [4] now at Bayer Schering Pharma AG, Nonclinical Drug Safety, Müllerstr. 178, 13353 Berlin, Germany [5] Institute for Pharmaceutical Chemistry, LiFF/ZAFES, Johann Wolfgang Goethe-University, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany [6] Chair for Computer-Assisted Drug Design, Institute of Pharmaceutical Sciences, Wolfgang-Pauli-Str. 10, Eidgenössische Technische Hochschule Zürich, 8093 Zürich, Switzerland

1 Introduction

We present methods and results of a successful prospective ligand-based virtual screening study [1] for novel agonists of the peroxisome proliferator-activated receptor γ (PPARγ [2], Figure 1), a nuclear receptor involved in lipid and glucose metabolism that is related to type-2 diabetes and dyslipidemia.

Key facts:

- Data set: 176 published PPARγ agonists [3]
- Descriptors: CATS2D [4], MOE 2D [5], Ghose-Crippen [6], ISOAK [7] (multiple kernel learning [8])
- Prediction: Gaussian process regression [9]
- Assay: Cell-based transactivation assay [10]

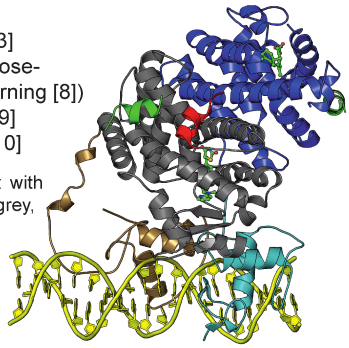


Figure 1: PPARγ-RXRα hetero-dimer in complex with DNA (PDBid 3dzy). Shown are PPARγ (LBD grey, DBD brown, activation function-2 red, ligand rosiglitazone green), RXRα (LBD blue, DBD cyan, ligand 9-cis-retinoic acid green), co-activator peptides (green), and, DNA (yellow). RXR = retinoid X receptor, LBD = ligand binding domain, DBD = DNA binding domain.

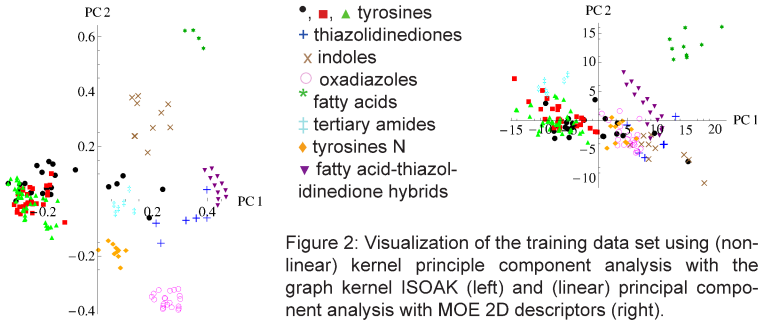


Figure 2: Visualization of the training data set using (non-linear) kernel principle component analysis with the graph kernel ISOAK (left) and (linear) principal component analysis with MOE 2D descriptors (right).

2 Machine learning

- The iterative similarity optimal assignment kernel (ISOAK [7], Figure 3) is a similarity measure defined directly on the annotated structure graph. It was designed specifically to compare small molecules, and is suitable for use with kernel-based machine learning methods.

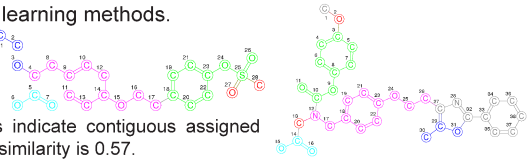


Figure 3: ISOAK assignment example of tesaglitazar (left) vs. rosiglitazar (right). Colors indicate contiguous assigned regions. Total normalized similarity is 0.57.

- ISOAK iteratively solves the non-linear system of equations

$$X_{i,j} = (1-\alpha)k_v(v_i, v'_j) + \alpha \max_{\pi} \frac{1}{|v'_j|} \sum_{v \in n(v_i)} X_{v,\pi(v)} k_e(\{v_i, v\}, \{v'_j, \pi(v)\})$$

- Based on the resulting matrix X of pairwise atom similarities, the atoms of the smaller molecule are assigned to atoms of the larger molecule. The (normalized) sum of the assigned atoms similarity values gives the final score.
- Gaussian process regression ([9], Figure 4) is a Bayesian kernel-based regression method. Its major advantage is that it provides confidence estimates for its predictions, i.e., it has built-in domain of applicability [11].

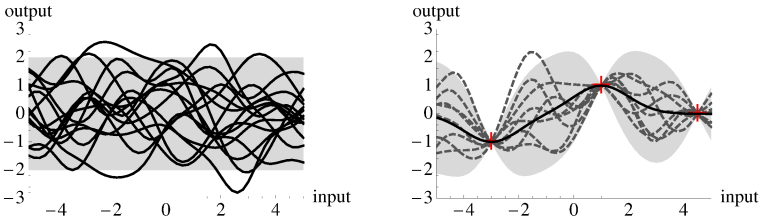


Figure 4: Gaussian process regression. Starting from the prior distribution (left), one conditions on the observed samples (red crosses). Mean (solid line) and variance (grey area) of the posterior (right) serve as predictor and confidence estimates, respectively.

3 Screening

- 16 regression models (Figure 5) were evaluated using 10 runs of leave-5-clusters-out cross-validation.
- 3 models were selected (Table 1) and checked via *y*-scrambling.
- The Asinex [12] Gold and Platinum libraries (360,150 compounds) were virtually screened. From the top 30 predictions of each of the three models, a total of 15 compounds were manually selected according to scaffold novelty.
- The selected compounds were tested in an in vitro cell-based transactivation assay ([10], Figure 6).

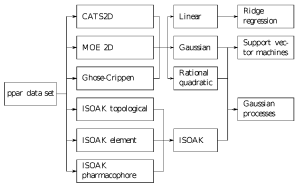


Figure 5: Investigated models.

Model	MAE	FI_{20}
CATS2D/RBF+RQ	0.66±0.09	0.27±0.14
ISOAK+all/RBF+all/RQ+W	0.66±0.07	0.32±0.15
ISOAK+allmkl/RBF+allmkl/RQ	0.71±0.12	0.21±0.09

Table 1: Performance of selected models. MAE = mean absolute error, FI_{20} = fraction of inactives in the 20 top-ranked compounds, all = all descriptors, allmkl = all descriptors with one kernel per descriptor, W = compounds weighted by activity.

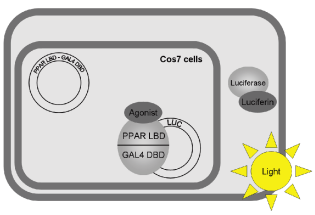
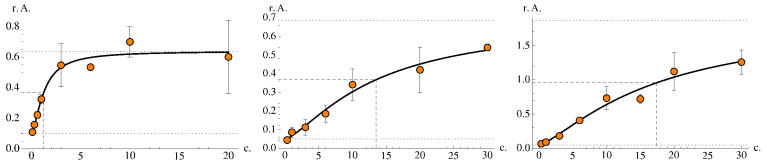


Figure 6: Cell-based transactivation assay for PPAR. Immortalized simian kidney cells (Cos7) are co-transfected with an expression plasmid for a Gal4 hybrid protein and a luciferase-encoding reporter plasmid. Upon receptor activation, the reporter gene is expressed and the resulting luminescence is measured.

4 Results

- From 15 selected compounds, 4 were active on PPAR (Figure 7).



hPPARα MA 65 ± 5%, EC_{50} = 1.25 ± 0.37 μM.

hPPARα MA 68 ± 24%, EC_{50} = 13.48 ± 8.53 μM.

hPPARγ MA 138 ± 2%, EC_{50} = 10.03 ± 0.2 μM.

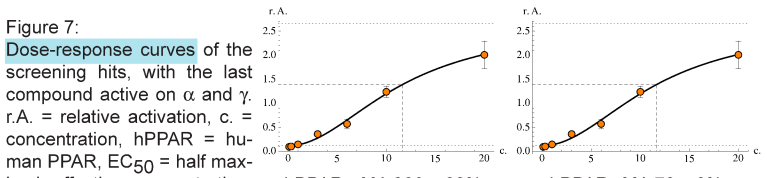
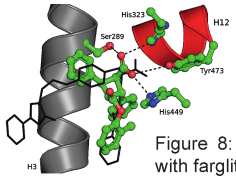


Figure 7: Dose-response curves of the screening hits, with the last compound active on α and γ. r.A. = relative activation, c. = concentration, hPPAR = human PPAR, EC_{50} = half maximal effective concentration, MA = maximum activation.

hPPARα MA 200 ± 32%, EC_{50} = 12.98 ± 4.21 μM.

hPPARγ MA 76 ± 3%, EC_{50} = 3.75 ± 0.2 μM.



- One of the hits (Figure 7, upper right corner) is a natural product
- Truxillic acid derivative, occurs in plant cell walls
- No cytotoxicity in cell-based assays
- Hypothetical binding pose using docking (Figure 8)

Figure 8: Hypothetical binding mode as determined using GOLD [13] with farglitazar (black lines model) as reference.

References

[1] Rupp, M., Schroeter, T., Steri, R., Zettl, H., Proschak, E., Hansen, K., Rau, O., Schwarz, O., Müller-Kuhrt, L., Schubert-Zsilavecz, M., Müller, K.-R., Schneider, G.: From Machine Learning to Natural Product Derivatives Selectively Activating Transcription Factor PPARγ, submitted, 2009.
[2] Michalik, L., Auwerx, J., Berger, J., Chatterjee, K., Glass, C., Gonzalez, F., Grimaldi, P., Kadowaki, T., Lazar, M., O'Rahilly, S., Palmer, C., Plutzky, J., Reddy, J., Spiegelman, B., Staels, B., Wahli, W.: International Union of Pharmacology. LXI. Peroxisome Proliferator-Activated Receptors, *Pharmacol. Rev.* 58(4): 726-741, 2006.
[3] Rücker, C., Scarsi, M., Meringer, M.: 2D QSAR of PPARγ Agonist Binding and Transactivation, *Bioorg. Med. Chem.* 14(15): 5178-5195, 2006.
[4] Schneider, G., Neidhart, W., Giller, T., Schmid, G.: "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening, *Angew. Chem. Int. Ed.* 38(19): 2894, 1999.
[5] Molecular Operating Environment, Chemical Computing Group, www.chemcomp.com, 2008.
[6] Viswanadhan, V., Ghose, A., Revankar, G., Robins, R.: Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4, *J. Chem. Inform. Comput. Sci.* 29(3): 163-172, 1989.
[7] Rupp, M., Proschak, E., Schneider, G.: Kernel Approach to Molecular Similarity Based on Iterative Graph Similarity, *J. Chem. Inf. Model.* 47(6): 2280-2286, 2007.
[8] Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large Scale Multiple Kernel Learning, *J. Mach. Learn. Res.* 7(7): 1531-1565, 2006.
[9] Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning, MIT Press, Cambridge, 2006.
[10] Rau, O., Wurglics, M., Paulke, A., Zitzkowski, J., Meindl, N., Bock, A., Dingermann, T., Abdel-Tawab, M., Schubert-Zsilavecz, M.: Carnosic Acid and Carnosol, Phenolic Diterpene Compounds of the Labiate Herbs Rosemary and Sage, are Activators of the Human Peroxisome Proliferator-Activated Receptor Gamma, *Planta Med.* 72(10): 881-887, 2006.
[11] Schroeter, T., Schwaighofer, A., Mika, S., Ter Laak, A., Sülzle, D., Ganzer, U., Heinrich, N., Müller, K.-R.: Estimating the Domain of Applicability for Machine Learning QSAR models: A study on Aqueous Solubility of Drug Discovery Molecules, *J. Comput. Aided Mol. Des.* 21(9): 485-498, 2007.
[12] Asinex, www.asinex.com, Geroev Panfilovtzev Str. 20, Building 1, 125480 Moscow, Russia.
[13] GOLD Protein-Ligand Docking Software, www.ccdc.cam.ac.uk, Cambridge Crystallographic Data Centre, Union Road 12, Cambridge CB2 1EZ, United Kingdom.

modlab http://www.modlab.de

FIRST Frankfurt International Research Graduate School for Translational Biomedicine

Beilstein Institut

GOETHE UNIVERSITÄT FRANKFURT AM MAIN