universität
wien

DISSERTATION

Titel der Dissertation

„Combined in silico/in vitro screening tools for identification of new insulin receptor ligands“

Verfasserin

Dipl.-Ing. (FH) Daniela Digles

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien, 2011

Studienkennzahl lt. Studienblatt:	A 091 490
Dissertationsgebiet lt. Studienblatt:	Dr.-Studium der Naturwissenschaften Molekulare Biologie
Betreuerin / Betreuer:	Univ.-Prof. Dr. Gerhard F. Ecker

Acknowledgements

This work would not have been possible without the help and support by a lot of people. First, I would like to thank my supervisor Prof. Ecker, who caught my fascination for pharmacoinformatics. He was always the source of good advise and new ideas, and encouraged me to present my work on conferences. I also want to thank my second supervisor Prof. Dirsch, for suggesting this interesting topic and providing me the opportunity to perform the experimental part of this work myself in her lab. Although both are very busy with their work, they always found time for discussions and meetings.

I had the luck to work not only in one, but in two fantastic groups, the Pharmacoinformatics Research Group at the Department of Medicinal Chemistry and the Molecular Targets Group at the Department of Pharmacognosy. I am very thankful to Dominik Kaiser and Michael Demel for their patience and help with all the questions I had at the beginning of my thesis and to Christoph Waglechner for supporting me while I was learning programming. As already during my diploma thesis, Elke Heiß was always available for questions and discussions regarding the cell culture work. Also, she and Marta Pinto gave me valuable feed-back on the manuscript of this thesis. Renate Baumgartner supported me at the beginning of the experimental work, especially on the topic of PTP1B. Matthias Kramer shared his knowledge about the handling and differentiation of adipocytes with me.

But the members of both groups did not only support me in scientific questions, but also provided the basis of a friendly environment where I found myself welcome. Here a special thank goes to Ishrat Jabeen and Yogesh Aher for a lot of discussions, cooking and insights into different cultures. Also to Andrea Schiesaro, with whom I shared the office during most of the time and all other neighbours I had in the last years.

Ein ganz besonderer Dank gilt meiner Familie für ihre Liebe und ihre persönliche, als auch finanzielle Unterstützung. Meiner Mutter Brigitte, für ihre Motivation und Zuversicht. Meinem Vater Günther, der mein Interesse an Computern geweckt und gefördert hat. Meinem Bruder Dominik, für unsere spätabendlichen Diskussionen über Programmiersprachen und andere Themen. Und ganz besonders Martin, der mir die ganze Zeit über zur Seite gestanden ist, mich nach Rückschlägen immer wieder aufgemuntert hat und mir ein Ruhepol in all der Hektik meiner Arbeit war und ist.

I am very thankful to the Austrian Academy of Sciences for supporting my work with a DOC-fFORTE-fellowship.

Contents

1 **Introduction** **1**
- 1.1 Diabetes mellitus . . . 1
- 1.2 The insulin receptor . . . 4
  - 1.2.1 Structure of the insulin receptor . . . 4
  - 1.2.2 Activation of the insulin receptor . . . 5
  - 1.2.3 Insulin receptor signalling . . . 8
  - 1.2.4 Insulin Resistance . . . 10
- 1.3 Small molecule modulators of the IR . . . 10
  - 1.3.1 Demethylasterriquinone B-1 . . . 11
  - 1.3.2 Other insulin receptor modulators . . . 13
  - 1.3.3 Other kinase activating compounds . . . 17
- 1.4 Virtual screening . . . 17
  - 1.4.1 Virtual screening for anti-diabetic compounds . . . 19
  - 1.4.2 Similarity methods . . . 21
  - 1.4.3 Self-organizing maps . . . 22
- 1.5 Aims . . . 25

2 **Methods and materials** **26**
- 2.1 Computational part . . . 26
  - 2.1.1 Databases . . . 26
  - 2.1.2 Descriptors . . . 27
  - 2.1.3 Self-organizing maps . . . 28
  - 2.1.4 Fingerprint similarity . . . 31
  - 2.1.5 Shape similarity . . . 35
  - 2.1.6 Compound selection . . . 36
  - 2.1.7 Docking . . . 37
- 2.2 Experimental part . . . 38
  - 2.2.1 Molecules for testing . . . 38
  - 2.2.2 Media and solutions . . . 38
  - 2.2.3 Technical equipment . . . 42
  - 2.2.4 Cultivation of the cells . . . 43

2.2.5 Western blotting 43
2.2.6 Crystal violet assay 45
2.2.7 Glucose uptake assay 46
2.2.8 PTP1B inhibition assay 47

3 Results 49
3.1 Computational part 49
3.1.1 Self-organizing maps 49
3.1.2 Fingerprint similarity 63
3.1.3 Shape similarity 79
3.1.4 Selection of compounds for testing 81
3.1.5 Docking 87
3.2 Experimental part 93
3.2.1 Preliminary studies with DMAQ-B1 93
3.2.2 Screening for active compounds 94
3.2.3 Further characterization of active compounds 99

4 Discussion 103
4.1 Computational methods 103
4.2 Biological investigations 107

5 Conclusions 110

Bibliography 111

A Appendix 123
Databases 123
Self-organizing maps 141
Abstract 148
Zusammenfassung 149
Curriculum Vitae 150

1

Introduction

1.1 Diabetes mellitus

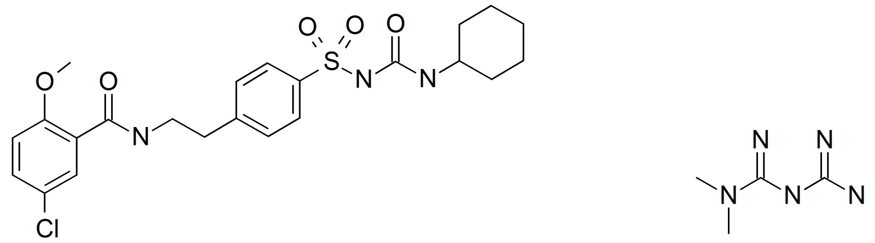
Insulin is a hormone involved in the maintenance of normal blood glucose levels. When blood glucose levels are elevated, for example due to the intake of a meal, insulin is secreted by the β-cells of the pancreas. Insulin then leads to the uptake of glucose into insulin sensitive tissue (liver, muscle and fat), thus reducing the blood glucose levels.

Diabetes mellitus is a chronic disease which is characterized by the lack of, or resistance to, insulin action and consequently elevated blood glucose levels. Health surveys carried out by Statistik Austria in 2006/2007 showed that 390 000 people in Austria are suffering from it, with 91% of them receiving treatment or medication.[1] The number of cases increases with age. From people older than 75 years, 23% of females and 19% of males had encountered diabetes, while the average is 6%. Worldwide, the number of cases in the year 2000 was estimated to be approximately 171 millions, and due to the increasing aging and urbanization of the population this number is expected to double until 2030.[2]

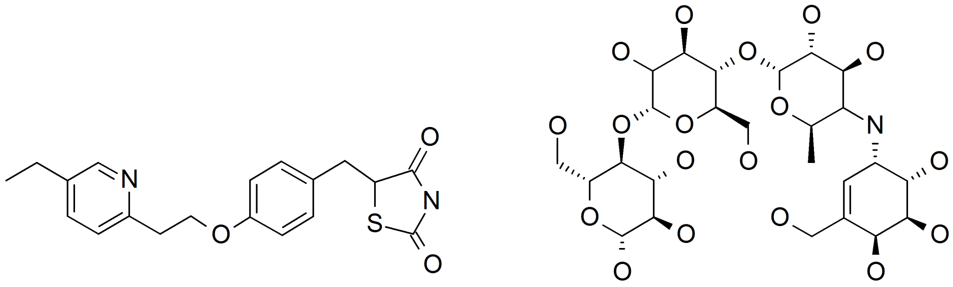
Diabetes mellitus is categorized in different types, with type 1 and 2 being the most prominent ones. Type 1 diabetes mellitus is characterized by an absolute deficiency of insulin. Here, the β-cells of the pancreas are destroyed by the immune system. This stops the production of insulin, which then needs to be provided from external sources. In contrast, type 2 diabetes mellitus (also non-insulin-dependent diabetes mellitus or mature onset diabetes) is characterized by a relative insulin deficiency. This relative deficiency can be caused by a decrease of insulin production, but more importantly by a resistance of the target cells to insulin. The resulting hyperglycemia increases the risk of microvascular damage such as retinopathy, nephropathy and neuropathy, as well as of macrovascular complications like ischaemic heart disease and stroke.

The reasons for diabetes are diverse and can include genetic as well as environmental causes. Major risk factors for type 2 diabetes mellitus are obesity and the lack of physical exercise, but also genetic factors have been shown to play a role in several subtypes of the disease.[3] A recent study showed that a high fat diet fed to male rats can lead to impaired insulin secretion and glucose tolerance in their female offspring, which could also show a role of epigenetics in type 2 diabetes.[4,5] But the exact mechanism in which insulin resistance evolves are not clear. Two main theories are currently available.[6] The first is that excess lipids can not be stored in fat tissue anymore and thus accumulate in muscle and liver cells instead, causing toxic effects in these cells. The other theory states that adipocytes release inflammatory cytokines, which then cause insulin resistance in other tissues.

Although lifestyle modifications such as a low-fat diet and increased physical exercise can already lead to a improved insulin sensitivity, additional medication is necessary in most of the cases. The classical treatments of type 2 diabetes mellitus include 4 main classes. The sulfonylureas increase the patient's insulin secretion from the pancreas by increasing the β-cell's glucose sensitivity. Representatives of this class are glibenclamide, gliclazide, glipizide and glimepiride. The biguanides (for example metformin and phenformin) reduce the hepatic glucose production. Thiazolidinediones (glitazones) are thought to be agonists of peroxisome proliferator-activated receptor-γ (PPARγ), thus enhancing the action of insulin. Examples for this class of compounds are pioglitazone and rosiglitazone. Rosiglitazone was recommended to be taken off the market by the European Medicines Agency in 2010 due to an increased risk of cardiovascular complications.[7,8] Acarbose is an inhibitor of α-glucosidase, which diminishes the blood glucose levels after meals by inhibiting the uptake of glucose in the gut. Structures of selected anti-diabetic drugs are shown in figure 1.1.



(a) Glibenclamide (sulfonylurea) (b) Metformin (biguanide)



(c) Pioglitazone (thiazolidinedione) (d) Acarbose

Figure 1.1: Examples for anti-diabetic compounds

Insulin and its analogues, which are generally used as treatment in type 1 diabetes, can also be used in some cases of type 2 diabetes.

Treatments with anti-diabetic compounds have sometimes severe side-effects, including weight gain, hypoglycemia, gastrointestinal problems, lactic acidosis, edema and anemia. Since type 2 diabetes mellitus is often associated with obesity, new approaches with a loss of weight, or at least no additional weight gain would be beneficial. Several newer targets are currently under investigation, leading already to some new drugs approved for the market. Among these are amylin analogs, peroxisome proliferator-activated receptor-α/γ (PPAR-α/γ) agonists, sodium-dependent glucose transporter inhibitors and fructose bisphosphatase inhibitors.[9,10] One target exemplified here in more detail are incretin mimetics and enhancers. Incretins are hormones increasing the insulin secretion, thus showing glucose lowering activity. They are said to aid the regeneration of insulin-secreting cells in the pancreas and to show heart protecting properties. Examples for incretin mimetics are exenatide, liraglutide, taspoglutide and lixisenatide, which are analogues of Glucagon-Like Peptide-1 (GLP-1). Endogenous incretins are rapidly degraded by dipeptidyl peptidase 4 (DPP-4). The gliptins (e.g. vildagliptin, sitagliptin, saxagliptin) are inhibitors of DPP-4 and thereby enhance the activity of the incretins.

1.2 The insulin receptor

The physiologic responses to the presence of insulin in the blood stream are mainly initiated by the binding of insulin to its receptor, which then leads to the activation of several signalling cascades. In the following sections, a brief overview on the structure and activation mechanism of the insulin receptor, the main downstream signalling pathways as well as possible reasons for insulin resistance is given.

1.2.1 Structure of the insulin receptor

The insulin receptor (IR, INSR) is a receptor protein-tyrosine kinase (EC 2.7.10.1). These enzymes pass on signals from their extracellularly bound ligands to the inside of the cell by transferring phosphate groups from donor molecules such as ATP to tyrosine residues of their substrates. Receptors belonging to the same subfamily as the insulin receptor are the insulin-like growth factor 1 receptor (IGF1R) and the insulin receptor-related protein (INSRR).[11] The human insulin receptor precursor (UniProt-ID: P06213) consists of a short signal peptide and two subunits of the insulin receptor, the α- and the β-chain. Numbering of the amino acids is varying, depending on whether the signal peptide is included or excluded. In the present work, the numbering without the signal peptide is used. To get the UniProt numbering, 27 has to be added to the amino acid number. Two different isoforms are produced by alternative splicing: the isoform Long (HIR-B) and the isoform Short (HIR-A), which misses 12 amino acids in the α-chain.

During the maturation process, the insulin receptor precursor is N-glycosylated and intra- and intermolecular disulfide bridges are formed in the endoplasmic reticulum. The α- and β-chains are subsequently cleaved at the trans-Golgi network and further glycosylation occurs before the mature receptor is finally transported to the plasma membrane. The functional form of the insulin receptor consists of two disulfide-linked α,β-dimers. The α-subunits are extracellular and contain the insulin binding site. The two β-monomers each have a single transmembrane helix. The C-terminal region is intracellular and contains the kinase domain which is responsible for the activity of the insulin receptor (figure 1.2)

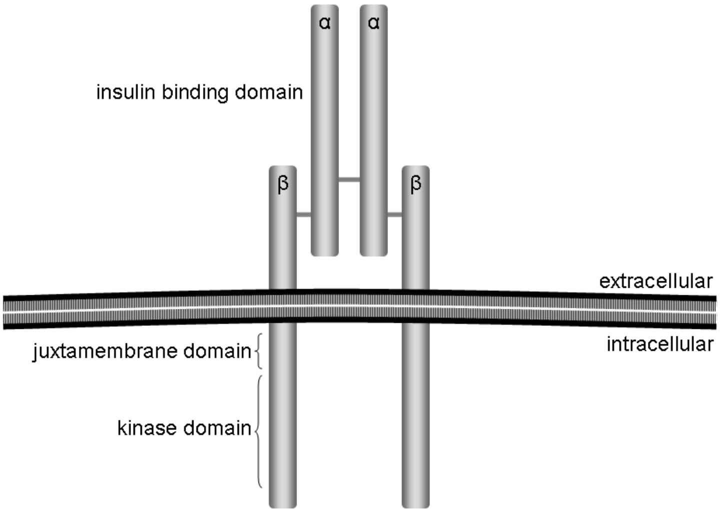


Figure 1.2: Schematic representation of the insulin receptor.

X-ray structures of the extracellular domain,[12] as well as the intracellular kinase domain[13–15] have been resolved. A structure of the whole insulin receptor was determined using electron microscopy.[16]

1.2.2 Activation of the insulin receptor

The first step of the activation of a receptor tyrosine kinase is in general the binding of its ligand to the extracellular domain of the receptor. In many cases this is thought to stabilize the dimerized state of two receptor monomers. Dimerization is necessary to bring the two kinase domains close to each other, so that trans-phosphorylation of the subunits can occur. But dimerization alone is not the only prerequisite for the activation of the receptor.[17] Binding of the ligand leads to a conformational change, which subsequently leads to the activation of the intracellular kinase domain. For the epidermal growth factor receptor (EGFR) it was shown, that it can exist as an inactive dimer on the cell surface without its ligand.[18] Binding of the ligand might lead to a movement in the transmembrane and juxtamembrane region, bringing the two kinase domains in the right distance for autophosphorylation. Another important structural feature in many kinases is the so called activation-loop. Structural rearrangement of this loop which is often associated with the phosphorylation of an amino acid residue can be necessary for activation of the intrinsic kinase.

In the case of the insulin receptor, the dimerization is not necessary as the subunits are already linked by disulfide bonds. Binding of insulin to the extracellular part of its receptor induces a conformational change followed by trans-phosphorylation of Tyr1158, Tyr1162 and Tyr1163 within the activation-loop. In the inactive state, the activation-loop sterically blocks the access for the protein substrate and the ATP binding pocket. Phosphorylation of the tyrosine residues leads to a conformational change of the activation-loop (see figure 1.3), which exposes the binding site and activates the tyrosine kinase domain.[13, 14]

Recent crystallographic studies identified a possible binding pocket for an insulin receptor activator binding to the intracellular domain.[19] They showed the possible role of an additional tyrosine residue (Tyr984) for the activation of the insulin receptor. This tyrosine, which is conserved in all insulin receptor proteins, seems to be important for the autoinhibition of the kinase domain. Tyr984 is positioned in the juxtamembrane region next to the kinase domain. Figure 1.4 shows the N-terminal lobe of the insulin receptor kinase domain in the active and inactive conformation. In a crystal structure of the inactive state, Tyr984 is located in a hydrophobic pocket of the tyrosine kinase domain, whereas in the active state it is not. The active and inactive structures of the insulin receptor also show a movement of the αC helix in the N-terminal lobe of the kinase domain. Stabilization

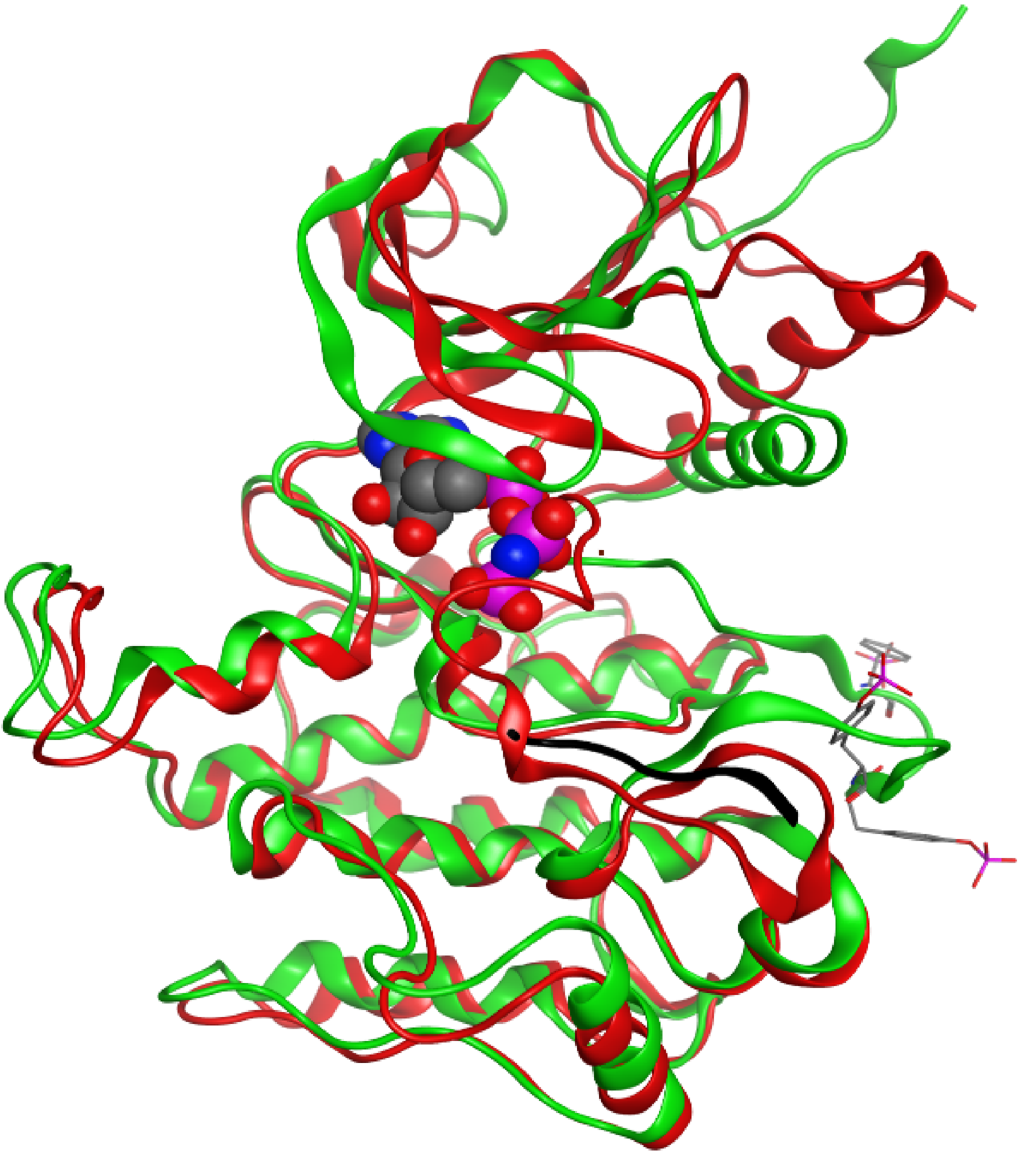


Figure 1.3: Comparison of the active (1IR3, green) and inactive (1IRK, red) conformation of the insulin receptor kinase domain. An ATP analogue and a substrate peptide bound to the activated state are depicted as space filling molecule and a black ribbon, respectively. Phosphorylated tyrosine residues on the activation loop are depicted in stick representation.

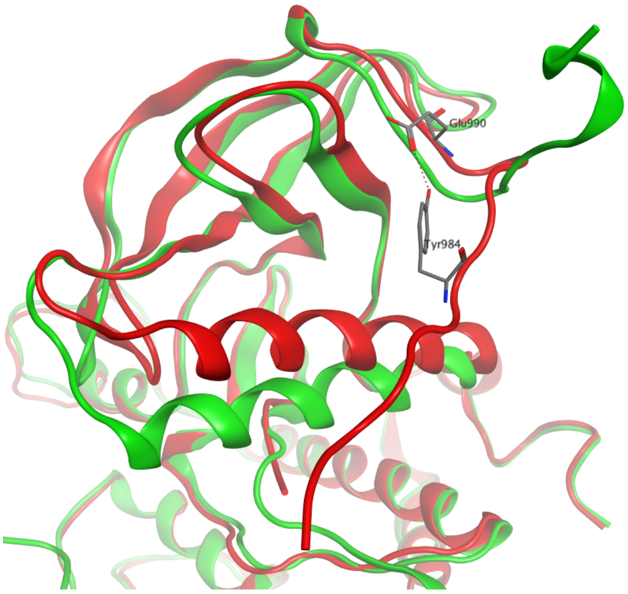


Figure 1.4: Comparison of the active (1IR3, green) and inactive (1P14, red) conformation of the N-terminal lobe of the insulin receptor kinase domain. Tyr984 of the inactive conformation forms a hydrogen bond to Glu990. Additionally, a movement of the αC helix can bee seen.

of the helix in its active position might be important for the positioning of amino acids in the catalytic site. It was proposed that the binding of small molecules to the hydrophobic pocket between the αC helix and the β-sheets can displace Tyr984, leading to insulin receptor activation.[19]

1.2.3 Insulin receptor signalling

Following the activation of the insulin receptor, several substrate proteins can bind to and are phosphorylated by the receptor, leading to different signalling pathways. Insulin receptor signalling has been the topic of several reviews.[20–27] The two main downstream signalling cascades are the phosphatidylinositol 3-kinase (PI3K)–Akt pathway and the Ras–MAP kinase pathway. A simplified overview of the insulin signalling can be seen in figure 1.5.

Different to other receptor tyrosine kinases such as the EGF receptor and the PDGF receptor, the insulin receptor does not provide a docking

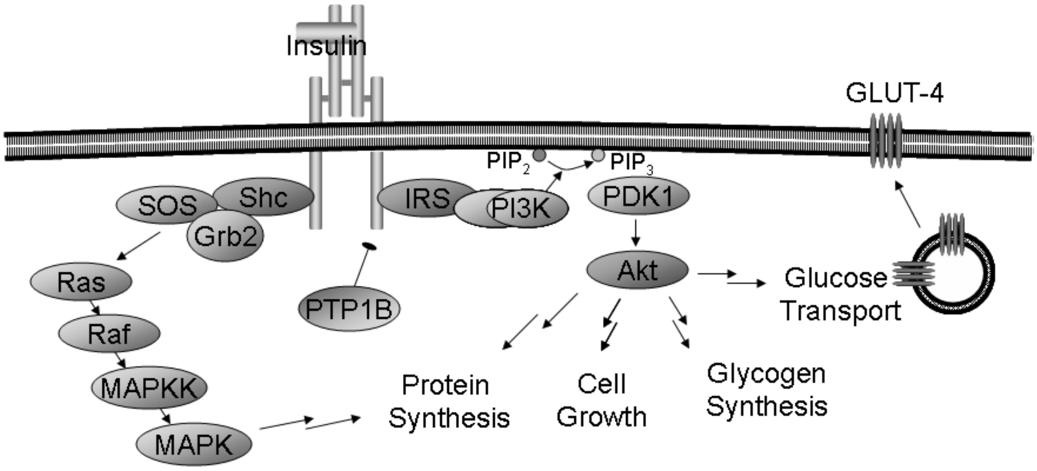


Figure 1.5: Overview of the main steps of the insulin signalling pathway. Adapted from reference 27.

site for signal transduction proteins itself, but uses additional substrate proteins instead. The most prominent of these direct substrates are the insulin receptor substrate proteins (IRS proteins),[26,28] with IRS1 and IRS2 being the most widely distributed ones. The IRS proteins have a pleckstrin-homology (PH) domain which mediates binding to membrane phospholipids and/or a phosphotyrosine-binding (PTB) domain with which they can bind to the phosphorylated insulin receptor. Additionally, they contain several tyrosine residues which can be phosphorylated by the insulin receptor, and are subsequently used as docking sites for proteins showing a Src-homology-2 (SH2) domain, such as the regulatory subunit of phosphatidylinositol 3-kinase (PI3K), growth-factor-receptor-bound protein-2 (Grb2) or SH2-domain containing tyrosine phosphatase-2 (SHP2). IRS proteins can be negatively regulated by serine phosphorylation.

Activation of PI3K signalling leads to many different insulin regulated metabolic effects, such as glucose uptake in muscle and adipose tissues, protein synthesis and glycogen synthesis. The catalytic subunit of PI3K generates phosphatidylinositol-3,4,5-triphosphate (PIP_3) from the 4,5-bisphosphate at the plasma membrane. Activation of 3-phosphoinositide-dependent protein kinase-1 (PDK1) and other kinases leads to the activation of Akt (PKB), which subsequently phosphorylates several targets. One example is the glycogen synthase kinase 3 (GSK3), which is inhibiting glycogen synthase. Akt-mediated phosphorylation inhibits GSK3 and thereby leads to the storage of glucose as glycogen.

The Ras–MAPK pathway leads to gene expression, cell growth, survival and differentiation. Ras is a small GTPase, which is activated following the association of the guanyl nucleotide-exchange factor son-of-sevenless (SOS) to Grb2, which is in turn bound to a insulin receptor substrate protein such as Shc (Src-homology-2-containing protein) or Gab1 (Grb2-associated binder-1). This initiates a cascade, where in turn Ras, Raf, mitogen-activated protein (MAP) kinase kinase (MAPKK) and MAP kinase (MAPK, ERK) are sequentially activated.[26,29]

1.2.4 Insulin Resistance

When the biologic response to insulin is smaller than normal, one speaks of insulin resistance. Insulin resistance can have many causes, which can be either intrinsic to the target cells, or can be external factors affecting the sensitivity of the target tissue.[30] Intrinsic defects can be mutations of the insulin receptor which are associated with syndromes of severe insulin resistance, such as type A syndrome, Rabson-Mendenhall syndrome and leprechaunism. Several external factors can influence the sensitivity of the target tissue to insulin, for example free fatty acids or insulin as well as physiologic states such as fasting, pregnancy or obesity. In the case of external factors, the insulin sensitivity is usually restored after removal of these factors. High levels of insulin (hyperinsulinemia) for example, down-regulate the receptor, thus rendering the cells less sensitive to it.[30] Obesity is one of the major risk factors for insulin resistance and type 2 diabetes, and weight loss was shown to improve insulin sensitivity.[31]

1.3 Small molecule modulators of the IR

As the insulin receptor plays a key role in the signalling of insulin, compounds activating the insulin receptor might be a possible treatment for both type

1 and 2 diabetes mellitus. A small molecule could be an orally available alternative to insulin, and a different activation mode might be able to activate insulin receptors insensitive to insulin. Currently known activators and modifiers of the insulin receptor as well as other kinases will be presented in the following sections.

1.3.1 Demethylasterriquinone B-1

In 1999, Zhang et al.[32] discovered a small molecule (**1**) from a fungal extract (L-783,281 or demethylasterriquinone B-1, DMAQ-B1, Merck L7, see figure 1.6) in a cell based screening with Chinese hamster ovary cells overexpressing the human insulin receptor (CHO.IR). After treatment of the cells, the insulin receptors were immunopurified and the activity of the tyrosine kinase domain was determined. In this assay, DMAQ-B1 activated the tyrosine kinase of the insulin receptor with an EC50 value of 3–6 µM. Additionally, it was able to enhance the action of insulin when administered in lower concentrations (0.6–2 µM). Unlike insulin, DMAQ-B1 was shown not to bind to the extracellular part of the receptor, but to the intracellular domain.[32] Two research groups state, based on unpublished results, that the activation of the insulin receptor can not be attributed to inhibition of phosphatases,[32] or more specifically, PTP1B,[33] the major negative regulator of IR phosphorylation.

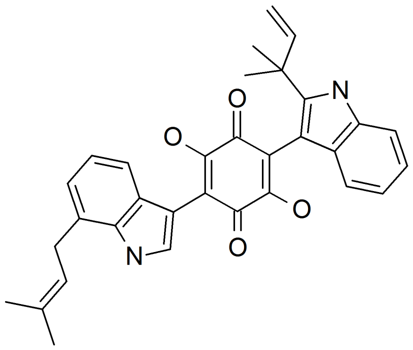


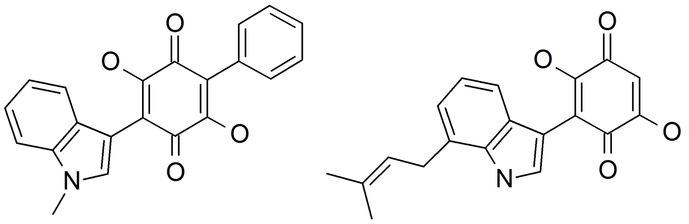
Figure 1.6: Demethylasterriquinone B-1 (**1**)

Compound **1** is not only able to activate the insulin receptor, but also leads to downstream effects of the insulin signalling pathway. Treatment with the compound led to increased PI3K activity and Akt phosphorylation in CHO.IR cells.[32] Glucose uptake was increased in rat primary adipocytes,[32] vascular smooth muscle cells,[34] 3T3L1 adipocytes[35] and mouse soleus muscle.[32] It showed glucose lowering effects in different mouse models of type 2 diabetes[32] and enhanced insulin induced IR autophosphorylation in cell models of insulin resistance, where the binding of insulin does not result in the necessary conformational change.[36]

Still, several differences of the action of insulin and **1** were reported. While insulin induces proliferation of vascular smooth muscle cells (VSMCs), DMAQ-B1 does not. This could be beneficial if used as an anti-diabetic agent as it may decrease the development of atherosclerosis.[34] In hIRcB fibroblasts, compound **1** showed a higher phosphorylation of Akt than of IR, PI3K, and ERK1 and 2, while insulin activated all of the kinases to a similar extent.[35] Similarities and differences of the action of insulin and **1** on gene expression in HepG2 hepatoma cells were also investigated using microarrays.[35] The expression of several genes was different, which could in some cases be explained by cytotoxic effects of DMAQ-B1. The higher activation of Akt by **1** could be explained by a down-regulation of a subunit of the phosphatase PP2A, which is dephosphorylating Akt and is up-regulated by insulin.[35] To identify additional targets of **1**, phage display cloning was performed with a biotinylated derivative. This led to the identification of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) as a target.[37]

In the following years, more than 300 derivatives of DMAQ-B1 were synthesized and tested for their ability to activate the insulin receptor. The activities of about 100 of these structures (compounds **2**–**102**) have been published.[33,35,38–46] A list of these molecules can be found in table A.1 on page 123.

It was shown that changes in the prenyl groups had less effect than changes in the dihydroquinone core, which resulted in a loss of activity.[40] With the aim to find regions of the molecule which were not crucial for activity and to enable the formation of a biotin conjugate as affinity reagent, methyl scanning was performed. Here, methyl groups were introduced to different positions of the molecule, showing that the 7-substituted indole, except for positions 1 and 2, and the OH groups did not tolerate the introduction of a methyl group without loss of activity.[33] Additionally, compounds with a simplified structure were found which still activate the insulin receptor. Compound **15** (**2h** in reference 38) is more active than DMAQ-B1, and has additionally a higher selectivity for the insulin receptor compared to homologous receptors.[38] Also, one of the indole rings can be omitted, as was done for example for compound **65** (ZL-196 in reference 43).



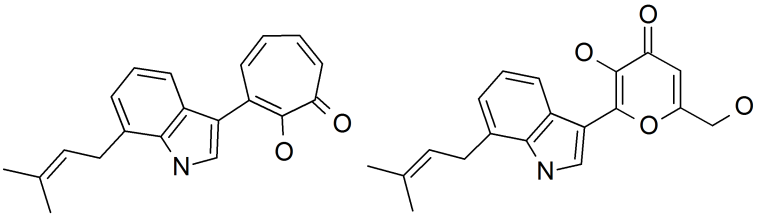
(a) **15** (b) **65**

Figure 1.7: Examples for active derivatives of DMAQ-B1 with a simplified structure.

Since all of the active derivatives at this point contained the quinone scaffold, which might be problematic when used chronically, effort has been made to find replacements for this structural feature (see figure 1.8). This led to the identification of an active kojic acid derivative of **65**, while a tropolone derivative was inactive.[45] Recently, a hydroxyfuroic acid derivative of **1** has been developed, which shows insulin receptor activation, is less cytotoxic than **1** and shows inhibition of epidermal growth factor.[46] These studies show that the replacement of the quinone scaffold is in principle possible.

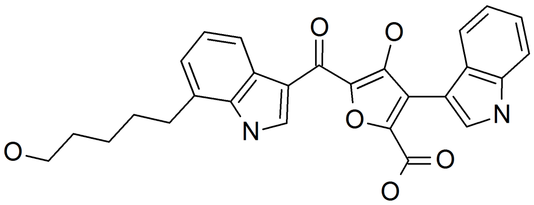
1.3.2 Other insulin receptor modulators

Several compounds able to mimic or increase the action of insulin, or to modify the activity of the insulin receptor have been identified in the last years. Glucose was shown to have the ability to bind to regions of the insulin receptor which resemble insulin.[47] It was shown to lower the binding of insulin to the receptor, although this effect might be due to glucose binding to insulin directly. The insulin receptor changes its conformation depending



(a) **99**: tropolone derivative (inactive)

(b) **100**: kojic acid derivative (active)



(c) **102**: hydroxyfuroic acid derivative (active)

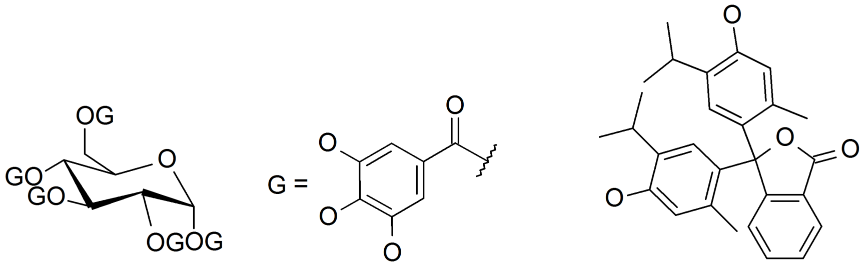
Figure 1.8: Attempts to replace the quinone substructure.

on the glucose concentration, as demonstrated using UV spectrophotometry. Still, this effect might also be due to interaction of glucose with GLUT 1, which then allosterically modifies the insulin receptor.[47] Pentagalloylglucose (PGG, fig. 1.9a), which is an ester of glucose with five galloyl groups, was shown to probably bind to the α-subunit of the insulin receptor to a site different to the insulin binding site. It is a partial agonist for glucose transport activity, but displaces insulin from its binding site.[48]

With the aim to identify molecules binding to the insulin binding site, thymolphthalein (fig. 1.9b) was identified as a weak activator of the insulin receptor. Derivatives, such as erythrosine and iodophenol blue, were found to compete stronger with insulin, but had no effect on or inhibited autophosphorylation of the receptor.[49]

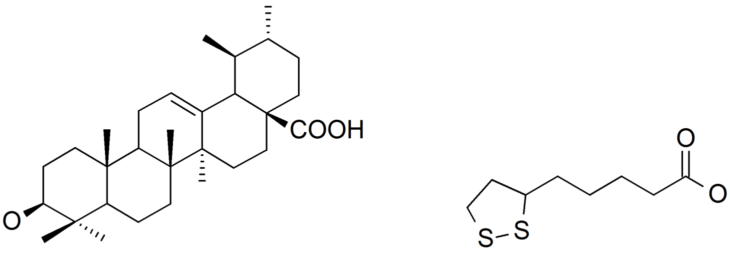
The role of the anti-diabetic drug metformin (fig. 1.1b) in the activation of the insulin receptor is controversial. Metformin was reported to activate the intracellular domain of the insulin receptor β-subunit to 20–30% at therapeutic concentrations.[50]

Ursolic acid (fig. 1.9c) was found to activate the insulin receptor and to increase the action of insulin.[51] It has been suggested that the compound



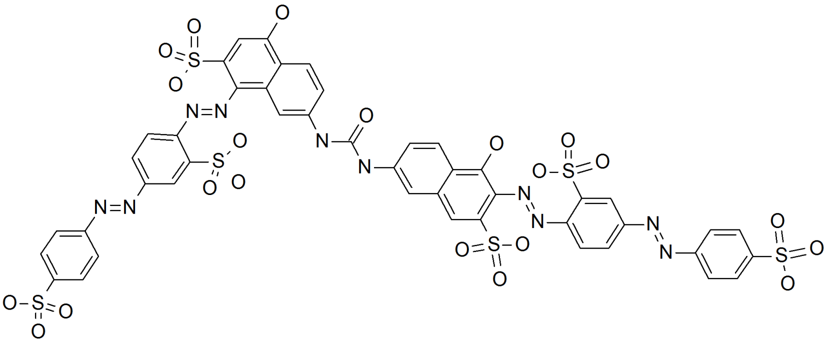
(a) α-Pentagalloylglucose (PGG)

(b) Thymolphthalein



(c) Ursolic acid

(d) α-Lipoic acid



(e) TLK16998

Figure 1.9: Examples for insulin receptor modulating compounds

acts by binding to the intracellular domain of the IR, but the mode of action has not been shown yet to be identical to DMAQ-B1. Besides, earlier studies showed that ursolic acid is an inhibitor of PTP1B.[52]

α-Lipoic acid (fig. 1.9d) is another molecule reported to directly activate the insulin receptor. It was shown to have anti-apoptotic effects in hepatocytes, probably due to activation of the PI3K–Akt pathway. Molecular dynamics studies of the insulin receptor kinase domain with α-lipoic acid, placed in a binding pocket different to the one proposed in reference 19, showed how the molecule could stabilize the activation loop in its active conformation.[53]

The insulin sensitizer TLK16998 (fig. 1.9e) and derivatives were shown to directly interact with the intracellular domain of the insulin receptor, but different to DMAQ-B1 they do not increase the phosphorylation of the insulin receptor in the absence of insulin.[54,55] Further studies showed more differences in the action of those two compound classes on different cellular models of insulin resistance.[36] Additionally, TLK16998 was shown to increase the IR autophosphorylation induced by compound **1**.[36] Recently, simpler aminonaphthalene-sulfonic acid derivatives with insulin sensitizing effects have been described.[56] Another derivative was described to activate the insulin receptor without insulin.[57]

Vanadate is also known to mimic the biological effects of insulin.[58] But the mode of action of vanadates is thought to be either inhibition of protein tyrosine phosphatases or by another mechanism not involving the phosphorylation of the insulin receptor.[59,60]

The activity of the insulin receptor can also be enhanced by alterations of the receptor itself. Besides the well known activation by phosphorylation of the tyrosine residues of the activation loop, also the oxidation of cystein residues has been reported to play a role. This modification of the insulin receptor by redox processes might be a necessary intermediate state between the inactive and the phosphorylated state.[61] Additionally, trypsination of the insulin receptor was observed to lead to the activation of the receptor.[62,63] It was proposed that trypsin cleaves the α-subunit, probably at the insulin binding site, and thereby leads to a conformational change similar to that induced by insulin.[64]

1.3.3 Other kinase activating compounds

A large proportion of the currently ongoing research on kinases aims at finding new inhibitors. With the identification of DMAQ-B1 as insulin receptor activator, a new focus has been set on finding allosteric activators, although they are still scarce. Currently known allosteric modulators of protein kinases, including activators of growth factor receptors, have been reviewed recently.[65] Work on two exemplary kinases will be sketched in the following.

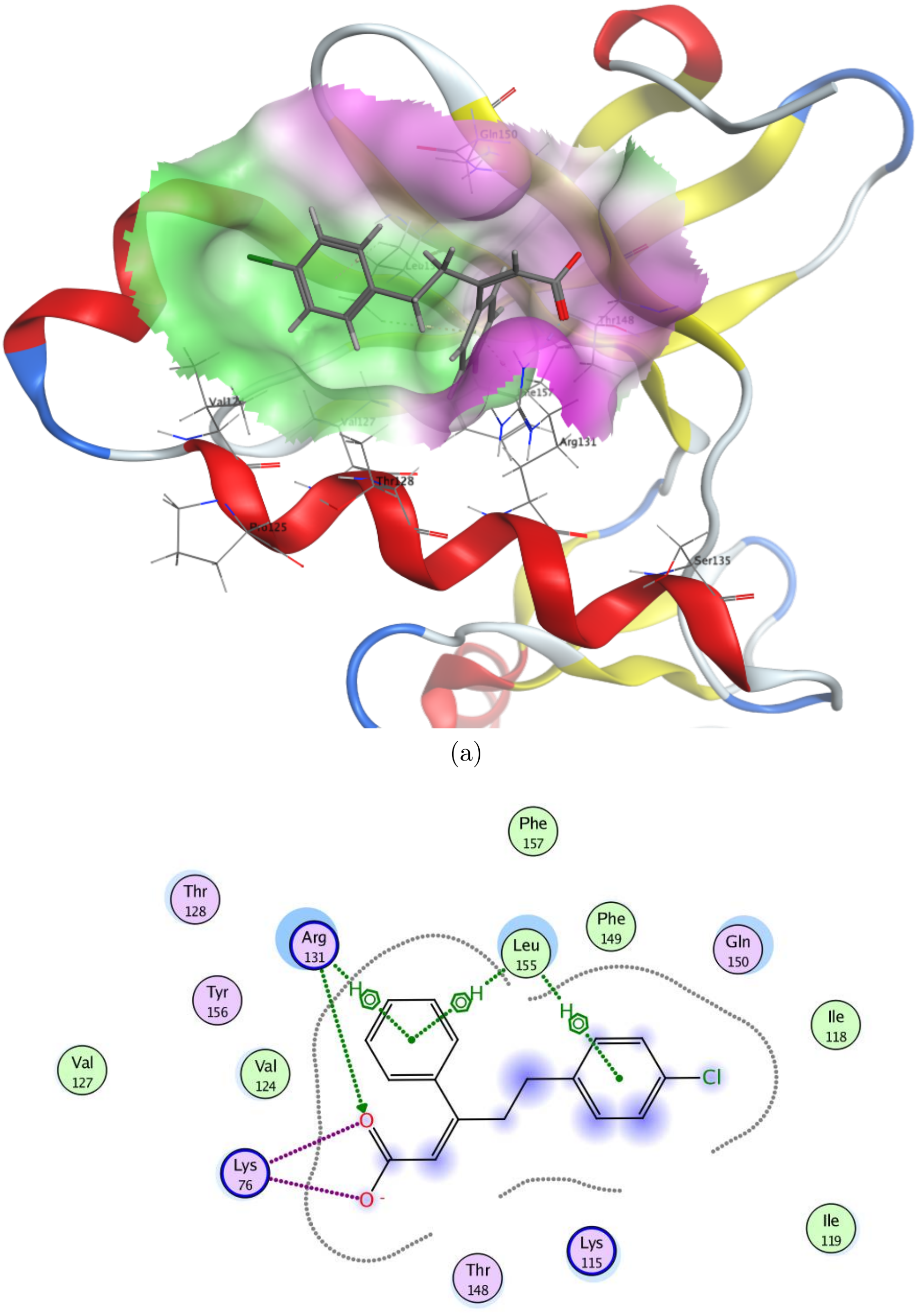
The activation of tropomyosin-related kinase receptor B (TrkB) by brain-derived neurotrophic factor (BDNF) promotes survival, differentiation and function of neurons. Massa et al.[66] used information about regions of BDNF known to be necessary for activating TrkB to build a pharmacophore model. Virtual screening of more than one million compounds with this pharmacophore, led to the identification of small molecules activating TrkB.

Sphingosine was found to activate phosphoinositide-dependent protein kinase 1 (PDK1).[67] Also, a crystal structure of a kinase with a small molecule activator was published by Hindie et al. in 2009 for PDK1.[68] The binding pocket corresponds to the pocket which was proposed as a possible binding site for insulin receptor kinase activating compounds (compare figure 1.10 and figure 1.4 on page 8). By linking small molecule fragments with disulfide bridges to the kinase, Sadowsky et al.[69] identified activators as well as inhibitors at this site.

1.4 Virtual screening

Virtual (or *in silico*) screening is a fast way to select molecules which have a probability to modify a given target with computational methods. It is a good alternative if classical high-throughput screening (HTS) is not available to identify new hits for a target.

In general, two main strategies can be followed. One is to use the information from available ligands of the target (ligand based methods), the other is



(a)

(b)

Figure 1.10: Crystal structure (a) and 2D ligand interactions (b) of PDK1 with a small molecule activator (3HRF).

to implement the information of the target directly (structure or target based methods). While “retrospective” analyses test the performance of the screening methods on known datasets, “prospective” studies include selection of new compounds with screening methods and subsequent experimental evaluation of the hits. This is usually done to identify new classes of molecules, which is also known as “scaffold hopping”. Ripphausen et al.[70] recently investigated the current literature of prospective virtual screening studies. They saw that structure based methods (mainly docking) are used more frequently than ligand based methods. Still, these docking studies are often combined with ligand based filters for the reduction of compounds to screen. Comparing the potency of hits, ligand based methods were in general more successful in finding higher active molecules. Docking into homology models found in general more potent hits than docking into X-ray structures.[70]

In the following sections, first some examples for virtual screening in anti-diabetic research will be given. This is followed by a general introduction to similarity methods and a more detailed description of the usage of self-organizing maps for screening purposes.

1.4.1 Virtual screening for anti-diabetic compounds

Some of the molecules mentioned in chapter 1.3.2 were identified using computational methods. Thymolphthalein was identified using the 3D information from three amino acids of insulin which were known to be important for activity. The 3DB Unity tool from Sybyl was used to find molecules fitting to this query. Subsequently, other molecules were selected using a substructure search.[49] QSAR modelling was performed for some derivatives of **1**,[43] but to our knowledge no *in silico* screening was performed to identify new molecules.

Apart from these studies, previous implementations of computational methods in anti-diabetic research have focused on targets other than the insulin receptor, as reviewed in reference 71. Studied targets include PTP1B, PPARγ, and others.

The phosphatase PTP1B regulates insulin signalling by dephosphorylation of the insulin receptor, thereby leading to its inhibition. Overexpression of PTP1B can therefore lead to insulin resistance. As the activity of the insulin receptor is closely connected with that of PTP1B, virtual screening efforts for this target will be discussed here in more detail.

Many inhibitors of PTP1B have been identified in the last years, some of them with the help of virtual screening. As several crystal structures of PTP1B with substrates and inhibitors are available, the main method of choice is currently docking. The first *in silico* screening was performed with docking using the DOCK program.[72] 150 000 molecules from the Available Chemicals Directory were screened using two different approaches, and the best scored 1000 poses each were manually examined to select 25 molecules for testing. For seven of these, inhibition was measurable at 100 µM. Doman et al. compared classical high-throughput screening with virtual screening using docking.[73] For the classical screening, 400 000 compounds from an in-house collection were tested, leading to 85 molecules with an IC_{50} value between 1 and 100 µM. Flexible docking of the ACD, BioSpecs and Maybridge databases and subsequent evaluation of 365 selected molecules led to the identification of 127 hits with an IC_{50} value below 100 µM. However, selection of the molecules from the top-scored poses for experimental validation included additional considerations, and the docking scores and IC_{50} values showed poor correlation. The hits from the two different approaches showed no structural overlap and the docking hits were in general more drug-like than the hits from HTS.[73] Recently, another application of docking for virtual screening was reported, leading to the identification of nine inhibitors in the range of 10 to 50 µM.[74]

Using a ligand based approach, Taha et al. have been successful in identifying new hits for PTP1B.[75] They built several pharmacophore models, and subsequently built a QSAR model using different descriptors and the fit values of the molecules to the pharmacophores. The best pharmacophore model was used to screen a database, and the hits were further refined, including the estimation of activity using the QSAR equation. From 60 selected hits, five were available for testing. All five molecules showed potent activity, the most active one with an IC_{50} value of 0.47 µM.[75]

In a comparison of ligand similarity based and docking based virtual screening on different targets, the ligand based ones were generally outperforming the docking approaches.[76] An exception was PTP1B, where the docking program Glide performed better than the other methods. In the case studies presented in this section, docking was used more frequently for screening. The ligand based approach[75] was more successful in identifying potent hits, but it was depending on activity data from the literature. A general statement on which type of method works best in the case of PTP1B can not be given though, due to contradictory results.

1.4.2 Similarity methods

Ligand based virtual screening is often done by identifying compounds which are 'similar' to a given query structure. Similarity can be defined in different ways,[77] for example by similarity in property/descriptor values, molecular graph similarity, similarity of fingerprints, or similarity in the three dimensional shape or the pharmacophores of the molecules. The assumption for this type of screening is that molecules which are similar to each other in some way, are also similar in bioactivity. Martin et al.[78] addressed the question whether structurally similar molecules show similar biological activity. Using Unity fingerprints, a Tanimoto similarity larger than 0.85 is generally regarded as having similar activity. According to Martin et al. with this threshold only 30% of the molecules similar to an active one are active themselves, but it gives an at least 30-fold increased enrichment compared to random screening.[78] Also, there are several examples known, where only small changes in the structure lead to dramatic changes in activity.[79] This is also the case for derivatives of **1**, where for example the introduction of a methyl group can lead to loss of activity.[33]

As many different similarity search methods are available, the selection of the method to use is an important step. Ideally, the method should be able to discriminate known active from inactive molecules, thus showing the possibility to identify actives. Additionally, several similarity methods were found to behave different to each other in the identification of active molecules.[80] Therefore, one should not rely on only one method, but use several to increase the chance of finding all actives in a database.

The methods used in our study are self-organizing maps using two different sets of descriptors, similarity of different types of fingerprints, and shape similarity.

1.4.3 Self-organizing maps[1]

In 1982, Teuvo Kohonen developed an artificial neural network able to map representations of signals onto a one- or two-dimensional array, while preserving the topology of the primary events.[82] This neural network is built up by a single layer of neurons placed on an ordered lattice (e.g. a two-dimensional array) with each having a defined set of neighbours. Each neuron contains a number of weights corresponding to the dimensions of the input objects. At the beginning of the training process, the weight matrix is generated randomly. Subsequently, each object is presented repeatedly to the network to determine the winning neuron, i.e. the neuron whose weights are most similar to the input values. The weights of the winning neuron and its neighbours are then adjusted to the input descriptors and subsequently the next object is presented to the network. During the training procedure, the size of the neighbourhood (the number of neurons near to the winning neuron which are additionally influenced by the training), as well as the degree to which the weights are adjusted to the input vectors is decreased. In the final step, each object is placed on the trained map, with objects being close to each other in high dimensional space, as represented by the input vector, ending up in the same or adjacent neurons.

10 years later, this self-organizing (feature) map (SOM, also called Kohonen map) found its first applications in the field of chemistry. Rose et al. used physicochemical parameters of antifilarial antimycin analogues as input vectors and were able to achieve a good separation of activity values, whereas principle component analysis failed due to structural outliers.[83] Gasteiger et al. used this self-organizing network to depict the 3D properties

---

[1] The main part of this section has been published in reference 81.

of a molecular surface on a topology preserving two-dimensional map.[84] A comprehensive overview of the early uses of self-organizing maps in chemistry and drug design can be found in a review by Anzali et al.[85]

In the past decade, self-organizing maps provided exciting results for datasets which are not showing a clear structure activity relationship, for example in classifying substrates for polyspecific targets such as P-Glycoprotein[86–88] or the hERG channel,[89–92] or to cluster compounds according to more complex principles such as toxic/non-toxic[93] or drug like/non-drug like.[94]

The major difference between supervised and unsupervised learning methods is the use of class values during the training process. Supervised methods, such as support vector machines, decision trees, binary QSAR and random forests include class values into the training process and thus allow building a model best suited to separate the given classes. As an unsupervised method, self-organizing maps organize objects solely on basis of the similarity of their input vectors. Thus, they represent a versatile tool for a rapid indicative check on the applicability of a given input vector (descriptor matrix) for the classification problem approached. Subsequently, this information can either be used by utilizing the self-organizing map directly as a screening tool, or by using the descriptors best suited for the separation of actives and inactives as input to other classification methods.[89,95–97]

When applying self-organizing maps for screening, principally two different approaches can be followed. The first one is to train a map with compounds of known activity and subsequently place compounds with unknown activity on the trained map. Although being very fast, this method has the drawback that while placing a new compound on the most similar neuron of the map, the distance might still be too large to gain meaningful information. It is therefore necessary to judge the applicability of the network for the respective screening library. The easiest way to do this is to rank the identified compounds according to the distance to the winning neuron and to examine only the top ranking ones.[98] A more sophisticated way is the usage of novelty detection with self-organizing maps as described by Hristozov et al.[99] In this method, the average distance of each neuron to its neighbours in the trained map is calculated to give a local accuracy. Each new compound placed on the map is then scored depending on its distance to the winning neuron and the local accuracy calculated for this neuron. If the distance is larger than the local accuracy, the compound is classified as novel and is therefore discarded. The distance threshold used to discard compounds for which the activity prediction is likely to be wrong is therefore selected automatically depending on each individual neuron.

The second approach is to train compounds with known and unknown activity together on a larger map. This approach is based on the basic principle that the compounds which are near in the multidimensional space will be placed in the same neuron. Assuming that the used descriptors are able to separate active from inactive compounds, molecules from the screening library that are placed in the same neuron as known actives should also be active. The separation of different classes of molecules while trained with compounds not belonging to these classes on a larger map was shown to be possible by Bauknecht et al.[100] They could separate dopamine agonists and benzodiazepine receptor agonists while training them together with more than 8000 compounds from a vendor library on a map of 40x30 neurons. Compounds with unknown activity which were co-localized with compounds of known activity were suggested as promising for biological testing. Kaiser et al.[88] were successful in implementing this method for the identification of new inhibitors of P-Glycoprotein. 131 propafenone analogues with known activity values were trained together with 134 767 compounds of the SPECS database on 250x250 and 360x360 sized self-organizing maps. Compounds co-localizing with the most active compounds but having a different type of scaffold were biologically tested, with only one out of seven being inactive. On the other hand, from the compounds co-localized with the most inactive compounds, only one of eight showed modest inhibitory activity. However, a clear drawback of this method is the increase in calculation time, which is due to the increased number of neurons as well as the higher number of training objects.

This type of screening with self-organizing maps was successful in our group,[88] but has so far not been used in other studies. We therefore chose this method to further study the abilities of self-organizing maps. While inhibitors of P-Glycoprotein are known to have diverse structures, the derivatives of **1** activating the insulin receptors show higher similarity. One of our interests was therefore to see whether self-organizing maps are able to find new structures under these conditions.

1.5 Aims

Although there are several drugs available for the treatment of type 2 diabetes mellitus, these drugs can show side effects and their efficacy can decrease during the progression of the disease. Failure of the β-cells makes additional injections of insulin necessary. Demethylasterriquinone B-1 (compound **1**) and derivatives were shown to directly activate the insulin receptor, which could be an interesting target for the treatment of diabetes.

Based on the structure of compound **1** and its published derivatives we want to identify new molecules able to activate the insulin signalling pathway. The main aim of this study is to investigate possible replacements for the asterriquinone scaffold to find new insulin mimetic compounds with computational methods. A focus is set on the performance of self-organizing maps and the comparison of this method to other, more established, virtual screening methods.

For experimental investigation of the results, some hit compounds should be selected for testing in biological assays.

2

Methods and materials

2.1 Computational part

2.1.1 Databases

Training database: A database with compound **1** and its derivatives (**2**-**101**) was compiled from the literature as training set.[32,33,35,38–45] The molecules were drawn by hand and energy minimized using MOE2007.10.[101]

As the method for determining the activity was different in most of the papers, we chose to divide the compounds into active and inactive according to the information available in the papers instead of using the given activity values directly. The activity of compounds tested in more than one publication was additionally used to compare the activity values of those papers and to set the thresholds accordingly.

All compounds from Wood et al.[40] were classified as active when having an activity value higher than 25% (compared to insulin at 100 nM) at a concentration of 10 µM. As comparison, **1** showed 75% activity under this conditions. For compounds published by Liu et al.[38] the threshold was set to $EC_{50} \leq 30$ µM. Compounds published by Pirrung et al.[33] in 2005 were classified as active when showing at least 40% activation of the insulin receptor at 30 µM as compared to insulin (50 ng/mL) and classified as inactive when showing less than 25% activation, as these compounds were stated to be significally different to **1** (58% activation) by statistical analysis. Compounds by Lin et al.[43] were treated as active when showing at least 30% activation of the insulin receptor at 30 µM (compared to 17 (or 8.3?) nM insulin).

This led to a total number of 43 active and 58 inactive compounds in the training set. A new active molecule (**102**) was published[46] after the building of the models (see figure 1.8 on page 14). This molecule was used for external validation. A list of all DMAQ-B1 derivatives collected from the literature can be found in table A.1 in the appendix on page 123.

Screening database: For virtual screening, a compound library provided by ChemDiv[102] was used. The two dimensional graph depictions of the molecules were converted to three dimensional structures using the rebuild 3D functionality of MOE2008.10[101] using the MMFF94x forcefield with a gradient of 0.1, followed by energy minimization with a gradient of 0.01. Ten entries which were containing more than one molecule each were excluded. Finally, a total of 620 225 molecules was used for screening. Using 2D autocorrelation descriptors, 6523 structures were excluded as the descriptors could not be calculated. This was mainly due to charges on the molecules.

To generate smaller subsets of the database, the euclidean distances to the training database molecules were calculated using the normalized VSA and 2D autocorrelation descriptors, respectively. With a distance cutoff of 4, 7126 compounds were selected using the VSA descriptors. A distance of 3 was chosen for the 2D autocorrelation descriptors, yielding a subset of 7320 molecules.

2.1.2 Descriptors

Subdivided surface area descriptors (VSA): The subdivided van der Waals surface area (VSA) descriptors[103] as implemented in MOE[101] describe the surface of a molecule in dependency on the atom-wise contributions to lipophilicity, molar refractivity and partial charges. There are three different descriptor sets, the SlogP-VSA, SMR-VSA and PEOE-VSA descriptors. Each of these descriptor sets uses a different property to choose the atoms which are used to calculate the surface area. The atomic contributions to the van der Waals surface area (in $Å^2$) are calculated using a connection table approximation. The descriptors therefore do not take into consideration the three dimensional structure of the molecules, but can be computed from the two dimensional graph.

2D-autocorrelation descriptors: As described in equation 2.1, the autocorrelation descriptor $A(d)$ is the sum of all products of the properties p of all pairs of atoms i and j, that have the distance d. This distance is determined by the number of intervening bonds d_{ij} between the atoms, using the shortest path. If d_{ij} is equal to d, δ_{ij} becomes 1, else it is 0.

$$A(d) = \frac{1}{2} \sum_{i,j} p_i p_j \delta_{ij} (d - d_{ij}) \tag{2.1}$$

Descriptors for different bond distances can be combined to an autocorrelation vector.

The descriptors were calculated using ADRIANA.*Code*[104] (Version 2.0) for the distances of zero to six bonds. The used properties were sigma charge, pi charge, total charge, sigma electronegativity, pi electronegativity, lone-pair electronegativity and polarizability. This lead to a total number of 49 descriptors. Since only uncharged molecules can be processed, the descriptors could not be provided for the entire screening library.

2.1.3 Self-organizing maps

Self-organizing maps are unsupervised neural networks, which can be used to place higher dimensional objects onto a two or three dimensional map using a non-linear projection. The main idea is that two objects which are placed in the same or adjacent neurons (fields) of the map, are also near to each other in the higher dimensional space.

Training of self-organizing maps: Each neuron of a self-organizing map contains a number of weights which have the same dimensionality as the input objects. In our case, the input objects are molecules, which are represented by

m descriptors. The descriptors were standardized by subtracting the mean value and then dividing the resulting number by the standard deviation. This was done using the script scale.svl by Dominik Kaiser. Mean values and standard deviations were saved, to allow the inclusion of new molecules using the original parameters. The training of the self-organizing maps was performed with SONNIA.[105]

During the training, the molecules are repeatedly presented to the network. The distances between the molecules $X(t)$ and each neuron W_j from a total of n neurons are calculated using formula 2.2 to determine the ‘winning neuron’. This is the neuron with the most similar weights to the descriptors, thus showing the minimum distance.

$$\sum_{i=1}^{m} [x_i(t) - w_{ji}]^2 \tag{2.2}$$

The next step is the adaptation of the weights of and near to the winning neuron c. The adaptation depends on the difference between the pattern vector $x_i(t)$ and the weight vector $w_{ji}(t)$ and on the neighbourhood function $h_{ji}(t,c)$. For each time step t (each presentation of a molecule) the weights of the neurons j are therefore adapted following equation 2.3:

$$w_{ji}(t+1) = w_{ji}(t) + h_{ji}(t,c)\big(x_i(t) - w_{ji}(t)\big) \tag{2.3}$$

The neighbourhood function $h_{ji}(t,c)$ (equation 2.4) is a combination of the learning-rate function $\eta(t)$ and the distance of the neuron to the winning neuron $(\varphi_j(t,c))$.

$$h_{ji}(t,c) = \eta(t) \cdot \varphi_j(t,c) \tag{2.4}$$

During the training process, the learning rate as well as the size of the neighbourhood is decreased, by multiplying the values with an adaptation factor. For the training with SONNIA,[105] the initial learning rate as well as the learning rate adaptation factor were kept at their default values of 0.9. For the initial learning span the default and maximum allowed values, width/2.0 and height/2.0, were used for the small maps. For the training with the screening database, width/10.0 and height/10.0 were used.

Analysis of the self-organizing maps: As the number of active and inactive molecules in the training set was not equal, the threshold to determine the activity of the neurons was set accordingly. A neuron was classified as active if all located compounds were active, or if the ratio of active to inactive compounds was larger than the ratio of total active to inactive molecules (43.0/58.0). A java script (Map.java) was written to change the colours of the .map files from SONNIA[105] according to this threshold.

The colouring was done using the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, white: empty. When trained together with the screening database, neurons containing only new molecules are coloured grey.

Numbering of the neurons was done according to the x- and y-axis, starting with zero for the first neuron and locating the origin to the upper left corner of the map. For example, 0/0 corresponds to the neuron in the upper left corner, 1/0 is the second neuron in the first row.

Additionally, the accuracy and precision values were calculated for the distribution of actives and inactives. For this, the number of correctly assigned active molecules (true positives, TP), of correctly assigned inactive molecules (true negatives, TN), of actives incorrectly placed into an inactive neuron (false negative, FN) and the number of inactives placed into an active neuron (false positive, FP) were determined. The total accuracy (equation 2.5) gives the fraction of correctly predicted molecules.

$$\text{Total accuracy (Acc)} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

The accuracy on actives (also called sensitivity) in equation 2.6 gives the ratio of all true actives to all actives in the dataset. Accordingly, the accuracy on inactives (specificity) shows the ratio of all correctly predicted inactives to the total number of inactives (equation 2.7).

$$\text{Accuracy on actives } (\text{Acc}_{\text{Act.}}) = \frac{TP}{TP + FN} \tag{2.6}$$

$$\text{Accuracy on inactives } (\text{Acc}_{\text{Inact.}}) = \frac{TN}{TN + FP} \tag{2.7}$$

Precision values give the ratio of correct classifications to all compounds which were assigned to that class. These values give an indication on how high the probability is, that a compound assigned to a class is really a member of it. The precision on actives and inactives is shown in equations 2.8 and 2.9, respectively.

$$\text{Precision on actives } (\text{Pre}_{\text{Act.}}) = \frac{TP}{TP + FP} \tag{2.8}$$

$$\text{Precision on inactives } (\text{Pre}_{\text{Inact.}}) = \frac{TN}{TN + FN} \tag{2.9}$$

2.1.4 Fingerprint similarity

The fingerprint of a molecule is a binary vector, indicating the presence or absence of a given set of features. The similarity of two molecules can be accessed by calculating the Tanimoto coefficient of their fingerprints:

$$T = \frac{N_{12}}{N_1 + N_2 - N_{12}} \tag{2.10}$$

Here, N_1 and N_2 represent the number of features present in the fingerprints of molecules 1 and 2, respectively, and N_{12} represent the number of features which are present in both fingerprints.

For calculating the similarity of the active compounds to the screening library, eight different fingerprint types were calculated with MOE 2008.10:[101]

1.) MACCS Structural Keys: The MACCS fingerprint of MOE is an implementation of the 166-bit MACCS fingerprint introduced by MDL.[106] This fingerprint was created primarily for substructure searching, and indicates the presence of 166 different substructures or atom types.

2.) Pharmacophore Atom Triangle (piDAPH3): This fingerprint is based on a 3-point pharmacophore using the 3D structure of the molecule. The fingerprint features are built up using the distances and the pharmacophoric features (combinations of “in pi system”, “is donor” or “is acceptor”) of three different atoms.

3.) Pharmacophore Graph Triangle (GpiDAPH3): The Pharmacophore Graph Triangle is similar to the Pharmacophore Atom Triangle, but it uses information from the 2D graph, instead of the 3D conformation. Distances between the atoms are the number of bonds of the shortest path between the atoms (graph distance).

4.) Pharmacophore Atom Quadruplet (piDAPH4): The Pharmacophore Atom Quadruplet uses the inter-atomic distances and pharmacophoric features (combinations of “in pi system”, “is donor” or “is acceptor”) of four different atoms.

5.) Typed Graph Distance (TGD): The features of the Typed Graph Distance fingerprint are built up of the atom types and the graph distance between two atoms. Atom types can be either acid, base, hydrogen bond donor, hydrogen bond acceptor, both hydrogen bond acceptor and donor, hydrophobic ore none of the others.

6.) Typed Graph Triangle (TGT): The Typed Graph Triangle fingerprint utilizes the atom types (either hydrogen bond donor or base, hydrogen bond acceptor or acid, both hydrogen bond acceptor and donor, or hydrophobic) and the graph distances between three atoms. The graph distances are binned in a way that there is a higher resolution at smaller distances and less resolution at distances larger than five bonds.

7.) Typed Atom Distance (TAD): This fingerprint is similar to the Typed Graph Distance, but uses binned distances of the atoms using the currently available 3D structure of the molecule.

8.) Typed Atom Triangle (TAT): The Typed Atom Triangle is similar to the Typed Atom Distance, but uses the atom types and binned distances of three atoms, using the 3D structure of the molecule.

Implementation: Similarity matrices of the active molecules of the training database against the compounds from the screening library were calculated using the command `ph4_SimilarityMatrix[‘mdb1’,‘mdb2’,‘fp_code’,‘similarity’,‘output.txt’]` in the MOE command line. The options `mdb1` and `mdb2` represent the input databases, with the first one being the larger screening database; `fp_code` specifies the fingerprint to use (e. g. FP:MACCS), `similarity` is the similarity metric (tanimoto), and `output.txt` determines the new output file. To produce a human readable output of this command, the file ph4addfp.svl was modified by Christoph Waglechner.

For subsequent analysis, the resulting matrices were individually imported into a MySQL database. The table, which is in the following called `sreeningdb`, contains three columns (`chemdiv` for the codes of the screening library, `actives` for the codes of the queries and `similarity` for the similarity values of each pair of compounds).

Identification of the most similar entries: To find for each entry of the screening library the most similar query structure, the following commands were used. First, for each compound the highest similarity value was searched and written into an intermediate table (`screeningdb_maxonly`).

```
insert into screeningdb_maxonly
select chemdiv, max(similarity) as maxsim
from screeningdb
group by chemdiv;
```

When used to calculate the similarity of the active database to itself, the entries where identical compounds are compared need to be excluded from the table before searching the highest similarity value for each compound. This was done with the command `delete from screeningdb where(chemdiv=`

`actives);`. All steps for importing and preprocessing of the data until here were included into a Java script (fingerprint.java).

Subsequently, the intermediate table can be merged with the original table (`sreeningdb`) to include the names of the active compounds.

```
select orig.chemdiv, orig.actives, maxsim
from screeningdb as orig
inner join screeningdb_maxonly as maxonly
on (orig.chemdiv = maxonly.chemdiv and maxsim=similarity);
```

Enrichment factors: To calculate enrichment factors, the similarity values of all active structures of the training database to each other and the similarity values of the queries to the screening library were combined.

```
(Select * from actives_maxonly)
union
(Select * from screeningdb_maxonly)
order by maxsim desc limit 100;
```

The enrichment factor (EF) was calculated for a given subset size as shown in equation 2.11.[107]

$$\mathrm{EF}_{\text{subset size}} = \frac{\text{fraction active in subset}}{\text{fraction active in library}} \tag{2.11}$$

For the first 1% of the screening hits the subset size is 6203, leading to a maximum possible enrichment factor of 100. For a subset of the first 100 compounds, the highest enrichment factor is 6203. A random enrichment would yield a value of 1.

Thresholds: To get an overview on how many compounds are identified at different similarity thresholds the following MySQL command was used. Additional similarity thresholds can be added analogously.

```
SELECT
sum(if(maxsim=1,1,0)) as '=1',
sum(if(maxsim>0.95,1,0)) as '>0.95',
sum(if(maxsim>0.90,1,0)) as '>0.90',
sum(if(maxsim>0.85,1,0)) as '>0.85',
sum(if(maxsim>0.80,1,0)) as '>0.80'
FROM screeningdb_maxonly;
```

All identified compounds together with the queries and the similarity values can be selected using `SELECT * FROM screeningdb where similarity > threshold;`, where `threshold` needs to be replaced with the chosen similarity value. Additionally the output can be sorted by adding `order by` and the wanted column name.

2.1.5 Shape similarity

Shape similarity search was performed with Phase 3.0 from Schrödinger.[108] There, the similarity of different alignments of a conformer of the screening library to the query structure is calculated based on overlapping hard-sphere volumes.

The 43 active molecules of the training database were used as shape queries. A MacroModel[109] conformational search was performed with default parameters. The used force field was OPLS_2005 and water was used as solvent. The minimum energy conformation was then used as input for the shape similarity search.

The screening library was prepared using the Manage 3D Database panel from Phase. The molecules were imported as 3D sd-file from MOE. A maximum number of 100 conformers per structure was generated using 10 steps per rotable bond using the MacroModel search method ConfGen. The rapid sampling option was chosen and conformations of amide bonds were varied. Preprocessing was skipped and no energy minimization was done for post-processing. Redundant conformers were eliminated using an RMSD cutoff value of 1 Å.

The shape similarity search itself was done using the Shape Screening panel from Phase. The volume scoring was done without taking into account the atom types. Only one alignment per compound was kept and conformers with a similarity below 0.65 were discarded. Further processing of the data was done as described for the fingerprint similarity search.

2.1.6 Compound selection

All compounds identified by the different methods were merged into a database. To compute the scaffolds of the compounds, the script sca.svl from the MOE exchange server was used.[110,111] Subsequently, similar scaffolds (for example exchanges of one atom in the scaffold) were merged by hand.

To identify the scaffolds which were selected more frequently than other scaffolds, the combined database (see table A.2 on page 130) was imported to MySQL.

```
load data infile 'identified_367.txt'
into table identified_367
fields terminated by '\t'
lines terminated by '\r\n'
ignore 1 lines
(name,SOM2d,SOMvsa,Shape,FP,Scaffold);
```

Subsequently, the counts for each scaffold were calculated:

```
select Scaffold, sum(SOM2d) as SOM2d, sum(SOMvsa) as SOMvsa,
sum(Shape) as Shape, sum(FP) as FP
from identified_367
group by (Scaffold);
```

Scaffolds were considered for further investigation, if they were either selected by two different methods, or selected by one method for a minimum of ten times. Representative structures of the scaffold clusters were chosen according to a consensus vote of selection by hand and selection of a representative structure (the medoid) using three different fingerprint sets. The fingerprint sets were the VSA descriptors, 2D autocorrelation descriptors and a set of 11 simple physicochemical descriptors (a_acc, a_don, b_rotN, logP(o/w), mr, PEOE_VSA_HYD, TPSA, vsa_acc, vsa_don, vsa_hyd and Weight calculated with MOE). The selection of the medoid was performed using the partitioning around medoids (pam) functionality of the cluster package[112] in R[113] using a script written by Michael Demel. Clusters were set according to the scaffold clusters, and one medoid per cluster was retrieved.

2.1.7 Docking

Docking studies were performed on a crystal structure of the active intracellular domain of the insulin receptor (PDB-ID: 1IR3) using MOE2008.10 (for preliminary studies) and MOE2009.10. Water molecules and ligands were deleted and missing side chains were modeled using Prime from Schrödinger.[114] Hydrogens and protonation states were calculated with the Protonate 3D functionality in MOE. The binding site was chosen between the αC helix and the nearby β-sheets as suggested in reference 19. Dummy atoms were placed in this binding pocket using the Site Finder tool of MOE. The docking poses presented in the result section were generated using Triangle Matcher as placement algorithm and London dG as scoring function. The obtained poses were refined using the LigX functionality of MOE2009.10, thus allowing the energy minimization of the ligand as well as the adjacent amino acids.

Common scaffold investigation was done to identify poses which were found for several ligands, allowing a comparison of activities with the interactions. Common scaffold clustering was implemented successfully in our group for P-Glycoprotein.[115] For this approach, only those molecules containing both indole rings were used. The common scaffold was defined by the SMILES string `n1cc(c2c1cccc2)C=1CC=C(CC=1)c1c2c(nc1)cccc2`, encoding the structure depicted in figure 2.1.

The structure of this scaffold was written in a new field of the MOE database using the script DMAQscaffold.svl, which was adapted from a script by Freya Klepsch and Lars Richter. Subsequently, an RMSD matrix was calculated comparing all poses with each other on basis of the scaffold. This calculation was performed using the script rmsd_matix.svl written by Lars

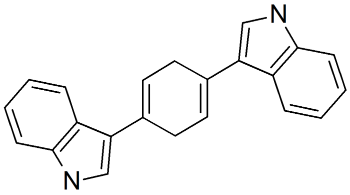


Figure 2.1: Common scaffold

Richter, which uses mol_rmsd.svl from the SVL Exchange site.[116]

2.2 Experimental part

2.2.1 Molecules for testing

Demethylasterriquinone B-1 was purchased as positive control from Biotrend (Cat. No.: BN0178). Compounds identified as hits from virtual screening were purchased from ChemDiv. All compounds which were soluble enough were dissolved in DMSO at a concentration of 100 mM and stored in aliquots at -20°C.

2.2.2 Media and solutions

Growth media

- HCC-1.2 cultivation medium: RPMI-1640 medium (without L-glutamine, Lonza Group Ltd.) was supplemented with 10% FBS (fetal bovine serum; Gibco®, Invitrogen), 2 mM L-glutamine and 1% penicillin/streptomycin mixture.

- HCC-1.2 starvation medium: Identical to HCC-1.2 cultivation medium, but without FBS.

- MEF cultivation medium: DMEM medium (without L-glutamine and phenol red, Lonza Group Ltd) was supplemented with 10% FBS, L-glutamine and 1% penicillin/streptomycin mixture.

- MEF starvation medium: Identical to MEF cultivation medium, but instead supplemented with 0.1% FBS.
- 3T3-L1 cultivation medium: DMEM medium (Lonza Group Ltd., without L-glutamine and phenol red) was supplemented with 10% NBS and L-glutamine.
- adipocyte cultivation medium: DMEM medium (Lonza Group Ltd., without L-glutamine and phenol red) was supplemented with 10% FBS and L-glutamine.
- 3T3-L1 differentiation medium: Adipocyte cultivation medium was supplemented with 1 µg/mL insulin, 500 nM dexamethasone and 50 µM IBMX. Medium was prepared and sterile filtered prior to usage.

Growth media were stored at 4°C and warmed to 37°C in a water bath before use.

Media supplements

- FBS: fetal bovine serum, Gibco®, Invitrogen
- NBS: newborn bovine serum, Lonza
- Penicillin/streptomycin mixture (Lonza Group Ltd.):
  10 000 U/mL potassium penicillin
  10 000 µg/mL streptomycin sulfate
- IBMX (3-Isobutyl-1-methylxanthine, Sigma): 50 mM stock solution in 0.5 N KOH.
- Insulin (Sigma): 10 mg/mL stock solution (1.7 mM) in 25 mM HEPES buffer.
- Dexamethasone (Sigma): 2.5 mM stock solution in ethanol.

Phosphate Buffered Saline (PBS): NaCl (36.0 g), Na_2HPO_4 (7.4 g) and KH_2PO_4 (2.15 g) dissolved in 5 L H_2O; pH 7.4, autoclaved.

Trypsin/EDTA: Trypsin (0.05%,GIBCO) and Na_2EDTA (0.02%) dissolved in PBS; sterile filtered.

Crystal violet assay

- Crystal violet solution
  0.5% crystal violet dissolved in 20% methanol; filtered.

- Sodium citrate solution
  0.05 M sodium citrate tribasic dihydrate (BioChemica) in 50% ethanol.

Western blotting

- RIPA lysis buffer: For the stock solution, Tris/HCl (50 mM, pH 7.4), NaCl (500 mM), NP40 (5.044 mM), Na-Deoxycholate (12.06 mM), SDS (3.47 mM) and NaN_3(7.7 mM) were dissolved in H_2O. Prior to use, Complete (4%, Roche), PMSF (1 mM), NaF (1 mM) and $NaVO_3$ (1 mM) were added to the stock solution.

- Bradford reagent: Roti®-Quant (Carl Roth) was diluted 1:5 in water before usage.

- 3x SDS sample buffer: For the stock solution, 37.5 mL Tris-HCl (0.5 M stock, pH 6.8), 6.0 g SDS, 30.0 mL glycerol and 15.0 mg bromophenol blue were mixed. Water was added to a total amount of 100 mL of buffer. 15% of 2-Mercaptoethanol were added to obtain the 3x solution.

- TBS-T: For the tris buffered saline 3.0 g of Tris-base and 11.1 g of NaCl were dissolved in 1 L of water. To obtain TBS-T, 1 mL of Tween 20 was added. The pH value of 8.0 was set with concentrated HCl.

- Resolving gel: The percentage of polyacrylamide in the gels was varied according to the protein in investigation. For gels showing the insulin receptor, 7.5% gels were used. Detection of Akt was done using 10% gels for separation. 1.875 and 2.5 mL of PAA solution (30% with 0.8% bisacrylamide) were used for the 7.5 and 10% gels, respectively. PAA was mixed with 1.875 mL of Tris-base (1.5 M, pH 8.0), 75 µL SDS (10%) and 3.675 mL water. The polymerisation was initiated using 7.5 µL TEMED and 37.5 µL APS (10%).

- Stacking gel: 640 µL PAA solution (30% with 0.8% bisacrylamide) was mixed with 375 µL of Tris-base (1.25 M, pH 6.8), 37.5 µL SDS (10%) and 2.62 mL water. The polymerisation was initiated using 3.75 µL TEMED and 18.8 µL APS (10%).

- Electrophoresis buffer: 30 g of Tris-Base, 144 g glycine and 10 g SDS were dissolved in 1 L of water to prepare the 10x solution. This stock was then diluted 1:10 for usage.

- Blotting buffer: For the 5x solution, 15.17 g of Tris-base and 72.9 g of glycine were dissolved in 1 L of water. For usage, 100 mL buffer were diluted with 100 mL methanol and 300 mL water.

- ECL: The enhanced chemiluminescence (ECL) solution was prepared directly before usage. Stock solutions (0.44 g luminol in 10 mL of DMSO and 0.15 g p-coumaric acid in 10 mL of DMSO) were stored at -20°C. For the final solution, 1 mL TRIS-base (1 M, pH 8.5), 50 µL luminol stock solution, 22 µL p-coumaric acid stock solution and 3 µL H_2O_2 (30%) were diluted in 9 mL of water.

- Antibodies:
  Anti-phospho-insulin receptor/insulin-like growth factor-1 receptor p-Tyr1158, 1162 and 1163 antibody (Sigma),
  Anti-IR-β antibody (New England Biolabs),
  Anti-phospho-Akt Ser473 antibody (Cell Signal),
  Anti-α-tubulin antibody (Santa Cruz),
  Anti-actin antibody (Santa Cruz),
  Anti-rabbit antibody (New England Biolabs),
  Anti-mouse antibody (Upstate).
  All antibodies were diluted to their working concentration in TBS-T.

Glucose uptake assay

- KRH buffer: Hepes (50 mM, from a 1 M stock, pH 7.4), NaCl (136 mM), KCl (23.5 mM), $MgSO_4$ (1.25 mM), $CaCl_2$ (1.25 mM) and 0.1% BSA were dissolved in water. The buffer was sterile filtered after completion.
- ^{3}H-Deoxyglucose (DOG): deoxy-D-glucose, 2-[1,2-^{3}H (N)]-, with a specific activity of 5-10 Ci/mmol and a concentration of 1 mCi/mL (Perkin Elmer).
- DOG solution: For a 12-well plate, 7.43 µL deoxyglucose (0.1 M) and 3.72 µL of ^{3}H-deoxyglucose were diluted in 789 µL KRH buffer.
- Scintillation cocktail: Ultima Gold™ (Perkin Elmer)

PTP1B inhibition assay

- PTP1B: Recombinant human PTP1B (R&D Systems) was dissolved in reconstitution buffer to 1 µg/µL and stored at -80°C.
- Reconstitution buffer: 10 mM HEPES, 0.1 mM EGTA, 0.1 mM EDTA, 1 mM dithiothreitol and 0.5 mg/mL BSA, pH 7.5.
- MOPS buffer: 3-(N-morpholino)-propanesulfonic acid (50 mM, pH 6.5) with or without 4 mM pNPP (para-Nitrophenylphosphate). The buffer with pNPP was prepared directly before the experiment by dissolving 11.14 mg pNPP in 7.5 mL of MOPS buffer and adding 15 µL DTT (1M).

2.2.3 Technical equipment

- Cell counter: Vi-Cell™ XR Cell Viability Analyzer (Beckman Coulter).
- Plate reader: Sunrise™ (Tecan).
- Western blot chemoluminescence detector: LAS-3000™ (Fujifilm). Images were detected using the Image Reader LAS-3000™ software (version 2.0) and analyzed using AIDA™ (Advanced Image Data Analyzer, version 4.06, raytest).

- Scintillation detector: TRI-CARB 2100TR Liquid Scintillation Analyzer (Packard)

2.2.4 Cultivation of the cells

Human hepatocytes (liver carcinoma cells, HCC-1.2) were grown in 75 cm^2 cell culture flasks (Greiner Bio-One) using 25 mL of HCC-1.2 cultivation media. Cells were passaged after one week, seeding them at a density of one million cells per flask. Medium was changed every two or three days.

Mouse embryonic fibroblasts (MEF) were grown in 75 cm^2 cell culture flasks using 13 mL of MEF cultivation media. Cells were passaged twice per week and seeded at a density of one million cells per flask.

3T3-L1 preadipocytes were seeded at a density of 0.415 x 10^6 cells in 175 cm^2 flasks using 45 mL of 3T3-L1 cultivation medium. Cells were passaged after three days. To avoid that the cells were loosing their ability to differentiate into adipocytes, they were used up to a maximum passage of 14, and only if the total amount of cells after passaging did not exceed ten million cells.

2.2.5 Western blotting

To detect the phosphorylation state of the insulin receptor and of Akt, the Western blot technique was used. For this, the cells were lysed after treatment and the proteins were seperated by SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis). Here, the proteins are first denaturated by SDS and heat. Additionally SDS leads to a uniform negative charge on the proteins, resulting in the separation of proteins according to their weight and not their initial charge while they are moving through the polyacrylamide gel to the anode. After the separation, proteins are transferred onto a membrane where the protein of interest can be detected using antibodies. Usually, the primary antibody against the target protein is applied first. The detection of the protein is then made possible by using a secondary antibody against the primary one, which is linked to a reporter enzyme. In our case, this enzyme was horseradish peroxidase, which can catalyze the oxidation of luminol, leading to chemoluminescence.

Sample preparation: Hepatocytes were seeded at a density of 150 000–200 000 cells per well in 4 mL of HCC-1.2 cultivation medium in 6-well plates. Cells were grown for two to four days. Two hours before the experiment, the medium was exchanged to 2 mL starvation medium. Mouse embryonic fibroblasts were seeded at a density of 400 000 cells per well in 2 mL cultivation medium in 6-well plates. After one day, the medium was exchanged to 2 mL starvation medium and the cells were grown for another day.

Cells were treated with the compounds at a final concentration of 0.5% DMSO. Insulin (10 nM) and 0.5% DMSO were used as positive and negative control. After the treatment time (usually 5 or 10 min), cells were washed with cold PBS and lysed with 100 µL RIPA buffer for 10 min on ice. Subsequently, cells were scraped out of the wells into reaction tubes and centrifuged at 13 000 rpm at 4°C for 15 min. Hepatocytes were additionally sonicated before the centrifugation step. The supernatant was then used further for protein quantification and Western blotting. The samples were stored at −20°C. For electrophoresis, the amount for 20 µg protein was mixed 2+1 with 3x SDS sample buffer. The samples were boiled for 5 min at 95°C to denature the proteins.

Protein quantification: To determine the amount of protein in the samples, the quantification according to Bradford was carried out. For this, 10 µL of a 1:10 dilution of the samples was pipetted on a 96-well plate in triplicate and 190 µL of Bradford reagent were added. The absorbance of the resulting colour was measured at 595 nm in a plate reader. The protein concentration was determined using a standard curve of BSA (bovine serum albumin) on the same plate.

Electrophoresis and Western blot: The prepared samples were pipetted into the chambers of the stacking gel. Proteins were separated in a Mini-PROTEAN® 3 Electrophoresis Cell (Bio-Rad) using 25 mA power supply per gel (PowerPac HC) until the blue colour of the sample buffer reached the end of the gel. To allow a better separation of Akt from neighbouring bands, the electrophoresis was performed longer in this case. Subsequently, proteins were transferred to a PVDF membrane in a Mini Trans-Blot® Electrophoretic Transfer Cell (Bio-Rad) at 100 V for two hours. To block non-specific binding of the antibodies, the membrane was then treated for one hour with 0.5% BSA in TBS-T.

The membrane was incubated with the first antibody overnight at 4°C, and with the secondary antibody for two to three hours at room temperature. In between and afterwards, the membrane was washed three times with TBS-T for ten minutes each. Chemoluminescence was initiated with ECL for one minute and the signal was then detected using the automatically determined exposure time. Before the application of the next primary antibody, the membrane was stripped using a solution of NaOH (0.5 M) for 10 min and washed with TBS-T.

2.2.6 Crystal violet assay

To assess the cytotoxicity of the tested compounds, the amount of remaining cells after treatment for 24 h was measured using the crystal violet assay. For quantification, the cells were coloured with crystal violet, a dye that stains proteins in an unspecific way. After washing away the excess dye, the remaining crystal violet is dissolved and analyzed photometrically. The optical density of this solution depends on the concentration of crystal violet and therefore on the amount of cells.

Mouse embryonic fibroblasts were seeded at a density of 5 000 cells per well in 198 µL cultivation medium in 96-well plates. Some wells were filled with medium only to determine its effect on the staining. Cells were grown near confluence for two days. Compound stock solutions were diluted with medium to a working solution in 50% DMSO. Cells were treated with 2 µL of the working solution, leading to a final DMSO concentration of 0.5%. All treatments and controls were performed in triplicate. After 24 h, the growth medium was discarded. Wells were filled with 100 µL crystal violet solution for 10 min. Then the staining solution was washed away with water. After air drying of the plates, the remaining crystal violet was dissolved in 100 µL of a sodium citrate solution and the absorption was measured with a plate reader at 550 nm. Values were normalized by dividing through the DMSO control value of each experiment.

2.2.7 Glucose uptake assay

To investigate the effect of selected compounds on a biological outcome more relevant to diabetes, glucose uptake was measured in adipocytes and myocytes. For this, differentiated cells were treated with the compounds after a period of starvation. Glucose uptake was then assessed using deoxyglucose with a radioactive label. The experiments with myocytes were carried out by Elke Heiss.

Differentiation of adipocytes: 3T3-L1 preadipocytes were seeded at a density of 1.5x10^5 cells in 12-well plates using 2 mL of 3T3-L1 cultivation medium per well. When the cells reached confluency after two to three days, medium was changed and the cells were grown for another two days. To initiate the differentiation, the medium was then exchanged for 3T3-L1 differentiation medium. After three days, the medium was replaced by adipocyte cultivation medium supplemented with 1 µg/mL insulin. Subsequently, the medium was replaced every second day and on the day before the experiment with adipocyte cultivation medium (without insulin). Experiments were carried out between day eight and ten after the initiation of differentiation.

Glucose uptake into adipocytes: To prepare the cells for the experiment, serum and glucose were withdrawn. First, the medium was replaced by 2 mL of DMEM supplemented with 0.1% BSA for four hours to remove the serum. Then, the medium was exchanged for 1 mL of KRH buffer for one hour to remove the glucose. Testing compounds and insulin were dissolved in KRH buffer and cells were stimulated with a total amount of 450 µL per well. After the stimulation time, 50 µL of the DOG solution were added. The cells were incubated for 5–15 minutes at 37°C. The glucose uptake was terminated by washing the cells with ice-cold PBS. The cells were lysed using 350 µL NaOH (0.05 M in PBS) at 4°C over night. After shaking, 250 µL lysate were mixed with 5 mL of scintillation cocktail and measured with the scintillation detector. The specific activity of the ^{3}H-DOG was assessed by measuring 50 µL of DOG solution with 5 mL of scintillation cocktail. The background value was detected using NaOH (0.05 M in PBS). The amount of protein was determined as described on page 44 with a 1:5 dilution of the samples.

2.2.8 PTP1B inhibition assay

The phosphatase PTP1B dephosphorylates the insulin receptor leading to the inactivation of the kinase. Inhibitors of PTP1B could therefore enhance the signal transduction by the insulin receptor. The inhibition of PTP1B was investigated directly using an enzymatic *in vitro* assay. The readout of this assay is based on the conversion of para-nitrophenylphosphate (pNPP) to para-nitrophenol. Addition of NaOH gives the yellow-coloured sodium-para-nitrophenolate, which can be measured photometrically at 405 nm.

Sodium orthovanadate (SOV) and ursolic acid (UA) are known inhibitors of PTP1B and were used as positive control. The solvent (DMSO, 1%) was used as negative control. Controls and test substances were diluted in MOPS buffer (without pNPP) and 50 µL per well were pipetted in a 96-well plate. For each test substance, 8 wells were used to allow the measurement with and without enzyme in quadruplicate each. Immediately before usage, an 1.4 µL aliquot of the 1 µg/µL stock solution of PTP1B was diluted with 280 µL of reconstitution buffer to a concentration of 0.005 µg/µL. To avoid loss of activity, it was kept in a closed tube on ice whenever possible. 5 µL of this dilution were pipetted to half of the wells. The other half was measured without enzyme to get a background value of the substances. Here, 5 µL of reconstitution buffer were added. Finally, 50 µL of MOPS buffer with pNPP were added to each well. The final concentrations per well were 0.025 µg of PTP1B and 2 mM pNPP. The reaction was then carried out for half an hour. During this time, the absorption was measured at 405 nm in the plate reader (11 cycles with a measurement every three minutes). The plate was shaken in between the cycles for five seconds, with two seconds rest before the next measurement. Before a final measurement, 25 µL NaOH (10 M) were added to intensify the yellow colour. A more detailed description of the assay procedure is given in reference 117.

The final measurement was used for further evaluation. The background mean values were subtracted from the values with the enzyme. This was done to minimize the influence of coloured substances. Mean values of the wells with enzyme were then divided by the negative control (DMSO, 1%) and multiplied with 100 to get the residual PTP1B activity in % control.

3

Results

3.1 Computational part

3.1.1 Self-organizing maps

Performance of the method

To judge the ability of the descriptors to separate active from inactive molecules, small self-organizing maps were trained using compounds with known activity only (**1-101**). Preliminary studies using different descriptor sets and network topologies were performed with a smaller training set (data not shown).

Statistics (*quantization error*, accuracy and precision) of the resulting maps are collected in table 3.1 for the VSA descriptors and table 3.2 for the 2D autocorrelation descriptors. The maps are named according to the following scheme: `widthxheight,epochs,topology`. Here `width` and `height` denote the number of neurons in the x- and y-axis, `epochs` specifies how often each molecule is presented to the network during the training and `topology` can be either `R` for rectangular or `T` for toroidal maps.

The *quantization error* decreases drastically when changing the training time from 100 epochs to 500 epochs. An additional increase of the training time to 1000 epochs has less effect. This can be explained when investigating the used parameters in the finished network file (.knet). SONNIA stopped

Size,Epochs,Top.	QE	$\text{Acc}_{\text{Act.}}$	$\text{Acc}_{\text{Inact.}}$	$\text{Pre}_{\text{Act.}}$	$\text{Pre}_{\text{Inact.}}$
5x5,100,R	302.9	0.72	0.84	0.78	0.80
5x5,500,R	161.6	0.84	0.79	0.75	0.87
5x5,1000,R	167.8	0.84	0.83	0.78	0.87
7x7,100,R	246.2	0.86	0.78	0.74	0.88
7x7,500,R	89.5	0.84	0.84	0.80	0.88
7x7,1000,R	96.1	0.86	0.83	0.79	0.89
10x10,100,R	162.8	0.88	0.86	0.83	0.91
10x10,500,R	42.1	0.86	0.90	0.86	0.90
10x10,1000,R	41.5	0.91	0.84	0.81	0.92
14x14,100,R	62.4	0.93	0.88	0.85	0.94
14x14,500,R	11.0	0.95	0.90	0.87	0.96
14x14,1000,R	8.4	0.93	0.88	0.85	0.94

Table 3.1: Small maps with VSA descriptors

Size,Epochs,Top.	QE	$\text{Acc}_{\text{Act.}}$	$\text{Acc}_{\text{Inact.}}$	$\text{Pre}_{\text{Act.}}$	$\text{Pre}_{\text{Inact.}}$
5x5,100,R	204.8	0.95	0.57	0.62	0.94
5x5,500,R	132.4	0.79	0.76	0.71	0.83
5x5,1000,R	126.9	0.79	0.78	0.72	0.83
7x7,100,R	156.1	0.88	0.83	0.79	0.91
7x7,500,R	69.3	0.91	0.84	0.81	0.92
7x7,1000,R	60.4	0.93	0.84	0.82	0.94
10x10,100,R	120.3	0.93	0.84	0.82	0.94
10x10,500,R	32.7	0.95	0.84	0.82	0.96
10x10,1000,R	40.3	0.95	0.84	0.82	0.96
14x14,100,R	11.0	0.95	0.93	0.91	0.96
14x14,500,R	7.1	0.98	0.93	0.91	0.98
14x14,1000,R	8.2	1.00	0.90	0.88	1.00

Table 3.2: Small maps with 2D autocorrelation descriptors

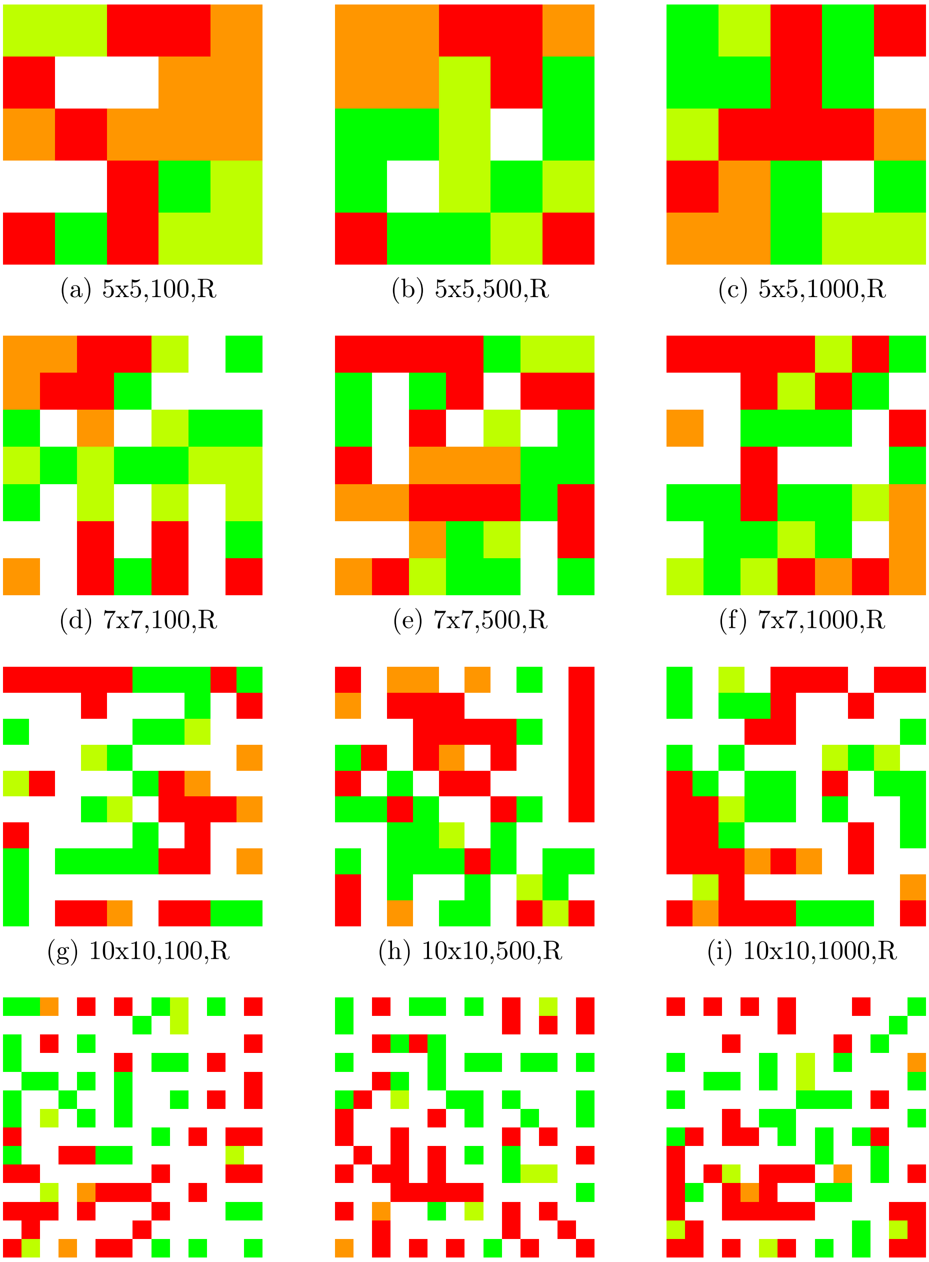
Size,Epochs,Top.	2D		VSA	
	random	real	random	real
5x5,100,R	0.69	0.73	0.71	0.79
5x5,500,R	0.70	0.77	0.69	0.81
5x5,1000,R	0.71	0.78	0.68	0.83
7x7,100,R	0.72	0.85	0.72	0.81
7x7,500,R	0.76	0.87	0.76	0.84
7x7,1000,R	0.79	0.88	0.73	0.84
10x10,100,R	0.84	0.88	0.79	0.87
10x10,500,R	0.85	0.89	0.83	0.88
10x10,1000,R	0.82	0.89	0.84	0.87
14x14,100,R	0.93	0.94	0.89	0.90
14x14,500,R	0.93	0.95	0.89	0.92
14x14,1000,R	0.93	0.94	0.91	0.90

Table 3.3: Comparison of total accuracy values between maps with real and random activity.

the training after approx. 250–300 epochs in both cases. When judging the quality of the maps according to the accuracy and precision values, the training time has in general less influence. Here, the size of the networks largely influences the results. However, this can be explained by a higher number of singletons, but it does not necessarily result in a better separation of actives and inactives. For this, a visual inspection of the maps is necessary. Pictures of the maps are collected in figure 3.1 for the VSA descriptors and in figure 3.2 for the 2D autocorrelation descriptors.

Some molecules produced conflicts even in the larger maps. Here, structural changes led only to a small or even no variation in descriptor values, but had a large influence on activity. This was mostly the case where molecules differed only in the position of small residues, such as methyl groups, on the indole ring.

Random shuffling of the activity values was performed to see if activity clusters are formed by chance. A comparison of the accuracy values resulting from the random and the real activity distribution is shown in table 3.3. The accuracy values are generally higher for the maps with the real activity. Only



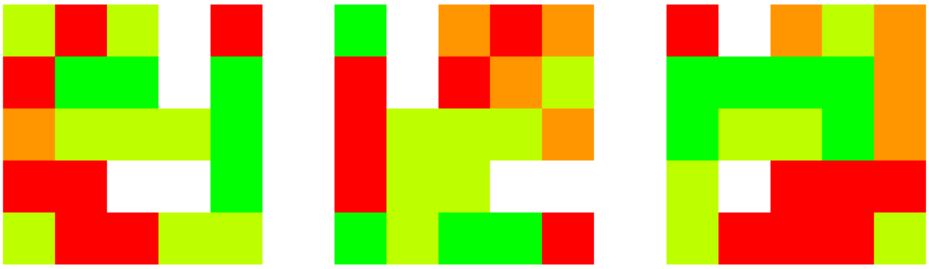
(a) 5x5,100,R (b) 5x5,500,R (c) 5x5,1000,R

(d) 7x7,100,R (e) 7x7,500,R (f) 7x7,1000,R

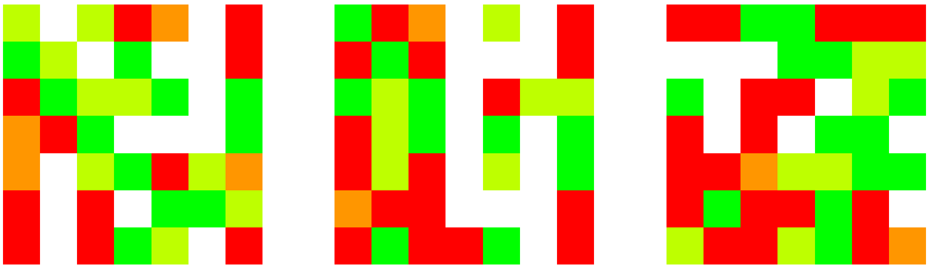
(g) 10x10,100,R (h) 10x10,500,R (i) 10x10,1000,R

(j) 14x14,100,R (k) 14x14,500,R (l) 14x14,1000,R

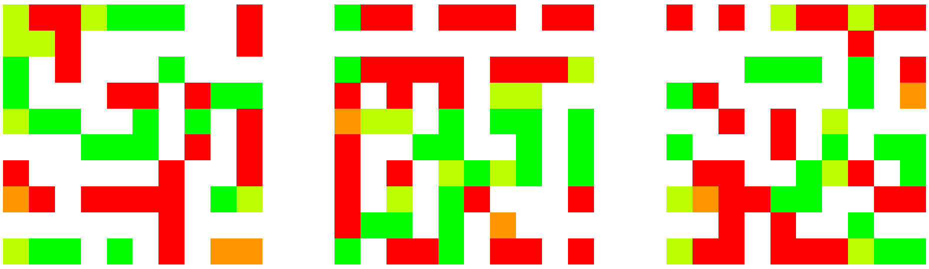
Figure 3.1: Self-organizing maps of compounds with known activity using VSA descriptors. Maps are coloured according to the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, white: empty.



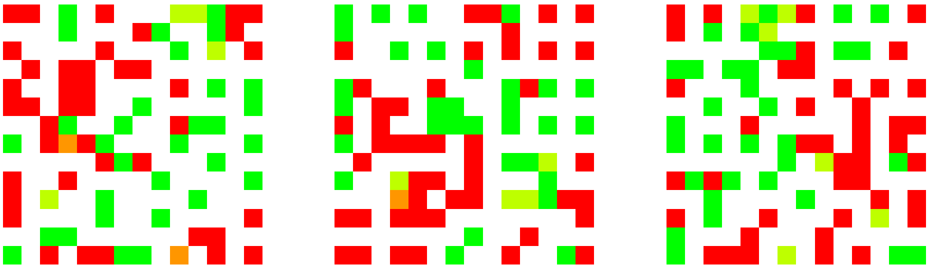
(a) 5x5,100,R (b) 5x5,500,R (c) 5x5,1000,R



(d) 7x7,100,R (e) 7x7,500,R (f) 7x7,1000,R



(g) 10x10,100,R (h) 10x10,500,R (i) 10x10,1000,R



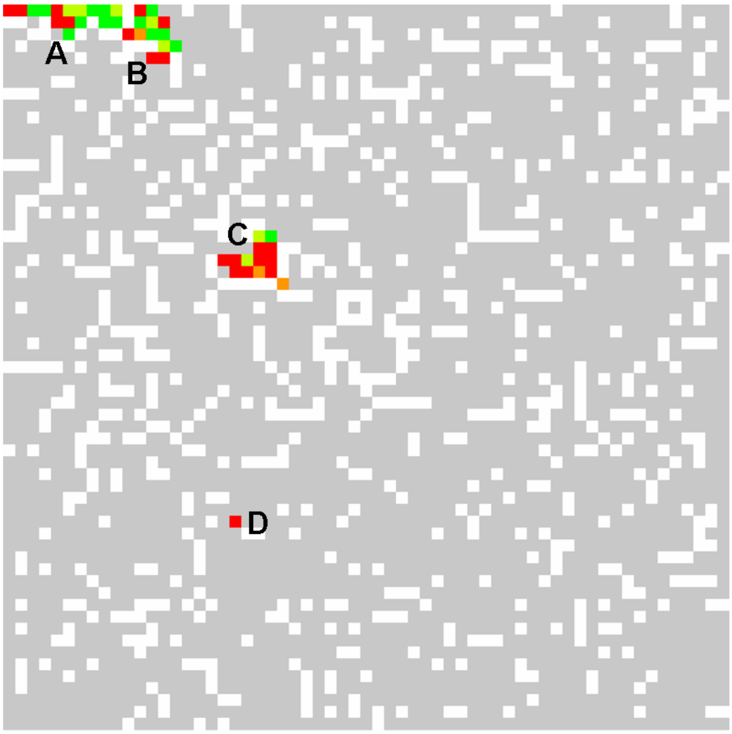
(j) 14x14,100,R (k) 14x14,500,R (l) 14x14,1000,R

Figure 3.2: Self-organizing maps of compounds with known activity using 2D autocorrelation descriptors. Maps are coloured according to the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, white: empty.

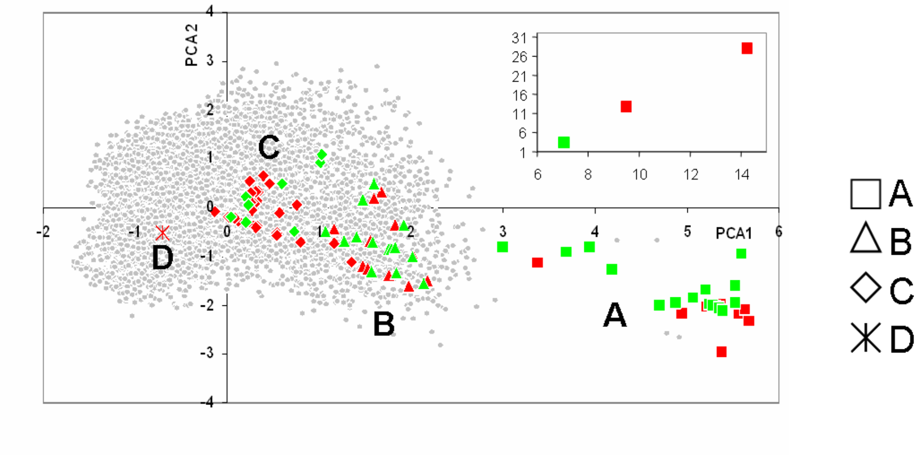
in one case (map 14x14,1000,R with 2D autocorrelation descriptors), the random shuffling leads to a 0.01 units higher accuracy value. The differences tend to be higher for the smaller networks, while for the 14x14 networks the differences are negligible. This effect can again be explained by a higher number of singletons on the larger maps. Here a rearrangement of the activity values has less influence on the accuracy values.

Before the training of the compounds together with the whole screening database, studies were performed with smaller subsets, allowing the calculation of a larger amount of maps in a reasonable time scale. These sets were selected as being similar to the training database according to their Euclidean distance using the 2D autocorrelation and the VSA descriptors, respectively. Comparison of one of the maps with the plot of the first two principal components of the 2D autocorrelation descriptors (explaining 48% of the variance) shows similar clustering of the compounds (see figure 3.3). Clusters B and C are in close vicinity to each other on the principal component plot, while clusters A and B seem to be nearer on the self organizing map. This can be explained by the ability of self-organizing maps to focus on regions where more objects are located. Using this type of visualization, one has to be careful that the distances between the neurons are not equal, as larger distances are generally not preserved in the two dimensional mapping. The distances between the neurons are usually larger in regions where no objects are located, and smaller in crowded regions. In the case of this map, cluster A is placed in a corner of the map and separated from the rest by empty neurons. The small distance allows to focus more on separating the crowded regions. The principle component plot on the other hand focuses more on this distance as it focuses on the variance of the dataset. In this way, outliers can influence the method and prevent it from showing a high resolution in the more crowded area.

Cluster A in the upper left corner of the map contains the molecules with both indole rings present. Compound **1** for example can be found in this cluster. Compounds **19** and **20**, which contain triflate (trifluoromethanesulfonate) groups on the quinone ring, are found in neurons 0/0 and 0/1, respectively. Using principal component analysis, these compounds end up



(a) Self-organizing map



(b) Principal component analysis

Figure 3.3: Comparison of SOM and PCA using 2D autocorrelation descriptors.

Size,Epochs,Top.	Size DB	$Acc_{Act.}$	$Acc_{Inact.}$	$Pre_{Act.}$	$Pre_{Inact.}$	Hits
43x43,100,R	7 418	0.78	0.82	0.76	0.84	0
43x43,100,T	7 418	0.85	0.79	0.74	0.88	0
43x43,500,R	7 418	0.85	0.77	0.73	0.88	0
43x43,1000,R	7 418	0.83	0.75	0.71	0.86	0
61x61,100,R	7 418	0.85	0.82	0.78	0.89	0
61x61,100,T	7 418	0.95	0.79	0.76	0.96	1
61x61,500,R	7 418	0.88	0.81	0.77	0.90	1
61x61,1000,R	7 418	0.95	0.79	0.76	0.96	0
392x392,100,R	613 803	0.81	0.79	0.74	0.85	87
392x392,100,T	613 803	0.86	0.79	0.76	0.88	58

Table 3.4: Maps with 2D autocorrelation descriptors

far away from the majority of the molecules. On the map, the molecules are separated from the screening library molecules by empty neurons. Compounds in cluster B contain scaffolds, where at least one indole ring is missing or replaced by a phenyl residue, or the quinone ring is replaced with a naphthoquinone as described by Lin et al.[43] All compounds in cluster C have a missing indole ring. Derivatives of this type are also found in cluster B, but there the indole ring has substituents with aromatic rings or conjugated double bonds, while the substituents in cluster C are in general smaller and contain one double bond at maximum. Compound **92**, a squaric acid derivative is separated from the rest of the training compounds in both the principle component plot and on the map (cluster D). As only 8 from the 39 molecules in cluster C are active, and the only compound in cluster D is inactive, the map allows to some extend to separate active from inactive molecules.

Table 3.4 gives an overview on the self organizing maps calculated with 2d autocorrelation descriptors, table 3.5 on those calculated with VSA descriptors using the subset as well as the whole screening database. Pictures of these maps are shown in the appendix (figures A.1–A.7 on pages 141–147). Accuracy and precision values were calculated, to judge the maps' abilities to separate active from inactive compounds. The hits of map 557x557,100,R were not included into the final list of identified compounds due to its long calculation time of over 2 months.

Size,Epochs,Top.	Size DB	$Acc_{Act.}$	$Acc_{Inact.}$	$Pre_{Act.}$	$Pre_{Inact.}$	Hits
43x43,100,R	7 227	0.86	0.81	0.77	0.89	3
43x43,500,R	7 227	0.84	0.81	0.77	0.87	0
43x43,1000,R	7 227	0.81	0.83	0.78	0.86	0
60x60,100,R	7 227	0.84	0.84	0.80	0.88	1
60x60,500,R	7 227	0.88	0.83	0.79	0.91	0
60x60,1000,R	7 227	0.81	0.88	0.83	0.86	0
85x56,100,R	7 227	0.84	0.86	0.82	0.88	2
85x56,500,R	7 227	0.88	0.83	0.79	0.91	0
85x56,1000,R	7 227	0.88	0.83	0.79	0.91	0
392x392,100,R	620 326	0.88	0.74	0.72	0.90	93
557x557,50,R	620 326	0.84	0.64	0.63	0.84	5
557x557,100,R	620 326	0.81	0.72	0.69	0.84	15

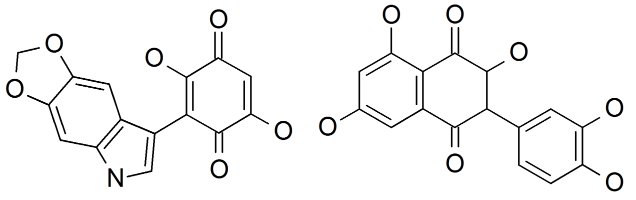
Table 3.5: Maps with VSA descriptors

The maps calculated with the subset of the screening library identify only a few hits. Most of the hits are identified with the networks with a size of 392x392 neurons, while increasing the size to 557x557 decreases the number of co-localizations.

Performance on a new active molecule: The new insulin mimetic molecule (compound **102**[46]) was placed on the trained maps to see if it would have been identified using this method. Indeed, it was placed in the same neuron as two active compounds (**93** and **96**) in some of the maps trained with the subset using 2D autocorrelation descriptors. It could not be identified on the maps trained with the whole database or using the VSA descriptors.

Identified compounds

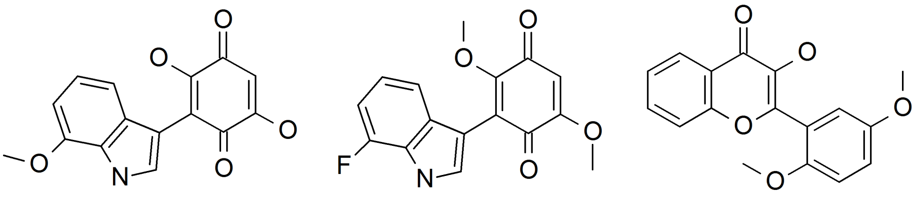
In the following, we will have a closer look on the compounds from the screening database which were placed into an active neuron. Using 2D autocorrelation descriptors to train a subset of the compounds leads to only two new compounds. The map 61x61,100,T identifies compound 000A-0047 (see figure 3.4). This compound is co-localized with **79**, which shows 69% activation of the insulin receptor at 30 µM, as compared to insulin at 17 nM.[43] Compound **79** therefore has comparable activity to the original compound **1**, which shows 73% activation under the same conditions.



(a) **79** (active) (b) 000A-0047 (new)

Figure 3.4: Co-localization in neuron 45/47 on map 61x61,100,T using 2D autocorrelation descriptors.

On map 61x61,500,R (see figure 3.5), compound 6877-0609 is co-localized with one active (**91**) and one inactive compound (**72**). While **72** only shows 6% activation at 30 µM, compound **91** shows 99% of the activity of insulin (17 nM) and is more active than compound **1**.[43]

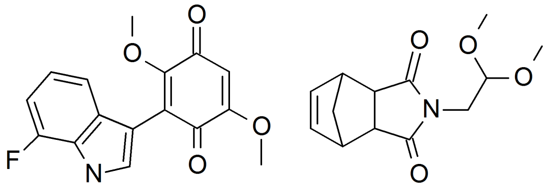


(a) **72** (inactive) (b) **91** (active) (c) 6877-0609 (new)

Figure 3.5: Co-localization in neuron 21/19 on map 61x61,500,R using 2D autocorrelation descriptors.

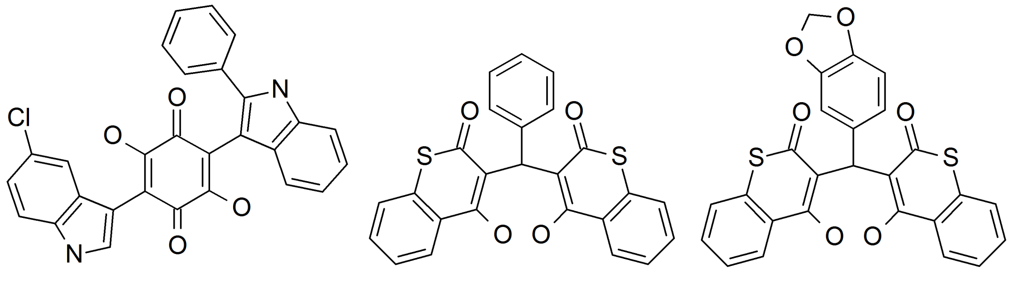
Using VSA descriptors with the selected subset, five molecules have been selected. One of them (5982-0159) was selected with two different maps, each using a different query molecule. Map 43x43,100,R shows co-localizations in two of its neurons (figure 3.6). The first identifies 4161-2736 using query molecule **91**. This very active molecule (99% activity) already identified another compound on map 61x61,500,R using 2D autocorrelation descriptors. When comparing the two identified compounds one can see how different the descriptors behave in identifying similar compounds to the same query (compare figures 3.5c and 3.6b). The second neuron with co-localizations identifies 5982-0100 (**105**) and 5982-0159 using **95** as query, a molecule comparable in activity to compound **1**.[44]

Neuron 8/15:



(a) **91** (active) (b) 4161-2736 (new)

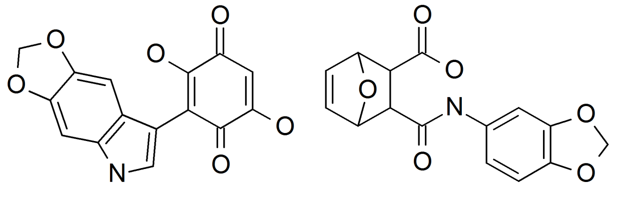
Neuron 36/7:



(c) **95** (active) (d) 5982-0100 (**105**, new) (e) 5982-0159 (new)

Figure 3.6: Co-localizations on map 43x43,100,R using VSA descriptors.

Compound 6049-2038 is placed in the same neuron on map 60x60,100,R as **79**, a medium active compound showing 38% activity at 30 µM as compared to insulin[43] (figure 3.7).

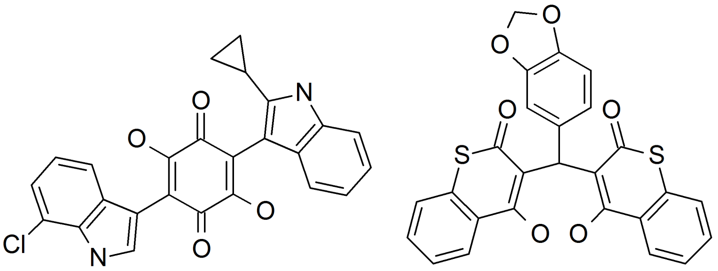


(a) **79** (active) (b) 6049-2038 (new)

Figure 3.7: Co-localization in neuron 29/15 on map 60x60,100,R using VSA descriptors.

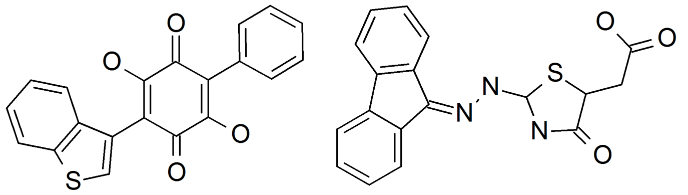
5982-0159, which was already identified using map 43x43,100,R is again identified in neuron 2/21 on map 85x56,100,R. Here, the query is compound **96**, a molecule structurally similar to the earlier query compound **95**, but having the chlorine residue at a different position and a cyclopropane group instead of a phenyl ring. Additionally, activities of both compounds are comparable to compound **1**.[44] In neuron 84/14 of map 85x56,100,R a structurally very different molecule (6231-0119) to the query (**32**) was placed.

Neuron 2/21:



(a) **96** (active) (b) 5982-0159 (new)

Neuron 84/14:



(c) **32** (active) (d) 6231-0119 (new)

Figure 3.8: Co-localizations on map 85x56,100,R using VSA descriptors.

As more than 100 molecules were placed in active neurons on the maps trained with the whole screening database, not all of them will be discussed here. Instead, we will focus on those structures identified with compound **1** as query structure, as well as those which were subsequently selected for biological testing.

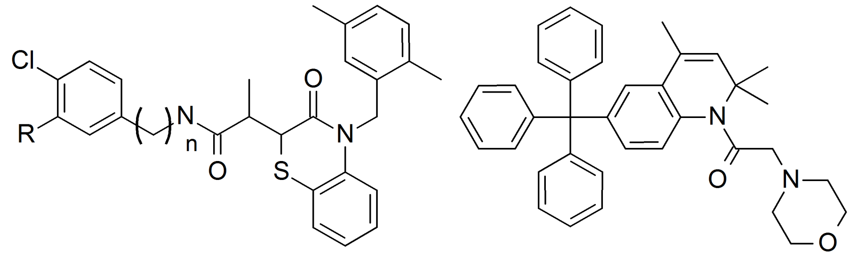
The molecules which were co-localized with compound **1** are shown in figure 3.9). On map 392x392,100,R using VSA descriptors, eight active and seven inactive molecules (compounds **1**–**14** and **97**) were placed in the same neuron (258/286) as two new compounds (C493-1072 and K788-5456). With

Nr.	ChemDiv ID	Queries
103	4204-0085	**67**
104	4451-0051	**44**, **66**, **67**, **68**, **69**
105	5982-0100	**95**
108	8014-1054	**27**
109	C073-3327	**25**
110	C090-0245	**25**
111	D159-0883	**67**
112	E938-0156	**30**, **32**, **33**, **34**
113	K788-0448	**91**

Table 3.6: Molecules selected for testing by self-organizing maps with their query structures.

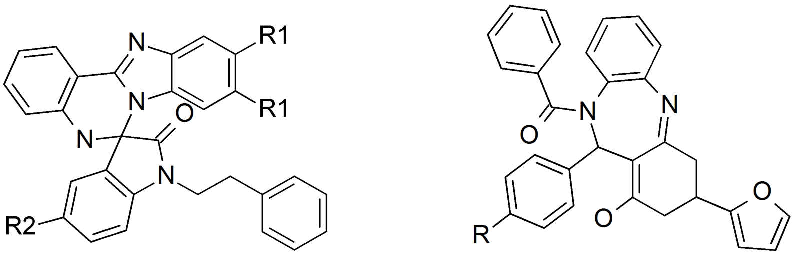
the 2D autocorrelation descriptors, ten active and ten inactive compounds (**1**–**14**, **17**, **18**, **21**, **22**, **23** and **27**) were localized together with six molecules from the screening library (3807-4416, 8017-6445, 8017-6446, F019-1000, F019-2195 and K786-3665) in neuron 11/124 of map 392x392,100,R. The corresponding toroidal map has four molecules (4052-4503, 5218-1214, 5218-1215 and 5218-1227) placed in the same neuron (7/382) as nine active and seven inactive molecules (compounds **1**, **4**–**14**, **21**, **24**, **27** and **97**). All these new molecules look structurally diverse to compound **1**, and the quinone substructure is no longer present.

As will be discussed later, several compounds were selected for experimental evaluation in biological assays. In the case of the self-organizing maps, all those molecules (with the exception of compound **105**) were initially selected from the maps trained with the whole screening database. Compounds **104** and **110** were selected on map 392x392,100,R using 2D autocorrelation descriptors, while compounds **103**, **111** and **112** were selected by the corresponding toroidal network. Using the VSA descriptors, compounds **108**, **109** and **113** were selected on map 392x392,100,R. All these compounds with their corresponding query molecules are summarized in table 3.6.



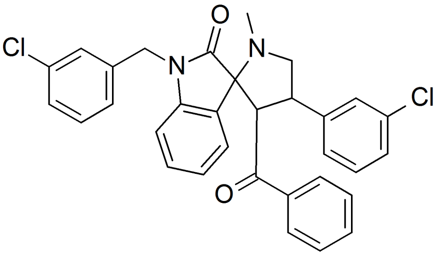
(a) C493-1072: n=2, R=H
K788-5456: n=1, R=Cl

(b) 3807-4416

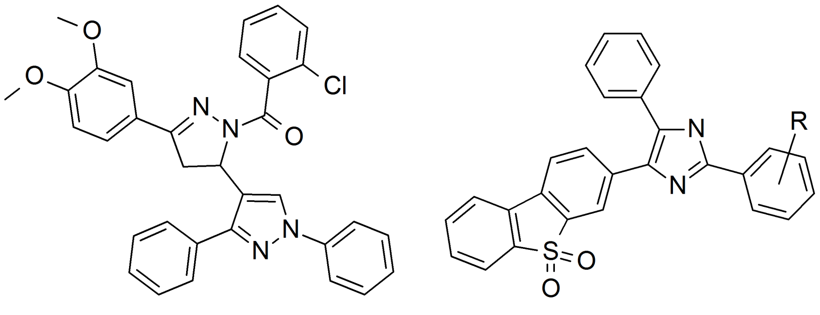


(c) 8017-6445: R1=H, R2=F
8017-6446: R1=methyl, R2=ethyl

(d) F019-1000: R=Cl
F019-2195: R=methoxy



(e) K786-3665



(f) 4052-4503

(g) 5218-1214: R=2,3-dimethoxy
5218-1215: R=3,5-dimethoxy-4-hydroxy
5218-1227: R=2,4,5-dimethoxy

Figure 3.9: Molecules co-localized with compound **1**.

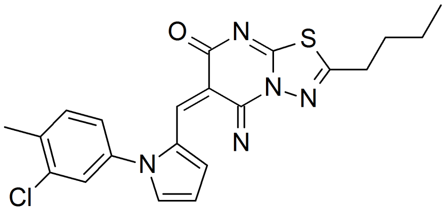
3.1.2 Fingerprint similarity

Performance of the method

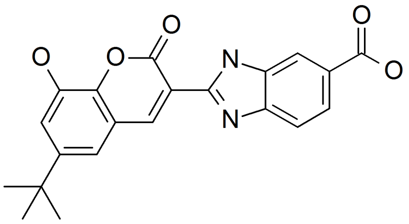
The performance of eight different fingerprint types in retrieving actives was compared to judge their applicability as a screening method in the present work. Figure 3.10 shows the most similar molecules to compound **1** according to these different fingerprints. All of the fingerprints identify a different molecule as the most similar one, with similarity values ranging from 0.50 to 0.89. To see which features are determining the similarity in the case of the MACCS fingerprint, the script ph4maccs_x.svl from the SVL Exchange site[118] was used to compare the keys of **1** and K026-0233. Compound **1** has 46 one-bits and K026-0233 shows 49 from a total of 166 possible features. Of those, 41 are identical in both molecules. The features in common are listed in table 3.7 and the corresponding structures are shown in figure 3.11.

These features are often overlapping, so that more than one feature account for a single structural motive. For example, the tertiary butyl group is found by features 66, 74, 112, 141, 149 and 160 and the benzimidazole group in K026-0233 as well as the indole group in **1** or its nitrogens are found with a total of 16 features (65, 83, 96, 105, 120, 121, 125, 131, 137, 142, 151, 156, 161, 162, 163 and 165).

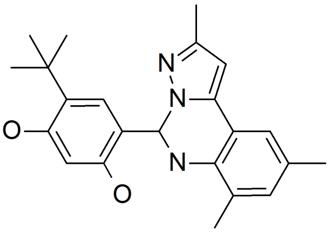
In general, a Tanimoto similarity value of 0.85 is regarded as a threshold for similarity. As the most similar compounds to **1** had very differing similarity values depending on which fingerprint was used, a general investigation of how many compounds are identified at a certain threshold was performed. The results are shown in figure 3.12. The typed atom/graph distance/triangle fingerprints find a higher number of compounds from the screening library, using the insulin mimetic compounds as queries, compared to the rest of the fingerprints at the same similarity threshold. Taking a threshold of 0.6 for example finds the majority of the compounds with TAD, TAT, TGD and TGT, but only a smaller fraction of the compounds using GpiDAPH3, MACCS, piDAPH3 and piDAPH4 fingerprints.



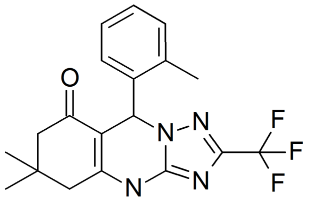
(a) C370-3764
GpiDAPH3: 0.55



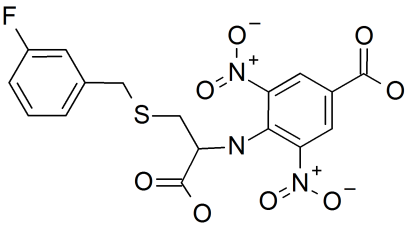
(b) K026-0233
MACCS: 0.76



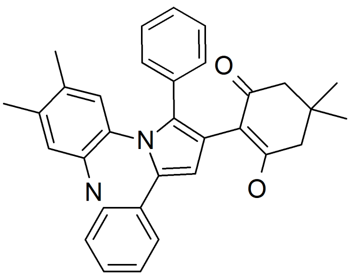
(c) 7210-2197
piDAPH3: 0.66



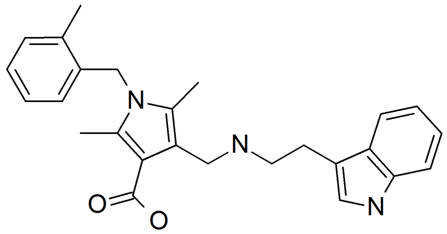
(d) K832-3461
piDAPH4: 0.50



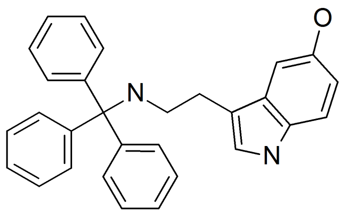
(e) K781-8112
TAD: 0.85



(f) 6364-0110
TAT: 0.77



(g) C879-0521
TGD: 0.89



(h) 0167-0107
TGT: 0.87

Figure 3.10: Molecules of the screening library most similar to DMAQ-B1 (compound **1**) using different fingerprints.

Nr.	Description
50:	C in C=C bonded to >= 3 C
62:	non-ring bonds that connect rings
65:	N in aromatic bonds with C
66:	CX4 bonded to >= 3 carbons
72:	O separated by 3 bonds
74:	dimethyl substituted atoms
76:	C in C=C bonded to >= 3 heavy atoms
83:	heteroatoms in 5 ring
96:	atoms in 5-rings
99:	C in C=C; hets. ring bonded to a 3-ring bond X
112:	atoms with coordination number >= 4
120:	key(137)-1 if key(137)>1; else 0 (comment: >= 2 heterocycle atoms in rings)
121:	N in rings
125:	Is # of aromatic rings > 1?
127:	key(143)-1 if key(143) > 1; else 0 (comment: >= 2 non ring O connected to a ring)
131:	het atoms with H
136:	Is there more than 1 O=?
137:	Total # ring heterocycle atoms
139:	OH groups
140:	key(164)-3 if key(164)>3; else 0 (comment: >= 4 oxygens)
141:	key(160)-2 if key(160)>2; else 0 (comment: >= 3 CH3 groups)
142:	key(161)-1 if key(161)>1; else 0 (comment: >= 2 N atoms)
143:	non ring O connected to a ring
144:	atoms separated by (!:):(!:)
145:	#6M ring >1
146:	key(164)-2 if key(164)>2; else 0 (comment: >= 3 oxygens)
149:	key(160)-1 if key(160)>1; else 0 (comment: >= 2 CH3 groups)
150:	#X separated by (!r)-r-(!r)
151:	NH
152:	C bonded to >=2 C and 1 O
154:	O in C=O
156:	XN where coord. # of X>=3
157:	O in C-O single bonds
159:	key(164)-1 if key(164)>1; else 0
160:	CH3 groups
161:	N
162:	aromatics
163:	atoms in 6 rings
164:	oxygens
165:	ring atoms

Table 3.7: MACCS fingerprint features shared by **1** and K026-0233.

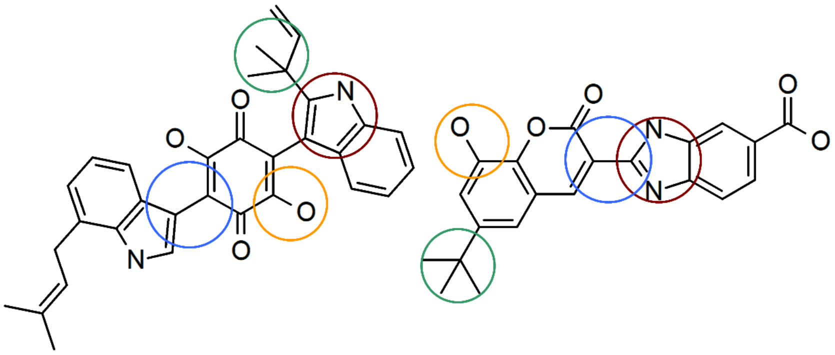


Figure 3.11: Compound **1** and K026-0233. Some selected common features are marked with coloured circles.

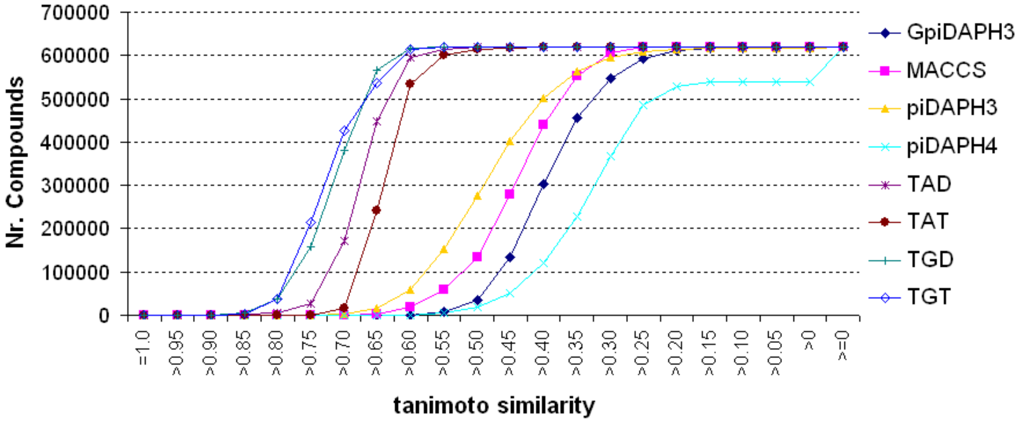


Figure 3.12: Number of identified compounds at different similarity thresholds.

Another description of this effect is given in table 3.8, which shows the similarity threshold which can be used to find 1% of the combined query and screening databases. These similarity values range from 0.54 to 0.84. Additionally, the enrichment factors of active compounds at this value are given. Although the enrichment factor values are very high, one needs to consider that the active molecules show a high structural similarity to each other. The values seem to be higher for those fingerprints where a lower similarity value is needed to find one percent of all compounds. This effect is even more pronounced when only the first 100 structures are observed.

FP	$SV_{1\%}$	$EF_{1\%}$	EF_{100}
GpiDAPH3	0.55	100	6203
MACCS	0.64	100	6058
piDAPH3	0.68	97.7	5914
piDAPH4	0.54	100	6058
TAD	0.80	83.7	3173
TAT	0.75	97.7	6058
TGD	0.84	95.3	5193
TGT	0.83	97.7	5770

Table 3.8: Similarity values ($SV_{1\%}$) and enrichment factors at 1% of the database ($EF_{1\%}$) and for the first 100 compounds (EF_{100}). The maximum enrichment factors are 100 and 6203, respectively.

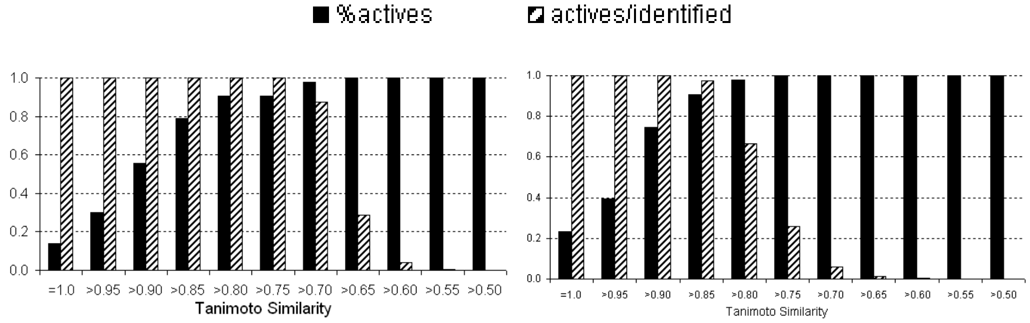
The behaviour of all the investigated fingerprints in retrieving actives as compared to the identified ratio of actives to inactives is shown in figure 3.13. While most of the fingerprints find the majority of the active compounds before the ratio of actives in the identified subset is decreased, the behaviour is different for the TAD fingerprint. This fingerprint reaches a low ratio of actives in the selected subset long before all actives are identified.

As the fingerprints are very different in their behaviour, we decided to set the similarity threshold for selecting new hits separately for each fingerprint. The threshold was set at the similarity step which identified a ratio of at least 60% actives. The chosen thresholds are summarized in table 3.9. TAT was excluded from the screening as no compounds were selected according to the selected threshold. As can be seen in figure 3.13f, no molecules from the screening library were identified with a threshold of 0.80, a similarity value at which nearly all active compounds have already been selected. The ratio of actives to inactives drops drastically at the next similarity step (> 0.75).

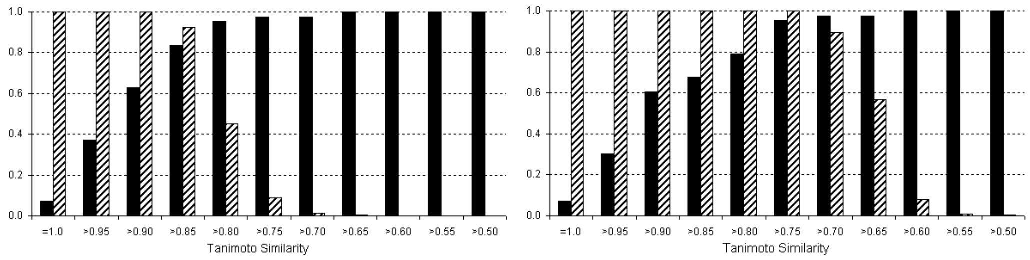
GpiDAPH3	MACCS	piDAPH3	piDAPH4	TAD	TGD	TGT
> 0.70	> 0.80	> 0.85	> 0.70	> 0.95	> 0.95	> 0.90

Table 3.9: Chosen thresholds for the fingerprints.

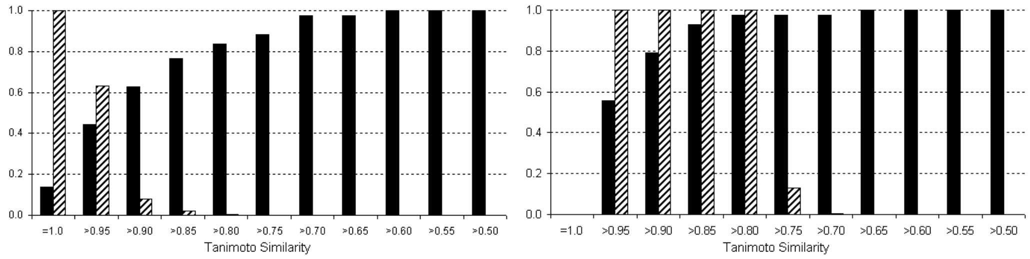
Performance on a new active molecule: Using compound **102** as external validation, none of the fingerprint methods would have identified this molecule using the defined thresholds. The most similar molecules to **102** using the different fingerprints and their corresponding similarity values are given in table 3.10. Only three of the fingerprints have the same query structure as the most similar one (**26**). The remaining fingerprints all have different queries as the most similar structure.



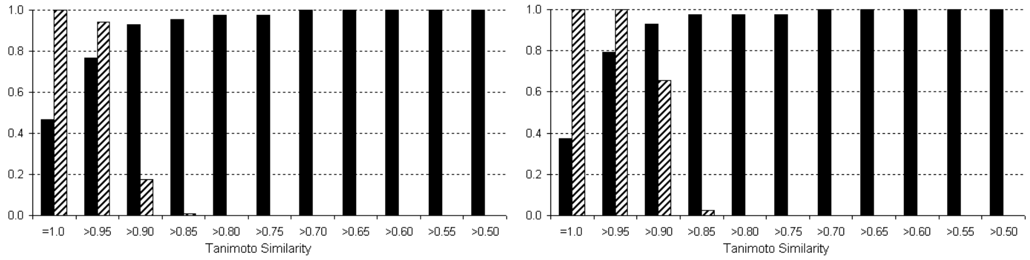
(a) GpiDAPH3 (b) MACCS



(c) piDAPH3 (d) piDAPH4



(e) TAD (f) TAT



(g) TGD (h) TGT

Figure 3.13: Performance of different fingerprints in retrieving actives.

fingerprint	query	similarity
MACCS	**79**	0.6981
GpiDAPH3	**25**	0.5684
piDAPH3	**69**	0.6365
piDAPH4	**81**	0.4794
TAD	**28**	0.7324
TAT	**26**	0.7367
TGD	**26**	0.8246
TGT	**26**	0.8384

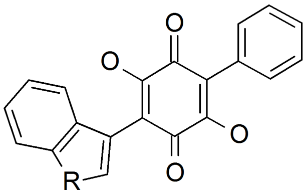
Table 3.10: Most similar queries to compound **102**

Identified compounds

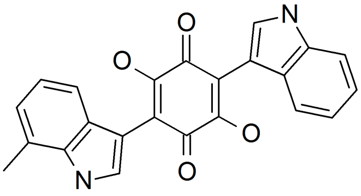
Using the thresholds defined in table 3.9, all selected molecules were found by using only ten active compounds as queries. These query structures are summarized in figure 3.14 on page 70. The original insulin mimetic (**1**) did not identify any new molecules at the selected thresholds. Instead, the successful queries showed simpler structures, without large residues on the indole rings and some with the indole ring substituted by a phenyl/naphthyl ring or missing at all. Compounds **15**, **31** and **32** are all more active than **1**, showing EC_{50} values of 0.3–1.5 µM as compared to the EC_{50} of 5 µM of the original compound.[38] Molecule **29** with an EC_{50} value of 7 µM is comparable to **1** in activity and structure, but lacks the large chains on the indole ring.[38] Compounds **33** and **34**, which differ only in the position where the rest of the molecule is connected on the naphthalene group, have a large difference in activity (EC_{50} of 6 and 30 µM, respectively).[38] While compounds **79**, **80** and **84** are only medium active showing 34–38% activity as compared to insulin at 30 µM, **91** shows 99% activity. As comparison, compound **1** has an activity of 73% under the same conditions.[43] In the following, an overview of the identified molecules using these queries is given.

The GpiDAPH3 fingerprint compares the graph distances of three features (aromatic system, donor or acceptor). Here, it identified six new compounds using three different queries with a threshold of 0.70. The structures are depicted in figure 3.15. When comparing the molecules to their query

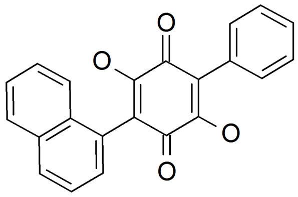
Queries:



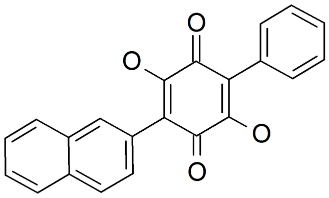
(a) **15**: R=NCH_3
31: R=O
32: R=S



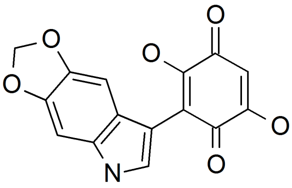
(b) **29**



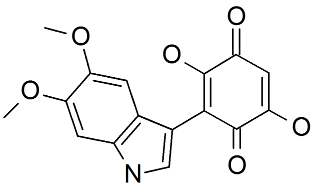
(c) **33**



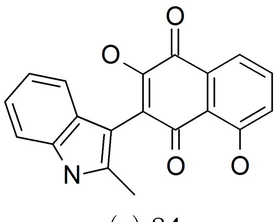
(d) **34**



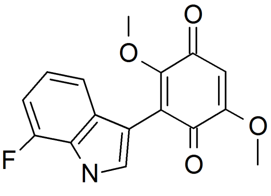
(e) **79**



(f) **80**



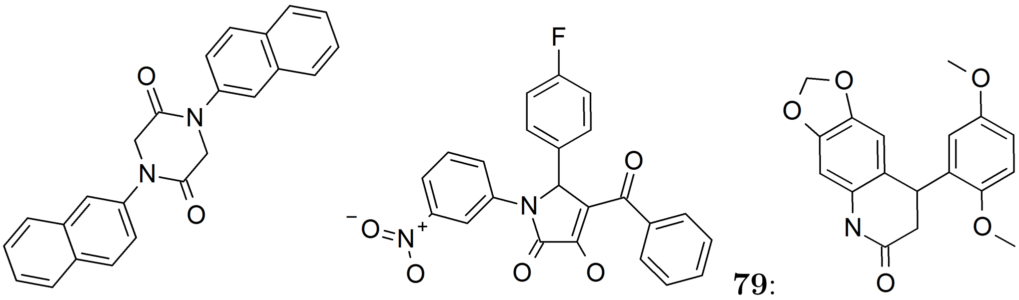
(g) **84**



(h) **91**

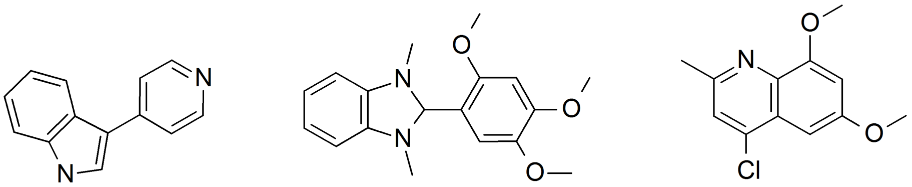
Figure 3.14: Fingerprint query structures used for the final selection.

34: **79**:



(a) 8001-9473 (b) 8009-8459 (c) C276-0105

91:



(d) 1683-6896 (e) 5696-0017 (f) 8017-3855

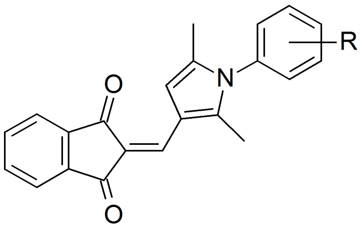
Figure 3.15: Compounds identified with GpiDAPH3

structures, they show in general a similar size and a corresponding distribution of the features. However, 8009-8459 looks different in respect to the arrangement of the rings. Instead of having the ring systems connected in a linear way, three rings are connected to a central part. Molecules showing this type of scaffold have also been identified using self-organizing maps with 2D autocorrelation descriptors.

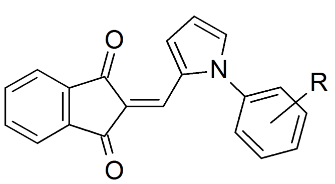
MACCS identified 21 compounds using eight query structures with a threshold of 0.80. Together with TGT, this fingerprint selects the highest number of molecules. However, some of this molecules show a high similarity to each other (see figures 3.16 and 3.17). Nearly all of the molecules selected with **15** and **31** share an indane-1,3-dione substructure. With a Tanimoto value of 0.89 to **80**, K815-0024 (figure 3.17e) shows the highest similarity to an active molecule achieved with MACCS fingerprint.

Using the piDAPH3 fingerprint at a threshold of 0.85, only three molecules were identified (figure 3.18 on page 74). Compound C270-0349 was identified with three different query structures (**15**,**31** and **32**) which are identical using this type of fingerprint. Compounds **33** and **34**, which have

15:

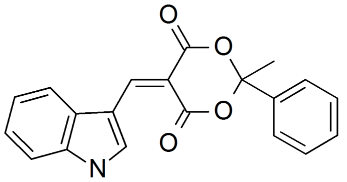


(a) 4281-2071: R=p-carboxy
4281-2127: R=m-carboxy
4764-3635: R=m,m-dicarboxy
6332-2060: R=o-methyl,p-carboxy



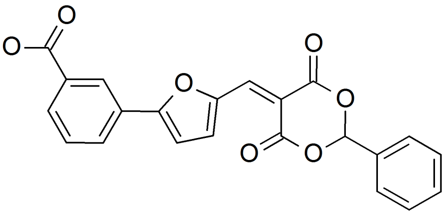
(b) 5775-0353 (**106**): R=m-carboxy
6463-4542: R=p-carboxy

29:

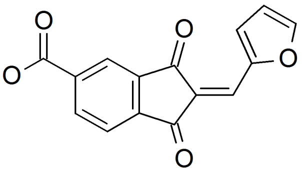


(c) 5650-0022

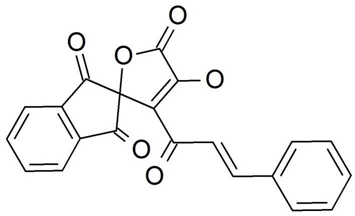
31:



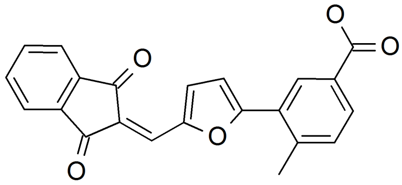
(d) 3553-1638



(e) 4057-0014



(f) 5340-2513



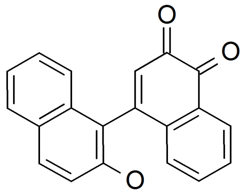
(g) 6332-1077

Figure 3.16: Compounds identified with MACCS (part 1)

nearly identical structures, identify two different types of scaffolds. 4659-0068 has a rhodanine substructure, while 5750-3148 has a structure which is more similar to the naphthoquinone derivatives found to be mostly inactive on the insulin receptor.[43]

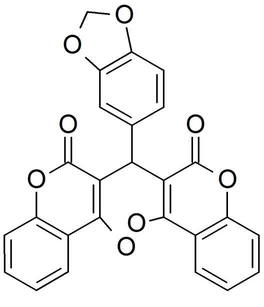
The 4-point pharmacophore fingerprint piDAPH4 finds five molecules with two of the query structures with a similarity higher than 0.70 (see figure 3.19 on page 75). Using query compound **33** both identified compounds show a rhodanine substructure, similar to the molecule identified with this query using the piDAPH3 fingerprint. Using compound **91** as query, three molecules with a thiazolidinedione substructure are identified. These mole-

33/**34**:

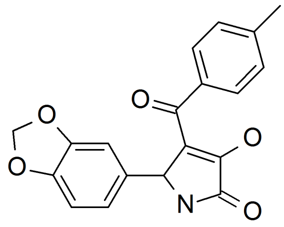


(a) 000A-0190

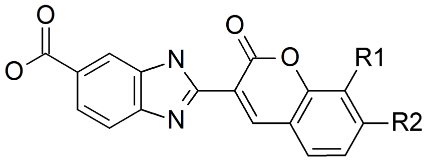
79:



(b) 3029-0578

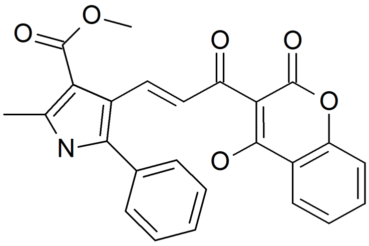


(c) 8009-3415



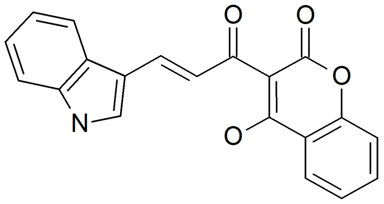
(d) K026-0216: R1=H, R2=OBn
K026-0228: R1=OH, R2=H
K026-0229: R1=H, R2=OH

79/**80**:



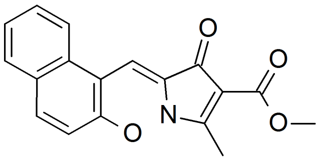
(e) K815-0024

79/**84**:

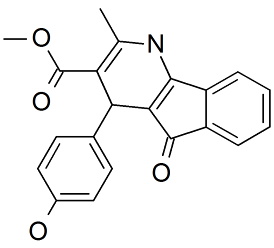


(f) K815-0023 (**114**)

80:



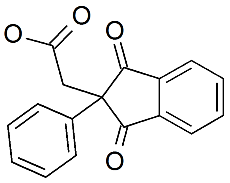
(g) 4693-1125



(h) 5408-1692

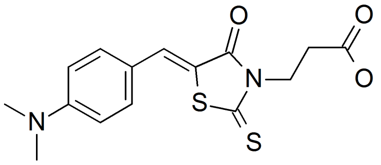
Figure 3.17: Compounds identified with MACCS (part 2)

15/31/32:



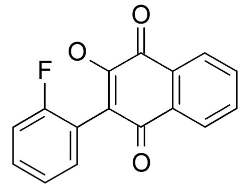
(a) C270-0349 (**121**)

33:



(b) 4659-0068

34:



(c) 5750-3148

Figure 3.18: Compounds identified with piDAPH3

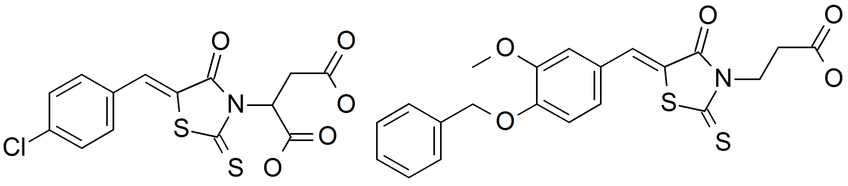
cules differ structurally from the molecules identified using GpiDAPH3 with the same query.

With two query structures only, the TAD fingerprint finds 11 new molecules at a threshold of 0.95 (see figure 3.20 on page 76). Although having very high similarity values to the queries, they do not contain any quinone substructures. Instead, the indole ring is more frequently found in structures identified by TAD.

Using a threshold of 0.95, TGD finds two new molecules, one of them with two of the query structures (see figure 3.21 on page 77). Here, the similarity of the query structures is also reflected in a high similarity of the identified structures to each other.

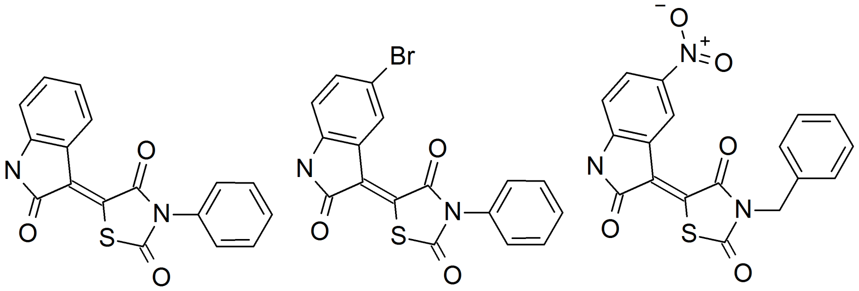
TGT selects 21 molecules using a threshold of 0.90. Of the selected molecules, 16 were identified using **91** as query (see figure 3.23). Nearly all of these molecules have an indole ring in their structure. The query structure has methoxy- instead of the hydroxyl-groups at the quinone ring. This might be reflected by a high number of ester and ether groups in the hit list. The remaining structures were identified by several queries simultaneously (see figure 3.22).

33:



(a) 1682-6957 (b) 3057-0993

91:

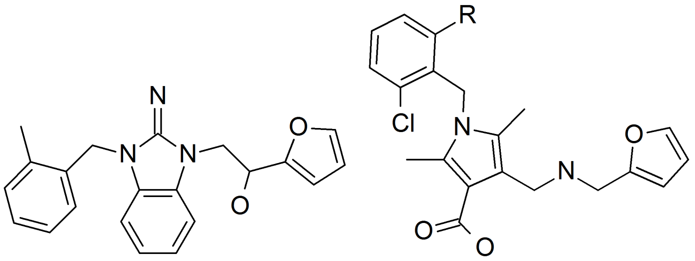


(c) 2110-0307 (d) 2110-0308 (e) 3232-1864

Figure 3.19: Compounds identified with piDAPH4

Overall, fingerprint similarity identifies a high number of interesting new structures. Different to self-organizing maps, the utilized selection procedure did not lead to molecules selected by compound **1** as query. Instead, simpler query structures were used. The identified molecules look quite diverse when the results from different fingerprint types, as well as different query structures are compared. This shows the importance of using a set of queries, even if there is redundancy in the input structures.

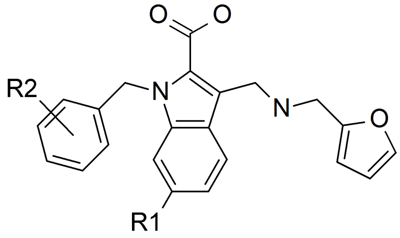
31:



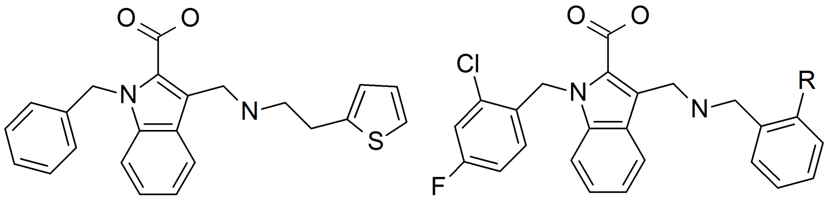
(a) 2950-0554

(b) C879-1278: R=F
G396-0972: R=H

33:

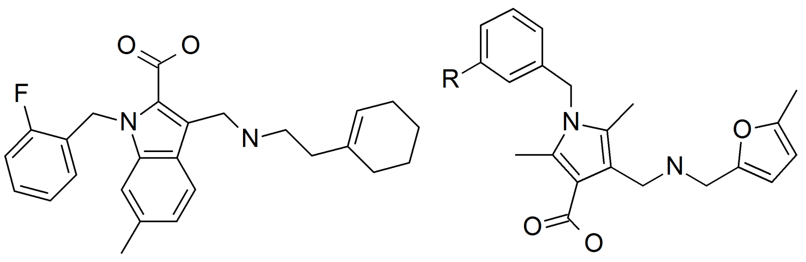


(c) C753-0198: R1=H, R2=o-F
E518-1612: R1=CH_3, R2=p-CH_3



(d) C753-1342

(e) E693-0068: R=Cl
E693-0476: R=F

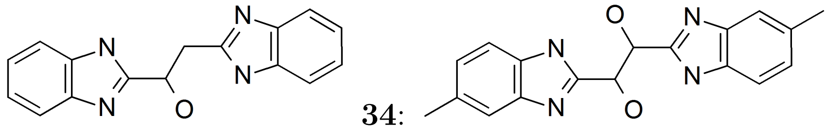


(f) E847-0220

(g) G396-0138: R=CH_3
G396-0426: R=F

Figure 3.20: Compounds identified with TAD

32/**33**:

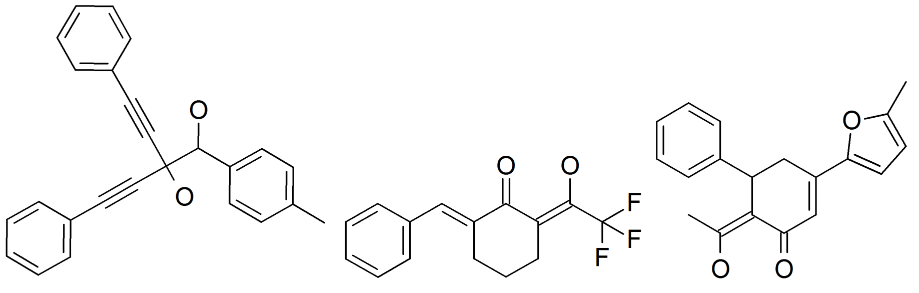


34:

(a) 4587-0405 (b) 6843-3207

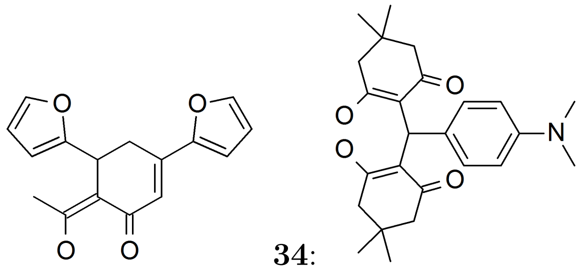
Figure 3.21: Compounds identified with TGD

15/**31**/**32**/**33**/**34**:



(a) 1302-0002 (b) 5634-0239 (c) 8012-0041

15/**31**/**32**/**33**:

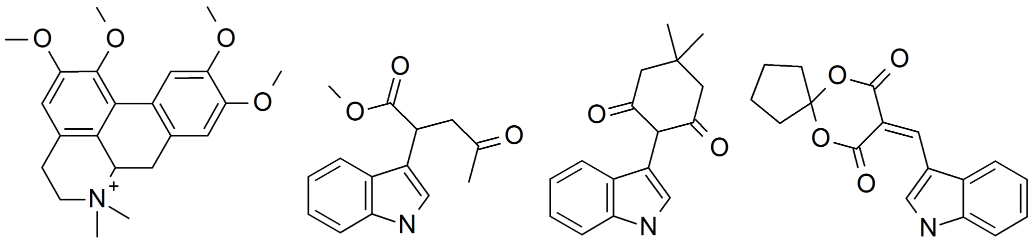


34:

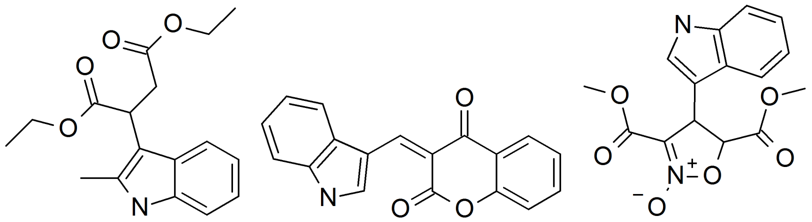
(d) 8012-0040 (e) 5547-0011

Figure 3.22: Compounds identified with TGT (part 1)

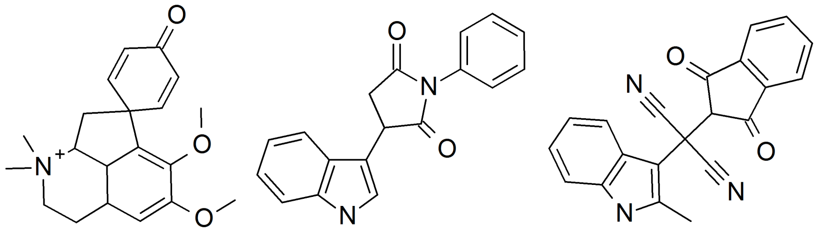
91:



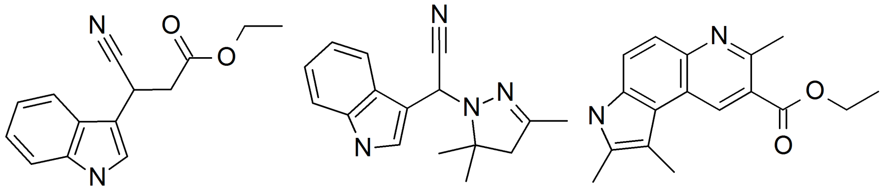
(a) 000A-0036 (b) 0392-0008 (c) 0883-0041 (d) 3270-0678



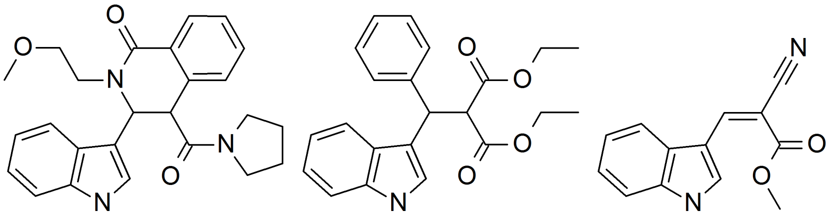
(e) 3630-0578 (f) 3989-0098 (g) 4513-0296



(h) 6944-0119 (i) 7244-0063 (**107**) (j) 8009-7265 (**120**)



(k) 8012-8948 (l) 8017-5369 (m) 8017-7046



(n) C294-0271 (o) D155-0032 (p) K781-0936

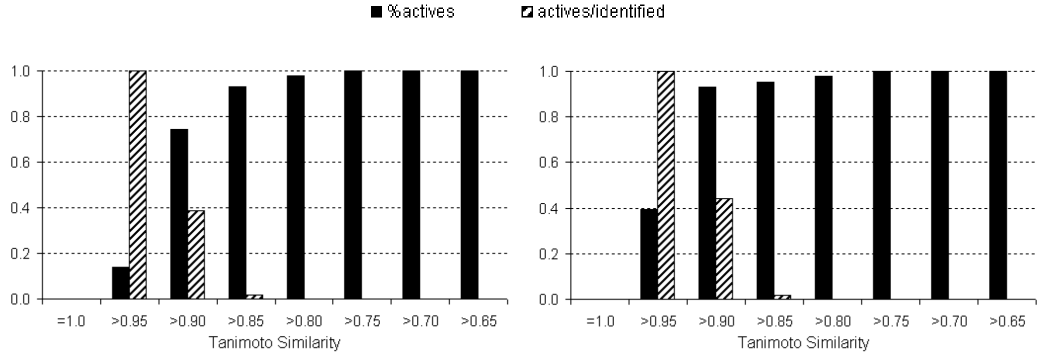
Figure 3.23: Compounds identified with TGT (part 2)

3.1.3 Shape similarity

Performance of the method

Shape similarity finds the first percent of the compounds at a similarity value of 0.83. The enrichment factor at this percentage is 97.7, which corresponds to finding 42 of the 43 active molecules.

Using the same definition of the threshold as done for the fingerprint similarity, no new compounds would have been selected by shape similarity. To include at least some hits from shape similarity, the similarity threshold was decreased by 5% from 0.95 to 0.90. At this threshold, the ratio of actives to identified compounds was approximately 40% (figure 3.24a).



(a) (b)

Figure 3.24: Performance of shape similarity in retrieving actives. The shown bars are excluding (a) or including (b) identical molecules.

Interestingly, this method does not identify any active compound at a similarity higher than 0.95, even if the active molecules are allowed to find themselves (figure 3.24b). It seems that the generated conformations do not include the original conformation of the queries.

Using 4 different conformations of compound **1** by rotating the ring systems leads to a diverse hit list. From the first 100 compounds, only eleven were identified by two different conformations, and only one (C499-0927) was identified by three conformations.

Performance on a new active molecule: As with the fingerprints, the new active molecule (**102**) would not have been identified using shape similarity at the selected threshold. The most similar query is compound **9**, showing a similarity of 0.71. The aligned structures of those two molecules are shown in figure 3.25.

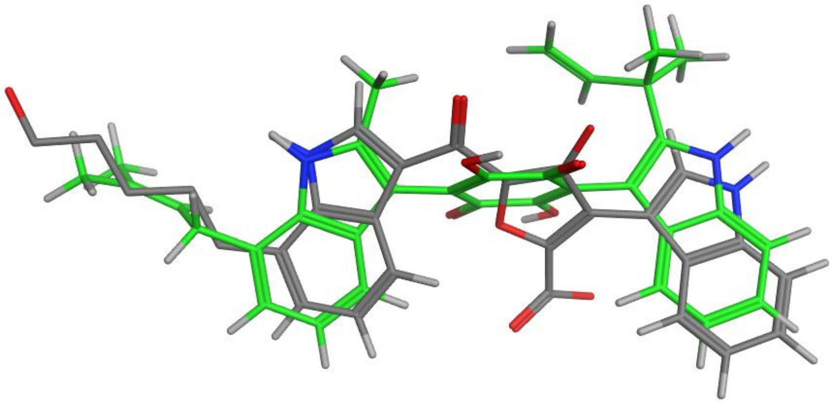
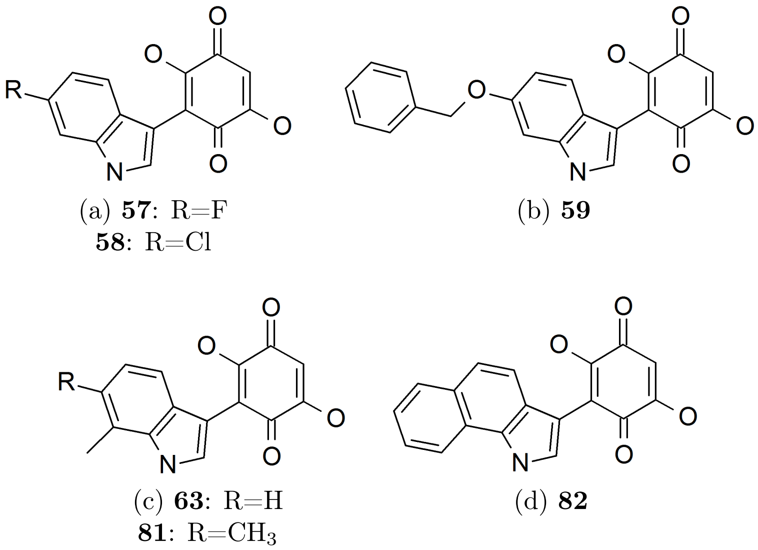


Figure 3.25: Shape similarity alignment of query structure **9** (green) with the new active compound **102** (grey).

Identified compounds

In total, 51 molecules of the screening database were selected with a similarity value higher than 0.90. Eleven of the query molecules were used for the selection of this new compounds. Five of them (compounds **33**, **79**, **80**, **84** and **91**) have been used as queries for fingerprint similarity as well. Those structures can be seen in figure 3.14 on page 70, the remaining compounds are shown in figure 3.26. The compounds most similar to the queries are depicted in figure 3.27. In many cases, the ring topology remains identical or very similar to the query structures. However, the quinone substructure is usually replaced by substituted benzene rings.

As only the shape, but not the properties of the molecules were taken into account using this screening method, the possibility of finding active molecules with this technique might be lower than that of the previous methods. Nevertheless, the hits were included for further processing to investigate the overlap with the other identified molecules.



(a) **57**: R=F
58: R=Cl

(b) **59**

(c) **63**: R=H
81: R=CH_3

(d) **82**

Figure 3.26: Shape query structures used for the final selection.

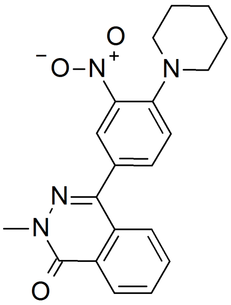
3.1.4 Selection of compounds for testing

One aim of this thesis was to evaluate the results of the *in silico* screening with biological experiments. For this, we sought a number of representative compounds in the hit-lists of the different methods.

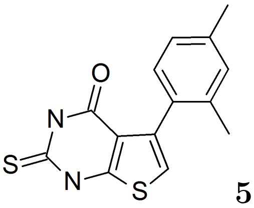
Combining the identified compounds by self-organizing maps (VSA or 2D autocorrelation descriptors) as well as the fingerprint and shape similarity approaches, a total number of 367 compounds was identified. A summary of all these compounds can be found in table A.2 on page 130.

None of the compounds was identified with more than one method. Therefore, the scaffolds of the identified molecules were further investigated to reduce the number of hits to an amount feasible for testing. In total, 112 scaffolds were retrieved, of which 37 were found more than once with one method or with at least two different methods. An overview of the number of compounds and scaffolds identified with the different methods is presented in table 3.11.

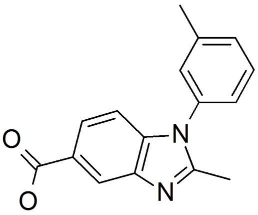
The number of scaffolds was further restricted to those which were found by at least two different methods and/or at least ten times with one of the methods. This was done to increase the possibility to find active molecules,

33: 

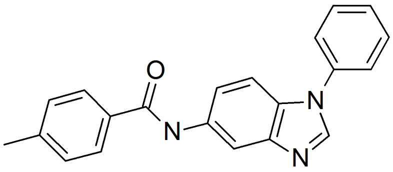
(a) 5498-2747

57: 

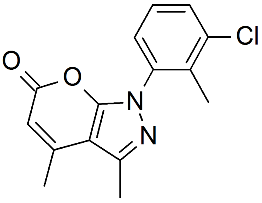
(b) 3237-1339

58: 

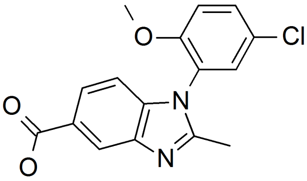
(c) R153-0142

59: 

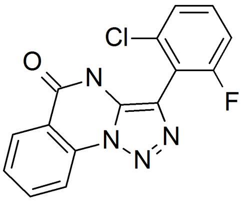
(d) 3546-0632

63: 

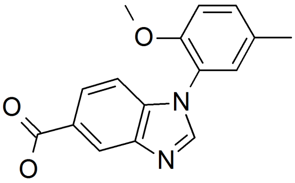
(e) 6266-1861

79: 

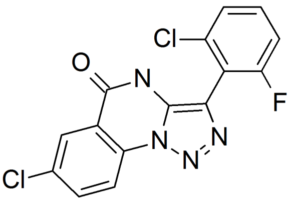
(f) R153-0159

80: 

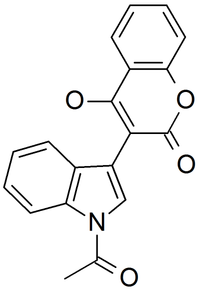
(g) C200-2391

81: 

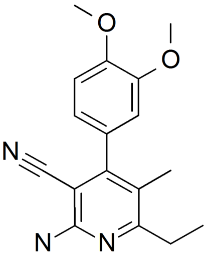
(h) R153-0213

82: 

(i) C200-4795

84: 

(j) 3347-1012

91: 

(k) 8017-7602

Figure 3.27: Most similar compounds to the queries according to shape.

Method	Nr. compounds	Nr. scaffolds
SOM 2D	145	51
SOM VSA	102	25
FP	69	26
Shape	51	25
Sum	367	127
Total	367	112

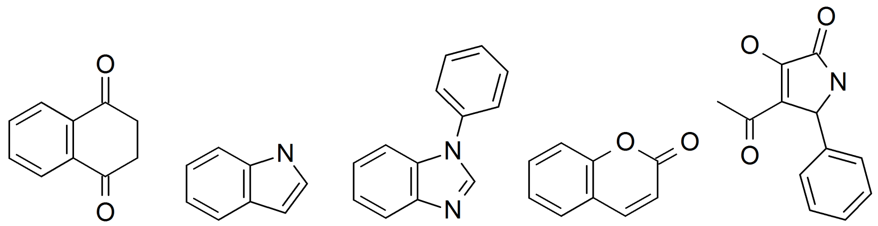
Table 3.11: Identified compounds and scaffolds

but also to have a number of derivatives for further investigations in case that an active molecule is identified. Structures of all scaffolds are shown in figure 3.28.

Representative structures of these clusters were determined by calculating the medoid according to three different types of descriptors, as described in section 2.1.6 on page 36. Independently, one molecule per cluster was selected as representative structure by hand. A consensus of the three medoids and the selection by hand was searched by counting the votes and the corresponding molecules were purchased for biological evaluation. The structures of these molecules are shown in figure 3.29. The results of the cellular assay are presented in section 3.2.2 on page 94.

In a follow-up study, derivatives of the active compound **112** were purchased (figure 3.30) to allow the identification of preliminary structure-activity relationships. These were mainly selected out of scaffold cluster 105, trying to find a representative overview of all molecules in this cluster. Additionally, some compounds bearing the same scaffold were chosen out of the whole screening library, to complement the selection. Compound **116** was picked as substitute for 3331-2182, a compound found by self-organizing maps with VSA descriptors, which has an additional fluorine at the para position and was not available for purchase.

An overview of all purchased compounds, as well as the method which was used to identify them, is given in table A.3 on page 140.



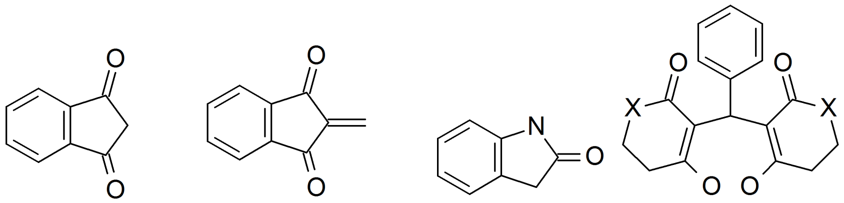
(a) 20: all methods

(b) 32: all methods

(c) 34: Shape

(d) 87: all methods

(e) 88: SOM (2D), FP

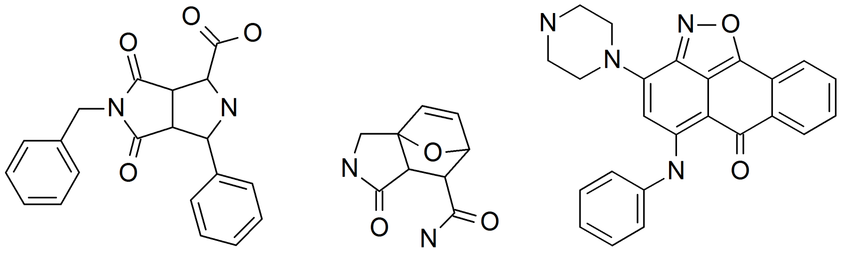


(f) 105: SOM (2D/VSA), FP

(g) 139: SOM (VSA), FP

(h) 150: SOM (2D)

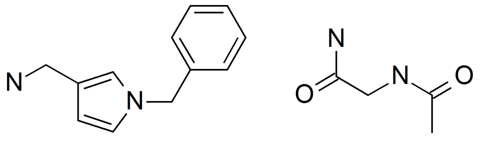
(i) 208: SOM (VSA), FP



(j) 291: SOM (VSA)

(k) 331: SOM (VSA)

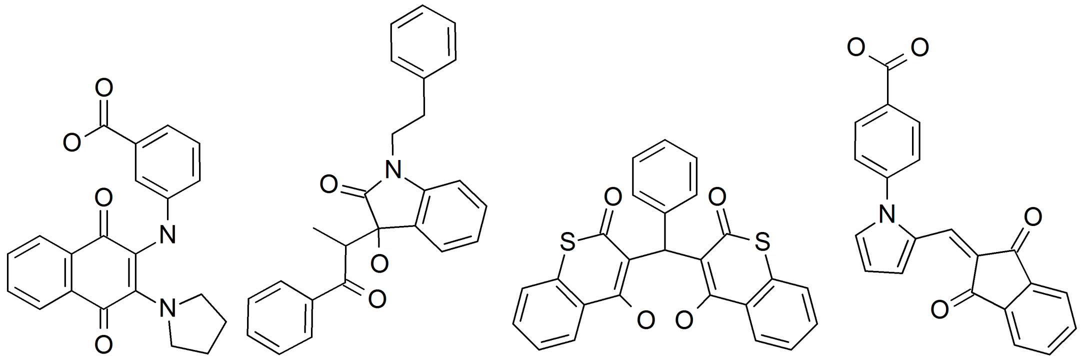
(l) 339: SOM (2D)



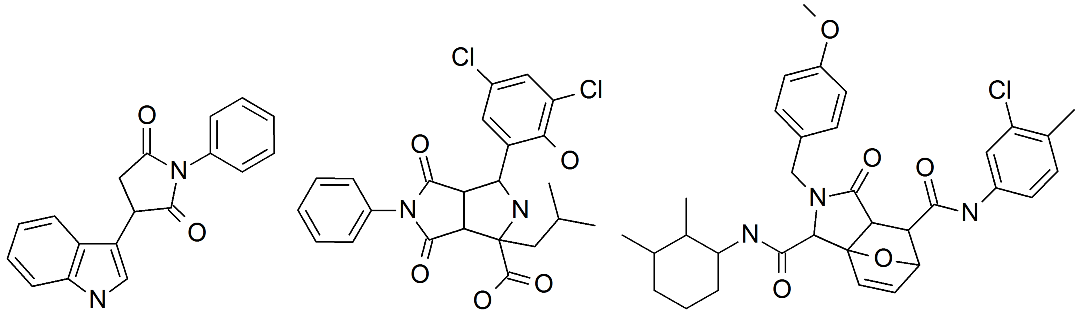
(m) 380: FP

(n) 385: SOM (2D/VSA)

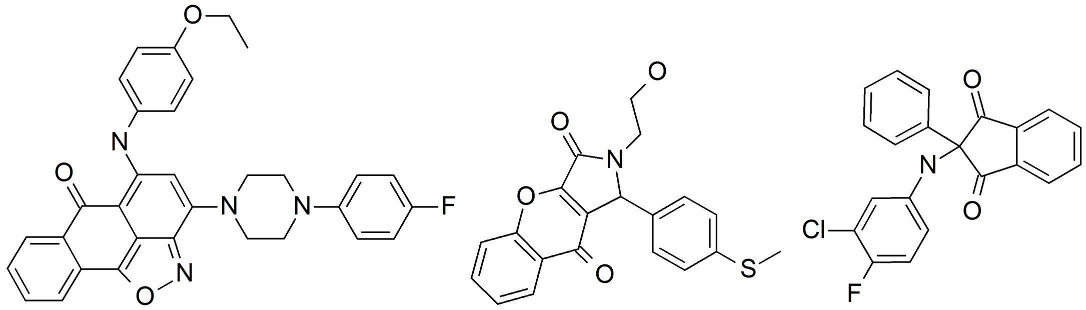
Figure 3.28: Final selection of the scaffolds. The scaffold ID corresponds to the numbers given in table A.2 on pages 130 ff.



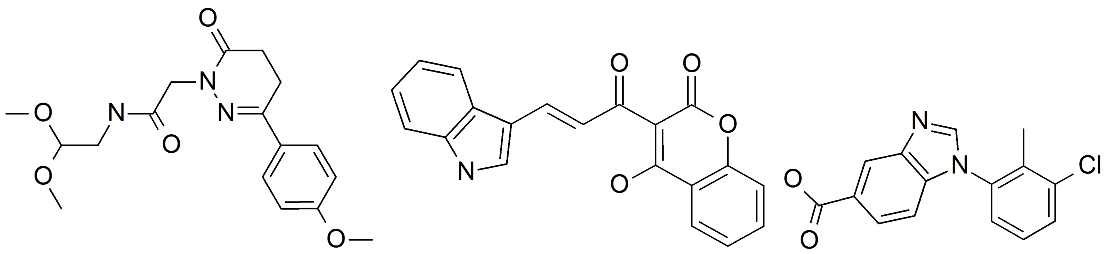
(a) 4204-0085 (**103**) (b) 4451-0051 (**104**) (c) 5982-0100 (**105**) (d) 6463-4542 (**106**)



(e) 7244-0063 (**107**) (f) 8014-1054 (**108**) (g) C073-3327 (**109**)



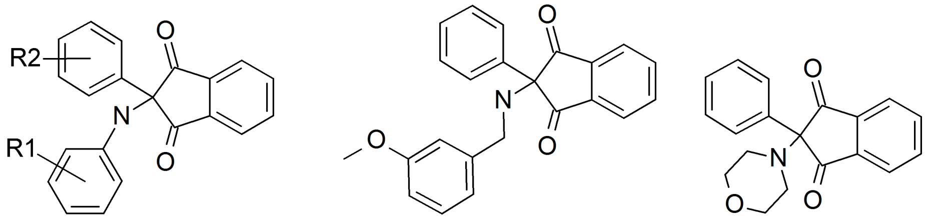
(h) C090-0245 (**110**) (i) D159-0883 (**111**) (j) E938-0156 (**112**)



(k) K788-0448 (**113**) (l) K815-0023 (**114**) (m) R153-0192 (**115**)

Figure 3.29: Molecules selected for biological investigation.

SOM (2D):

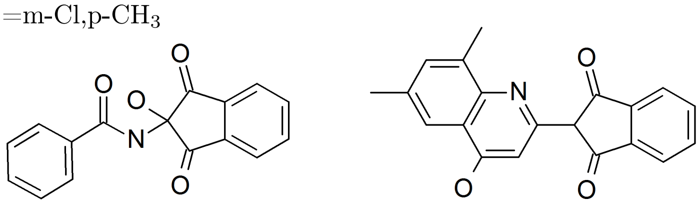


(a) 8008-8508 (**119**): R1=p-phenyl
C301-5215 (**123**): R1=m-Cl,p-OCH_3
C301-5408 (**124**): R1=m-F, R2=p-OCH_3
E938-0003 (**126**): R1=2-OCH_3,5-CH_3
E938-0021 (**127**): R1=m-Cl,p-CH_3

(b) E938-0051 (**129**)

(c) E938-0077 (**130**)

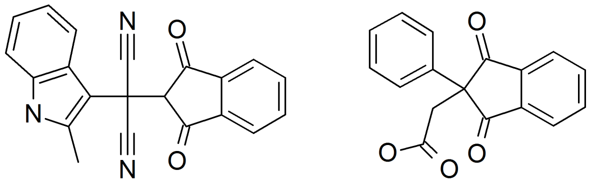
SOM (VSA):



(d) 6623-0410 (**117**)

(e) 7165-0402 (**118**)

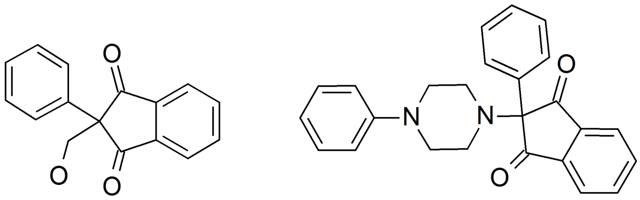
FP:



(f) 8009-7265 (**120**)

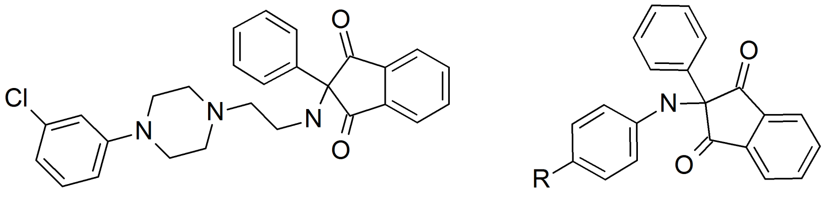
(g) C270-0349 (**121**)

hand-picked:



(h) 0095-0198 (**116**)

(i) C301-4948 (**122**)



(j) C301-5428 (**125**)

(k) E938-0036 (**128**): R=methyl ester
E938-0078 (**131**): R=carboxylic acid

Figure 3.30: Selected derivatives of compound **112**.

3.1.5 Docking

Beside the ligand based screening described in the previous sections, we have additionally applied docking studies to the activated kinase domain of the insulin receptor (PDB-ID: 1IR3). Aim of this docking study was to investigate the potential of **1** and its derivatives to interact with previously proposed binding sites on the kinase and to judge if this method would be applicable for screening.

First, docking studies of the active compounds to the proposed binding site at the N-terminal lobe of the kinase domain[19] were performed. A surface representation of this binding site is shown in figure 3.31. Potential H-bonding partners are depicted in purple, mild polar regions are shown in blue and hydrophobic areas are coloured green. Amino acids of the binding sites which could be important for binding are shown in figure 3.32 using the same perspective. Arg1041 as well as Lys1052 can form cation/π interactions to the indole rings of the molecules. Using a different rotamer of Arg1041 could even result in poses where both interactions can be seen simultaneously in one molecule, when the molecule is placed in parallel to the αC helix. An-

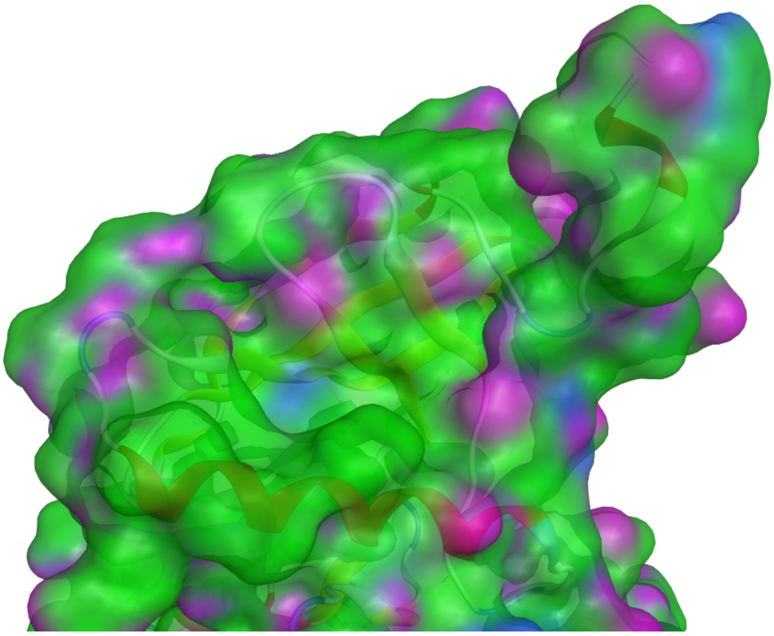


Figure 3.31: Surface representation of a proposed binding site.

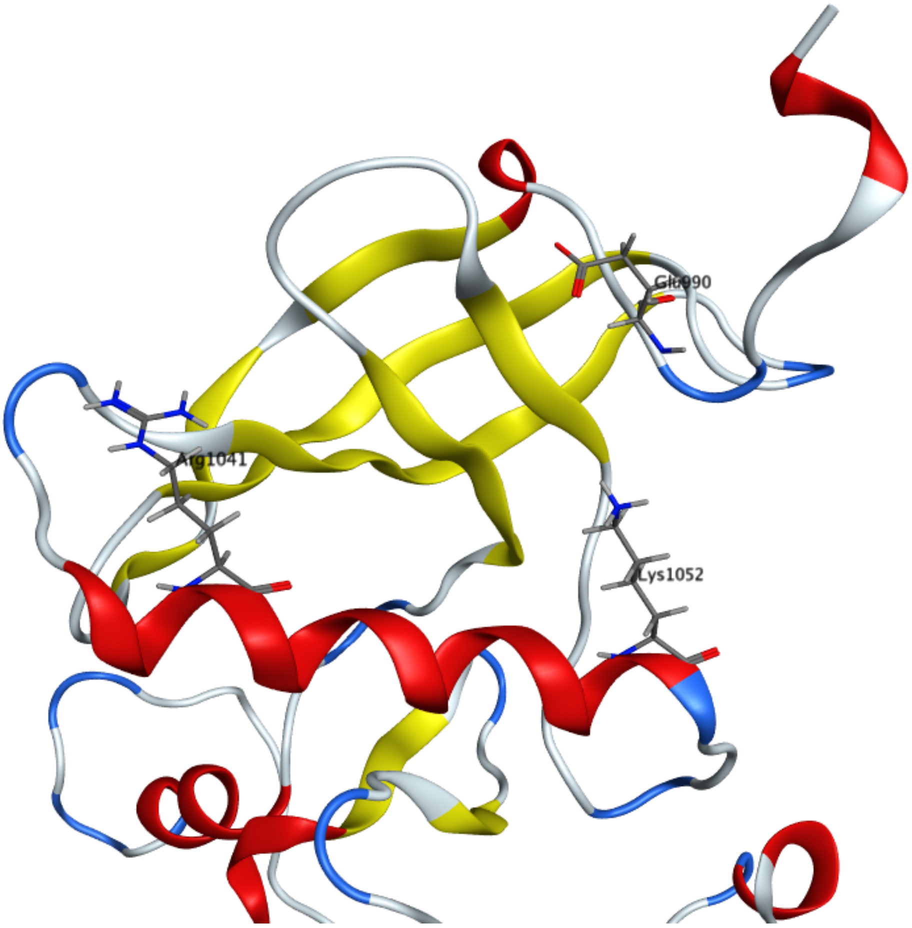


Figure 3.32: Selected amino acids of the proposed binding site.

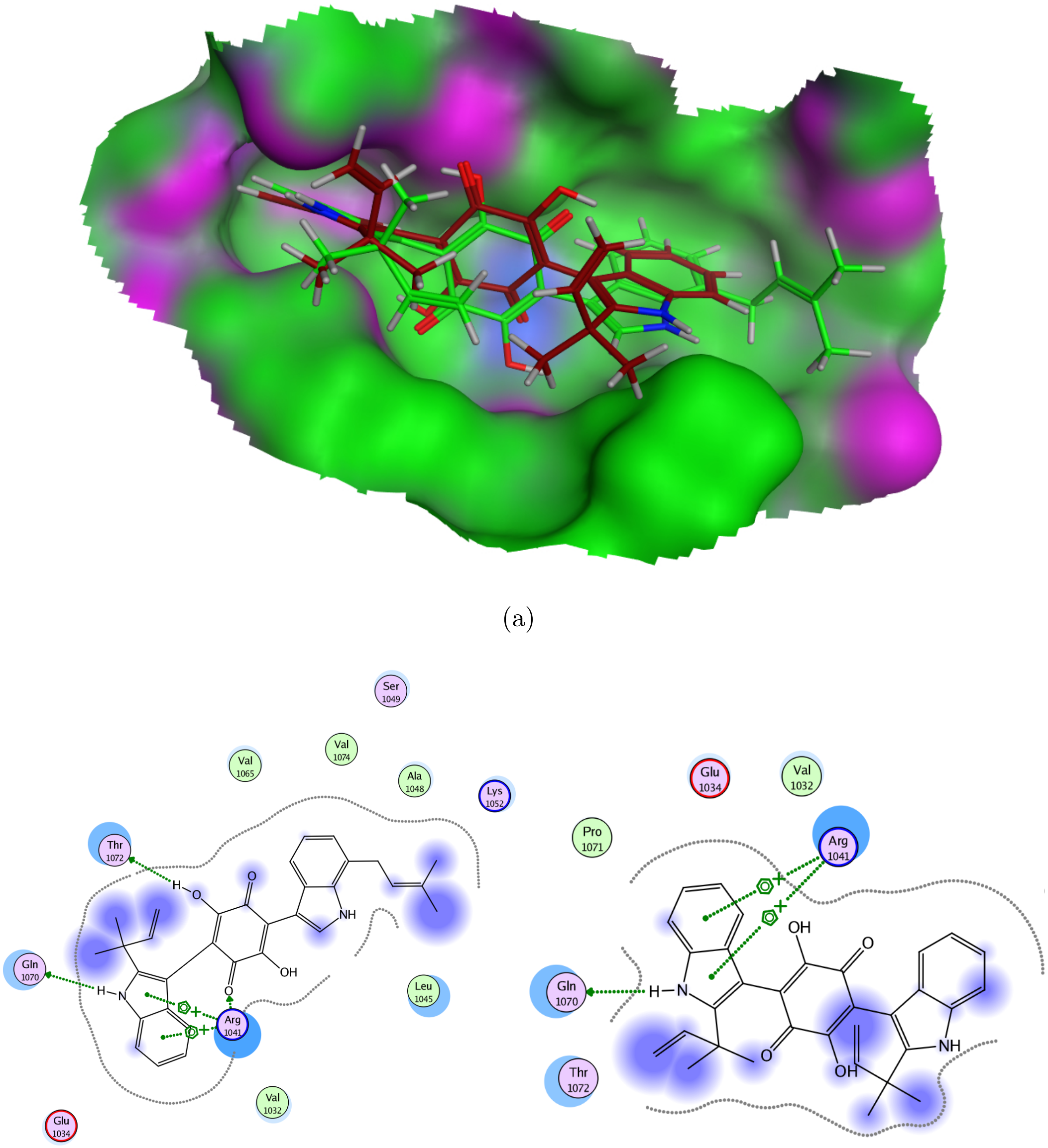
other possibility is to place the molecules in an upright position with the quinone ring forming a hydrogen bond to Glu990. At the same time, one of the indole rings could be stabilized by a cation/π interaction to Lys1052.

To our knowledge, no information is available in the literature on the importance of any amino acids for the interaction of compound **1** or its derivatives with the receptor kinase. Therefore, only the information derived from the molecules can be used to prioritize one of the docking poses. For this, only compounds having both indole rings were used to identify binding poses with the scaffold in a similar orientation. This led to the identification of a cluster showing poses of all investigated molecules. These poses show cation/π interactions of Arg1041 to one of the indole rings and a possible hydrogen bond of the indole-NH to Gln1070. The latter interaction can only be seen after energy refinement of the binding pocket. The quinone ring can form hydrogen bonds to Arg1041 as well as to Thr1072. A comparison of the active compound **1** and the inactive compound **2** is shown in figure 3.33. While a large residue on position 7 of the right indole ring fits well into the pocket, this is not the case for substituents on position 2. This could lead to a displacement of the indole ring, as well as the adjacent quinone, moving it outwards of the binding pocket. Methyl substitutions on other positions at the indole ring as seen in several inactive derivatives of **1** would point to the inside of the binding pocket.

Compound **3** has a methyl group on the indole nitrogen. The decreasing activity could be explained by the lost possibility of forming a hydrogen bond to Gln1070 (figure 3.34a). Still, not all of the activity changes can be easily explained by this proposed binding mechanism. For example, compounds **23** and **27** differ only by the presence or absence of the methyl group on the indole nitrogen. But in this case, the molecule with the methylation is the more active one.[40]

We also investigated a binding site which was proposed for α-lipoic acid by Diesel et. al (see figure 3.35).[53] The Site-finder tool of MOE was not able to detect this binding pocket. As compound **1** is larger than α-lipoic acid (compare figure 1.9d on page 15) we were not able to place the compound into this binding site. Docking of the compounds might only be possible after structural rearrangement of the binding site.

While docking to the binding site on the N-terminal lobe could propose many possible ways in which the molecules could interact with the binding site residues, the exact binding mode could not be determined with certainty. A complicating factor is that this site is exposed to solvent, thus increasing the possible modes in which a molecule could fit into the pocket. Additionally, the crystal structure only includes the intracellular domain of the insulin receptor. In the whole receptor, the N-terminal part would be connected to the transmembrane domain. This, as well as the dimerisation of the receptor, could alter the properties of the binding site, leading to different interaction possibilities. We therefore decided not to include docking for screening.

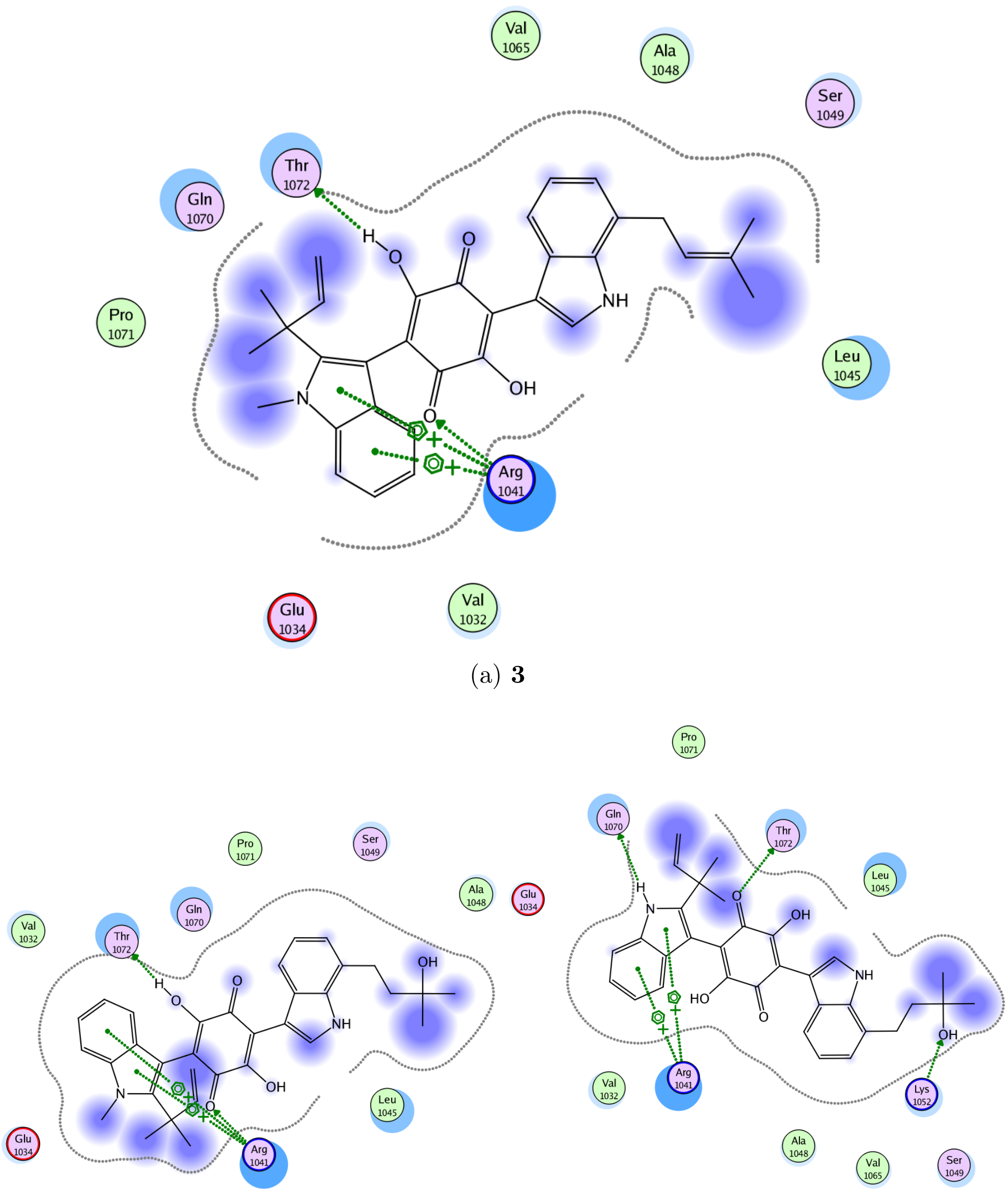


(a)

(b) **1**

(c) **2**

Figure 3.33: Comparison of the docking pose of compound **1** (active, in green) with compound **2** (inactive, in red).



(a) **3**

(b) **23**

(c) **27**

Figure 3.34: Docking poses of compounds **3**, **23** and **27**.

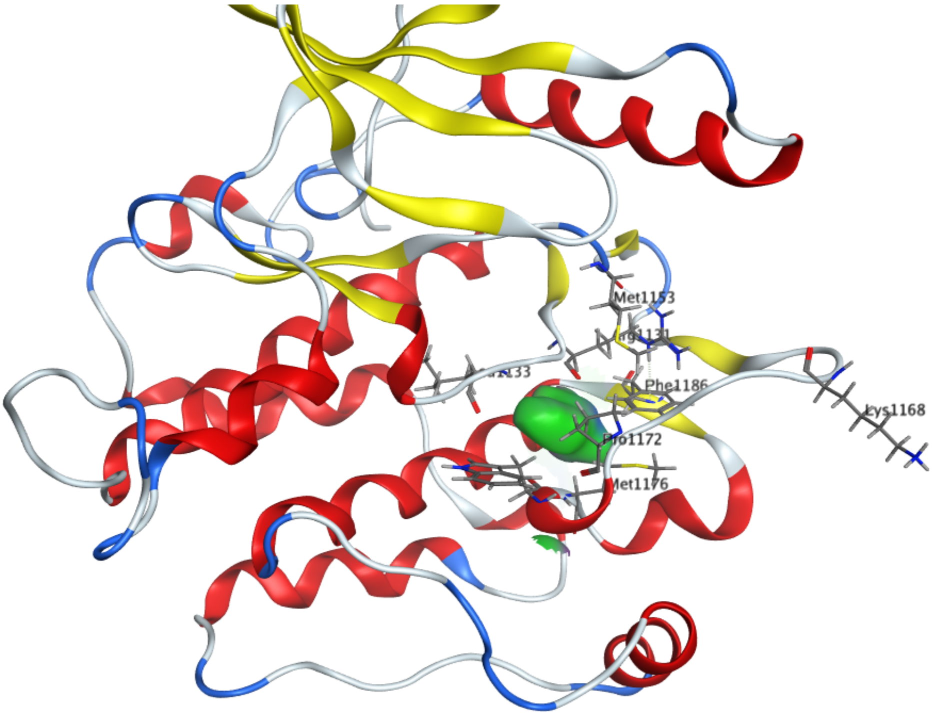


Figure 3.35: Binding site proposed in reference 53. Amino acids mentioned in the publication are depicted in stick representation, the binding site is shown as a surface.

3.2 Experimental part

3.2.1 Preliminary studies with DMAQ-B1

To find the best conditions for the screening, preliminary studies were conducted with demethylasterriquinone B1 (compound **1**). First, its ability to trigger insulin receptor phosphorylation directly was assessed in hepatocytes (HCC-1.2). Although different concentrations (figure 3.36) and incubation times (figure 3.37) were tried, compound **1** only elicited a weak increase in IR phosphorylation. Interestingly, higher concentrations of compound **1** decreased the total number of the insulin receptor β-subunit. The strongest signal was achieved after 5 minutes. Using a concentration of 30 µM, compound **1** led to an approximately five-fold increase of insulin receptor phosphorylation as compared to solvent control, while insulin (10 nM) increased the signal 15-fold under the same conditions (figure 3.37).

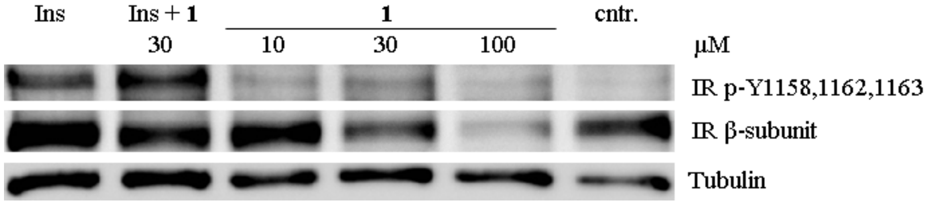
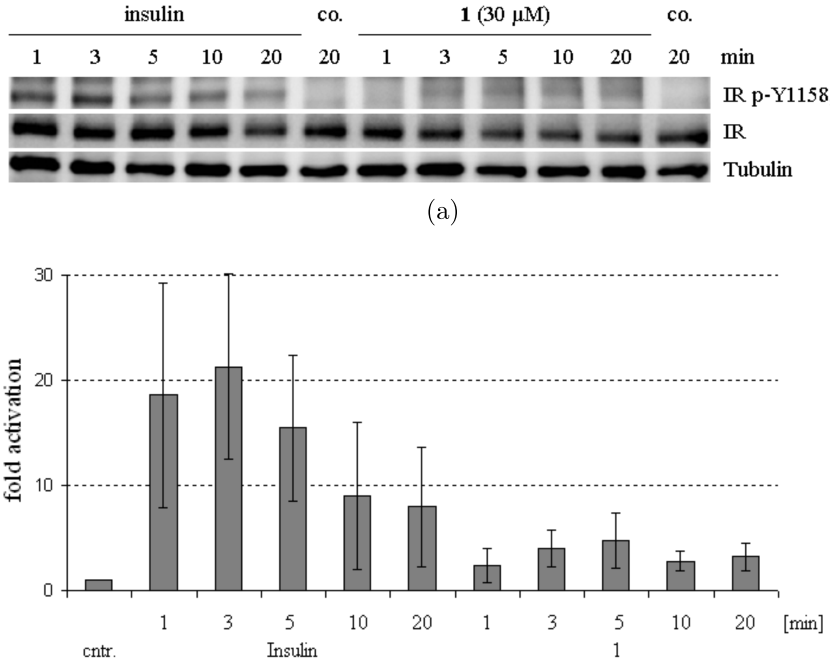


Figure 3.36: Phosphorylation of the insulin receptor by stimulation with insulin and compound **1** at different concentrations for 10 minutes in HCC-1.2 cells (single experiment).

Since this would not leave much room for the detection of new molecules which are less active than compound **1**, a read-out for screening was selected. Phosphorylation of Akt on Ser473, which is a downstream target of insulin signaling, turned out to give a stronger signal, especially when tested in mouse embryonic fibroblasts (MEF). Detection of Akt Ser473 was therefore subsequently used to identify active hits.



(a)

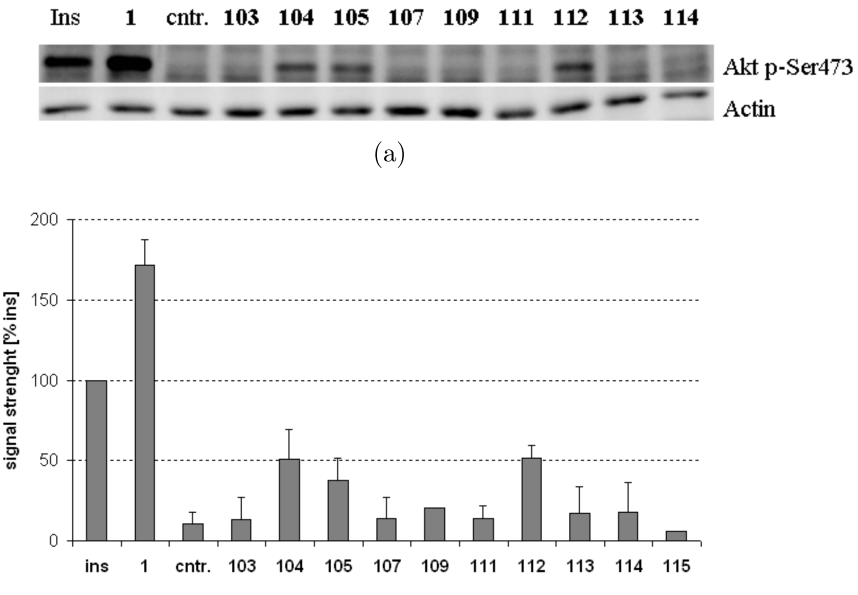
(b)

Figure 3.37: Time course of insulin receptor phosphorylation upon exposure to insulin and compound **1** (30 µM, 0.5% DMSO) in HCC-1.2 cells. (a) Representative Western blot; (b) mean values (± SD) of four independent experiments.

3.2.2 Screening for active compounds

The purchased compounds were screened for inducing Akt Ser473 phosphorylation in MEF at 30 µM (figure 3.38). Of the 13 purchased compounds, ten were soluble in DMSO at a concentration of 100 or 50 mM. One of the molecules, compound **109** precipitated when diluted in cell culture medium. Of the tested compounds, three molecules were found to be active by leading to markedly increased Akt phosphorylation in MEF (compounds **104**, **105** and **112**).

Derivatives of compound **112** were subsequently tested to allow the investigation of basic structure activity relationships. The outcome of this experiment is shown in figure 3.39 and the corresponding structures are depicted in figure 3.40. Different substitutions of the phenyl rings (compounds **123**,



(a)

(b)

Figure 3.38: Akt activation by selected hit compounds (30 µM, 0.5% DMSO).(a) Representative Western blot; (b) mean values ± standard deviation relative to the signal strenght of insulin (n=1 for compounds **109** and **115**, n=3 for remaining compounds).

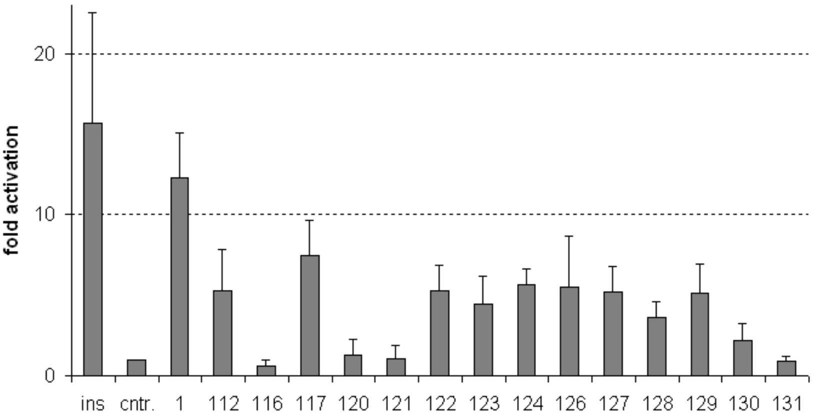
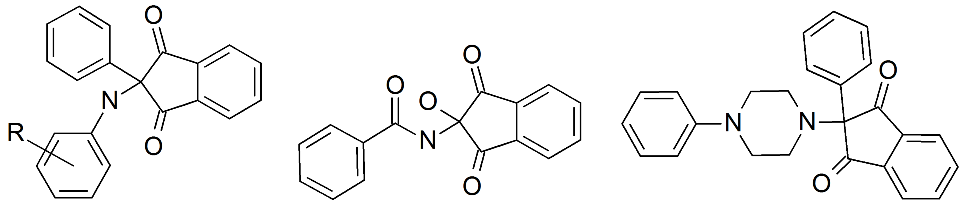


Figure 3.39: Akt activation by derivatives of compound **112** (30 µM). Shown are the mean values ± standard deviation of the intensities relative to solvent control (0.5% DMSO) of three independent experiments.

124, **126** and **127**) were generally well tolerated. The only exception found, was a carboxylic acid group at the para position of the aniline ring (compound **131**). This compound was inactive, while the corresponding methyl ester (**128**) was medium active. Replacement of the whole aniline ring by a morpholine group (compound **130**) decreases the activity, but replacement of the nitrogen by a piperazine group (**122**) does not affect the activity. Not fully explored is the influence of omitting one of the rings. Compounds **116** and **121** have a carboxy and a carboxylic acid moiety instead of the aniline ring, respectively, which leads to a complete loss of activity. On the other hand, compound **117** has an carboxy group instead of the second phenyl ring and is one of the most active compounds identified in this study. But this compound also has an amide instead of the nitrogen, thus it is not clear which feature is finally responsible for the activity.

Dose response studies with compound **117**, one of the most active new molecules, were performed to show the concentration dependency of the activity and to compare it to that of compound **1** (figure 3.41). Although **1** was more active at 30 µM, the activity was comparable to that of the original compound at 10 µM.

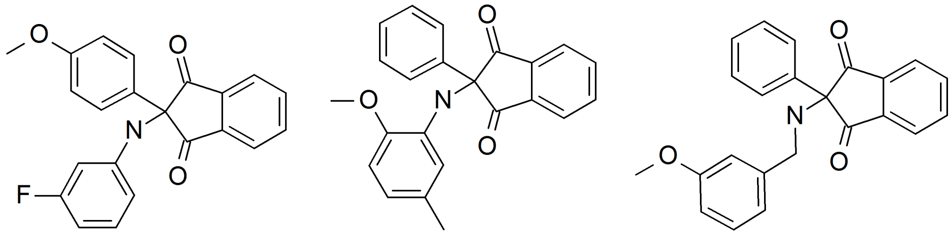
Active compounds:



(a) **112**: R=m-Cl,p-F
123: R=m-Cl,p-OCH$_3$
127: R=m-Cl,p-CH$_3$

(b) **117**

(c) **122**

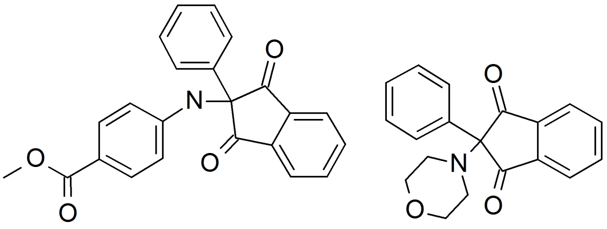


(d) **124**

(e) **126**

(f) **129**

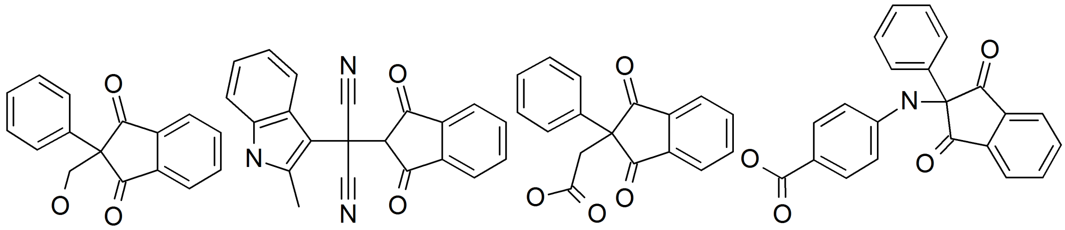
Medium active compounds:



(g) **128**

(h) **130**

Inactive compounds:



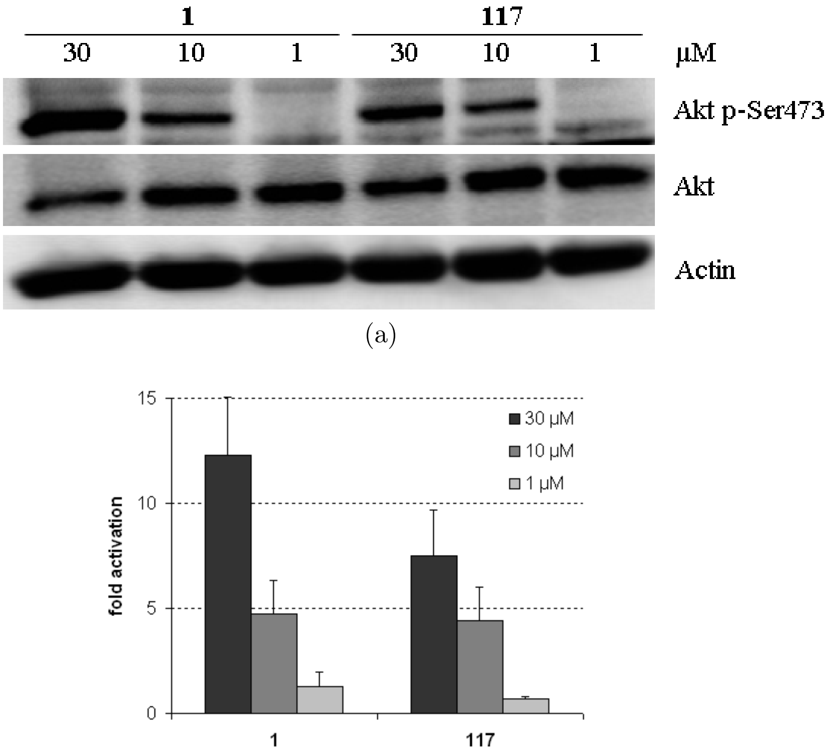
(i) **116**

(j) **120**

(k) **121**

(l) **131**

Figure 3.40: Structure activity relationships of derivatives of compound **112**.



(a)

(b)

Figure 3.41: Comparison of Akt Ser473 phosphorylation elicited by different concentrations of compounds **1** and **117** in MEFs. (a) Representative Western blot; (b) Mean values (± SD) of at least two independent experiments.

3.2.3 Further characterization of active compounds

Cytotoxicity: As one of the goals was to identify substances less cytotoxic than the original compound **1**, the influence of the compounds on cell mass was investigated using the crystal violet assay. While compounds **104** and **112** show no cytotoxic effects at 30 µM, nearly all cells detach from the well when treated with compound **105**. Of the derivatives of compound **112**, only **117** shows cytotoxic effects comparable to those of compound **1**. The results of the crystal violet assay are summarized in figure 3.42.

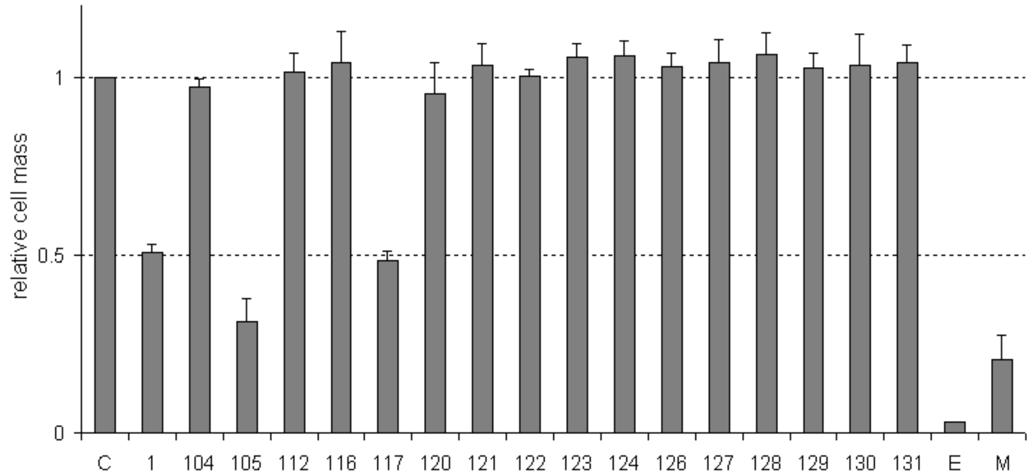


Figure 3.42: Cell mass after treatment with compounds at 30 µM for 24 h relative to solvent control values, ± standard deviation (n≥3). C: solvent control (DMSO 0.5%); E: empty wells; M: wells filled with media only.

PTP1B inhibition: The signals of compound **1** observed with the antibody against the phosphorylated insulin receptor resembled those seen with sodium orthovanadate, a known inhibitor of PTP1B (data not shown). We therefore decided to test the action of **1** against PTP1B. Indeed, compound **1** showed an IC_{50} of 7.3 µM *in vitro*, as measured by Renate Baumgartner.[117] Subsequently, all compounds from the first screening round were tested at 100 µM to investigate their activity on PTP1B inhibition. The results are shown in figure 3.43, with especially compound **105** showing a high inhibitory activity. Compounds **103** and **114** were active in the PTP1B assay, but in-

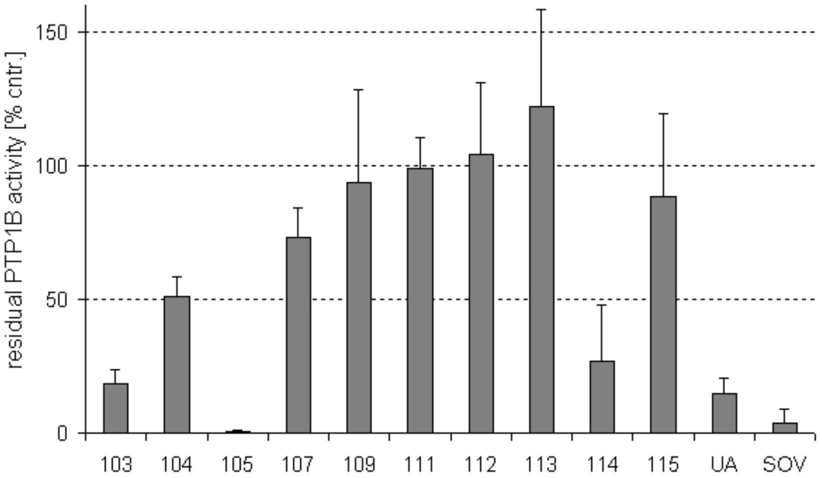


Figure 3.43: PTP1B inhibition by the indicated compounds (100 µM). Values are given as residual PTP1B activity as compared to solvent control (1% DMSO), ± standard deviation (n≥2). UA: ursolic acid (30 µM), SOV: sodium orthovanadate (10 µM)

active in phosphorylating Akt in the cellular assay. Both however showed increased background values in the PTP1B assay and might be false positives. Especially compound **114** was showing a yellow/orange colour itself, and thus might interfere with the readout of the assay, which measures the presence of a yellow coloured product of PTP1B cleavage (see page 47).

Selected derivatives of **112** were tested for PTP1B inhibition by Sophie Bartenstein (figure 3.44). Although all three compounds are activating Akt, only compound **117** shows high inhibitory activity against PTP1B.

Glucose uptake: The three active compounds (**104**, **105** and **112**) were tested in a single glucose uptake experiment in myocytes to get a first overview on their activity. Figure 3.45 shows the result of this experiment, where only compound **112** increased the glucose uptake at a concentration of 100 µM. Compounds **1**, **104** and **105** decreased the uptake of glucose, but led also to a markedly lower protein concentration compared to DMSO-treated control cells, suggesting a cytotoxic effect.

Dose response studies (30 and 100 µM) of selected compounds are shown in figure 3.46. Compound **1** led to a dose-dependent decrease in cellular

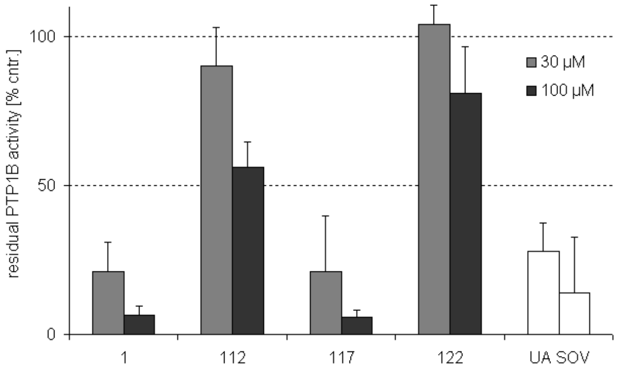


Figure 3.44: PTP1B inhibition by the indicated compounds (30 and 100 µM). Values are given as residual PTP1B activity as compared to solvent control (1% DMSO), ± standard deviation (n≥2). UA: ursolic acid (30 µM), SOV: sodium orthovanadate (10 µM)

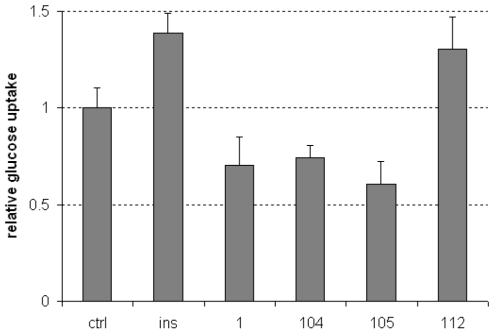


Figure 3.45: Glucose uptake in myocytes (triplicate from single experiment). Cells were treated with insulin (100 nM) and compounds (100 µM) for 60 min before incubation with ^{3}H-DOG (10 min).

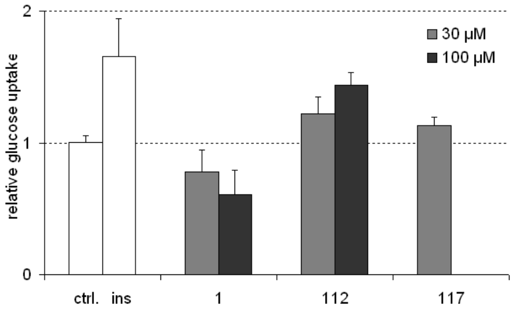


Figure 3.46: Glucose uptake in myocytes relative to solvent control (mean values ± standard deviation of at least three independent experiments).

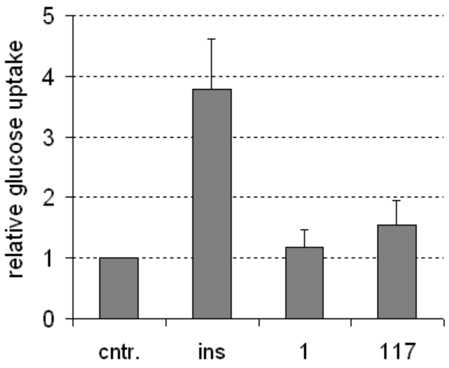


Figure 3.47: Glucose uptake in adipocytes relative to solvent control (0.1% DMSO). Cells were treated with insulin (100 nM, 30 min) or the indicated compounds (30 µM, 60 min) before incubation with ^{3}H-DOG (5 min). Glucose uptake values are mean ± standard deviation, n=3.

glucose uptake, possibly due to cytotoxic effects. Of the tested compounds, only compound **112** led to a dose-dependent increase of glucose uptake. Compound **117** showed a small increase of glucose uptake as well, but was too cytotoxic to be tested at 100 µM.

Contrary to previous reports in the literature, we could not see a relevant increase of the glucose uptake rate in adipocytes by compound **1** in the tested concentrations and incubation times (30 and 100 µM, 30 and 60 min, data not shown), which is consistent with our data obtained in myocytes. Compound **117** was able to increase the glucose uptake (see figure 3.47), but not in a statistically significant way as assessed by Student's t-test (p=0.148).

4

Discussion

4.1 Computational methods

According to Seifert and Lang, the aim of virtual screening is to “enable the initiation of a medicinal chemistry program with a reasonable probability for identifying a lead compound”.[119] A virtual screening approach should therefore find at least one active molecule which can then be further used in medicinal chemistry. Following this definition, the presented screening was successful, leading to the identification of an active molecule (**112**), of which the activity and cytotoxicity can be modified by slight structural changes. Another prerequisite for a successful virtual screening method is that the same result could not have been acquired with much simpler techniques.[70] As the training of large self-organizing maps with the screening database is a time consuming process, this is a valid objection. The performance of the different methods which were used in this project is therefore discussed in more detail in the following. The performance of the different methods in retrieving active hits is shown in table 4.1 for the initial screen, as well as for compound **112** and its derivatives. The number of compounds soluble in the stock solutions are indicated in brackets in case they are dissimilar to the total number.

All initially identified active compounds were found using self-organizing maps. Additionally, self-organizing maps were the only investigated method

Method	active/tested	
	screening	derivatives
SOM (2D autocorr.)	2/5(4)	5/7(6)
SOM (VSA descr.)	1/4(3)	1/2(1)
Fingerprints	0/3(2)	0/2
Shape	0/1	0/0
Picked by hand	0/0	1/5(4)
Total	3/13(10)	7/16(13)

Table 4.1: Number of active vs. amount of purchased (tested) compounds for the initial screening and for the tested derivatives of compound **112**.

able to identify compound **102**,[46] an active molecule published after the building of the models. The compound was identified on several maps trained with 2D autocorrelation descriptors using the subset of the screening library.

Self-organizing maps in general performed well in separating the active and inactive compounds of the training set. Still, they were not able to distinguish between molecules of different classes in all cases. Especially when slight structural changes like the position of a methyl group on an indole ring were causing a change in activity. Although it might have been possible to increase the performance of the self-organizing maps to separate active from inactive derivatives by using different descriptors and feature selection methods we decided to use more general sets of descriptors. With this, we intended to allow a separation between molecules similar and dissimilar to the training set, rather than a separation between slight structural changes seen to be responsible for activity. Additionally, one has to take into account that self-organizing maps are unsupervised machine learning methods. They do not use the class information for the training process, and are thus not primarily classification tools. Instead, our aim was to identify molecules which are near to the training compounds in chemical space.

For this, we also investigated whether first selecting a subset of molecules according to their descriptor similarity would influence the resulting maps. Of the 145 compounds which were in total selected with self-organizing maps using 2D autocorrelation descriptors, 47 were already in the subset which was selected according to the Euclidean distance. The self-organizing maps with the VSA descriptors find in total 102 compounds, with 40 of them being already in the distance based subset. From the nine purchased molecules selected by self-organizing maps in the first round, five were already in the subsets, among them all three active compounds. Setting a similarity threshold based on the Euclidean distance to the query compounds might thus be an additional filter to refine the hits found by a self-organizing map. But training the compounds with the subset only, would not have led to the identification of the active hits, as they were localized in active neurons on the large maps only. Instead of using the Euclidean distance to the training compounds as a threshold, novelty detection with self-organizing maps[99] could be performed to use the distance to the winning neuron as assessment for the applicability domain. Additionally, increasing the size of the network could decrease the number of erroneous co-localizations, but this leads to a significant prolongation of the calculation time, an effect not wanted in virtual screening approaches.

Fingerprint similarity, which is a very fast method, identified several molecules with interesting structures. However, all tested molecules identified with this method were found to be inactive. For example, compound **107** has an indole and a phenyl ring connected by a quinone-like structure (see figure 3.29e). It therefore resembles the query compounds more than the identified actives, but was nevertheless inactive in our test system. Still, as only a small number of the identified compounds has been tested, we might have missed actives identified with this method.

After knowing the activities of the purchased molecules, we investigated if changes in our selection criteria for the fingerprints would have led to the identification of the active compounds. An alternative to setting an overall threshold for each fingerprint would be to select the most similar molecule to each of the queries. Indeed, compound **105** was the most similar molecule of the screening library to compound **32**, showing a similarity value of 0.77. The chosen threshold for the MACCS fingerprint was 0.80, thus not allowing the identification of the molecule with this method.

The used criterion for selecting molecules with fingerprints led to a different amount of identified compounds for the different types of fingerprints. Especially molecules identified with MACCS and TGT fingerprints are therefore overrepresented in the dataset, which is, especially for the MACCS fingerprint, in part compensated by a higher similarity of the hits to each other. When considering the enrichment rates of the different fingerprints for the first 100 compounds (table 3.8 on page 66), TGT performs worse than TAT, for which no hits have been selected at all using our thresholds.

Shape similarity: The final selection of compounds for testing included only one molecule identified by shape similarity. Although this compound was then inactive in our test system, the usability of this method can not be judged by one instance only. As the identified hits of this method heavily depend on the 3D conformation of the query structures it might not be appropriate to use one query conformation only. Instead, one could use a different software which allows conformational flexibility of the query as well as the screening database. A different possibility would be to use the ligand conformation derived from docking studies, which could be the topic of future work.

Docking was used to investigate whether compound **1** and its derivatives could bind to one of the binding sites on the kinase domain of the insulin receptor proposed in references 19 and 53. While the latter binding site seems to be too small for the compounds, they should be able to bind to the pocket between the αC helix and the β-sheets of the N-terminal lobe of the kinase domain. This binding pocket is large enough in size and contains several amino acids which could contribute to the binding of the compounds. It is also equivalent to the binding pocket of activators of PDK1, which is to our knowledge the only published kinase co-crystallized with an activator.[68,69]

4.2 Biological investigations

In our cell systems we could observe that while compound **1** was able to activate Akt to similar levels as insulin, it was much less able to activate the insulin receptor under the same conditions. This is comparable to observations made previously in reference 35. We therefore chose Akt phosphorylation as our screening readout, identifying three active hits. Since, however, Akt is activated by many different signalling pathways,[120] we aimed at showing and confirming the activity of the identified compounds in an assay more relevant to diabetes. As one of the main goals of anti-diabetic drugs is lowering of blood glucose levels, the ability of selected molecules to stimulate glucose uptake in fat- and muscle cells was investigated. Indeed, compound **112** stimulated glucose uptake in myocytes, while the others (including compound **1**) decreased the uptake. Additionally, glucose uptake induced by compound **1** in 3T3L1 adipocytes as reported in the literature[35] could not be reproduced in our system. This could maybe be explained by different handling of the cells, or that our cells were more susceptible to the cytotoxic effects of the compound.

Different to previous reports in the literature,[32, 33] we measured inhibition of PTP1B by compound **1**. None of the cited references reported the assays used to assess possible inhibitory effects of compound **1**. It is therefore difficult to judge what led to these different results. Also some of our new compounds (**105** and **117**) showed this inhibition of PTP1B. Interestingly, molecules active in the PTP1B assay were also those showing cytotoxic effects and compounds activating Akt were not necessarily inhibiting PTP1B.

Pan assay interference compounds (PAINS): Baell and Holloway[121] published a list of frequent hitters in high throughput screening. They identified 2 062 compounds which were found in at least four of their assays and 362 which were found with all six assays. From these structures they defined substructures which are potentially problematic. The filters for these pan assay interference compounds (PAINS) were implemented in a KNIME workflow recently.[122]

Using these filters, already most of the compounds of the training database would have been identified due to their p-quinone structure. From the active molecules, only compounds **100**, **101** and **102** were not recognized by the filters. Quinones were already reported to be protein-reactive and compound **1** was found to be active against additional targets than the insulin receptor.

Many of the structure types reported to have the possibility to be PAINS have been identified with our computational methods. Of the final 367 identified compounds, 133 failed the check using the KNIME workflow with the Indigo nodes, 91 using the RDKit nodes. One example are rhodanine-like structures. Using piDAPH3 and piDAPH4 fingerprints, several molecules containing a thiazolidinedione or a rhodanine substructure have been identified. Some of the currently known anti-diabetic compounds belong to the class of the thiazolidinediones. One of our rhodanine hits, compound 3057-0993 (figure 3.19b on page 75), was identified by Choi et al. as having a similar scaffold as active hits in a virtual screening for PPAR-γ agonists (compound SP1802 in reference 123). This compound however only showed little activity, having only 11.71% PPARγ binding activity and 1.21 fold transcriptional activation of PPAR-γ in cells.[123] Similar structures were also identified by virtual screening using docking for PTP1B inhibition.[73,74] Other thiazolidinedione derivatives were shown to act both as inhibitor of PTP1B as well as activator of PPAR-γ .[124]

Even some of our purchased molecules were identified as possible PAINS with these filters. Compound **103** was identified due to its p-quinone moiety. This compound showed PTP1B inhibition, but was inactive in the cellular system. Compounds **106**, **108** and **110** were identified by the filtering, but could not be tested due to poor solubility in the stock solutions. But also the active molecule **112** and its derivatives were identified because of their 1,3-indandione substructures (keto_keto_beta_A group in reference 121). Still, **112** seemed to be less cytotoxic than the other initial hits, which was one of the reasons to choose it for further investigation of derivatives. Indeed, the cytotoxicity of the compounds did not seem to correlate with the activity of the molecules. But for one of the molecules it could be true that it is a pan assay interference compound. Compound **117** is one of the most active molecules identified in this study. But it additionally inhibits PTP1B, and also shows cytotoxic effects. This molecule (PubChem SID: 4242461) was reported before to be an inhibitor of other phosphatases. It has an IC_{50} of 14.2 µM against the mitogen-activated protein kinase (MAPK) phosphatase-1 (MKP-1).[125] Furthermore it was reported to have an IC_{50} of 4.3 µM against human cell division cycle 25 protein B (Cdc25B), but an IC_{50} larger than 50 µM for MKP-1 and an IC_{50} of 48.7 µM against MKP-3.[126] To determine if the mode of action of the compound against Cdc25B involves oxidation, the strong reducing agent dithiothreitol (DTT, 1 mM) in the assay was replaced by β-mercaptoethanol (1 mM), reduced glutathione (1 mM) or DTT (25 mM). Additionally, catalase (100 U) was added to degrade the produced H_2O_2 with 1 and 25 mM DTT. Compound **117** had an IC_{50} larger than 50 µM in all these conditions. In a redox cycling H_2O_2 generation assay in the presence of 0.5 mM DTT the compound showed a 50% activation concentration (AC_{50}) value of 28.6 µM. These results indicate that the inhibition of Cdc25B by compound **117** is due to the generation of reactive oxygen species (ROS).[126] As PTP1B is susceptible to oxidation as well, this might be the cause of the compound's activity against this phosphatase.

5

Conclusions

Using our computational screening methods and testing only a few selected molecules in biological assays, we were successful in identifying possible insulin mimetic compounds. These molecules were shown to activate Akt, a downstream target of the insulin receptor. One of the compounds was shown to increase the glucose uptake in myocytes.

Still, only a small proportion of the identified molecules has been tested. To conclude which of the used methods is superior to the others, a higher number of tested molecules would be necessary, which was not feasible in the course of this project. In addition, the other scaffold clusters possibly also contained some active molecules. It has been shown that there is a high chance of missing the activity in a set of similars, if only one molecule out of this set is tested.[78] Further research will also be necessary to confirm whether the mode of action of the new molecules is direct activation of the insulin receptor, as was done for the original compounds.

Given that the identification of compound **1** needed the cell-based screening of over 50 000 samples,[32] the recent work was a successful start to the investigation of new insulin receptor activators, showing many possible topics for future research.

Bibliography

[1] Statistik Austria, Gesundheitsbefragung 2006/07. Available from `http://www.statistik.at/web_de/statistiken/gesundheit/index.html`.

[2] S. Wild, G. Roglic, A. Green, R. Sicree and H. King, Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030. *Diabetes Care* **2004**; 27(5): 1047–1053, doi:doi:10.2337/diacare.27.5.1047.

[3] I. Barroso, Genetics of Type 2 diabetes. *Diabet Med* **2005**; 22(5): 517–535, doi:10.1111/j.1464-5491.2005.01550.x.

[4] S.-F. Ng, R. C. Y. Lin, D. R. Laybutt, R. Barres, J. A. Owens and M. J. Morris, Chronic high-fat diet in fathers programs β-cell dysfunction in female rat offspring. *Nature* **2010**; 467(7318): 963–966, doi:10.1038/nature09491.

[5] M. K. Skinner, Metabolic disorders: Fathers' nutritional legacy. *Nature* **2010**; 467(7318): 922–923, doi:10.1038/467922a.

[6] G. Taubes, Insulin resistance. Prosperity's plague. *Science* **2009**; 325(5938): 256–260, doi:10.1126/science.325_256.

[7] European Medicines Agency recommends suspension of Avandia, Avandamet and Avaglim. Press release, 23 September 2010. Available from `www.ema.europa.eu/docs/en_GB/document_library/Press_release/2010/09/WC500096996.pdf`.

[8] E. Blind, K. Dunder, P. A. de Graeff and E. Abadie, Rosiglitazone: a European regulatory perspective. *Diabetologia* **2011**; 54(2): 213–218, doi:10.1007/s00125-010-1992-5.

[9] S. Akkati, K. G. Sam and G. Tungha, Emergence of promising therapies in diabetes mellitus. *J Clin Pharmacol* **2011**; 51(6): 796–804, doi:10.1177/0091270010376972.

[10] G. Nicholson and G. M. Hall, Diabetes mellitus: new drugs for a new epidemic. *Br J Anaesth* **2011**; 107(1): 65–73, doi:10.1093/bja/aer120.

[11] D. R. Robinson, Y.-M. Wu and S.-F. Lin, The protein tyrosine kinase family of the human genome. *Oncogene* **2000**; 19(49): 5548–5557, doi:10.1038/sj.onc.1203957.

[12] N. M. McKern, M. C. Lawrence, V. A. Streltsov, M.-Z. Lou, T. E. Adams, G. O. Lovrecz, T. C. Elleman, K. M. Richards, J. D. Bentley, P. A. Pilling, P. A. Hoyne, K. A. Cartledge, T. M. Pham, J. L. Lewis, S. E. Sankovich, V. Stoichevska, E. D. Silva, C. P. Robinson, M. J. Frenkel, L. G. Sparrow, R. T. Fernley, V. C. Epa and C. W. Ward, Structure of the insulin receptor ectodomain reveals a folded-over conformation. *Nature* **2006**; 443(7108): 218–221, doi:10.1038/nature05106.

[13] S. R. Hubbard, L. Wei, L. Ellis and W. A. Hendrickson, Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* **1994**; 372(6508): 746–754, doi:10.1038/372746a0.

[14] S. R. Hubbard, Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J* **1997**; 16(18): 5572–5581, doi:10.1093/emboj/16.18.5572.

[15] J. H. Till, A. J. Ablooglu, M. Frankel, S. M. Bishop, R. A. Kohanski and S. R. Hubbard, Crystallographic and solution studies of an activation loop mutant of the insulin receptor tyrosine kinase: Insights into kinase mechanism. *J Biol Chem* **2001**; 276(13): 10049–10055, doi:10.1074/jbc.M010161200.

[16] R. Z.-T. Luo, D. R. Beniac, A. Fernandes, C. C. Yip and F. P. Ottensmeyer, Quaternary structure of the insulin-insulin receptor complex. *Science* **1999**; 285(5430): 1077–1080, doi:10.1126/science.285.5430.1077.

[17] G. Jiang and T. Hunter, Receptor signaling: when dimerization is not enough. *Curr Biol* **1999**; 9(15): R568–R571, doi:10.1016/S0960-9822(99)80357-1.

[18] T. Moriki, H. Maruyama and I. N. Maruyama, Activation of preformed EGF receptor dimers by ligand-induced rotation of the transmembrane domain. *J Mol Biol* **2001**; 311(5): 1011–1026, doi:10.1006/jmbi.2001.4923.

[19] S. Li, N. D. Covino, E. G. Stein, J. H. Till and S. R. Hubbard, Structural and biochemical evidence for an autoinhibitory role for tyrosine 984 in the juxtamembrane region of the insulin receptor. *J Biol Chem* **2003**; 278(28): 26007–26014, doi:10.1074/jbc.M302425200.

[20] M. F. White and C. R. Kahn, The insulin signaling system. *J Biol Chem* **1994**; 269(1): 1–4.

[21] B. Cheatham and C. R. Kahn, Insulin action and the insulin signaling network. *Endocr Rev* **1995**; 16(2): 117–142, doi:10.1210/edrv-16-2-117.

[22] J. Avruch, Insulin signal transduction through protein kinase cascades. *Mol Cell Biochem* **1998**; 182(1-2): 31–48, doi:10.1023/A:1006823109415.

[23] A. R. Saltiel and C. R. Kahn, Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **2001**; 414(6865): 799–806, doi:10.1038/414799a.

[24] A. R. Saltiel and J. E. Pessin, Insulin signaling pathways in time and space. *Trends Cell Biol* **2002**; 12(2): 65–71, doi:10.1016/S0962-8924(01)02207-3.

[25] P. Cohen, The twentieth century struggle to decipher insulin signalling. *Nat Rev Mol Cell Biol* **2006**; 7(11): 867–873, doi:10.1038/nrm2043.

[26] C. M. Taniguchi, B. Emanuelli and C. R. Kahn, Critical nodes in signalling pathways: insights into insulin action. *Nat Rev Mol Cell Biol* **2006**; 7(2): 85–96, doi:10.1038/nrm1837.

[27] J. F. Youngren, Regulation of insulin receptor function. *Cell Mol Life Sci* **2007**; 64(7-8): 873–891, doi:10.1007/s00018-007-6359-9.

[28] M. F. White, The IRS-signalling system: a network of docking proteins that mediate insulin action. *Mol Cell Biochem* **1998**; 182(1-2): 3–11, doi:10.1023/A:1006806722619.

[29] B. P. Ceresa and J. E. Pessin, Insulin regulation of the Ras activation/inactivation cycle. *Mol Cell Biochem* **1998**; 182(1-2): 23–29, doi:10.1023/A:1006819008507.

[30] D. E. Moller and J. S. Flier, Insulin resistance–mechanisms, syndromes, and implications. *N Engl J Med* **1991**; 325(13): 938–948, doi:10.1056/NEJM199109263251307.

[31] K. Choi and Y.-B. Kim, Molecular mechanism of insulin resistance in obesity and type 2 diabetes. *Korean J Intern Med* **2010**; 25(2): 119–129, doi:10.3904/kjim.2010.25.2.119.

[32] B. Zhang, G. Salituro, D. Szalkowski, Z. Li, Y. Zhang, I. Royo, D. Vilella, M. T. Díez, F. Pelaez, C. Ruby, R. L. Kendall, X. Mao, P. Griffin, J. Calaycay, J. R. Zierath, J. V. Heck, R. G. Smith and D. E. Moller, Discovery of a small molecule insulin mimetic with antidiabetic activity in mice. *Science* **1999**; 284(5416): 974–977, doi:10.1126/science.284.5416.974.

[33] M. C. Pirrung, Y. Liu, L. Deng, D. K. Halstead, Z. Li, J. F. May, M. Wedel, D. A. Austin and N. J. G. Webster, Methyl scanning: total synthesis of demethylasterriquinone B1 and derivatives for identification of sites of interaction with and isolation of its receptor(s). *J Am Chem Soc* **2005**; 127(13): 4609–4624, doi:10.1021/ja044325h.

[34] M. A. Weber, A. Lidor, S. Arora, G. M. Salituro, B. B. Zhang and A. N. Sidawy, A novel insulin mimetic without a proliferative effect on vascular smooth muscle cells. *J Vasc Surg* **2000**; 32(6): 1118–1126, doi:10.1067/mva.2000.111280.

[35] N. J. G. Webster, K. Park and M. C. Pirrung, Signaling effects of demethylasterriquinone B1, a selective insulin receptor modulator. *ChemBioChem* **2003**; 4(5): 379–385, doi:10.1002/cbic.200200468.

[36] M. Li, J. F. Youngren, V. P. Manchem, M. Kozlowski, B. B. Zhang, B. A. Maddux and I. D. Goldfine, Small molecule insulin receptor activators potentiate insulin action in insulin-resistant cells. *Diabetes* **2001**; 50(10): 2323–2328, doi:10.2337/diabetes.50.10.2323.

[37] H. Kim, L. Deng, X. Xiong, W. D. Hunter, M. C. Long and M. C. Pirrung, Glyceraldehyde 3-phosphate dehydrogenase is a cellular target of the insulin mimic demethylasterriquinone B1. *J Med Chem* **2007**; 50(15): 3423–3426, doi:10.1021/jm070437i.

[38] K. Liu, L. Xu, D. Szalkowski, Z. Li, V. Ding, G. Kwei, S. Huskey, D. E. Moller, J. V. Heck, B. B. Zhang and A. B. Jones, Discovery of a potent, highly selective, and orally efficacious small-molecule activator of the insulin receptor. *J Med Chem* **2000**; 43(19): 3487–3494, doi:10.1021/jm000285q.

[39] S. A. Qureshi, V. Ding, Z. Li, D. Szalkowski, D. E. Biazzo-Ashnault, D. Xie, R. Saperstein, E. Brady, S. Huskey, X. Shen, K. Liu, L. Xu, G. M. Salituro, J. V. Heck, D. E. Moller, A. B. Jones and B. B. Zhang, Activation of insulin signal transduction pathway and anti-diabetic activity of small molecule insulin receptor activators. *J Biol Chem* **2000**; 275(47): 36590–36595, doi:10.1074/jbc.M006287200.

[40] H. B. Wood, R. Black, G. Salituro, D. Szalkowski, Z. Li, Y. Zhang, D. E. Moller, B. Zhang and A. B. Jones, The basal SAR of a novel insulin receptor activator. *Bioorg Med Chem Lett* **2000**; 10(11): 1189–1192, doi:10.1016/S0960-894X(00)00206-7.

[41] G. M. Salituro, F. Pelaez and B. B. Zhang, Discovery of a small molecule insulin receptor activator. *Recent Prog Horm Res* **2001**; 56: 107–126.

[42] V. D. H. Ding, S. A. Qureshi, D. Szalkowski, Z. Li, D. E. Biazzo-Ashnault, D. Xie, K. Liu, A. B. Jones, D. E. Moller and B. B. Zhang, Regulation of insulin signal transduction pathway by a small-molecule insulin receptor activator. *Biochem J* **2002**; 367(1): 301–306, doi:10.1042/BJ20020708.

[43] B. Lin, Z. Li, K. Park, L. Deng, A. Pai, L. Zhong, M. C. Pirrung and N. J. G. Webster, Identification of novel orally available small molecule insulin mimetics. *J Pharmacol Exp Ther* **2007**; 323(2): 579–585, doi:10.1124/jpet.107.126102.

[44] M. C. Pirrung, Z. Li, E. Hensley, Y. Liu, A. Tanksale, B. Lin, A. Pai and N. J. G. Webster, Parallel synthesis of indolylquinones and their cell-based insulin mimicry. *J Comb Chem* **2007**; 9(5): 844–854, doi:10.1021/cc070062m.

[45] M. C. Pirrung, L. Deng, B. Lin and N. J. G. Webster, Quinone replacements for small molecule insulin mimics. *ChemBioChem* **2008**; 9(3): 360–362, doi:10.1002/cbic.200700597.

[46] H. J. Tsai and S.-Y. Chou, A novel hydroxyfuroic acid compound as an insulin receptor activator – structure and activity relationship of a prenylindole moiety to insulin receptor activation. *J Biomed Sci* **2009**; 16: 68, doi:10.1186/1423-0127-16-68.

[47] R. Root-Bernstein and J. Vonck, Glucose binds to the insulin receptor affecting the mutual affinity of insulin and its receptor. *Cell Mol Life Sci* **2009**; 66(16): 2721–2732, doi:10.1007/s00018-009-0065-8.

[48] Y. Li, J. Kim, J. Li, F. Liu, X. Liu, K. Himmeldirk, Y. Ren, T. E. Wagner and X. Chen, Natural anti-diabetic compound 1,2,3,4,6-penta-O-galloyl-D-glucopyranose binds to insulin receptor and activates insulin-mediated glucose transport signaling pathway. *Biochem Biophys Res Commun* **2005**; 336(2): 430–437, doi:10.1016/j.bbrc.2005.08.103.

[49] M. Schlein, S. Ludvigsen, H. B. Olsen, A. S. Andersen, G. M. Danielsen and N. C. Kaarsholm, Properties of small molecules affecting insulin receptor function. *Biochemistry* **2001**; 40(45): 13520–13528, doi:10.1021/bi015672w.

[50] B. J. Stith, K. Woronoff and N. Wiernsperger, Stimulation of the intracellular portion of the human insulin receptor by the antidiabetic drug metformin. *Biochem Pharmacol* **1998**; 55(4): 533–536, doi:10.1016/S0006-2952(97)00540-6.

[51] S. H. Jung, Y. J. Ha, E. K. Shim, S. Y. Choi, J. L. Jin, H. S. Yun-Choi and J. R. Lee, Insulin-mimetic and insulin-sensitizing activities of a pentacyclic triterpenoid insulin receptor activator. *Biochem J* **2007**; 403(2): 243–250, doi:10.1042/BJ20061123.

[52] W. Zhang, D. Hong, Y. Zhou, Y. Zhang, Q. Shen, J.-Y. Li, L.-H. Hu and J. Li, Ursolic acid and its derivative inhibit protein tyrosine phosphatase 1B, enhancing insulin receptor phosphorylation and stimulating glucose uptake. *Biochim Biophys Acta* **2006**; 1760(10): 1505–1512, doi:10.1016/j.bbagen.2006.05.009.

[53] B. Diesel, S. Kulhanek-Heinze, M. Höltje, B. Brandt, H.-D. Höltje, A. M. Vollmar and A. K. Kiemer, α-Lipoic acid as a directly binding activator of the insulin receptor: protection from hepatocyte apoptosis. *Biochemistry* **2007**; 46(8): 2146–2155, doi:10.1021/bi602547m.

[54] V. P. Manchem, I. D. Goldfine, R. A. Kohanski, C. P. Cristobal, R. T. Lum, S. R. Schow, S. Shi, W. R. Spevak, E. Laborde, D. K. Toavs, H. O. Villar, M. M. Wick and M. R. Kozlowski, A novel small molecule that directly sensitizes the insulin receptor in vitro and in vivo. *Diabetes* **2001**; 50(4): 824–830, doi: 10.2337/diabetes.50.4.824.

[55] C. Pender, I. D. Goldfine, V. P. Manchem, J. L. Evans, W. R. Spevak, S. Shi, S. Rao, S. Bajjalieh, B. A. Maddux and J. F. Youngren, Regulation of insulin receptor function by a small molecule insulin receptor activator. *J Biol Chem* **2002**; 277(46): 43565–43571, doi:10.1074/jbc.M202426200.

[56] L. Robinson, S. Bajjalieh, N. Cairns, R. T. Lum, R. W. Macsata, V. P. Manchem, S. J. Park, S. Rao, S. R. Schow, S. Shi and W. R. Spevak, 5-Substituted isophthalamides as insulin receptor sensitizers. *Bioorg Med Chem Lett* **2008**; 18(12): 3492–3494, doi:10.1016/j.bmcl.2008.05.031.

[57] R. T. Lum, M. Cheng, C. P. Cristobal, I. D. Goldfine, J. L. Evans, J. G. Keck, R. W. Macsata, V. P. Manchem, Y. Matsumoto, S. J. Park, S. S. Rao, L. Robinson, S. Shi, W. R. Spevak and S. R. Schow, Design, synthesis, and structure–activity relationships of novel insulin receptor tyrosine kinase activators. *J Med Chem* **2008**; 51(19): 6173–6187, doi:10.1021/jm800600v.

[58] C. E. Heyliger, A. G. Tahiliani and J. H. McNeill, Effect of vanadate on elevated blood glucose and depressed cardiac performance of diabetic rats. *Science* **1985**; 227(4693): 1474–1477, doi:10.1126/science.3156405.

[59] Y. Shechter, J. Li, J. Meyerovitch, D. Gefel, R. Bruck, G. Elberg, D. S. Miller and A. Shisheva, Insulin-like actions of vanadate are mediated in an insulin-receptor-independent manner via non-receptor protein tyrosine kinases and protein phosphotyrosine phosphatases. *Molecular and Cellular Biochemistry* **1995**; 153(1-2): 39–47, doi:10.1007/BF01075917.

[60] S. García-Vicente, F. Yraola, L. Marti, E. González-Muñoz, M. J. García-Barrado, C. Cantó, A. Abella, S. Bour, R. Artuch, C. Sierra, N. Brandi, C. Carpéné, J. Moratinos, M. Camps, M. Palacín, X. Testar, A. Gumàă, F. Albericio, M. Royo, A. Mian and A. Zorzano, Oral Insulin-Mimetic Compounds That Act Independently of Insulin. *Diabetes* **2007**; 56(2): 486–493, doi:10.2337/db06-0269.

[61] E. Schmid, J. E. Benna, D. Galter, G. Klein and W. Dröge, Redox priming of the insulin receptor β-chain associated with altered tyrosine kinase activity and insulin responsiveness in the absence of tyrosine autophosphorylation. *FASEB J* **1998**; 12(10): 863–870.

[62] S. Tamura, Y. Fujita-Yamaguchi and J. Larner, Insulin-like effect of trypsin on the phosphorylation of rat adipocyte insulin receptor. *J Biol Chem* **1983**; 258(24): 14749–14752.

[63] J. W. Leef and J. Larner, Insulin-mimetic effect of trypsin on the insulin receptor tyrosine kinase in intact adipocytes. *J Biol Chem* **1987**; 262(30): 14837–14842.

[64] S. Clark, G. Eckardt, K. Siddle and L. C. Harrison, Changes in insulin-receptor structure associated with trypsin-induced activation of the receptor tyrosine kinase. *Biochem J* **1991**; 276(1): 27–33.

[65] R. Eglen and T. Reisine, Drug discovery and the human kinome: recent trends. *Pharmacol Ther* **2011**; 130(2): 144–156, doi:10.1016/j.pharmthera.2011.01.007.

[66] S. M. Massa, T. Yang, Y. Xie, J. Shi, M. Bilgen, J. N. Joyce, D. Nehama, J. Rajadas and F. M. Longo, Small molecule BDNF mimetics activate TrkB signaling and prevent neuronal degeneration in rodents. *J Clin Invest* **2010**; 120(5): 1774–1785, doi:10.1172/JCI41356.

[67] C. C. King, F. T. Zenke, P. E. Dawson, E. M. Dutil, A. C. Newton, B. A. Hemmings and G. M. Bokoch, Sphingosine is a novel activator of 3-phosphoinositide-dependent kinase 1. *J Biol Chem* **2000**; 275(24): 18108–18113, doi:10.1074/jbc.M909663199.

[68] V. Hindie, A. Stroba, H. Zhang, L. A. Lopez-Garcia, L. Idrissova, S. Zeuzem, D. Hirschberg, F. Schaeffer, T. J. D. Jørgensen, M. Engel, P. M. Alzari and R. M. Biondi, Structure and allosteric effects of low-molecular-weight activators on the protein kinase PDK1. *Nat Chem Biol* **2009**; 5(10): 758–764, doi:10.1038/nchembio.208.

[69] J. D. Sadowsky, M. A. Burlingame, D. W. Wolan, C. L. McClendon, M. P. Jacobson and J. A. Wells, Turning a protein kinase on or off from a single allosteric site via disulfide trapping. *Proc Natl Acad Sci USA* **2011**; 108(15): 6056–6061, doi:10.1073/pnas.1102376108.

[70] P. Ripphausen, B. Nisius, L. Peltason and J. Bajorath, Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J Med Chem* **2010**; 53(24): 8461–8467, doi:10.1021/jm101020z.

[71] P. V. Bharatam, D. S. Patel, L. Adane, A. Mittal and S. Sundriyal, Modeling and informatics in designing anti-diabetic agents. *Curr Pharm Des* **2007**; 13(34): 3518–3530.

[72] M. Sarmiento, L. Wu, Y. F. Keng, L. Song, Z. Luo, Z. Huang, G. Z. Wu, A. K. Yuan and Z. Y. Zhang, Structure-based discovery of small molecule inhibitors targeted to protein tyrosine phosphatase 1B. *J Med Chem* **2000**; 43(2): 146–155, doi:10.1021/jm990329z.

[73] T. N. Doman, S. L. McGovern, B. J. Witherbee, T. P. Kasten, R. Kurumbail, W. C. Stallings, D. T. Connolly and B. K. Shoichet, Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* **2002**; 45(11): 2213–2221, doi:10.1021/jm010548w.

[74] H. Park, B. R. Bhattarai, S. W. Ham and H. Cho, Structure-based virtual screening approach to identify novel classes of PTP1B inhibitors. *Eur J Med Chem* **2009**; 44(8): 3280–3284, doi:10.1016/j.ejmech.2009.02.011.

[75] M. O. Taha, Y. Bustanji, A. G. Al-Bakri, A.-M. Yousef, W. A. Zalloum, I. M. Al-Masri and N. Atallah, Discovery of new potent human protein tyrosine phosphatase inhibitors via pharmacophore and QSAR analysis followed by in silico screening. *J Mol Graph Model* **2007**; 25(6): 870–884, doi:10.1016/j.jmgm.2006.08.008.

[76] G. B. McGaughey, R. P. Sheridan, C. I. Bayly, J. C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.-F. Truchon and W. D. Cornell, Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model* **2007**; 47(4): 1504–1519, doi:10.1021/ci700052x.

[77] N. Nikolova and J. Jaworska, Approaches to measure chemical similarity – a review. *QSAR Comb Sci* **2003**; 22(9-10): 1006–1026, doi:10.1002/qsar.200330831.

[78] Y. C. Martin, J. L. Kofron and L. M. Traphagen, Do structurally similar molecules have similar biological activity? *J Med Chem* **2002**; 45(19): 4350–4358, doi:10.1021/jm020155c.

[79] H. Kubinyi, Similarity and dissimilarity: a medicinal chemist's view. *Perspect Drug Discov* **1998**; 9–11: 225–252, doi:10.1023/A:1027221424359.

[80] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick and J. W. Davies, How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J Chem Inf Model* **2009**; 49(1): 108–119, doi:10.1021/ci800249s.

[81] D. Digles and G. F. Ecker, Self-organizing maps for in silico screening and data visualization. *Mol Inf* **2011**; 30(10): 838–846, doi:10.1002/minf.201100082.

[82] T. Kohonen, Self-organized formation of topologically correct feature maps. *Biol Cybern* **1982**; 43: 59–69.

[83] V. S. Rose, I. F. Croall and H. J. H. MacFie, An application of unsupervised neural network methodology (Kohonen topology-preserving mapping) to QSAR analysis. *Quant Struct-Act Relat* **1991**; 10: 6–15, doi:10.1002/qsar.19910100103.

[84] J. Gasteiger, X. Li and A. Uschold, The beauty of molecular surfaces as revealed by self-organizing neural networks. *J Mol Graph* **1994**; 12(2): 90–97, doi:10.1016/0263-7855(94)80073-1.

[85] S. Anzali, J. Gasteiger, U. Holzgrabe, J. Polanski, J. Sadowski, A. Teckentrup and M. Wagener, The use of self-organizing neural networks in drug design. *Perspect Drug Discov Des* **1998**; 9-11: 273–299, doi:10.1023/A:1027276425268.

[86] K. V. Balakin, Y. A. Ivanenkov, N. P. Savchuk, A. A. Ivashchenko and S. Ekins, Comprehensive computational assessment of ADME properties using mapping techniques. *Curr Drug Discovery Technol* **2005**; 2(2): 99–113, doi:10.2174/1570163054064666.

[87] Y.-H. Wang, Y. Li, S.-L. Yang and L. Yang, Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J Chem Inf Model* **2005**; 45(3): 750–757, doi:10.1021/ci050041k.

[88] D. Kaiser, L. Terfloth, S. Kopp, J. Schulz, R. de Laet, P. Chiba, G. F. Ecker and J. Gasteiger, Self-organizing maps for identification of new inhibitors of P-glycoprotein. *J Med Chem* **2007**; 50(7): 1698–1702, doi:10.1021/jm060604z.

[89] O. Roche, G. Trube, J. Zuegge, P. Pflimlin, A. Alanine and G. Schneider, A virtual screening method for prediction of the HERG potassium channel liability of compound libraries. *ChemBioChem* **2002**; 3(5): 455–459, doi:10.1002/1439-7633(20020503)3:5<455::AID-CBIC455>3.0.CO;2-L.

[90] S. Ekins, K. V. Balakin, N. Savchuk and Y. Ivanenkov, Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. *J Med Chem* **2006**; 49(17): 5059–5071, doi:10.1021/jm060076r.

[91] D. S. Chekmarev, V. Kholodovych, K. V. Balakin, Y. Ivanenkov, S. Ekins and W. J. Welsh, Shape signatures: new descriptors for predicting cardiotoxicity in silico. *Chem Res Toxicol* **2008**; 21(6): 1304–1314, doi:10.1021/tx800063r.

[92] S. Hidaka, H. Yamasaki, Y. Ohmayu, A. Matsuura, K. Okamoto, N. Kawashita and T. Takagi, Nonlinear classification of hERG channel inhibitory activity by unsupervised classification method. *J Toxicol Sci* **2010**; 35(3): 393–399, doi:10.2131/jts.35.393.

[93] M. von Korff and T. Sander, Toxicity-indicating structural patterns. *J Chem Inf Model* **2006**; 46(2): 536–544, doi:10.1021/ci050358k.

[94] M. L. Lee and G. Schneider, Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries. *J Comb Chem* **2001**; 3(3): 284–289, doi:10.1021/cc000097l.

[95] J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design.* Second edn., Wiley-VCH, Weinheim, New York, Chichester, Brisbane, Singapore, Toronto **1999**.

[96] G. Espinosa, A. Arenas and F. Giralt, An integrated SOM-fuzzy ARTMAP neural system for the evaluation of toxicity. *J Chem Inf Comput Sci* **2002**; 42(2): 343–359, doi:10.1021/ci010329j.

[97] J. Gasteiger, A. Teckentrup, L. Terfloth and S. Spycher, Neural networks as data mining tools in drug design. *J Phys Org Chem* **2003**; 16(4): 232–245, doi:10.1002/poc.597.

[98] T. Noeske, D. Trifanova, V. Kauss, S. Renner, C. G. Parsons, G. Schneider and T. Weil, Synergism of virtual screening and medicinal chemistry: identification and optimization of allosteric antagonists of metabotropic glutamate receptor 1. *Bioorg Med Chem* **2009**; 17(15): 5708–5715, doi:10.1016/j.bmc.2009.05.072.

[99] D. Hristozov, T. I. Oprea and J. Gasteiger, Ligand-based virtual screening by novelty detection with self-organizing maps. *J Chem Inf Model* **2007**; 47(6): 2044–2062, doi:10.1021/ci700040r.

[100] H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski and J. Gasteiger, Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J Chem Inf Comput Sci* **1996**; 36(6): 1205–1213, doi:10.1021/ci960346m.

[101] *Molecular Operating Environment (MOE)*, 2010.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2010**.

[102] ChemDiv, Inc., 6605 Nancy Ridge Drive, San Diego, CA 92121 USA (`http://www.chemdiv.com`).

[103] P. Labute, A widely applicable set of descriptors. *J Mol Graph Model* **2000**; 18(4-5): 464–477, doi:10.1016/S1093-3263(00)00068-1.

[104] ADRIANA.*Code*, Molecular Networks GmbH, Erlangen, Germany (`http://www.mol-net.de`).

[105] SONNIA, Molecular Networks GmbH, Erlangen, Germany (`http://www.mol-net.de`).

[106] MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, California 94577.

[107] M. Stahl, Modifications of the scoring function in FlexX for virtual screening applications. *Perspect Drug Discov Des* **2000**; 20(1): 83–98, doi:10.1023/A:1008724921888.

[108] Phase, version 3.0, Schrödinger, LLC, New York, NY, 2008.

[109] MacroModel, version 9.5, Schrödinger, LLC, New York, NY, 2007.

[110] Scaffold-Based Classification Approach (SCA) script from SVL Exchange (`http://svl.chemcomp.com`). Chemical Computing Group, Inc., 2006.

[111] J. Xu and J. Stevenson, Drug-like index: a new approach to measure drug-like compounds and their diversity. *J Chem Inf Comput Sci* **2000**; 40(5): 1177–1187, doi:10.1021/ci000026+.

[112] M. Maechler, P. Rousseeuw, A. Struyf and M. Hubert, Cluster Analysis Basics and Extensions **2005**, unpublished.

[113] R Development Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria **2011**, ISBN 3-900051-07-0.

[114] Prime, version 1.6, Schrödinger, LLC, New York, NY, 2007.

[115] F. Klepsch, P. Chiba and G. F. Ecker, Exhaustive sampling of docking poses reveals binding hypotheses for propafenone type inhibitors of P-glycoprotein. *PLoS Comput Biol* **2011**; 7(5): e1002036, doi:10.1371/journal.pcbi.1002036.

[116] G. Kirsten, Dock RMSD Calculator script from SVL Exchange (`http://svl.chemcomp.com`). Chemical Computing Group, Inc., 2006.

[117] R. Baumgartner, *Virtual and real screening of natural products to find effective modulators of protein tyrosine phosphatase PTP1B.* Ph.D. thesis, University of Vienna **2010**.

[118] C. Williams, MACCS keys display script from SVL Exchange (`http://svl.chemcomp.com`). Chemical Computing Group, Inc., 2006.

[119] M. H. J. Seifert and M. Lang, Essential factors for successful virtual screening. *Mini Rev Med Chem* **2008**; 8(1): 63–72.

[120] I. Hers, E. E. Vincent and J. M. Tavaré, Akt signalling in health and disease. *Cell Signal* **2011**; 23(10): 1515–1527, doi:10.1016/j.cellsig.2011.05.004.

[121] J. B. Baell and G. A. Holloway, New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* **2010**; 53(7): 2719–2740, doi:10.1021/jm901137j.

[122] S. Saubern, R. Guha and J. B. Baell, KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Mol Inf* **2011**; 30(10): 847–850, doi:10.1002/minf.201100076.

[123] J. Choi, Y. Ko, H. S. Lee, Y. S. Park, Y. Yang and S. Yoon, Identification of (β-carboxyethyl)-rhodanine derivatives exhibiting peroxisome proliferator-activated receptor γ activity. *Eur J Med Chem* **2010**; 45(1): 193–202, doi: 10.1016/j.ejmech.2009.09.042.

[124] B. R. Bhattarai, B. Kafle, J.-S. Hwang, S. W. Ham, K.-H. Lee, H. Park, I.-O. Han and H. Cho, Novel thiazolidinedione derivatives with anti-obesity effects: Dual action as PTP1B inhibitors and PPAR-γ activators. *Bioorg Med Chem Lett* **2010**; 20(22): 6758–6763, doi:10.1016/j.bmcl.2010.08.130.

[125] P. A. Johnston, C. A. Foster, T. Y. Shun, J. J. Skoko, S. Shinde, P. Wipf and J. S. Lazo, Development and implementation of a 384-well homogeneous fluorescence intensity high-throughput screening assay to identify mitogen-activated protein kinase phosphatase-1 dual-specificity protein phosphatase inhibitors. *Assay Drug Dev Technol* **2007**; 5(3): 319–332, doi:10.1089/adt.2007.066.

[126] P. A. Johnston, C. A. Foster, M. B. Tierno, T. Y. Shun, S. N. Shinde, W. D. Paquette, K. M. Brummond, P. Wipf and J. S. Lazo, Cdc25B dual-specificity phosphatase inhibitors identified in a high-throughput screen of the NIH compound library. *Assay Drug Dev Technol* **2009**; 7(3): 250–265, doi: 10.1089/adt.2008.186.

[127] M. Z. Strowski, Z. Li, D. Szalkowski, X. Shen, X.-M. Guan, S. Jüttner, D. E. Moller and B. B. Zhang, Small-molecule insulin mimetic reduces hyperglycemia and obesity in a nongenetic mouse model of type 2 diabetes. *Endocrinology* **2004**; 145(11): 5259–5268, doi:10.1210/en.2004-0610.

Appendix A

Databases

Insulin mimetics: Table A.1 shows demethylasterriquinone B1 and all its derivatives from the literature used in this study. It gives the numbering used in this work as well as all previously published numbers and codes and the activity class which the compounds were assigned to. The structures of the molecules are given in SMILES format.

Nr.	Name in literature	Activity	SMILES
1	L-783,281,[32,39] compound **1**,[38–40] demethylasterriquinone B-1,[41] demethylasterriquinone B1,[33,35] demethylasterriquinone-B1,[43] DMAQ-B1,[41] DAQ B1,[33,35] DAQ-B1[43]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
2	L-767,827[32,41]	0	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1C(C=C)(C)C)cccc2)c1c2c([nH]c1C(C=C)(C)C)cccc2
3	26{*1*}[33]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c(n(C)c1C(C=C)(C)C)cccc2
4	26{*2*}[33]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)c(ccc2)C

Table A.1: DMAQ-B1 and its derivatives

Continued on next page...

Nr.	Name in literature	Activity	SMILES
5	26{*3*}[33]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cc(cc2)C
6	26{*4*}[33]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2cc(ccc2[nH]c1C(C=C)(C)C)C
7	26{*5*}[33]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2C
8	26{*6*},[33] **8**[40]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c(n(c1)C)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
9	26{*7*}[33]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1C)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
10	26{*8*}[33]	0	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2C)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
11	26{*9*}[33]	0	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(cc(c2)C)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
12	26{*10*}[33]	0	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(CC=C(C)C)c(cc2)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
13	26{*11*}[33]	0	O(C)C1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1C(C=C)(C)C)cccc2)c1c2c([nH]c1)c(ccc2)CC=C(C)C
14	26{*12*},[33] **3**[40]	0	O(C)C1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
15	**2h**,[38] compound **2**,[39,42] CPD2[127]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c(n(c1)C)cccc2)c1ccccc1
16	**2c**,[38] compound **3**[39,42]	0	O(C)c1ccc(cc1)C=1C(=O)C(O)=C(C(=O)C=1O)c1ccccc1
17	**2**[40]	0	O(C)C1=C(C(=O)C(OC)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
18	**4**[40]	0	O=C1C(=C(NC)C(=O)C(=C1NC)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2

Table A.1: DMAQ-B1 and its derivatives

Continued on next page...

Nr.	Name in literature	Activity	SMILES
19	**5**[40]	0	S(OC1=C(C(=O)C(OS(=O)(=O)C(F)(F)F)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2)(=O)(=O)C(F)(F)F
20	**6**[40]	0	S(OC1=C(C(=O)C(N)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2)(=O)(=O)C(F)(F)F
21	**7**[40]	0	OC1=C(C(=O)C(N)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
22	**9**[40]	0	OC1=C(C(=O)C(O)=C(C1=O)c1c2c(n(c1)C)c(ccc2)CC=C(C)C)c1c2c(n(C)c1C(C=C)(C)C)cccc2
23	**11**[40]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CCC(O)(C)C)c1c2c(n(C)c1C(C=C)(C)C)cccc2
24	**13**[40]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C)c1c2c([nH]c1C(CC)(C)C)cccc2
25	**14**[40]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CCC(C)C)c1c2c([nH]c1C(CC)(C)C)cccc2
26	**15**[40]	1	FC(F)(F)C(OC(CCc1c2[nH]cc(c2ccc1)C=1C(=O)C(O)=C(C(=O)C=1O)c1c2c([nH]c1C(C=C)(C)C)cccc2)(C)C)=O
27	**16**[40]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CCC(O)(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
28	**17**[40]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)c(ccc2)CC(O)C(O)(C)C)c1c2c([nH]c1C(C=C)(C)C)cccc2
29	**2a**[38]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)cccc2)c1c2c([nH]c1)c(ccc2)C
30	**2b**[38]	1	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1)cccc2)c1ccccc1
31	**2d**[38]	1	o1cc(c2c1cccc2)C=1C(=O)C(O)=C(C(=O)C=1O)c1ccccc1
32	**2e**[38]	1	s1cc(c2c1cccc2)C=1C(=O)C(O)=C(C(=O)C=1O)c1ccccc1

Table A.1: DMAQ-B1 and its derivatives

Continued on next page...

Nr.	Name in literature	Activity	SMILES
33	**2f**[38]	1	OC1=C(C(=O)C(O)=C(C1=O)c1ccccc1)c1c2c(ccc1)cccc2
34	**2g**[38]	1	OC1=C(C(=O)C(O)=C(C1=O)c1ccccc1)c1cc2c(cc1)cccc2
35	ZL-I-197[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)cccc2
36	ZL-I-186[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1C)cccc2
37	ZL-III-244[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1CC)cccc2
38	ZL-I-184[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1C1CC1)cccc2
39	ZL-I-185[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1C(C)C)cccc2
40	LD-I-205[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1C1(CC1)C)cccc2
41	ZL-II-205[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1C(C)(C)C)cccc2
42	ZL-III_254[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1C1(CCCCC1)C)cccc2
43	ZL-I-207[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1-c1ccccc1)cccc2
44	ZL-202,[43] **31**[33]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1C(C=C)(C)C)cccc2
45	LD-I-217[43]	0	Fc1c2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)ccc1
46	LD9B[43]	0	Clc1c2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)ccc1
47	LD11B[43]	0	Brc1c2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)ccc1
48	LD19B[43]	0	O(C)c1c2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)ccc1
49	ZL-III-249[43]	0	O(Cc1ccccc1)c1c2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)ccc1
50	LD13B[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)cccc2C
51	LD-I-204[43]	0	Fc1cc2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)cc1
52	ZL-III-255[43]	0	Clc1cc2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)cc1

Table A.1: DMAQ-B1 and its derivatives

Continued on next page...

Nr.	Name in literature	Activity	SMILES
53	ZL-III-257[43]	0	Brc1cc2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)cc1
54	LD-5B[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2cc(O)ccc2[nH]c1
55	ZL-III-248[43]	0	O(Cc1ccccc1)c1cc2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)cc1
56	LD20B[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2cc(ccc2[nH]c1)C
57	LD-I-210[43]	1	Fc1cc2[nH]cc(c2cc1)C=1C(=O)C(O)=CC(=O)C=1O
58	ZL-III-253[43]	1	Clc1cc2[nH]cc(c2cc1)C=1C(=O)C(O)=CC(=O)C=1O
59	LD-1-214[43]	1	O(Cc1ccccc1)c1cc2[nH]cc(c2cc1)C=1C(=O)C(O)=CC(=O)C=1O
60	ZL-III-250[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)cc(cc2)C
61	LD-I-207[43]	0	Clc1c2[nH]cc(c2ccc1)C=1C(=O)C(O)=CC(=O)C=1O
62	LD-I-216[43]	0	Brc1c2[nH]cc(c2ccc1)C=1C(=O)C(O)=CC(=O)C=1O
63	ZL-I-175[43]	1	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(ccc2)C
64	LD-I-215[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(ccc2)CCC
65	ZL-196[43,45]	1	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C
66	LD25B[43]	1	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(ccc2)C\C=C(\CCC=C(C)C)/C
67	LD26B[43]	1	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(ccc2)C\C=C(\CC\C=C(\CCC=C(C)C)/C)/C
68	LD-I-219[43]	1	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(ccc2)Cc1ccccc1
69	LD-I-218[43]	1	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(ccc2)Cc1ccccc1C
70	LD22B[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(ccc2)C(C)(C)C
71	LD-I-143[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(ccc2)-c1ccccc1

Table A.1: DMAQ-B1 and its derivatives

Continued on next page...

Nr.	Name in literature	Activity	SMILES
72	ZL-III-256[43]	0	O(C)c1c2[nH]cc(c2ccc1)C=1C(=O)C(O)=CC(=O)C=1O
73	LD-17[43]	1	O(Cc1ccccc1)c1c2[nH]cc(c2ccc1)C=1C(=O)C(O)=CC(=O)C=1O
74	ZL-I-199[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2cc(ccc2[nH]c1C)C
75	ZL-I-192[43]	0	O(C)c1cc2c([nH]c(C)c2C=2C(=O)C(O)=CC(=O)C=2O)cc1
76	ZL-III-243[43]	0	Clc1cc2c([nH]c(C)c2C=2C(=O)C(O)=CC(=O)C=2O)cc1
77	LD-I-209[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1C)cc(cc2)C
78	LD15B[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1C)c(ccc2)C
79	LD-I-125[43]	1	O1c2c(OC1)cc1[nH]cc(c1c2)C=1C(=O)C(O)=CC(=O)C=1O
80	ZL-III-251[43]	1	O(C)c1cc2c([nH]cc2C=2C(=O)C(O)=CC(=O)C=2O)cc1OC
81	LD-I-208[43]	1	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c(C)c(cc2)C
82	LD-I-213[43]	1	OC1=C(C(=O)C(O)=CC1=O)c1c2c([nH]c1)c1c(cc2)cccc1
83	ZL194[43]	0	OC1=C(C(=O)C(O)=CC1=O)c1c2c(n(c1)C)cccc2
84	ZL-III-168-II[43]	1	OC1=C(c2c3c([nH]c2C)cccc3)C(=O)c2c(cccc2O)C1=O
85	ZL-III198[43]	0	OC1=C(C(=O)c2c(cccc2O)C1=O)c1c2c([nH]c1)c(ccc2)C
86	ZL-III199[43]	0	OC1=C(C(=O)c2c(cccc2O)C1=O)c1c2c([nH]c1)c(ccc2)C(C)(C)C
87	ZL-III200[43]	0	OC1=C(C(=O)c2c(cccc2O)C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C
88	ZL-III202[43]	0	OC1=C(C(=O)c2c(cccc2O)C1=O)c1c2c([nH]c1)c(ccc2)CCC
89	ZL-III213[43]	0	OC1=C(C(=O)c2c(cccc2O)C1=O)c1c2c(n(c1)C)cccc2
90	ZL-III214[43]	0	O(C)c1c2[nH]cc(c2ccc1)C1=C(O)C(=O)c2c(C1=O)c(O)ccc2
91	LD-I-206[43]	1	Fc1c2[nH]cc(c2ccc1)C=1C(=O)C(OC)=CC(=O)C=1OC

Table A.1: DMAQ-B1 and its derivatives

Continued on next page...

Nr.	Name in literature	Activity	SMILES
92	ZLV-212[43]	0	OC1=C(C(=O)C1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C
93	**14**[44]	1	Clc1cc2[nH]cc(c2cc1)C=1C(=O)C(O)=C(C(=O)C=1O)c1c2c([nH]c1C(C)(C)C)cccc2
94	**15**[44]	1	Clc1cc2[nH]cc(c2cc1)C=1C(=O)C(O)=C(C(=O)C=1O)c1c2c([nH]c1CC)cccc2
95	**16**[44]	1	Clc1cc2c([nH]cc2C=2C(=O)C(O)=C(C(=O)C=2O)c2c3c([nH]c2-c2ccccc2)cccc3)cc1
96	**17**[44]	1	Clc1c2[nH]cc(c2ccc1)C=1C(=O)C(O)=C(C(=O)C=1O)c1c2c([nH]c1C1CC1)cccc2
97	DAQ A1[35]	0	OC1=C(C(=O)C(O)=C(C1=O)c1c2c(n(c1)C(C=C)(C)C)cccc2)c1c2c(n(c1)C(C=C)(C)C)cccc2
98	KP-271-1[35]	0	OC1=C(C(=O)C(O)=C(C1=O)c1c2c([nH]c1C)cccc2)c1c2c([nH]c1C)cccc2
99	**4**[45]	0	OC1=C(C=CC=CC1=O)c1c2c([nH]c1)c(ccc2)CC=C(C)C
100	**6**[45]	1	O1C(=C(O)C(=O)C=C1CO)c1c2c([nH]c1)c(ccc2)CC=C(C)C
101	**8**[45]	1	OC1=C(N(C)C(=CC1=O)CO)c1c2c([nH]c1)c(ccc2)CC=C(C)C
Used for external validation:			
102	D-410639[46]	1	o1c(C(O)=O)c(-c2c3c([nH]c2)cccc3)c(O)c1C(=O)c1c2c([nH]c1)c(ccc2)CCCCCO

Table A.1: DMAQ-B1 and its derivatives

Identified compounds: Table A.2 summarizes all 367 potentially active compounds identified with self-organizing maps, fingerprint similarity or shape similarity.

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
000A-0009	0	1	0	0	20
000A-0036	0	0	0	1	21
000A-0047	1	0	0	0	20
000A-0190	0	0	0	1	25
0099-0308	1	0	0	0	29
0392-0008	0	0	0	1	32
0457-0021	0	0	1	0	34
0669-0071	0	1	0	0	40
0682-0067	0	0	1	0	32
0828-0235	0	1	0	0	40
0883-0041	0	0	0	1	32
1302-0002	0	0	0	1	49
1306-0027	1	0	0	0	50
1345-2374	1	0	0	0	52
1348-1605	1	0	0	0	53
1574-1707	1	0	0	0	55
1682-6957	0	0	0	1	57
1683-6896	0	0	0	1	32
1773-0151	0	0	1	0	34
2110-0307	0	0	0	1	57
2110-0308	0	0	0	1	57
2367-1224	1	0	0	0	75
2509-0025	1	0	0	0	78
2820-0981	1	0	0	0	81
2950-0554	0	0	0	1	85
3029-0578	0	0	0	1	87/208
3042-5045	1	0	0	0	88
3057-0993	0	0	0	1	57
3093-0115	1	0	0	0	91
3232-1864	0	0	0	1	57
3237-1339	0	0	1	0	95
3254-3796	1	0	0	0	96
3257-2499	0	0	1	0	97
3270-0678	0	0	0	1	32
3296-0057	0	1	0	0	103
3331-2182	0	1	0	0	105

Table A.2: 367 identified compounds *Continued on next page...*

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
3347-1012	0	0	1	0	87
3365-7324	1	0	0	0	107
3379-3178	1	0	0	0	108
3406-0397	0	1	0	0	110
3454-1905	0	1	0	0	114
3480-0282	1	0	0	0	115
3505-6187	1	0	0	0	116
3505-6189	1	0	0	0	116
3546-0632	0	0	1	0	120
3546-0641	0	0	1	0	120
3553-1638	0	0	0	1	121
3555-0175	0	0	1	0	123
3630-0578	0	0	0	1	32
3807-4416	1	0	0	0	130
3902-0345	0	1	0	0	134
3902-0852	0	1	0	0	134
3966-0592	1	0	0	0	20
3989-0098	0	0	0	1	32
4052-4503	1	0	0	0	138
4057-0014	0	0	0	1	139
4076-0245	1	0	0	0	96
4137-1358	1	0	0	0	141
4161-2736	0	1	0	0	142
4204-0085	1	0	0	0	20
4281-2071	0	0	0	1	139
4281-2127	0	0	0	1	139
4333-2466	0	1	0	0	40
4340-0101	1	0	0	0	88
4340-1467	1	0	0	0	88
4451-0051	1	0	0	0	150
4451-0078	1	0	0	0	150
4459-0077	0	0	1	0	123
4478-7661	1	0	0	0	157
4513-0296	0	0	0	1	32
4513-0429	0	0	1	0	20
4522-0096	0	1	0	0	160
4533-0061	0	1	0	0	139
4546-0033	0	0	1	0	163
4587-0405	0	0	0	1	165
4608-0005	1	0	0	0	20

Table A.2: 367 identified compounds *Continued on next page...*

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
4608-0014	0	1	0	0	20
4659-0068	0	0	0	1	57
4663-2462	1	0	0	0	150
4673-0543	1	0	0	0	170
4693-1125	0	0	0	1	137
4725-0413	1	0	0	0	150
4764-3635	0	0	0	1	139
4883-0016	1	0	0	0	150
5119-1121	0	1	0	0	185
5181-0618	0	1	0	0	188
5218-0873	1	0	0	0	190
5218-1214	1	0	0	0	191
5218-1215	1	0	0	0	191
5218-1227	1	0	0	0	191
5241-0101	1	0	0	0	150
5340-2513	0	0	0	1	195
5408-1692	0	0	0	1	198
5408-2494	0	0	1	0	200
5408-2497	0	0	1	0	200
5441-1081	1	0	0	0	202
5498-2747	0	0	1	0	203
5547-0011	0	0	0	1	208
5634-0239	0	0	0	1	209
5650-0022	0	0	0	1	32
5683-0379	0	1	0	0	40
5696-0017	0	0	0	1	212
5735-0003	0	0	1	0	213
5750-3148	0	0	0	1	20
5775-0353	0	0	0	1	139
5910-0153	1	0	0	0	218
5977-0726	0	1	0	0	32
5982-0100	0	1	0	0	208
5982-0159	0	1	0	0	208
6049-2038	0	1	0	0	160
6173-0173	1	0	0	0	218
6231-0119	0	1	0	0	444
6253-0655	0	1	0	0	105
6266-1861	0	0	1	0	232
6332-1077	0	0	0	1	139
6332-1304	0	1	0	0	139

Table A.2: 367 identified compounds *Continued on next page...*

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
6332-2060	0	0	0	1	139
6463-4542	0	0	0	1	139
6623-0410	0	1	0	0	105
6711-0013	0	0	1	0	200
6725-3881	0	0	1	0	243
6843-3207	0	0	0	1	165
6877-0609	1	0	0	0	229
6944-0119	0	0	0	1	244
6957-0024	1	0	0	0	245
7165-0402	0	1	0	0	105
7217-0005	0	1	0	0	20
7244-0063	0	0	0	1	32
7287-0703	1	0	0	0	252
7407-0808	1	0	0	0	150
7491-0197	0	0	1	0	255
7790-2913	1	0	0	0	258
8001-8409	1	0	0	0	260
8001-9473	0	0	0	1	261
8003-1969	0	1	0	0	134
8003-9632	0	1	0	0	134
8004-2156	1	0	0	0	107
8005-3554	1	0	0	0	267
8006-9874	1	0	0	0	105
8007-3924	0	1	0	0	270
8007-8134	0	0	1	0	271
8008-6172	0	0	1	0	272
8008-8508	1	0	0	0	105
8009-1091	1	0	0	0	88
8009-1933	0	1	0	0	139
8009-3415	0	0	0	1	88
8009-4794	1	0	0	0	279
8009-6199	1	0	0	0	282
8009-7265	0	0	0	1	105
8009-8459	0	0	0	1	88
8010-2269	1	0	0	0	88
8010-2417	0	0	1	0	288
8010-3415	1	0	0	0	289
8010-7159	0	1	0	0	291
8010-8873	1	0	0	0	88
8011-9619	0	1	0	0	291

Table A.2: 367 identified compounds *Continued on next page...*

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
8011-9622	0	1	0	0	291
8012-0040	0	0	0	1	296
8012-0041	0	0	0	1	296
8012-4701	1	0	0	0	299
8012-5686	0	0	1	0	300
8012-6056	1	0	0	0	20
8012-6900	1	0	0	0	299
8012-8657	0	1	0	0	291
8012-8742	1	0	0	0	20
8012-8948	0	0	0	1	32
8012-9497	0	1	0	0	291
8012-9499	0	1	0	0	291
8013-0077	0	1	0	0	291
8013-0132	0	1	0	0	291
8013-0156	0	1	0	0	291
8013-0193	1	0	0	0	20
8013-0195	1	0	0	0	20
8013-0282	0	1	0	0	291
8013-0567	1	0	0	0	306
8013-0795	0	1	0	0	291
8013-0806	0	1	0	0	291
8013-1367	0	1	0	0	291
8013-1663	1	0	0	0	20
8013-1838	0	1	0	0	291
8013-1855	0	1	0	0	291
8013-1885	0	1	0	0	291
8013-5362	1	0	0	0	88
8013-5366	1	0	0	0	88
8013-5371	1	0	0	0	88
8013-5372	1	0	0	0	88
8014-1054	0	1	0	0	291
8014-1240	0	1	0	0	291
8014-2596	1	0	0	0	150
8014-8765	0	1	0	0	291
8014-8856	0	1	0	0	291
8014-8857	0	1	0	0	291
8014-9065	0	1	0	0	291
8015-2473	1	0	0	0	32
8015-2557	1	0	0	0	218
8015-4205	1	0	0	0	32

Table A.2: 367 identified compounds *Continued on next page...*

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
8017-3855	0	0	0	1	321
8017-5369	0	0	0	1	32
8017-5551	1	0	0	0	32
8017-6445	1	0	0	0	150
8017-6446	1	0	0	0	150
8017-7046	0	0	0	1	328
8017-7602	0	0	1	0	329
C073-3020	0	1	0	0	331
C073-3316	0	1	0	0	331
C073-3327	0	1	0	0	331
C073-3329	0	1	0	0	331
C073-3341	0	1	0	0	331
C073-3342	0	1	0	0	331
C073-3358	0	1	0	0	331
C073-3393	0	1	0	0	331
C073-3426	0	1	0	0	331
C073-3613	0	1	0	0	331
C073-3676	0	1	0	0	331
C073-3687	0	1	0	0	331
C073-3688	0	1	0	0	331
C073-3692	0	1	0	0	331
C073-3700	0	1	0	0	331
C073-3701	0	1	0	0	331
C073-3710	0	1	0	0	331
C073-3718	0	1	0	0	331
C073-3723	0	1	0	0	331
C073-3731	0	1	0	0	331
C090-0051	1	0	0	0	339
C090-0052	1	0	0	0	339
C090-0058	1	0	0	0	339
C090-0059	1	0	0	0	339
C090-0241	1	0	0	0	339
C090-0245	1	0	0	0	339
C090-0250	1	0	0	0	339
C090-0251	1	0	0	0	339
C090-0315	1	0	0	0	339
C090-0316	1	0	0	0	339
C090-0323	1	0	0	0	339
C090-0327	1	0	0	0	339
C090-0328	1	0	0	0	339

Table A.2: 367 identified compounds *Continued on next page...*

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
C090-0329	1	0	0	0	339
C090-0334	1	0	0	0	339
C090-0335	1	0	0	0	339
C090-0387	1	0	0	0	339
C090-0388	1	0	0	0	339
C090-0395	1	0	0	0	339
C200-2391	0	0	1	0	346
C200-4795	0	0	1	0	346
C202-0180	0	1	0	0	291
C226-0908	1	0	0	0	347
C226-1592	1	0	0	0	347
C229-0098	0	1	0	0	291
C229-0639	0	1	0	0	291
C229-0783	0	1	0	0	291
C229-0793	0	1	0	0	291
C270-0349	0	0	0	1	105
C276-0105	0	0	0	1	352
C294-0271	0	0	0	1	32
C301-0535	1	0	0	0	354
C301-0544	1	0	0	0	354
C301-0545	1	0	0	0	354
C301-1174	0	0	1	0	355
C301-4173	0	1	0	0	356
C301-5215	1	0	0	0	105
C301-5270	1	0	0	0	105
C301-5408	1	0	0	0	105
C301-5547	1	0	0	0	105
C350-0372	1	0	0	0	360
C350-0374	1	0	0	0	360
C350-0376	1	0	0	0	361
C350-0386	1	0	0	0	361
C350-0702	1	0	0	0	360
C350-0717	1	0	0	0	360
C448-1136	1	0	0	0	363
C493-1072	0	1	0	0	365
C493-1090	0	1	0	0	365
C547-0761	0	1	0	0	87
C607-0621	1	0	0	0	87
C620-0630	1	0	0	0	339
C620-0634	1	0	0	0	339

Table A.2: 367 identified compounds *Continued on next page...*

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
C620-0640	1	0	0	0	339
C620-0641	1	0	0	0	339
C620-0647	1	0	0	0	339
C620-0665	1	0	0	0	339
C620-0673	1	0	0	0	339
C620-0678	1	0	0	0	339
C651-0450	1	0	0	0	376
C753-0198	0	0	0	1	380
C753-1342	0	0	0	1	380
C756-0078	0	1	0	0	385
C879-1278	0	0	0	1	380
D052-0102	0	1	0	0	389
D052-0107	0	1	0	0	389
D143-0013	0	0	1	0	390
D155-0032	0	0	0	1	32
D159-0883	1	0	0	0	88
D177-1129	0	1	0	0	389
D252-0128	0	0	1	0	393
E015-0904	0	1	0	0	394
E518-1612	0	0	0	1	380
E693-0068	0	0	0	1	380
E693-0476	0	0	0	1	380
E847-0220	0	0	0	1	380
E938-0003	1	0	0	0	105
E938-0009	1	0	0	0	105
E938-0011	1	0	0	0	105
E938-0021	1	0	0	0	105
E938-0041	1	0	0	0	105
E938-0045	1	0	0	0	105
E938-0046	1	0	0	0	105
E938-0051	1	0	0	0	105
E938-0077	1	0	0	0	105
E938-0096	1	0	0	0	105
E938-0112	1	0	0	0	105
E938-0127	1	0	0	0	105
E938-0129	1	0	0	0	105
E938-0156	1	0	0	0	105
F019-0045	1	0	0	0	410
F019-1000	1	0	0	0	299
F019-2195	1	0	0	0	299

Table A.2: 367 identified compounds *Continued on next page...*

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
G396-0138	0	0	0	1	380
G396-0426	0	0	0	1	380
G396-0972	0	0	0	1	380
G856-6183	1	0	0	0	88
G856-6192	1	0	0	0	88
K026-0216	0	0	0	1	87
K026-0228	0	0	0	1	87
K026-0229	0	0	0	1	87
K089-0089	1	0	0	0	417
K235-0065	0	0	1	0	346
K405-3512	0	0	1	0	425
K405-3521	0	0	1	0	425
K780-0810	0	0	1	0	200
K780-0823	0	0	1	0	200
K781-0936	0	0	0	1	32
K784-3063	1	0	0	0	429
K784-4408	0	1	0	0	331
K786-3665	1	0	0	0	150
K786-6896	0	1	0	0	385
K786-6899	0	1	0	0	385
K786-6904	0	1	0	0	385
K786-9821	0	1	0	0	385
K786-9822	0	1	0	0	385
K786-9823	0	1	0	0	385
K786-9832	0	1	0	0	385
K788-0444	0	1	0	0	385
K788-0448	0	1	0	0	385
K788-0705	0	1	0	0	385
K788-4172	1	0	0	0	385
K788-5456	0	1	0	0	366
K788-6130	0	1	0	0	385
K788-6634	0	1	0	0	385
K788-7022	0	1	0	0	385
K815-0023	0	0	0	1	87
K815-0024	0	0	0	1	87
R153-0142	0	0	1	0	34
R153-0143	0	0	1	0	34
R153-0151	0	0	1	0	34
R153-0152	0	0	1	0	34
R153-0159	0	0	1	0	34

Table A.2: 367 identified compounds *Continued on next page...*

ChemDiv ID	SOM 2D	SOM VSA	Shape	FP	Scaffold ID
R153-0162	0	0	1	0	34
R153-0168	0	0	1	0	34
R153-0176	0	0	1	0	34
R153-0185	0	0	1	0	34
R153-0186	0	0	1	0	34
R153-0188	0	0	1	0	34
R153-0192	0	0	1	0	34
R153-0196	0	0	1	0	34
R153-0207	0	0	1	0	34
R153-0213	0	0	1	0	34
R153-0221	0	0	1	0	34

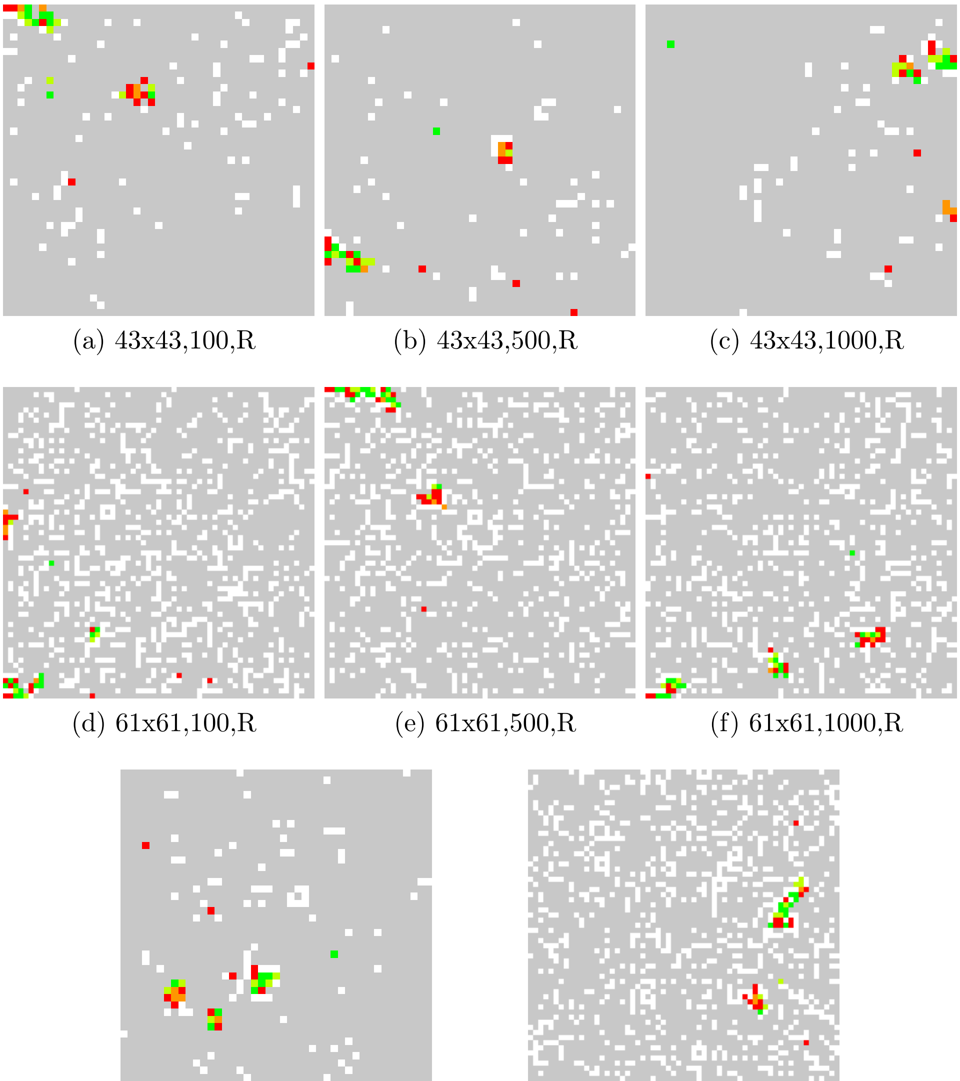
Table A.2: 367 identified compounds

Purchased compounds: The given codes are the numbering as used in the present work, the internal code during the testing of the compounds and the ID used in the ChemDiv[102] library. Molecules with grey background were not soluble in the stock solution. Structures of the molecules can be found on pages 85 and 86.

Nr.	Internal Code	ChemDiv ID	Method
103	CD1	4204-0085	SOM (2D)
104	CD2	4451-0051	SOM (2D)
105	CD3	5982-0100	SOM (VSA)
106	CD4	6463-4542	FP
107	CD5	7244-0063	FP
108	CD6	8014-1054	SOM (VSA)
109	CD7	C073-3327	SOM (VSA)
110	CD8	C090-0245	SOM (2D)
111	CD9	D159-0883	SOM (2D)
112	CD10	E938-0156	SOM (2D)
113	CD11	K788-0448	SOM (VSA)
114	CD12	K815-0023	FP
115	CD13	R153-0196	Shape
116	CD14	0095-0198	hand-picked
117	CD15	6623-0410	SOM (VSA)
118	CD16	7165-0402	SOM (VSA)
119	CD17	8008-8508	SOM (2D)
120	CD18	8009-7265	FP
121	CD19	C270-0349	FP
122	CD20	C301-4948	hand-picked
123	CD21	C301-5215	SOM (2D)
124	CD22	C301-5408	SOM (2D)
125	CD23	C301-5428	hand-picked
126	CD24	E938-0003	SOM (2D)
127	CD25	E938-0021	SOM (2D)
128	CD26	E938-0036	hand-picked
129	CD27	E938-0051	SOM (2D)
130	CD28	E938-0077	SOM (2D)
131	CD29	E938-0078	hand-picked

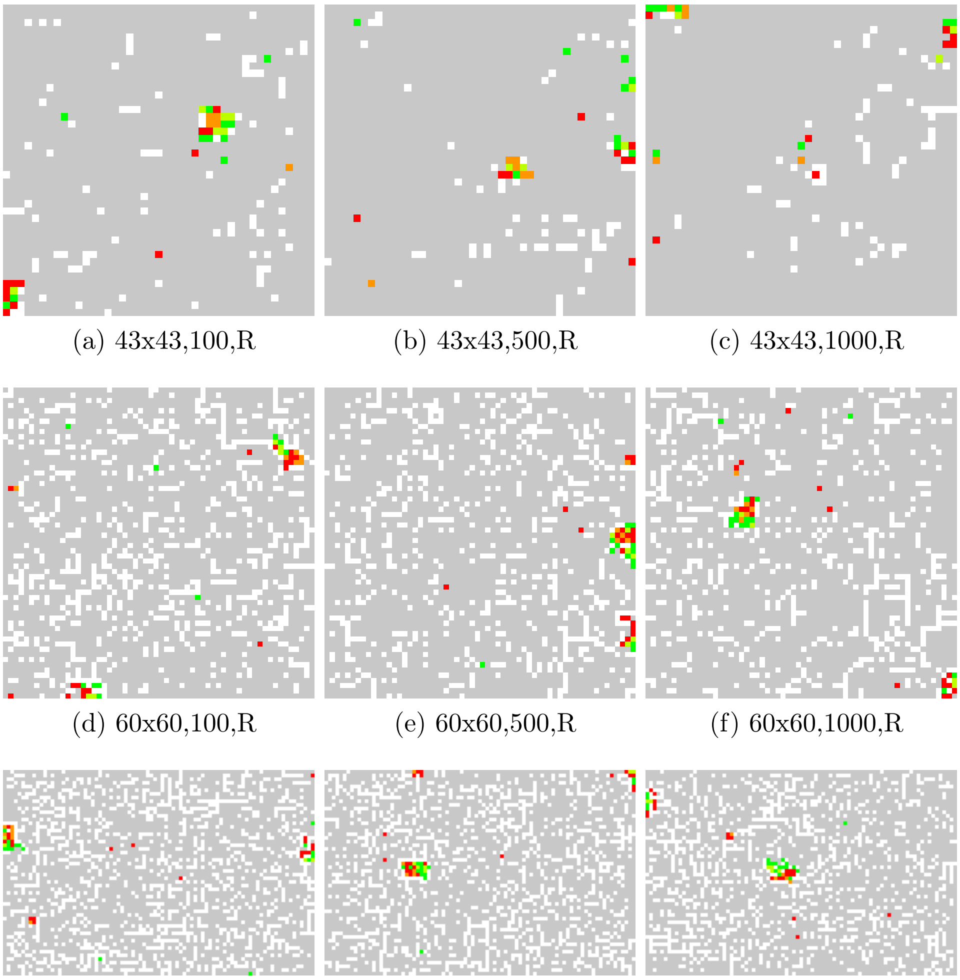
Table A.3: Purchased compounds.

Self-organizing maps



(a) 43x43,100,R
(b) 43x43,500,R
(c) 43x43,1000,R
(d) 61x61,100,R
(e) 61x61,500,R
(f) 61x61,1000,R
(g) 43x43,100,T
(h) 61x61,100,T

Figure A.1: Self-organizing maps of compounds with known activities trained together with a subset of 7418 compounds of the screening database. Maps are coloured according to the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, grey: screening database, white: empty.

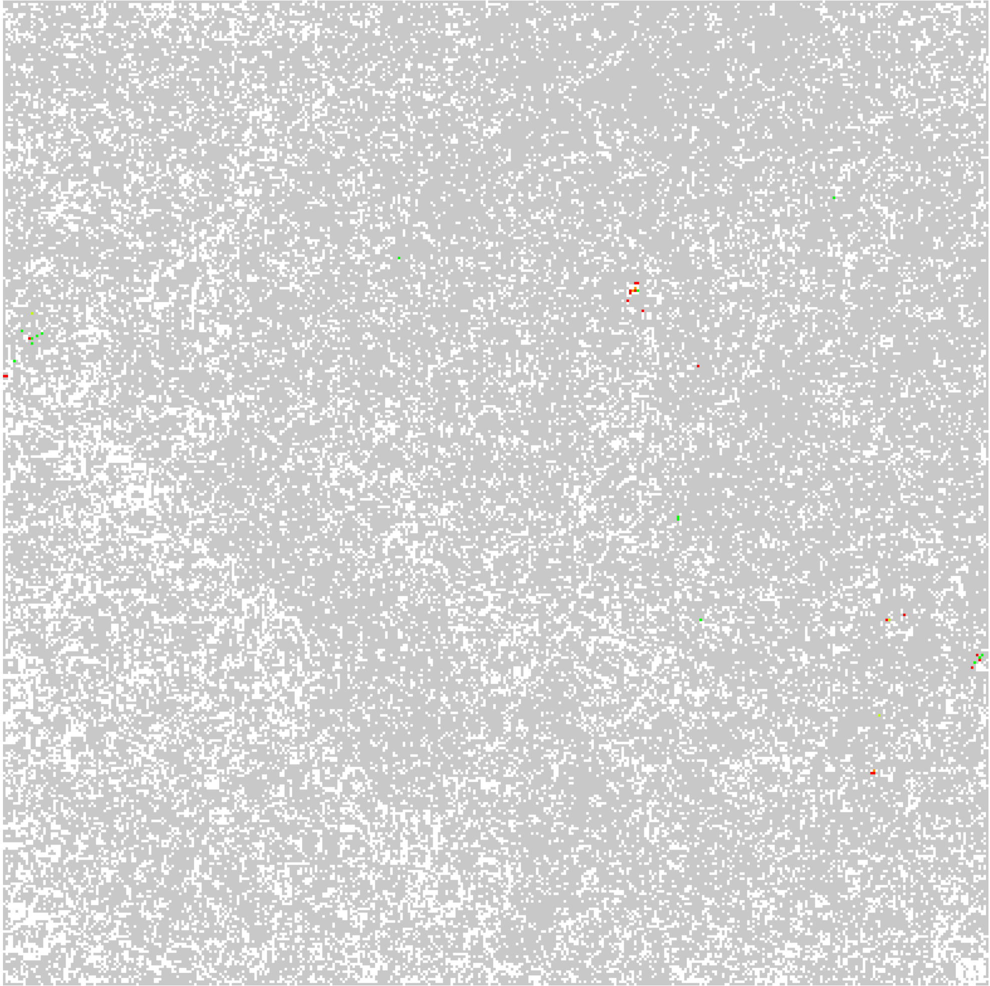


(a) 43x43,100,R (b) 43x43,500,R (c) 43x43,1000,R

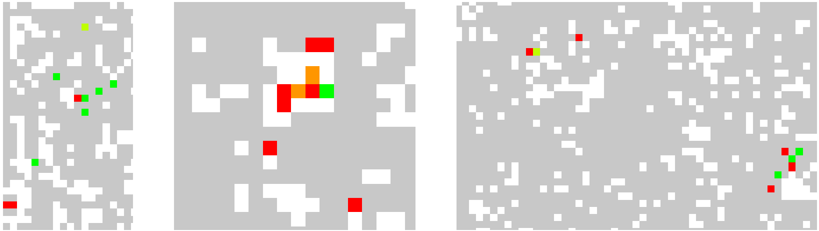
(d) 60x60,100,R (e) 60x60,500,R (f) 60x60,1000,R

(g) 85x56,100,R (h) 85x56,500,R (i) 85x56,1000,R

Figure A.2: Self-organizing maps of compounds with known activities trained together with a subset of 7227 compounds of the screening database. Maps are coloured according to the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, grey: screening database, white: empty.

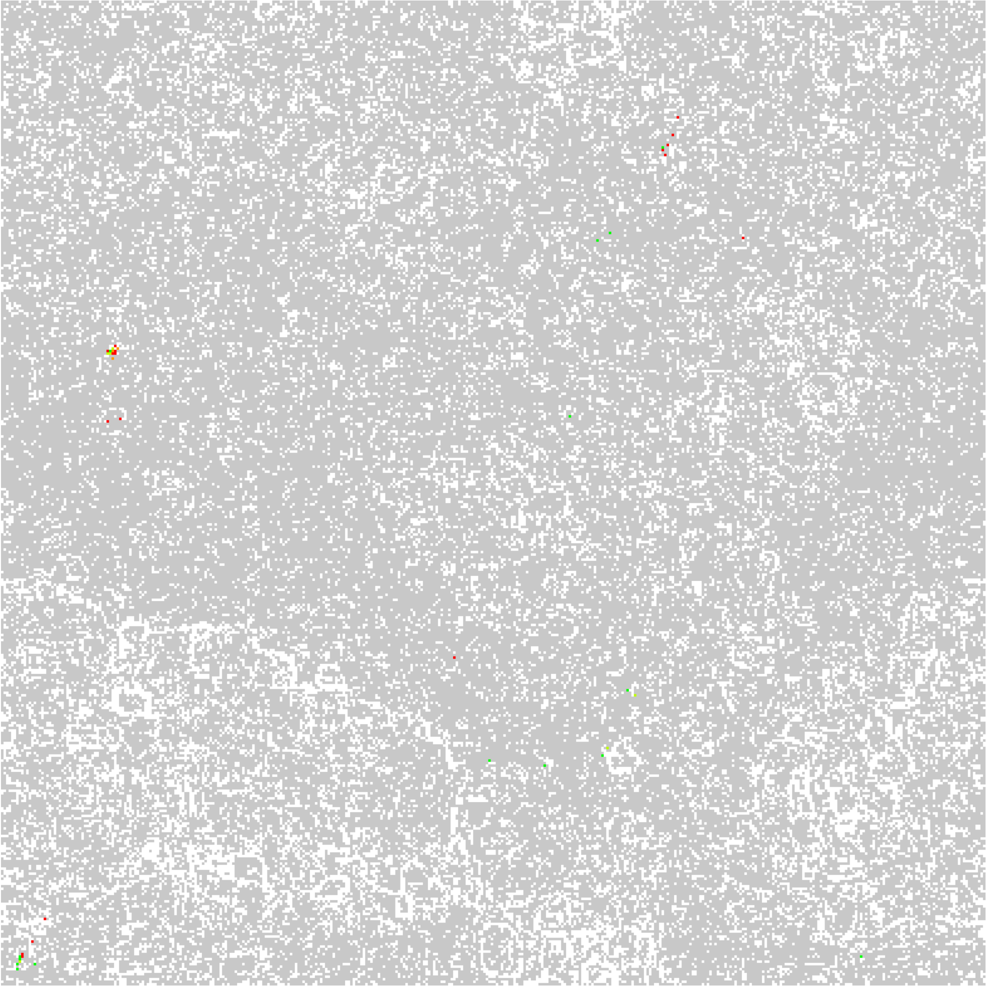


(a) 2d,392x392,100,R

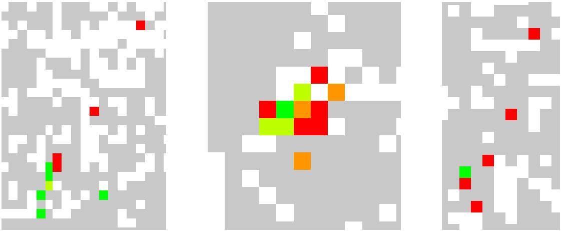


(b) details of 392x392,100,R

Figure A.3: Self-organizing map of compounds with known activities trained together with the screening database. Maps are coloured according to the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, grey: screening database, white: empty.

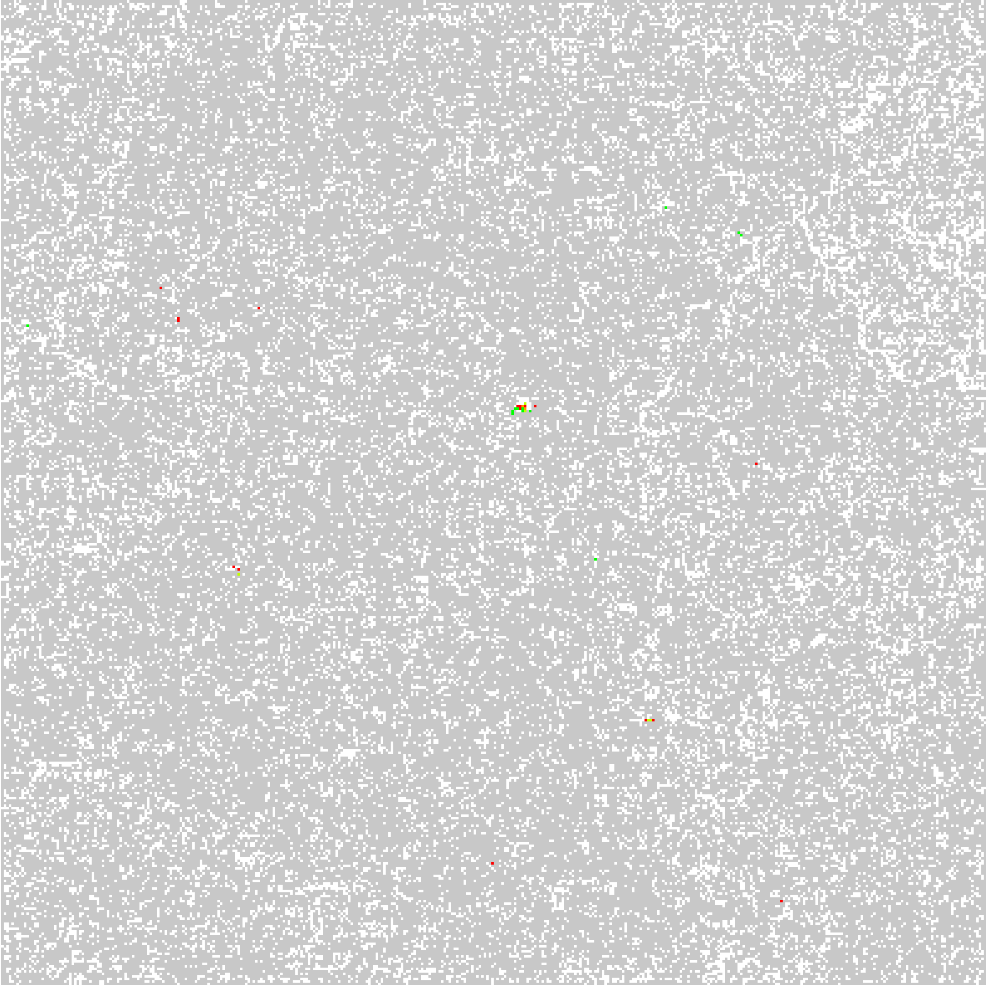


(a) 2d,392x392,100,T

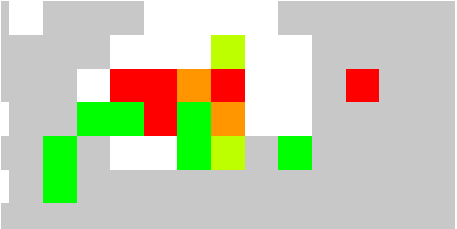


(b) details of 392x392,100,T

Figure A.4: Self-organizing map of compounds with known activities trained together with the screening database. Maps are coloured according to the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, grey: screening database, white: empty.

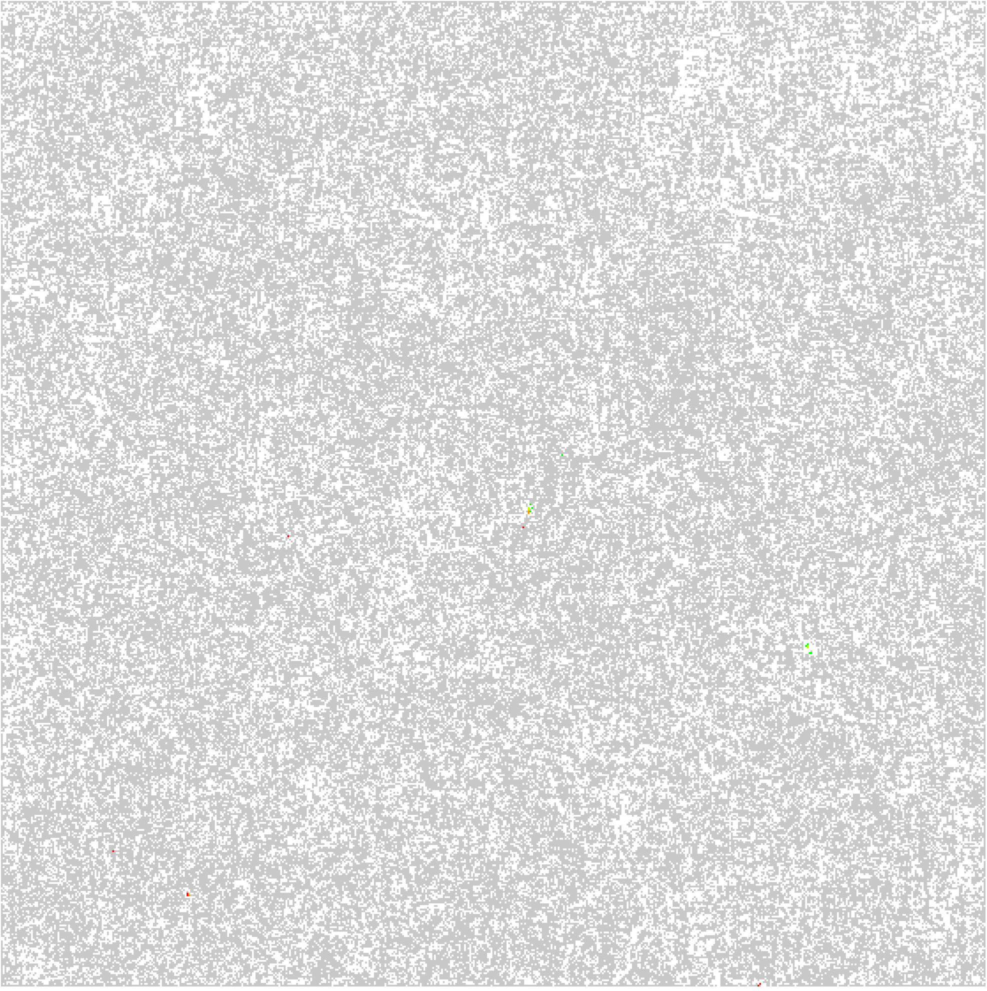


(a) vsa,392x392,100,R

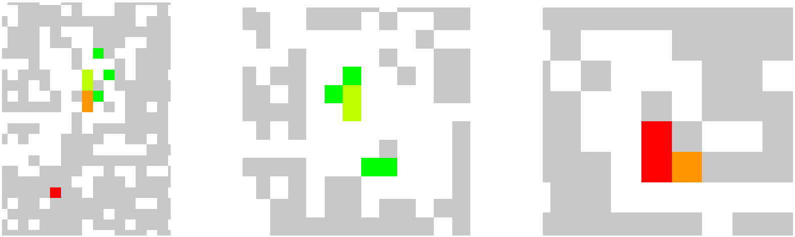


(b) detail of 392x392,100,R

Figure A.5: Self-organizing map of compounds with known activities trained together with the screening database. Maps are coloured according to the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, grey: screening database, white: empty.

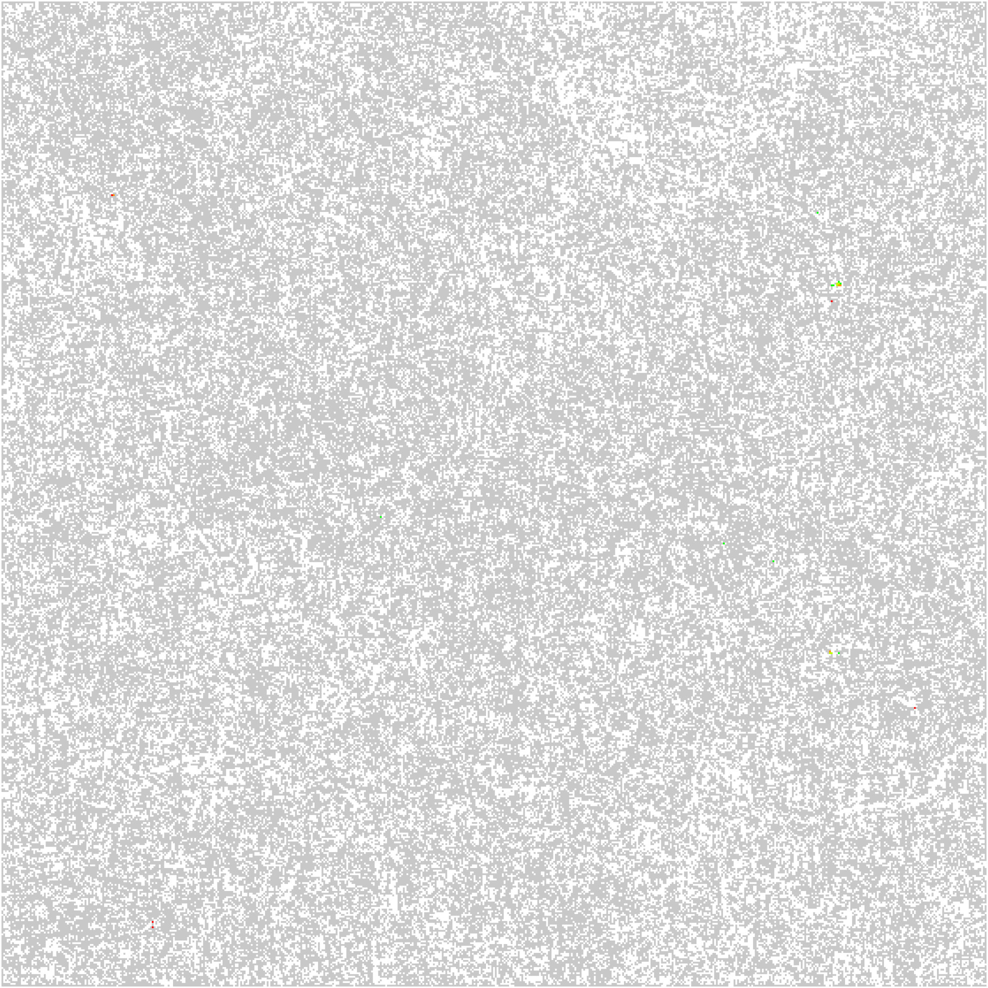


(a) vsa,557x557,50,R

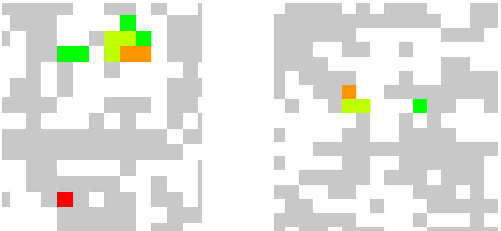


(b) details of 557x557,50,R

Figure A.6: Self-organizing map of compounds with known activities trained together with the screening database. Maps are coloured according to the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, grey: screening database, white: empty.



(a) vsa,557x557,100,R



(b) details of 557x557,100,R

Figure A.7: Self-organizing map of compounds with known activities trained together with the screening database. Maps are coloured according to the following scheme: red: inactive compounds only, orange: inactive neuron, light green: active neuron, green: active compounds only, grey: screening database, white: empty.

Abstract

The binding of insulin to the extracellular part of the insulin receptor is a key step in the insulin signalling pathway. Upon binding, the receptor is autophosphorylated and the intracellular tyrosine kinase is activated. In 1999, Zhang et al. published a small molecule identified from a fungal extract, which activates the human insulin receptor by binding directly to the intracellular domain of its beta-subunit. This compound (demethylasterriquinone B-1, DMAQ-B1) was shown to lower blood glucose levels in mouse models of type 2 diabetes mellitus. During the last years, structures and activities of approximately 100 derivatives of this compound have been published. Most of these structures contained a quinone substructure, which might cause toxic side effects. Since treatment of type 2 diabetes includes long-term administration of anti-diabetic compounds, it would be beneficial to find compounds with a different type of structure which activate the insulin receptor.

The aim of this dissertation was to build computational models which can be used to screen for new insulin-mimetic compounds and subsequent validation of the models by testing some of the obtained hits in relevant biological (i.e. cell-based) experiments. Three different ligand based computational methods, namely self-organizing maps, fingerprint similarity and shape similarity, have been used to screen a large vendor database for potential insulin receptor activating compounds. By testing 13 representative compounds from the identified scaffolds we found three compounds which are able to activate Akt kinase, an important downstream target of the activated insulin receptor.

One of the compounds increased glucose uptake in muscle cells. Derivatives of these compounds were further investigated to gain information on structure activity relationships. Additionally, the toxicity of the compounds in cells was assessed to show that the insulin-mimetic activity of our identified molecules is not correlated with toxic effects.

Zusammenfassung

Die Interaktion von Insulin mit dem extrazellulären Teil des Insulinrezeptors ist ein entscheidender Schritt des Insulin-Signalweges. Der Insulinrezeptor wird daraufhin auto-phosphoryliert und die intrazelluläre Tyrosinkinasedomäne wird aktiviert. Im Jahr 1999 publizierten Zhang et al. einen Wirkstoff der in einem Pilzextrakt gefunden wurde und den humanen Insulinrezeptor aktivieren kann, indem er direkt mit der intrazellulären Domäne der beta-Subeinheit interagiert. Diese Substanz (Demethylasterriquinone B-1, DMAQ-B1) ist in der Lage den Blutzuckerspiegel in Mausmodellen für Typ-2 Diabetes zu senken. In den letzten Jahren wurden Strukturen und Aktivitätswerte zu ca. 100 Derivaten dieser Substanz publiziert. Die meisten dieser Verbindungen enthalten eine Quinon-Substruktur, die zu toxischen Nebenwirkungen führen könnte. Da die Behandlung von Typ-2 Diabetes die Langzeittherapie mit Antidiabetes-Medikamenten beinhaltet, wäre es vorteilhaft, Insulinrezeptor aktivierende Wirkstoffe aus einer anderen Strukturklasse zu finden.

Das Ziel dieser Dissertation war die Entwicklung von Computermodellen, die zur Identifizierung von neuen, Insulin- imitierenden Wirkstoffen führen können, sowie die anschließende Validierung der Modelle in biologischen (zellbasierten) Experimenten. Drei unterschiedliche ligandenbasierte Methoden, nämlich Self-organizing Maps, Fingerprint- sowie Shape-ähnlichkeit, wurden verwendet um in einer großen kommerziellen Datenbank nach potenziellen Insulinrezeptor aktivierenden Wirkstoffen zu suchen. Durch die Testung von 13 repräsentativen Verbindungen der identifizierten Substanzklassen konnten wir drei Strukturen identifizieren, die Akt, eine downstream Kinase des aktivierten Insulinrezeptors aktivierten.

Eine dieser Substanzen war in der Lage die Glukoseaufnahme in Muskelzellen zu verstärken. Derivate dieser Struktur wurden untersucht, um weiterführende Informationen über Struktur-Aktivitätsbeziehungen zu erhalten. Zusätzlich wurde die Zytotoxizität der Substanzen getestet, um zu zeigen, dass die Insulin imitierende Aktivität der identifizierten Moleküle nicht mit toxischen Effekten korreliert.

Curriculum Vitae

Personal data

Daniela Digles
Date of birth: 13.09.1984 in Vienna, Austria.

Education

since 10/2007	Doctoral studies of natural science/molecular biology University of Vienna, Austria. Thesis: “Combined in silico/in vitro screening tools for identification of new insulin receptor ligands” supervised by Univ.-Prof. Dr. Gerhard F. Ecker and Univ.-Prof. Dr. Verena M. Dirsch.
09/2003–09/2007	University of Applied Sciences for Biotechnology FH-Campus Wien, Vienna, Austria. Specialisation: Chemistry of active substances. Thesis: “Structural optimization of biaryl compounds regarding the growth inhibition of human endothelial cells” supervised by Ao. Univ.-Prof. Dr. Thomas Erker and Univ.-Prof. Dr. Verena M. Dirsch. Diploma examination passed with excellence. Earned degree: Dipl.-Ing. (FH)
09/1995–06/2003	High school Auf der Schmelz, Vienna, Austria. Final examination (Matura) passed with excellence.

Awards/Fellowships

03/2009–02/2011	DOC-fFORTE-fellowship (Austrian Academy of Sciences).
04/2009	3rd place of the poster competition at the 21st Scientific Congress of the Austrian Pharmaceutical Society.
2005/2006	Merit grant (FH-Campus Wien).

Teaching

2008: Rational Drug Design Laboratory (tutor)
FH-Campus Wien.

since 2008: Arzneistoffanalytik (drug analytics practical course)
University of Vienna.

since 2009: Computational Drug Design (lecture and exercises)
University of Vienna.

Languages

German (mother tongue), English, French

Publications

Papers

D. Digles and G. F. Ecker, Self-Organizing Maps for In Silico Screening and Data Visualization. *Mol Inf* **2011**; 30(10): 838–846, doi: 10.1002/minf.201100082.

Conference contributions

D. Digles, V. Dirsch and G. F. Ecker, Classification of insulin receptor activators with self-organizing maps. Poster presentation at the XXth International Symposium on Medicinal Chemistry, Aug. **2008**, Vienna. Abstract in: *Drugs Fut.* 33(Suppl A):134 (2008).

D. Digles, V. Dirsch and G. F. Ecker, Combined in silico/in vitro screening tools for identification of new insulin receptor ligands. Poster presentation at the StipendiatInnenwochenende of the Austrian Academy of Sciences, Jan. **2009**, Vienna.

D. Digles, V. M. Dirsch and G. F. Ecker, Ligand-based screening tools for insulin receptor activating compounds. Poster presentation at the 21st Scientific Congress of the Austrian Pharmaceutical Society, Apr. **2009**, Vienna. Abstract in: *Sci Pharm* 77:202 (2009).

D. Digles, V. M. Dirsch and G. F. Ecker, In silico screening for insulin receptor activating compounds. Poster presentation at the 7th European Workshop in Drug Design, May **2009**, Siena.

D. Digles, V. M. Dirsch and G. F. Ecker, Docking studies on the insulin receptor. Poster presentation at the Joint Meeting on Medicinal Chemistry, Jun. **2009**, Budapest.

D. Digles, V. M. Dirsch and G. F. Ecker, Ligand based screening tools for identification of insulin receptor activating compounds. Posterpresentation at the 239th American Chemical Society National Meeting & Exposition, Mar. **2010**, San Francisco.

D. Digles, Identification of new insulin mimetic compounds using computational methods. Posterpresentation at the European School on Medicinal Chemistry, Jul. **2011**, Urbino.

D. Digles, E. H. Heiss, V. M. Dirsch and G. F. Ecker, Ligand based screening for insulin mimetic compounds. Posterpresentation at the Joint Meeting of the Austrian and German Pharmaceutical Societies, Sep. **2011**, Innsbruck.

Vienna, December 12, 2011