DOI: 10.1002/cmdc.200900469

From Machine Learning to Natural Product Derivatives that Selectively Activate Transcription Factor PPARγ

Matthias Rupp,[a] Timon Schroeter,[b] Ramona Steri,[c] Heiko Zettl,[c] Ewgenij Proschak,[a] Katja Hansen,[b] Oliver Rau,[c] Oliver Schwarz,[d] Lutz Müller-Kuhrt,[d] Manfred Schubert-Zsilavecz,[c] Klaus-Robert Müller,[b] and Gisbert Schneider*[a, e]

Peroxisome proliferator-activated receptors (PPARs) are nuclear proteins that act as transcription factors. They represent a validated drug target class involved in lipid and glucose metabolism as well as inflammatory response regulation.[1] We combined state-of-the-art machine learning methods including Gaussian process (GP) regression,[2] multiple kernel learning,[3] the ISOAK molecular graph kernel,[4] and a novel loss function to virtually screen a large compound collection for potential PPAR activators; 15 compounds were tested in a cellular reporter gene assay. The most potent PPARγ-selective hit (EC_{50} = 10 ± 0.2 μM) is a derivative of the natural product truxillic acid. Truxillic acid derivatives are known to be anti-inflammatory agents,[5] potentially due to PPARγ activation. Our study underscores the usefulness of modern machine learning algorithms for finding potent bioactive compounds and presents an example of scaffold-hopping from synthetic compounds to natural products. We thus motivate virtual screening of natural product collections as a source of novel lead compounds. The results of our study suggest that pharmacophoric patterns of synthetic bioactive compounds can be traced back to natural products, and this will be useful for "de-orphanizing" the natural bioactive agent.[6]

PPARs are present in three known isoforms: PPARα, PPARβ (δ), and PPARγ, with different expression patterns according to their function.[7] PPAR activation leads to an increased expression of key enzymes and proteins involved in the uptake and metabolism of lipids and glucose.[8] Unsaturated fatty acids and eicosanoids such as linoleic acid and arachidonic acid are physiological PPAR activators. Owing to their central role in glucose and lipid homeostasis, PPARs represent attractive drug targets for the treatment of diabetes and dyslipidemia.[9] Glitazones (thiazolidinediones) such as pioglitazone and rosiglitazone act as selective activators of PPARγ and are used as therapeutics for diabetes mellitus type 2.[10] In addition to synthetic activators, herbs are traditionally used for treatment of metabolic disorders, and some herbal ingredients have been identified as PPARγ activators, for example, carnosol and carnosic acid, as well as several terpenoids and flavonoids.[11,12]

We used several machine learning methods, with synthetic PPAR agonists as input, to find common pharmacophoric patterns for virtual screening in both synthetic and natural product derived substances. We focused on GP models, which originate from Bayesian statistics. Their original applications in cheminformatics were aimed at predicting aqueous solubility,[13] blood–brain barrier penetration,[14] hERG (human ether-á-go-go-related gene) inhibition,[14,15] and metabolic stability.[16] A particular advantage of GPs is that they provide error estimates with their predictions.[2]

In GP modeling of molecular properties, one defines a positive definite kernel function to model molecular similarity. Compound information enters GP models only via this function, so relevant (context-dependent) physicochemical properties must be captured. This is done by computing molecular descriptors (physicochemical property vectors), or by graph kernels that are defined directly on the molecular graph. From a family of functions that are potentially able to model the underlying structure–activity relationship ("prior"), only functions that agree with the data are retained (Figure 1). The weighted average of the retained functions ("posterior") acts as predictor, and its variance as an estimate of the confidence in the predic-

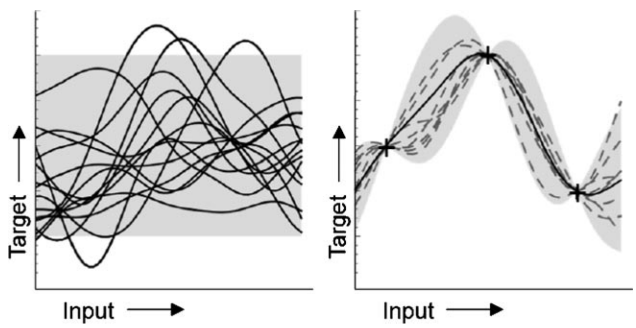


Figure 1. Nonlinear Bayesian regression with Gaussian processes starts with a family of functions that map input data to activity (target) values; 25 randomly selected functions are shown (left). This prior is then combined with measured data (crosses, right); only functions close to the observed data are retained. Averaging over the remaining functions yields the final predictor (solid line) and its variance (shaded area) as confidence estimate (domain of applicability).

[a] M. Rupp, Dr. E. Proschak, Prof. Dr. G. Schneider
Goethe University, Chair for Chem- and Bioinformatics, LIFF, ZAFES
Siesmayerstr. 70, 60323 Frankfurt a.M. (Germany)

[b] Dr. T. Schroeter, K. Hansen, Prof. Dr. K.-R. Müller
Technical University of Berlin, Chair for Machine Learning
Franklinstr. 28/29, 10587 Berlin (Germany)

[c] R. Steri, Dr. H. Zettl, Dr. O. Rau, Prof. Dr. M. Schubert-Zsilavecz
Goethe University, Chair for Pharmaceutical Chemistry, LIFF, ZAFES
Max-von-Laue-Str. 9, 60438 Frankfurt a.M. (Germany)

[d] Dr. O. Schwarz, Dr. L. Müller-Kuhrt
AnalytiCon Discovery GmbH
Hermannswerder Haus 17, 14473 Potsdam (Germany)

[e] Prof. Dr. G. Schneider
Present address: ETH Zürich, Institute of Pharmaceutical Sciences
Wolfgang-Pauli-Str. 10, 8093 Zürich (Switzerland)
E-mail: gisbert.schneider@pharma.ethz.ch

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cmdc.200900469: assay conditions, preparation of plant extracts, structure determination, machine learning theory and results, potential binding mode of compound **8**.

tion. Variance is small near reference data, that is, for molecules similar to known ligands, and increases with growing distance. Predictions and confidence estimates can be calculated analytically.

A total of 144 published PPARγ ligands and their pK_d values,[17] obtained by scintillation proximity assays,[18] served as training data. The Asinex Gold and Platinum collections[19] (version 11/2007; 360 000 compounds after removal of duplicates and preprocessing) were screened. Compounds were represented by descriptor vectors, structure graphs, and combinations thereof: 1) CATS2D[20] topological pharmacophore descriptor (210-dimensional correlation vector), 2) 184 MOE2D descriptors,[21] and 3) Ghose–Crippen fragment counts[22] (109 substructures). Structure graphs were used un-annotated (i.e., only graph topology), with element and bond type annotation, and with pharmacophore type (lipophilic, negative, positive, hydrogen bond acceptor and donor) annotation.[4] For descriptor vectors, the radial basis function (RBF) kernel and the rational quadratic kernel were used.[2] For structure graphs, the iterative similarity optimal assignment kernel (ISOAK)[4] was used (see Supporting Information).

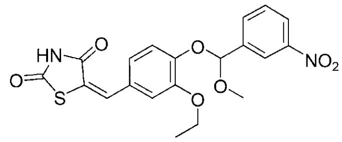
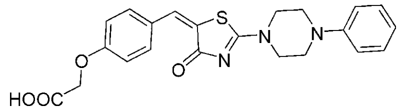
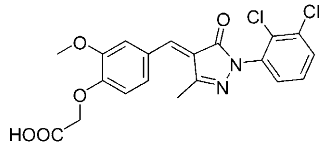
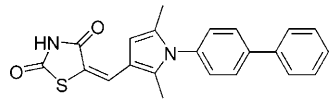
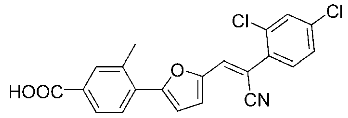
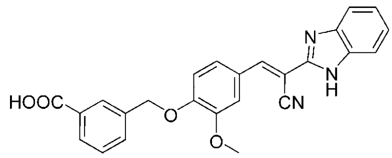
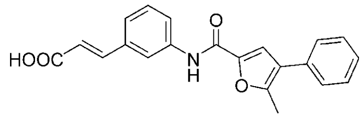
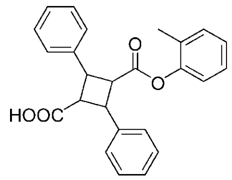
We statistically evaluated 16 prediction models on the PPARγ data and retained three models for prospective application after rigorous model testing and statistical evaluation (Supporting Information). Among other measures, we used the fraction of inactives among the top-20-ranked compounds (FI_{20}) as performance measure to address the early recognition problem.

Retrospective validation results (Supporting Information table S2) indicate that: a) PPARγ activation is a nonlinear function of molecular structure; b) GC fragment descriptors perform worse than CATS2D and MOE2D descriptors for this task; c) for ISOAK, the un-annotated structure graph performed best; d) compound weighting by activity did not improve performance; and e) using separate kernels for the vector descriptors in multiple kernel learning models slightly increased the mean absolute error (MAE) but lowered FI_{20}.

The three selected models were: 1) GP with ISOAK and three RBF and rational quadratic kernels on the vectorial descriptors (best FI_{20} score), 2) GP with ISOAK and one RBF and rational quadratic kernel on all descriptors (best MAE), and 3) GP with RBF and rational quadratic kernel on CATS2D descriptor (fast and simple model, i.e., only one vector representation). Screening compounds were ranked according to these models. From the 30 top-ranked compounds of each of the three lists, 15 were manually selected according to presumed activity and novelty of the structural scaffold. These were tested in a cellular PPAR activation assay. Ten of the candidates originated from the model with the best FI_{20} score.

Eight compounds exhibit agonistic activity toward PPARα, PPARγ, or both; all of them originated from the model with highest FI_{20} score, thereby corroborating this performance measure (Table 1). Compounds **4** (PPARα $EC_{50} = 1.3 \pm 0.4$ μM), **5** (PPARα $EC_{50} = 13 \pm 4$ μM), and **7** (PPARα $EC_{50} = 13 \pm 9$ μM) activated PPARα, whereas compound **8** is a selective PPARγ agonist with $EC_{50} = 10 \pm 0.2$ μM.

Table 1. Percent activation of PPARs by compounds from virtual screening.[a]

Compd	Structure	Activation [%] PPARα	Activation [%] PPARγ[a]
1		31 ± 16	36 ± 11
2		17 ± 8	inactive
3		23 ± 1	inactive
4		70 ± 10	inactive
5		92 ± 17	123 ± 12
6		30 ± 11	inactive
7		34 ± 8	32 ± 4
8		inactive	73 ± 17

[a] Activation is expressed relative to the effect of 1 μM GW7647 (PPARα), pioglitazone (PPARγ), or L165,04 (PPARδ) at a test compound concentration of 10 μM; all compounds were inactive at PPARδ; $n \geq 3$.

All active compounds contain an acid group, which most likely interacts with the activation function-2 helix of PPARs. While compounds **1**–**7** possess an elongated shape similar to known PPAR agonists that could be expected from the structural bias of the training data, the scaffold of compound **8** is most surprising. It may be considered a dimer of cinnamic acid and a monoester of cinnamic acid (Figure 2; see the Supporting Information for a potential receptor binding mode).

The stereochemical configuration of **8** (2,4-diphenyl-3-(*o*-tolyloxycarbonyl)cyclobutanecarboxylic acid) was obtained from ^1H NMR spectroscopy and NOE/ROESY analysis (Supporting Information). In summary, we determined the structure as a pair of *trans*-phenyl enantiomers, in which three protons (CH-COOR, CH-COOH and one CH-Ph) are situated on one side of the cyclobutane ring, and the remaining proton (CH-Ph) is located in the *trans* position (Figure 2). Enantioselective HPLC confirmed the presence of a pair of enantiomers (Supporting Information).

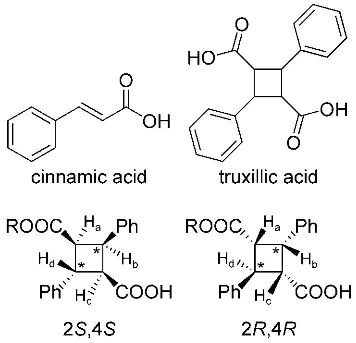


Figure 2. Top: truxillic acid derivatives such as compound **8** can be obtained from cinnamic acid monomers.[5] Such phenolic acids are widely distributed in plants.[24] Bottom: compound **8** (*trans*-phenyl, R = *o*-tolyl) was determined to be present as a pair of enantiomers.

From a drug design perspective it may seem counterintuitive to start with potent synthetic compounds and end up with a moderately active natural product derivative. While it is clear that this "inverse" strategy will not immediately result in new drug candidates, it certainly helps to: 1) understand the relationship between synthetic drugs and natural products, 2) define relevant pharmacophoric features in drugs and natural products, 3) find nature-inspired molecular scaffolds of innovative lead structures, and 4) attribute known pharmacological modes of action to herbal medicines. Consequently, we encourage the use of virtual screening not only for hit and lead finding but also for understanding structural and functional relationships between bioactive agents.

Some herbals have been shown to help ameliorate diabetic disorders in animal studies, but no molecular mechanism is known ("orphan" medicines) that could explain their effects on blood glucose levels. One example source is *Cynodon dactylon* (family *Poaceae*), the aqueous extract of which has been reported to possess anti-diabetic potential and to decrease hyperlipidemia in rats.[23] This plant contains high quantities of substituted truxillic acids.[24] We prepared both aqueous and lipophilic extracts, but were unable to observe significant PPARγ activation in vitro (data not shown). Further analysis of the extracts for the presence of truxillic acid derivatives is required before rendering a definitive conclusion. The reported effect on blood glucose levels by *Cynodon dactylon* extracts might also be caused through other molecular mechanisms; it is possible that our extract preparation from dried plants did not yield bioactive truxillic acid derivatives, or only insufficient amounts for PPARγ activation. Irrespective of this particular finding, we could demonstrate that truxillic acid derivative **8** has the ability to activate PPARγ, which explains the previously described anti-inflammatory activities of this compound class.[5]

We conclude that advanced machine learning methods are suited for modeling nonlinear structure–activity relationships. For predicting PPAR activation, linear models turned out to be insufficient, but several novel PPAR agonists were identified by using nonlinear prediction models. In particular, GP models based on combinations of graph kernels and established molecular descriptors allow the detection of novel lead compounds. Notably, 2D molecular representations proved to be sufficient and well suited for the task of "scaffold hopping".[20]

Experimental Section

Cell culture and assays: As described previously,[11] Cos7 cells were co-transfected with an expression plasmid for a Gal4 fusion hybrid of PPARγ and a luciferase-encoding reporter plasmid. After incubation with test compounds, cells were assayed for reporter gene activity by luminescence measurements.

Compound 8: ^{1}H NMR (250 MHz, [D$_6$]DMSO): $\delta = 7.40$ (m, 10 H), 7.15 (s, 1 H), 7.05 (m, 2 H), 6.04 (d, $J = 4.8$, 1 H), 4.56 (t, $J = 6.5$, 1 H), 4.36 (t, $J = 6.5$, 1 H), 4.14 (t, $J = 6.5$, 1 H), 3.82 (t, $J = 6.5$, 1 H), 1.71 ppm (s, 3 H).

Acknowledgements

This work was supported in part by the FP7-ICT Programme of the European Community (PASCAL2 Network of Excellence, ICT-216886), DFG grant MU 987/4-1, the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, and the LOEWE Lipid Signaling Forschungszentrum Frankfurt (LiFF). M.R. gratefully acknowledges support from the Frankfurt International Research Graduate School for Translational Biomedicine (FIRST); R.S. from the Else-Kroener-Fresenius-Stiftung and FIRST. We thank Peter Vascovic and Fabian Rathke for implementing interfaces between the GP toolbox and ISOAK code, and Prof. Dr. Michael Lämmerhofer (Vienna University) for enantioselective analytics.

Keywords: drug design · machine learning · natural products · NMR · virtual screening

[1] a) T. Chen, W. Xie, M. Agler, M. Banks, *Assay Drug Dev. Technol.* **2003**, *1*, 835–842; b) F. Chang, L. A. Jaber, H. D. Berlie, M. B. O'Connell, *Ann. Pharmacother.* **2007**, *41*, 973–983.

[2] C. Rasmussen, C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, 2006.

[3] a) K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, *IEEE Neural Netw.* **2001**, *12*, 181–201; b) S. Sonnenburg, G. Rätsch, C. Schäfer, B. Schölkopf, *J. Mach. Learn. Res.* **2006**, *7*, 1531–1565.

[4] M. Rupp, E. Proschak, G. Schneider, *J. Chem. Inf. Model.* **2007**, *47*, 2280–2286.

[5] Y. M. Chi, M. Nakamura, T. Yoshizawa, X. Y. Zhao, W. M. Yan, F. Hashimoto, J. Kinjo, T. Nohara, S. Sakurada, *Biol. Pharm. Bull.* **2005**, *28*, 1776–1778.

[6] a) R. Mohan, R. A. Heyman, *Curr. Top. Med. Chem.* **2003**, *3*, 1637–1647; b) Y. Landry, J. P. Gies, *Fundam. Clin. Pharmacol.* **2008**, *22*, 1–18.

[7] B. P. Kota, T. H. Huang, B. D. Roufogalis, *Pharmacol. Res.* **2005**, *51*, 85–94.

[8] a) P. Gervois, I. P. Torra, J. C. Fruchart, B. Staels, *Clin. Chem. Lab. Med.* **2000**, *38*, 3–11; b) S. Kersten, *PPAR Res.* **2008**, 132 960.

[9] O. Rau, H. Zettl, L. Popescou, D. Steinhilber, M. Schubert-Zsilavecz, *ChemMedChem* **2008**, *3*, 206–221.

[10] N. Cho, Y. Momose, *Curr. Top. Med. Chem.* **2008**, *8*, 1483–1507.

[11] a) O. Rau, M. Wurglics, A. Paulke, J. Zitzkowski, N. Meindl, A. Bock, T. Dingermann, M. Abdel-Tawab, M. Schubert-Zsilavecz, *Planta Med.* **2006**, *72*,

881–887; b) D. Poeckel, C. Greiner, M. Verhoff, O. Rau, L. Tausch, C. Hörnig, D. Steinhilber, M. Schubert-Zsilavecz, O. Werz, *Biochem. Pharmacol.* **2008**, *76*, 91–97.
[12] a) J. Liu, H. Li, S. H. Burstein, R. B. Zurier, J. D. Chen, *Mol. Pharmacol.* **2003**, *63*, 983–992; b) S. Ulrich, S. M. Loitsch, O. Rau, A. von Knethen, B. Brüne, M. Schubert-Zsilavecz, J. M. Stein, *Cancer Res.* **2006**, *66*, 7348–7354.
[13] a) A. Schwaighofer, T. Schroeter, S. Mika, J. Laub, A. Ter Laak, D. Sülzle, U. Ganzer, N. Heinrich, K.-R. Müller, *J. Chem. Inf. Model.* **2007**, *47*, 407–424; b) T. Schroeter, A. Schwaighofer, S. Mika, A. Ter Laak, D. Sülzle, U. Ganzer, N. Heinrich, K.-R. Müller, *J. Comput. Aided Mol. Des.* **2007**, *21*, 485–498.
[14] O. Obrezanova, G. Csányi, J. Gola, M. Segall, *J. Chem. Inf. Model.* **2007**, *47*, 1847–1857.
[15] K. Hansen, F. Rathke, T. Schroeter, G. Rast, T. Fox, J. Kriegl, S. Mika, *J. Chem. Inf. Model.* **2009**, *49*, 1486–1496.
[16] A. Schwaighofer, T. Schroeter, S. Mika, K. Hansen, A. Ter Laak, P. Lienau, A. Reichel, N. Heinrich, K.-R. Müller, *J. Chem. Inf. Model.* **2008**, *48*, 785–796.
[17] C. Rücker, M. Scarsi, M. Meringer, *Bioorg. Med. Chem.* **2006**, *14*, 5178–5195.
[18] J. Nichols, D. Parks, T. Consler, S. Blanchard, *Anal. Biochem.* **1998**, *257*, 112–119.
[19] Asinex Europe BV, Laan van Vredenoord 33, 2289 DA Rijswijk (The Netherlands), 2008: http://www.asinex.com (accessed December 14, 2009).
[20] G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem. Int. Ed.* **1999**, *38*, 2894–2896; *Angew. Chem.* **1999**, *111*, 3068–3070.
[21] MOE: Molecular Operating Environment, version 2007.09, Chemical Computing Group, 2007: http://www.chemcomp.com (accessed December 14, 2009).
[22] a) A. Ghose, G. Crippen, *J. Comp. Chem.* **1986**, *7*, 565–577; b) A. Ghose, G. Crippen, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35; c) A. Ghose, A. Pritchett, G. Crippen, *J. Comp. Chem.* **1988**, *9*, 80–90; d) V. Viswanadhan, A. Ghose, G. Revankar, R. Robins, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.
[23] S. K. Singh, A. N. Kesari, R. K. Gupta, D. Jaiswal, G. Watal, *J. Ethnopharmacol.* **2007**, *114*, 174–179.
[24] R. D. Hartley, W. H. Morrison III, F. Balza, G. H. N. Towers, *Phytochemistry* **1990**, *29*, 3699–3703.

Received: November 15, 2009
Revised: December 10, 2009
Published online on December 30, 2009

