Realia et naturalia

DISSERTATIONES
CHIMICAE
UNIVERSITATIS
TARTUENSIS
149

GIRINATH G. PILLAI

Computational Modelling of Diverse Chemical, Biochemical and Biomedical Properties

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS
149

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS
149

GIRINATH G. PILLAI

Computational Modelling of Diverse Chemical, Biochemical and Biomedical Properties

UNIVERSITY OF TARTU
PRESS
1632

Institute of Chemistry, Faculty of Science and Technology, University of Tartu, Estonia

This Dissertation is accepted for the commencement of the Degree of Doctor of Philosophy in Chemistry on June 18, 2015 by the Doctoral Committee of the Institute of Chemistry, University of Tartu.

Supervisors: Prof. Mati Karelson, Institute of Chemistry, University of Tartu, Tartu, Estonia

Ph.D. Kaido Tämm, Institute of Chemistry, University of Tartu, Tartu, Estonia

The Late Prof. Alan Roy Katritzky (Deceased 2014) University of Florida, Gainesville, USA

Opponent: Dr. Eric F. V. Scriven, Department of Chemistry University of Florida, USA

Commencement: October 07, 2015 at 12:00, Ravila 14a (Chemicum), Room 1021.

This work has been partially supported by Graduate School "Functional materials and technologies" receiving funding from the European Social Fund under project 1.2.0401.09-0079 in University of Tartu, Estonia and by Florida Center for Heterocyclic Compounds, Kenan Foundation, University of Florida, Gainesville, United States of America.

ISSN 1406-0299
ISBN 978-9949-32-919-9 (print)
ISBN 978-9949-32-920-5 (pdf)

University of Tartu Press
www.tyk.ee

To Professor Alan Roy Katritzky,
(1928–2014)

TABLE OF CONTENTS

LIST OF PUBLICATIONS

The presented thesis is based on the five articles listed below.

I. Biswas, S.; Kayaleh, R.; **Pillai, G. G**.; Seon, C.; Roberts, I.; Popov, V.; Alamry, K. A.; Katritzky, A. R. Long-Range Chemical Ligation from N→N Acyl Migrations in Tryptophan Peptides via Cyclic Transition States of 10- to 18-Members. *Chem. Euro. J.,* **2014**, *20* (26), 8189–8198.

II. Oliferenko, P. V.; Oliferenko, A. A.; Poda, G. I.; Osolodkin, D. I.; **Pillai, G. G**.; Bernier, U. R.; Tsikolia, M.; Agramonte, N. M.; Clark, G. G.; Linthicum, K. J.; Katritzky, A. R. Promising *Aedes aegypti* Repellent Chemotypes Identified through Integrated QSAR, Virtual Screening, Synthesis, and Bioassay. *PLoS One* **2013**, *8*, e64547.

III. Jabeen, F.; Oliferenko, P. V.; Oliferenko, A. A.; **Pillai, G. G**.; Ansari, F. L.; Hall, C. D.; Katritzky, A. R. Dual inhibition of the α-glucosidase and butyrylcholinesterase studied by Molecular Field Topology Analysis. *Eur. J. Med. Chem.* **2014**, *10*(80), 228-242

IV. **Pillai, G. G**.; Sikk, L.; Tamm, T.; Karelson, M.; Burk, P.; Tämm, K. Theoretical Modeling of HPV: QSAR and Novodesign with Fragment Approach. *Curr. Comput. Aided. Drug Des.* **2014**, *10* (4), 303–314.

V. Berhanu, W. M.; **Pillai, G. G**.; Oliferenko, A. A.; Katritzky, A. R. Quantitative Structure-Activity/Property Relationships: The Ubiquitous Links between Cause and Effect. *ChemPlusChem* **2012**, *77*, 507–517.

Author's contribution

Publication I: The author is responsible for the data preparation, calculations, and interpretation of the results including preparation of QSAR section in the manuscript.

Publication II: The author is responsible for the data preparation, QSAR & Docking analysis, result data analysis and reviewed parts of the manuscript.

Publication III: The author is responsible for the data sets, calculations, and preparation of link between Diabetes and Alzheimer's disease part in the manuscript.

Publication IV: The author is responsible for project methodology, data sets, calculations and preparation of the manuscript.

Publication V: The author is responsible for the preparation of QSAR limitations, Similarity Analysis, Fragment based approaches in the manuscript.

LIST OF ABBREVIATIONS

AChe	AcetylCholinesterase
ACh	AcetylCholine
AIDS	Acquired Immune Deficiency Syndrome
AM1	Austin Model 1
ANN	Artificial Neural Network
BChE	ButyrylCholinEsterase
BMLR	Best Multiple Linear Regression
CDK	Chemical Development Kit
CPSA	Charged Partial Surface Area
CODESSA	COmprehensive DEscriptors for Structural and Statistical Analysis
CoMFA	Comparative Molecular Field Analysis
EC_{50}	Effective Concentration at 50% value
ECHA	European Chemical Agency
EPA	Environmental Protection Agency
ESP	Electrostatic Potential
F	Fischer criterion
FBDD	Fragment Based Drug Design
FDA	Food and Drug Administration
GA	Genetic Algorithm
HAT	Measure of leverage
HIV-RT	Human Immunodeficiency Virus Reverse Transcriptase
HPV	Human Papilloma Virus
HTS	High Throughput Screening
IC_{50}	Inhibition Constant at 50% value
K_x	Total correlation in the model predictors (K = multivariate correlation index)
K_y	Total correlation in Response
LBDD	Ligand Based Designing
LMO	Leave-Many-Out cross-validation
LOO	Leave-One-Out cross-validation
MED	Minimum Effective Dosage
MFTA	Molecular Field Topological Analysis
MLR	Multiple Linear Regression
MM	Molecular Mechanics
MOPAC	Molecular Orbital PACkage
MSg	Molecular Supergraph
NCL	Native Chemical Ligation
NNRTI	Non-Nucleoside Reverse Transcriptase Inhibitors
OPLS	Optimized Potential for Ligand Simulations
PCA	Principal Component Analysis
PDB	Protein Data Bank

PEOE	Partial Equalization of Orbital Electronegativities
PETA	People for the Ethical Treatment of Animals
PLS	Partial Least Squares
PM3/6	Parameterized Model number 3/6
PRESS	Prediction Sum of Squares
Q^2	Estimation of the criterion R^2 obtained by cross-validation
QSAR	Quantitative Structure – Activity Relationship(s)
QSLR	Quantitative Structure – Ligation Relationship
R^2	coefficient of determination
R^2_{cv}	cross validation coefficient of determination
RMSD	Root-mean Squared Deviation
RMSE	Root-mean Squared Error
RMSPE	Root-mean Squared Error of Prediction
S or sE	Standard error of multiple linear regression
SBDD	Structure Based Drug Designing
SMILES	Simplified Molecular Input Line Entry Specification
t	student's test
T2DM	Type 2 Diabetes Mellitus

1. INTRODUCTION

Drug discovery is an empirical field of science, which identifies a molecule that can be a prescribed medicine to create simplified and reproducible biological solution. Nowadays, the initial step for identifying a molecule as a potential drug candidate is to create a simplified computational model for prediction of biological activities and pharmacophoric properties.[1]

Computer aided drug design (CADD) methods are mostly categorized into (i) ligand-based (LBDD) and (ii) structure-based (SBDD) methods. Ligand-based methods generally emphasize on comparative analysis of the structural similarity or diverse pharmacophore descriptors of experimentally known active ligands. The method does not rely on the protein target structural information, therefore a knowledge of experimentally characterized active compounds is important to the success of ligand-based methods.[2] On the contrary, SBDD methods do not exclusively depend on experimentally active compounds, but rather identify new molecules that are corresponding to the protein active site. Molecular docking, uses several binding pocket identification algorithms to predict the binding mode and the affinity of a given compound towards a target receptor, forms the basic outline in receptor-based virtual screening procedures and in lead discovery approaches. This allows to considerable savings in resources and material costs as only a small number of molecules of the complete library need to be tested experimentally. SBDD has demonstrated to be more effective in understanding the molecular basis of a disease and utilizes 3D structural data of the biological target. This ensures the reliability of proposing new drug chemical more rapidly and cost-efficiently.[3] Ligand based *de-novo* design approaches do not require the receptor information but exclusively depend on the known active ligands serving as a reference to generate a novel chemical entity. The molecules generated by SBDD or LBDD often challenge the synthetic feasibility. This problem has limited the success of *de-novo* packages, as only a small percentage of molecules are synthesizable with reasonable time and cost.[4] Many of the recently developed *de-novo* tools address this problem by employing fragment-based drug design methods (FBDD) and using linking rules to guide the assembly of building blocks or clusters. Most of the ligand-based and fragment-based methods need the reference fragment structure to initiate the design of novel molecule.[5] Novel molecules are generated by the building blocks from data base. These building blocks can be curated from drug-like molecules with a set of reaction linking rules and retro-synthetic pathways.[6]

As mentioned above, ligand-based methods are suitable for challenging biological problems, and do not require protein or receptor target information. These methods are mainly focused on developing 3D QSAR or pharmacophore based models of active, moderately active and inactive molecules by detecting their similar or diverse molecular and pharmacophoric features. Recent trend in QSAR shows an increased demand for consensus models combining the

predictive power of multiple individual approaches: e.g. the linear and non-linear QSAR utilizing different descriptor types is a common practice.[7] The schematic representation of different CADD approaches is shown in Figure **1**.

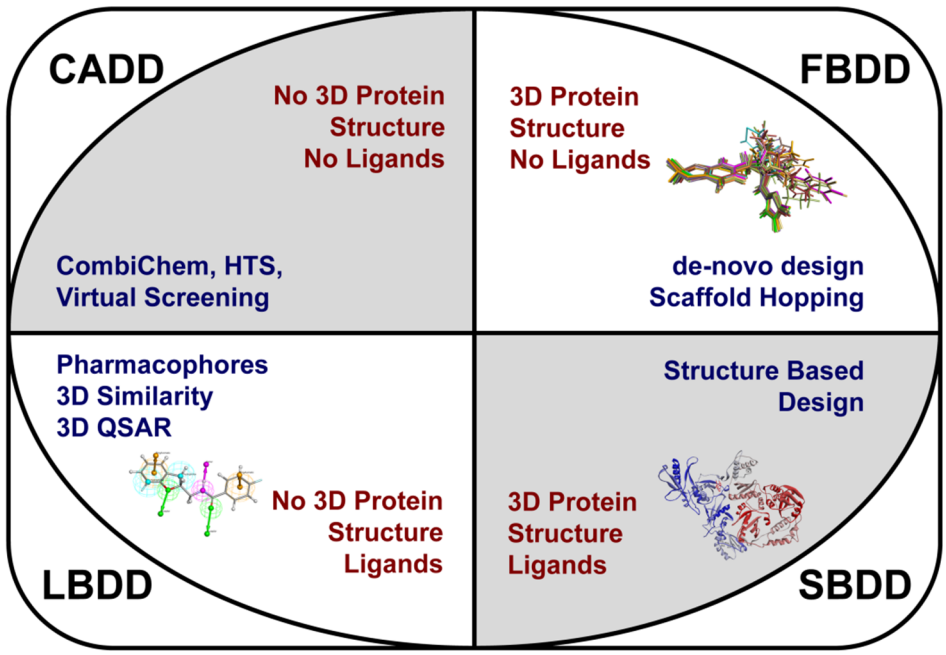

Figure 1. Schematic representation of computer aided drug design and virtual screening methods with reference to different scenarios of ligand and protein target availability.

This Ph.D. thesis provides an overview of the comprehensive and fragment based QSAR methodologies. It also summarizes work done on the chemical ligation, mosquito repellence, and modelling of dual inhibitors and HPV anti-viral agents.

2. AIMS OF THE STUDY

The main focus of this thesis is the development and implementation of the molecular modelling techniques and statistical modelling algorithms to process the biochemical and biomedical data in drug discovery. The thesis is organized into three main parts, (i) literature overview of chemical, biochemical and biomedical applications, (ii) computational methods and modelling techniques where one wants to obtain highly active molecules for a given experimental budget, and (iii) summary of the research findings. Specific highlights of this thesis include:

I. Developing for the first time, a QSLR statistical model for the prediction of relative abundance in chemical ligation from N→N Acyl migrations of tryptophan peptides and its chemical synthesis *(Article I)*.
II. Proposing and validating the hit expansion approach to identify diverse mosquito (*Aedes aegypti)* repellent chemotypes using virtual screening, QSAR and experimental approach *(Article II)*.
III. Evaluating the dual inhibition activity on diverse pharmacological properties and validating the predictions of link between Type 2 Diabetes Mellitus and Alzheimer's disease using Molecular Field Topology Analysis *(Article III)*.
IV. Designing of novel antiviral agents for Human Papilloma Virus (Type 6) inhibitors using customized fragment based QSAR approach *(Article IV)*.
V. Assessment and overview of QSAR in various areas of research *(Article V)*.

3. LITERATURE OVERVIEW

3.1 Native Chemical Ligation

Native chemical ligation (NCL), is the process of convergent synthesis of peptides which was first reported by Wieland in 1953[8] and further developed by Kent *et. al*.[9] NCL is the most widely used form of chemical ligation involving a chemo-selective reaction, usually in aqueous solution. A region-selective reaction of a thioester mediated covalent linking of unprotected peptide sub-divisions at a cysteine residue of an adjacent peptide gives a native amide bond at the ligation site over a rapid S-N acyl transfer *via* a cyclic transition state (TS).[10–12] NCL has overcome the limitations of classical synthetic organic chemistry into the total synthesis of proteins, and enables the routine total or semi-synthesis of protein molecules.[13] NCL process has contributed to build biologically active molecules as potential therapeutics in the synthesis of the cancer protein NY-ESO-1,[14] cytochrome b562,[15] dendrimers, and monodisperse macromolecules.[16] The major challenge was to control the intrinsic dual reactivity of bifunctional Cys-peptide-thioester because of low abundance of Cys and steric hindrance. This problem was overcome by developing thiol auxiliary groups.[17–19] Therefore, an improved new ligation method which performed the reversibility of the first step, the thiol(ate)–thioester exchange reaction was developed. Due to irreversibility, high yields of the final ligation product was obtained, even in the presence of internal Cys residues, under the reaction conditions of the second (S-to-N acyl shift) amide-forming step.[20] The intramolecular N→N acyl migration of Z-alanine to the N terminus to form native peptide is shown in Figure **2**. To rationalize and predict the relative abundance for native chemical ligation for the first time, full conformational analysis and statistical modelling is required to reduce the cost of trials in synthesis.

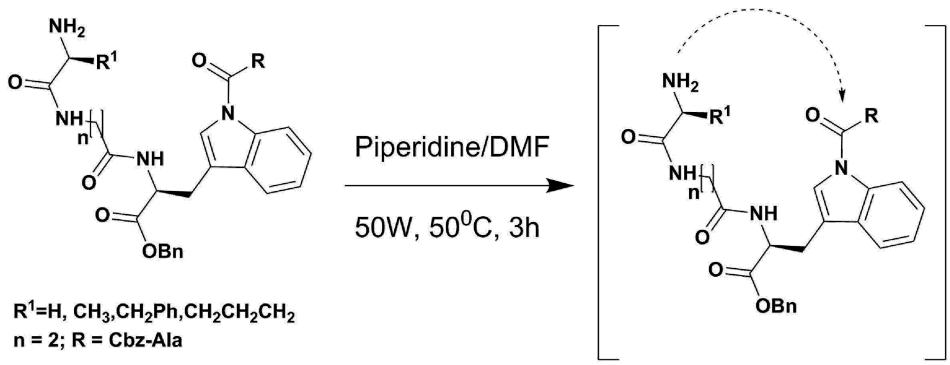

Figure 2. Scheme on chemical ligation of N-acyl isopeptides through 14-membered transition states.

3.2 Mosquito Repellents

Natural resources, such as lemon eucalyptus oil, lavender, cinnamon oil, thyme oil, Greek catnip oil, soybean oil, basil, gum, and aroma plant based smoke, have been used for years as mosquito repellents and are still utilized today throughout the tropical region.[21] We still need more effective, non-toxic to humans, long-lasting and water-resistant repellents because of more than one million cases of malaria and yellow fever are reported per annum in South Africa, India and Southern Americas. The most effective wide spectrum synthetic repellent is N,N-diethyl-3-methylbenzamide (DEET) discovered in 1952.[22] Although DEET is considered as the standard for insect repellents, it has drawbacks: (i) limited efficacy against the species *Anopheles albimanus*,[23] less tolerant on variants of *Aedes aegypti*,[24] and other vectors[25] (ii) skin irritation; (iii) possible neurotoxic effect;[26] and (iv) high cost. Other repellents such as the piperidine derivatives KBR 3023 (picaridin) and AI3-37220 are considered almost as effective as DEET, and will remain effective for a longer duration. The repellent diethyl phenyl acetamide (DEPA) is also as effective as DEET and can be produced at about half the cost of DEET. The ethyl ester of 3-[N-butyl-N-acetyl]-aminopropionic acid (IR3535), has few severe side effects but is less effective than DEET since its development in 1975.[27–29] Currently, identification of chemotypes of effective mosquito repellents with few severe side effects is necessary for the affected population in tropical regions. Computer aided molecular design provides relief to the identification of novel repellents.[30]

3.3 Link between T2DM and AD – Dual Inhibition

α-Glucosidase is a carbohydrase enzyme which catalyzes the release of α-D-glucopyranose located in the striated border of the small intestine by acting upon 1,4-α bonds.[31–33] The inhibition of its catalytic activity leads to the hindrance of glucose absorption and a decrease in postprandial blood glucose level leading to type 2 diabetes mellitus (T2DM).[34] Recently it was found that acarbose is efficient in patients with impaired glucose tolerance and could prevent or delay the development T2DM.[35–37]

Acetylcholinesterase (AChE), and butyrylcholinesterase (BChE) belong to the class of cholinesterases, that hydrolyze neurotransmitter acetylcholine (ACh) within cholinergic synapses of the brain and nervous system**.** BChE is considerably less active in ACh hydrolysis than AChE at low concentrations of the substrate and at the same time it is highly efficient at higher levels of ACh, when AChE becomes substrate-inhibited.[38] Suppression of the cholinergic transmission in synapses results in severe neuro-degenerative disorders such as Alzheimer's disease (AD). AD is considered as a loss of neurons caused by the formation of β-amyloid plaques and neurofibrillary tangles in brain nerve cells. Simultaneous depletion of AChE and some increase of BChE activity shifts the balance of ACh regulation.[39] Based on this observation, AChE inhibitors

reducing the rate of ACh cleavage such as tacrine, rivastigmine, galantamine, and donepezil have traditionally been used for symptomatic treatment of AD.[40]

According to the epidemiological and pathogenic studies, patients with T2DM have a higher tendency of dementia and AD and *vice versa*. Over the last decade research has attempted to understand the mechanisms of AD and T2DM. Thus, experimental evidence was found that the impairment of insulin might be a mechanistic link between both conditions since insulin (and leptin) have been shown to regulate neuronal and synaptic functions in brain.[41–45] At the same time, BChE may be involved in parthenogenesis of T2D through suppression of amyloid formation.[38,46] For multi target drug discovery, it is important to recognize the link between the T2DM and AD dual inhibition in order to avoid the off-target mode of action.[47] Comparative analysis is the necessity to evaluate the dual inhibition using pharmacophore and QSAR modelling approaches.

3.4 Human Papilloma Virus Inhibitors

High risk Human Papilloma Viruses (HPV) types 16 and 18 are the most common sexually transmitted carcinogenic infections.[48] HPVs preferentially occur in a latent life cycle, and wide variety of different types can be detected at random sites of healthy skin of humans.[49] The viruses infect and replicate in the cutaneous or mucosal epithelia. HPV type 6-E1 helicase ATPase is also responsible for the majority of genital warts. Antiviral agents inhibiting HPV replication could play a vital role in the treatment of the disease, but there are no effective agents present at this time.[50] Recent progress towards the discovery and characterization of specific molecular targets affords prospectus for efficient HPV antiviral compounds.[51] QSARs and other molecular modeling tools are widely used for discovery of novel and potentially active compounds against HPV.[52]

4. COMPUTATIONAL METHODS

Computer aided molecular and drug design as a process rarely occurs in one step. In addition to the deployment of computational methods, the data also needs to be prepared, analysed and validated. The following section of the thesis takes a closer look on the steps used and work done in individual articles. In Article I, methods like native chemical ligation with pre-organized conformational analysis and QSLR were employed. Article II & III employ ligand-based approaches by following MFTA, QSAR similarity search, and docking. Article IV presents newly developed fragment-based drug design (reverse) QSAR methodology.

4.1 Molecular Field Topology Analysis

Molecular Field Topology Analysis (MFTA) is a method for the analysis of structurally similar chemical compounds that is similar to the Comparative molecular field analysis (CoMFA) in 3D space.[53] MFTA does a structural alignment in two-dimensional grid and 2D molecular graphs are superimposed to make "molecular supergraph" (MSg) as shown in Figure **3.**[54] The MSg vertices and edges corresponding to atoms and bonds are characterized with values of local atomic descriptors. These form a rectangular atom descriptor matrix, which is processed by the PLS (Partial Least Square) method to link chemical or biological activity to molecular structure. PLS reduces the dimensionality of the descriptor matrix down to few sensible factors. Therefore using the number of factors (NF) in PLS is more common than usage of descriptors. In MFTA, a factor is presented as a linear combination over all selected descriptors. The basic MFTA descriptor space includes: atomic charges, van der Waals radii, electronegativity, hydrogen bond parameters, and lipophilicity. The quality of the prediction of a model is characterized by the statistical parameters such as squared correlation coefficient, R^2, and the cross-validation coefficient $Q^2(n)$, where n is a user-defined parameter for the number of structures in each leave-many-out (LMO) cross-validation procedure. MFTA has been successfully applied to several medicinal chemistry problems such as (i) discovery of new CX chemokine receptor-4 antagonists,[55] (ii) modelling of anticholinesterase activity of o-phosphorylated oximes,[56] and (iii) design of GABAA receptor selective ligands.[57] Articles II and III of the present Thesis employ MFTA to analyze and understand the pharmacophore sites in structure activity relationship to identify new promising candidates.

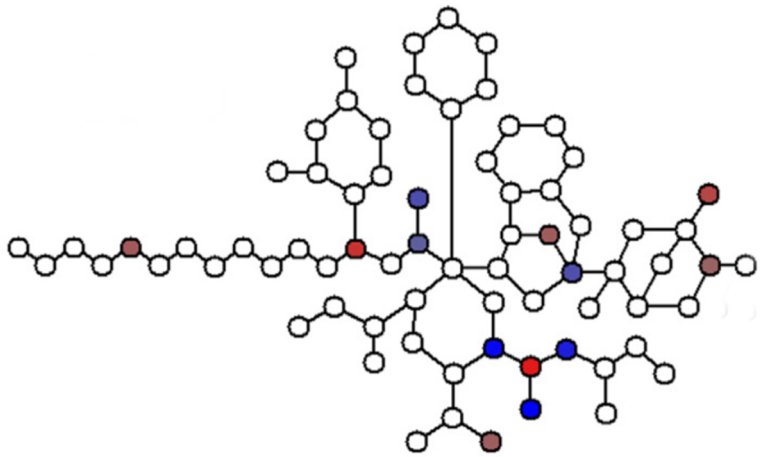

Figure 3. Molecular supergraph for MFTA model

4.2 Molecular Docking

Molecular docking is a widely used procedure in computer-aided drug design to explore and predict the predominant binding mode(s) of a ligand within a target 3D protein. The docking search methods examine interaction points in the binding pockets and utilize scoring functions like Dock Score, Glide Score, Chem Score, etc., to rank ligand dockings efficiencies.[58,59] Docking can be used to perform virtual screening on large chemical space, rank the docked poses, and recommend structural hypotheses on the mechanism of ligands inhibiting the protein target, which is crucial in "hit to lead" optimization. The input preparation of both protein and ligand structures for the docking is as important as the docking search algorithm parameters, and interpretation of the results can sometimes be ambiguous.[60] Molecular docking studies are sometimes used in QSAR to generate the conformers of the ligands (inhibitors/molecules) within the protein binding site to generate 3D and 4D molecular descriptors (using frozen conformer in semiempirical parameterization).[61] In this approach, AutoDock[62] and Glide[63] programs have been used for molecular docking and virtual screening studies. A protein-ligand interaction in 3D and 2D depiction is shown in Figure **4**. In article III of this Thesis, the molecular docking studies helped to understand the mode of action of repellents with odorant binding protein of *Aedes aegypti*. This led to the identification of new chemotypes through the virtual screening process.

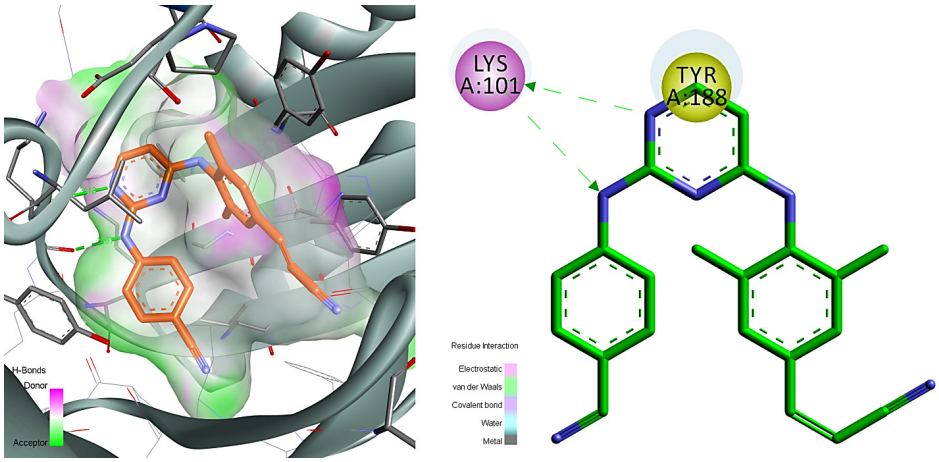

Figure 4. For example: Ligand Rilviparine binds to the pocket of HIV-1 reverse transcriptase protein target PDBID : 3MEE a) 3D representation of protein-ligand complex. b) 2D depiction of protein-ligand interaction with key amino acids taking part in hydrogen bond.

4.3 2D Similarity Search

Similarity search is a method to find chemical structures that are similar to the reference structure. The similarity is measured by comparing the molecular features (molecular descriptors) or fingerprints of chemical structures. The use of molecular fingerprints for chemical similarity search has made the examination of large databases much easier by encoding 2D sub-structural fragments in a molecule (hashed fingerprints, and binary fingerprints).[64] In the similarity search, the compounds are ranked by different metric functions and weightings such as Tanimoto, Euclidean, Tversky, Substructure, and Superstructure.[65] These metric functions can also be combined to increase the effectiveness of finding similar structures. The similarity search has been extensively used for finding homogenous molecules and drug like structures for subsequent QSAR modelling.[66,67] In this thesis, Instant JChem[68] was used for 2D similarity search and Accelrys Discovery Studio[69] for 3D overlap analysis. The illustration of 2D chemical similarity search is given in Figure **5**. In articles I, II, III and IV, the chemical similarity search played a vital role to understand the diversity and closeness of molecular structures to be applicable for QSAR studies.

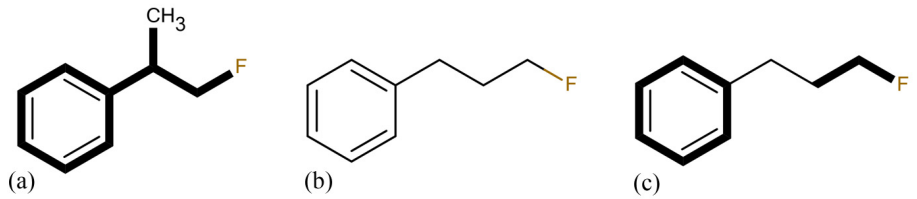

Figure 5. Similarity search: a) Query b) Target c) Search result with similarity score of 59.75%.

4.4 QSAR Modelling

The aim of QSAR is to correlate biological activities of chemical structures with the molecular descriptors, which are solely calculated from molecular structure. The process of constructing a QSAR model includes the following steps; i) selection of a data set; ii) generation of molecular structural data; iii) optimization of the 3D geometry by an appropriate method (to generate 3D/4D descriptors); iv) generation of various structural descriptors; v) application of variable selection or/and data reduction methods on the calculated descriptors; vi) regression analysis; and finally, vii) evaluation of the validity and predictability of the developed QSAR models using external datasets.[70]. QSAR modeling is a useful technique for accelerating development of drugs, agro and fine chemicals, materials, and toxicology predictions. The QSAR approach is under permanent scrutiny by the community to improve and enhance robustness by minimizing predictive errors and over-training.[71] The simplified QSAR approach is shown in Figure **6**.

In terms of methodology improvements, a new trend is to integrate QSAR with adjacent computational methods such as virtual screening and molecular dynamics to justify the predictive capacity of models with mechanism of action. Such synergy offers unique opportunities to overcome the limitation of modelling global QSAR models.[72,73]

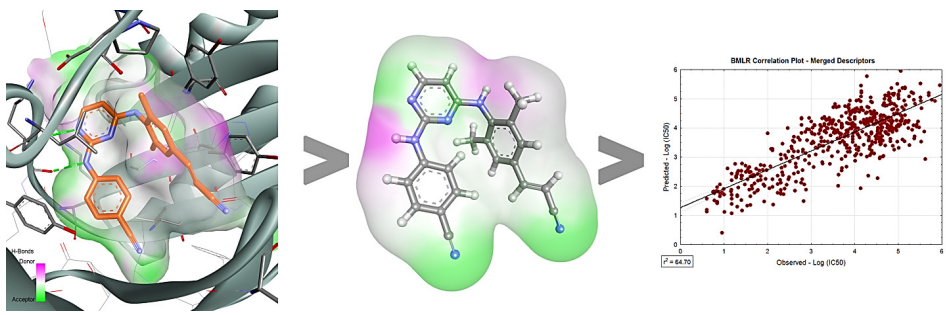

Figure 6. Generation of QSAR (3D) models a) Analysis of 3D conformer, b) Elucidation of molecular features to calculate descriptors c) Generate statistical models.

4.5 Molecular Descriptors

Molecular descriptors map the structure of the compound into a set of numerical or binary values representing various molecular features that are important for explaining the activity or property of the molecule. The descriptors establish a link between the molecular structure and the corresponding activities.[74–76]

Molecular descriptors are mainly classified as

a) Constitutional descriptors: the most simple and frequently used class of descriptors, reflecting the chemical composition of a compound without any information about its molecular geometry or atom connectivity. Some constitutional descriptors are; molecular weight (MW), number of atoms (nAT), number of bonds (nb), number of rings (nr), number of Hydrogen atoms (nH), number of Carbon atoms (nC), number of Nitrogen atoms (nN), number of Oxygen atoms (nO), number of halogen atoms (nX).[77]
b) Topological descriptors: consider the topology of a molecule. These are 2D descriptors which consider the internal atomic arrangement of compounds, and encode molecular size, shape, branching, presence of heteroatoms and multiple bonds information in numerical form. Some topological descriptors are; Wiener index, Balaban's index, Kier and Hall valence connectivity indices, Structural information content index, Topological electronic indices.[78]
c) Geometrical descriptors: characterize the shape and extent of the molecule in terms of its 3D Cartesian coordinates. As a result, accurate coordinates are required and so the structure must be geometry optimized before these descriptors can be calculated. Currently for biological activity, 3D conformers of the target protein binding site are used for meaningful geometrical descriptor generation.[77,79]
d) Electronic (Charge) descriptors: calculated from atomic charges, which can be calculated using semi-empirical methods based on the 2D topological structure of the molecule or a quantum chemical wave function of the molecule.[80–82]
e) Quantum chemical descriptors: describe electrostatic and electronic properties of a molecule. These descriptors are calculated using molecular orbital energies and wave functions of electronic motion in a molecule obtained by solving the respective time-independent Schrödinger equation. The semi-empirical AM1/PM3/PM6 parameterizations used in MOPAC/AMPAC programs are widely used to derive charges, dipole moments, and bond lengths. The computed quantum chemical descriptors include the partial atomic charges, HOMO and LUMO energy levels, dipole moment, polarizability, etc., as well as the derivative descriptors from them.[75,83]

The software reporting different molecular descriptors used for the studies in articles are listed in Table **1**. A detailed modelling schema for the elucidation of structural features to calculate molecular descriptors is shown in Figure **7**.

Criteria for Molecular Descriptors to be used in QSAR:

a) Physical significance and structural interpretation are very important.
b) Key molecular descriptors should have a correlation with the property/ activity.
c) 3D descriptors should discriminate isomers.
d) Co-linearity among the descriptors are trivial.

Table 1. List of software tools used for calculating molecular descriptors

Name	Developers	No. of Descriptors	Platform / License
CODESSA III	SemiChem Inc.	≈720	Win/Linux/Mac Commercial
CODESSA-Pro	Univ. of Florida / Univ. of Tartu	≈590	Windows Commercial
CDK GUI	Dr. Rajshri Guha	≈120	All Plaforms GPL, Freeware
EDRAGON	Virtual Computational Chemistry Laboratory	≈3000	All Platforms Online Server
PADEL	National University of Singapore	≈380	All Platforms GPL, Freeware
Indigo	GGA Software	≈50	All Platforms GPL, Opensource
RDKit	Greg Landrum	≈220	All Platforms GPL, Opensource

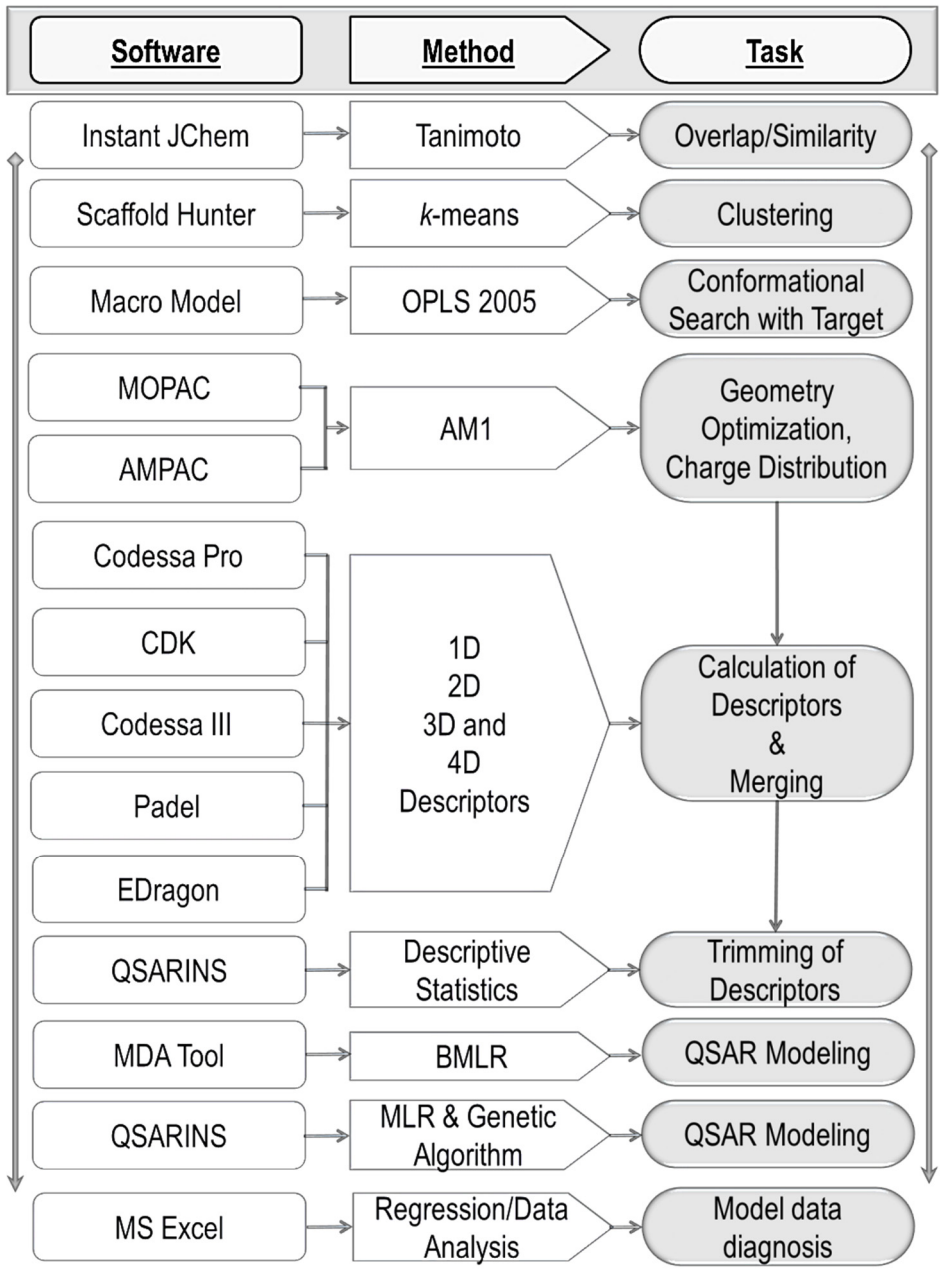

Figure 7. QSAR modelling schema for the elucidation of structural features to calculate molecular descriptors from different resources. (Chart style adapted from *J. Chem. Inf. Model., 2008, 48 (11), pp 2207–2213 and QSAR & Comb. Sci.,2009, 28, pp 811–814*)

4.6 Approaches in Data Treatment and Modelling

4.6.1 Data Source and Quality

The selection of compounds in a dataset is based on the molecular similarity search with known active drugs.[84] In QSAR the data clean-up plays important role as the quality of the structures and the respective biological data should be verified. The verification is required particularly, if the data is collected from different research groups. QSAR models can also be used to correct erroneous biological data associated with chemical compounds.[85–87]

4.6.2 Data Standardization

The data pre-processing step transforms the original dependent and independent variables into a new set of variables suitable for QSAR analysis.

i) Transformation of the dependent variable.
Biological data is often provided in units that are unsuitable for QSAR analysis. Furthermore, the collected experimental data (EC_{50}, LD_{50}, IC_{50}, MED, etc.) may not be normally distributed. In most cases the bioassay data is reported in different units like nm/ml, μg/L, etc., but for modelling purpose molar units are used. As the Pearson Product Moment Correlation (frequently used measure of model performance in QSAR) requires normally distributed data, a transformation of the original endpoint values is often necessary.[88]

ii) Transformation of the independent variables.
Being defined by unique mathematical expressions, all molecular descriptors certainly cover vastly different ranges of numerical values. For example, the molecular volume (expressed as $Å^3$) usually takes values in the range of a hundred to several thousand units, whereas the partial atomic charge of a C atom may vary from 0.010 to 0.199*e* units. If these two descriptors are used in a QSAR equation it would be extremely difficult to determine their relative impact on the modelled endpoint. Hence, it is desirable to use normalization or standardization procedures to bring all descriptors in proportion with one another.

4.6.3 Data Modelling Techniques

A plethora of supervised and unsupervised data processing algorithms are widely used for data modelling. Although methods, such as SVM (Support Vector Machines), kNN (k-Nearest Neighbors), GA (Genetic Algorithms), DT (Decision Trees), RF (Random Forests), ANN (Artificial Neural Networks) have become increasingly popular during the past decade, classical methods like MLR (Multiple Linear Regression), PCA (Principle Component Analysis) and

PLS (Partial Least Squares) are still preferred due to their simplicity, predictive capacity and easy interpretability of the generated models.

a) BMLR

The Best Multi-Linear Regression method (BMLR) was used to correlate the descriptors with the activities. The BMLR method is based on the (i) selection of the orthogonal descriptor pairs, (ii) extension of the correlation (saved on the previous step) with the addition of new descriptors until the *F*-criteria becomes less than that of the best 2-parameter correlation.[70] The best *N* correlations (by R^2) are saved. The method successfully solves the initial selection problem by reducing the number of pairs of descriptors in the "starting set". The major limitations are the pairwise selection on the first step and the low consistence of the presentation of the upper (according to the selected criteria) segment of the search ($N \approx 200$) due to the small size of the correlation selection.[89,90]

b) Genetic Algorithm

Genetic Algorithm (GA) is a stochastic optimization machine learning technique that simulates natural selection principles and its advantages have been proven in several QSAR studies.[91] The genetic algorithm used in this study was presented for the first time by Leardi *et al*.[92] The fitness function in the QSARINS program[93] is the leave-one-out (LOO) cross-validation correlation coefficient (Q^2). GA method is used for the selection of descriptors and rank the best model based on the applicability domain which depends on the William's plot, internal validation, external validation, and relevance of the descriptors' physical meaning to the inhibitors.[94] Since the models are described by several parameters, the major goal is the extraction of relevant information, together with the exclusion of redundant and noisy information. In regression modelling, the most relevant variables with respect to the specific problem of interest are searched for by different selection strategies. GAs perform this selection by considering populations of models generated through a reproduction process and optimised according to a defined objective function related to model quality.[95] The genetic algorithm functions and parameters used in this Thesis are defined in Appendix B.

4.6.4 Model Validation

A robust QSAR modelling workflow is required to generate models, validate and predict activities for new datasets. The fitting ability of the model is verified by internal validation on the leave-one-out (LOO) cross-validation and leave-many-out (LMO) cross-validation techniques.[96] In the LMO cross-validation technique, ≈20% of training set compounds are obliterated in different cycles based on outliers and heterogeneity of the compounds in the dataset. For all iterations, the biological activities of the excluded compounds are then predicted using the model developed with the corresponding dataset of com-

pounds.[97] Training sets are further divided into multiple sets of descriptive training and test sets of different size, i.e., based on descriptor similarity using tanimoto method and structure similarity using overlap analysis. The external predictive ability of the model is assessed based on the predictions of the test set and external validation set compounds followed by the calculation of the $Q^2_{LOO\ \&\ LMO}$ parameter.[98] A randomization procedure aimed at testing models for potential chance correlations is the so-called Y-scrambling. This procedure randomizes the dependent variable vector, by assigning endpoint values to compounds to which they do not belong. The most stringent form of validation is the external validation. The true external validation uses compounds never used in the model development, the prediction is then carried out and R^2_{ext} for the external test set is compared to the R^2 for the model. In case of data collected from different sources significant differences between these two R^2 can be expected.[99]

4.6.5 Identification of outliers

Outliers are compounds with deviating endpoint values, which do not fit in a QSAR model. This usually happens when: i) a compound acts by a different mechanism, ii) interacts with the target in a different mode, iii) it is affected by a random or a systematic experimental error having little effect on the other chemicals. In statistical terms, an outlier is a data point, which has a high absolute standardized residual compared to the other compounds in the data set. As the coefficients and the intercept of the regression are highly sensitive to the presence of outliers, such points may be removed from training set.[100]

4.7 Domain of Applicability

Once a QSAR model is generated and properly validated, it can then be used to predict the activities/properties of a novel chemical entity. When the datasets used to generate and validate the models have limited structural diversity, it is expected that the model's applicability for the prediction of new compounds is also limited. Hence, reliable predictions are usually confined to chemicals that are structurally similar to the training set. The chemical space for the reliable predictions is defined as Applicability Domain (AD). A defined AD provides the following benefits: i) identifies the type of compounds for which reliable predictions can be obtained, ii) determines the degree of generalization of a QSAR model and iii) gives an idea about the interpolation and extrapolation power of a model (the extrapolation often limited to 30% beyond the minimum and the maximum values of the data used in the model development).[101,102]

4.7.1 QSAR Model selection

QSAR models are statistical constructs (models) of different mathematical form that describe the relationship between the independent and dependent variables. Although, a large number of models may be statistically relevant, only a fraction of them would be biologically/chemically significant. Thus, various criteria have to be considered when selecting a final model among the many possible alternatives:[103,104]

i) The model should be biologically and/or chemically relevant. For example, models that contain hydrophobicity related descriptors should be considered more relevant comparing to those that are difficult to interpret. Hydrophobicity manages several biological processes, such as distribution, transport and metabolism of biological system, molecular recognition etc. Therefore, the understanding of a parameter that defines the activity of molecules into polar and nonpolar stages is vital to predict the transport and activity of drugs.[105]

ii) Preference for parsimonious models. The principle introduced by William of Occam's razor states that among a set of equally good explanations for a given phenomenon, the simplest explanation tends to be the right one. In the context of QSAR, the models should have as few parameters as possible and should be trimmed down until they are minimally adequate.[106]

iii) Models with superior predictive power. As the majority of QSAR is generated for prediction, models able to predict external data correctly should be preferred.

4.7.2 Interpretation

In general, there are two types of QSAR models: i) models that are built with the sole purpose of predicting an endpoint of interest to fill gaps in the data and ii) models built to pursue a better understanding of the underlying biochemical/chemical/physico-chemical phenomena.[107] When dealing with the second type of QSAR models, the first and major step in their interpretation is the ability to interpret the individual descriptors.[108] It is important as arbitrary interpretation of descriptors may lead to irrelevant interpretation of the QSAR.[103] The workflow for statistical data modelling of QSAR paradigm is given in Figure **8**.

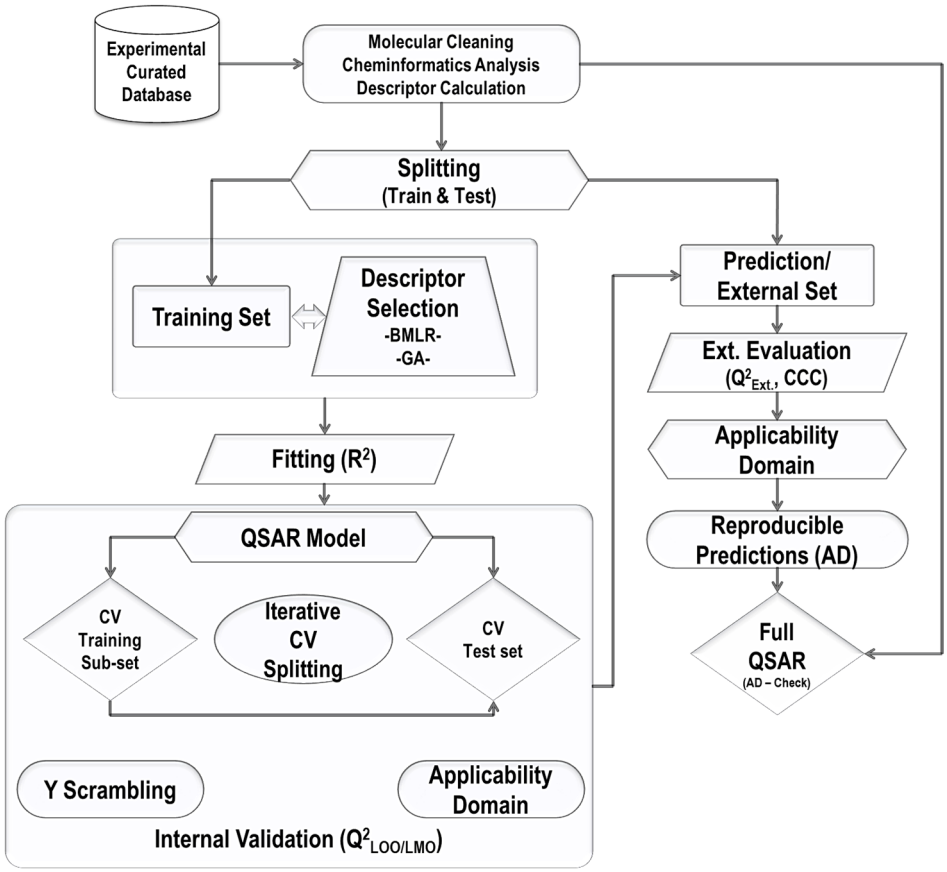

Figure 8. Overall workflow of QSAR (statistical data) modelling procedure. (Style adapted from QSARINS)

4.8 Fragment Based Scaffold Hopping

Fragment-based drug discovery (FBDD) method has been developed to generate new potential lead compounds.[109] The FBDD starts with the identification of fragments that generally bind with weak affinity to the target scaffold of interest. The fragments that form high quality interactions are then optimized to *lead* compounds with high affinity and selectivity.[110,111] The main idea of the fragment based QSAR (FQSAR) is the division of compound structures into appropriate fragments for which the fragment descriptors can be calculated.[112] FQSAR is not widely used in activity data modelling due to requirement of homogeneity and identical core of molecules in the dataset including limited number of descriptors.[113] Schematic representation of fragmentation for the FQSAR is shown in Figure **9**.

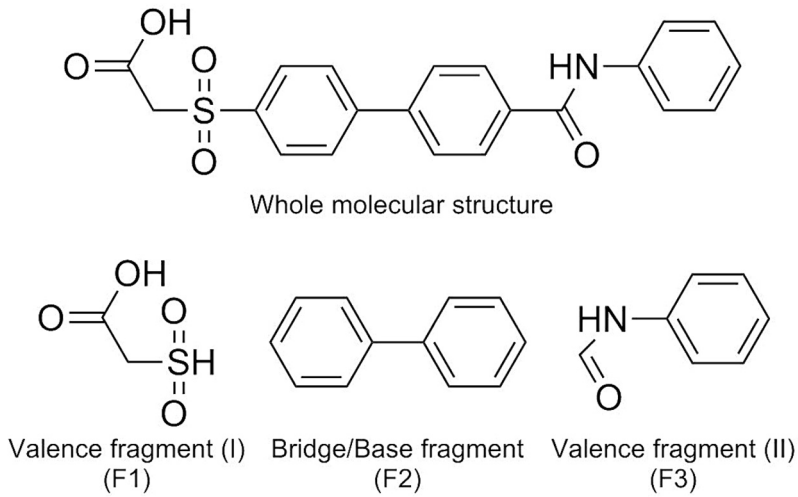

Figure 9. Illustration of custom fragmentation of molecular structures in FQSAR.

Pharmacophore modelling has become one of the major steps in drug discovery after the creation of drug or ligand databases with biological activity data..[114] In spite of the successes, pharmacophore approaches have not reached their expected full capacity, particularly in facing the demand for reducing the overall high cost associated with drug discovery and development. To overcome the shortcomings, scaffold hopping was introduced to find the fragments to substitute one part of a molecule with another, retaining their pharmacophoric interaction points.[114] In order to consider the interaction and spatial constraints, field point technology with XED forcefields was used to generate novel compounds with similar chemical activity. Bio-isostere replacement method[115] was used to perform scaffold hopping using field point constraints to generate novel compounds from fragment databases.[116,117] Schematic workflow for the fragment based (scaffold hopping) approach is shown in Figure **10**.

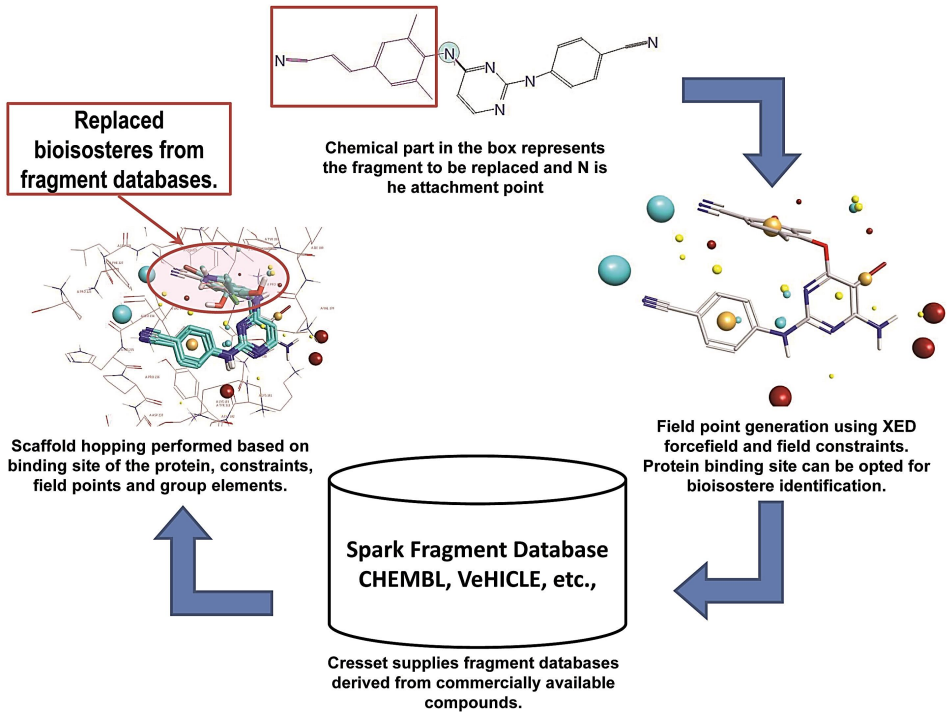

Figure 10. Schematic workflow for the fragment based scaffold hopping (bioisostere replacement) approach. Field-based template containing a single docked conformation of a chemical compound was considered on their 3D field point patterns to generate novel compounds.

5. SUMMARY OF ORIGINAL PUBLICATIONS

5.1 Application of QSLR in chemical ligation

Title: "Long-Range Chemical Ligation from N→N Acyl Migrations in Tryptophan Peptides via Cyclic Transition States of 10- to 18-Members"

The N→N acyl migration for the synthesis of native peptides has not been thoroughly explored, therefore we discovered the first examples of successful chemo-selective N→N acyl migration involving Trp-containing isopeptides via 10-, 11-, and 12-membered cyclic transition states.[I] However, this methodology still needed to be fully developed and explored by examining the following factors: 1) the range of cyclic transition states, 2) the best conditions for the ligation step, and 3) the effects of substituents in the amino acid residue and rationalization of the relative abundance of ligated product. This novel methodology was achieved without using Cys/Ser/Tyr residues or an auxiliary group at the ligation site. To rationalize the chemical ligation, a full conformation search was performed using MMX forcefield in PC Model,[118] considering both rotatable bonds and the phenyl rings. The bond distance - b(N-C) for twenty-one compounds were measured by generating the best pre-organized conformer for each compound. A statistical QSLR model was generated to predict the feasibility of ligation by considering the relative abundance as the activity data. The model generated using BMLR and Genetic Algorithm was further validated and with the experimental ligation data. The QSLR model equation was used to predict relative abundance of 6 more compounds and the model predictions were experimentally validated by measuring the relative abundance of the selected 3 compounds.[I] Given that there is an increasing number of studies involving the synthesis of longer peptides and iso-peptides, we believe this new ligation approach with QSLR represents a significant development in the field.

5.2 Identification of *Aedes aegypti* repellent chemotypes.

Title: "Promising *Aedes aegypti* Repellent Chemotypes Identified through Integrated QSAR, Virtual Screening, Synthesis, and Bioassay"

The repellent chemical library consisted of 43 carboxamides[119] together with 27 compounds for which the repellency was evaluated for this study. In this study, repellent activity measurements were carried out by USDA-ARS and the repellency was characterized by a minimum effective dosage (MED, μmol/cm^2). MED is defined as the minimum surface concentration of a compound that is required to produce a repellent effect. A QSAR (Quantitative Structure-Activity Relationships) pharmacophore model predicted the most favourable amide structure to consist of an aliphatic moiety and an aromatic hydrophobic moiety separated by a highly polar carboxyl group.[II] Another 3D (three dimensional) QSAR model defined an optimal structural pattern that consists of two oxygen

atoms (one of which belongs to an amide group) positioned a certain distance from each other and joined by a lipophilic moiety. Predictive models have also been derived by using multi-linear QSAR based on experimental and theoretical descriptors. Protection times of a large set of carboxamides and N-acylpiperidines had been qualitatively analysed using artificial neural networks and multiple linear regression.[119–121] The repellents in this study were classified as early spatial, late spatial, and contact. It found that few chemical bonds separating the hydroxyl and the hydrophobic fragments are beneficial for increasing repellent activity. Until very recently, no valid information on putative molecular targets was available. Analysis by QSAR revealed molecular determinants of repellent action against *Aedes aegypti*, and this knowledge was translated into search queries for a scaffold hopping step. Molecular docking using Glide software[63] against the *Aedes aegypti* OBP1 protein structure helped to identify highly promising scaffolds and individual compounds possessing mosquito repellent activity. From computational approaches, 27 assorted compounds containing hydroxyl, ether, ester, amine, nitro, and halogen functionalities were purchased and tested for measuring the MED.

5.3 Dual inhibition studies of Type 2 Diabetes Mellitus and Alzheimer's disease

Title: "Dual inhibition of the α-glucosidase and butyrylcholinesterase studied by Molecular Field Topology Analysis"

α-Glucosidase and BChE inhibitory activities were obtained for 42 and 65 compounds, respectively, of which 30 compounds had overlapping dual inhibition data. The compounds included assorted heterocyclic compounds: 27 alkyl and phenyl substituted triazoles, 20 benzothiazepines, 18 phenyl steryl ketones (chalcones). The whole library was synthesized and experimentally tested by the collaborators. This dual inhibitors subset predominantly consisted of 1,4-disubstituted-1,2,3-triazoles, whose specific structural features responsible for the poly-pharmacological activity were identified by MFTA.[III] As T2DM is a risk factor to AD, dual mode drugs acting on both of them are highly promising. The IC_{50} values for α-glucosidase vary from 11.9 to 6756.7 μM while those for BChE lie between 3.97 and 585 μM. The highest bi-target activity was found for two compounds, with IC_{50} values equal to 12 μM for α-glucosidase and 14 μM for BChE, respectively. The quantitative structure activity relationships and the common pharmacophore pattern identified in this work will help to design better drug candidates to counteract those two debilitating conditions.

5.4 Generation of QSAR models using fragment based approach

Title: "Theoretical Modeling of HPV: QSAR and Novodesign with Fragment Approach"

Antiviral agents capable of specifically inhibiting Human Papilloma Virus (HPV) replication could play an important role in the treatment of these diseases, but unfortunately no such antiviral agents are yet available. The recent progress toward the identification and characterization of specific molecular targets offers the prospect of effective HPV antiviral compounds.[122] Both standard and fragment based Quantitative Structure-Activity Relationships ((F)QSAR) methodology has been used to the analysis of HPV inhibitors, and is based on the experimental work done by White *et. al.* on a series of small molecules inhibiting the ATPase (Adinosine Tri-Phosphatase) activity of HPV6-E1 helicase.[50] E1 is the most highly conserved HPV protein that possesses enzymatic activity.[48] Thus, the E1 helicase has been considered the most attractive molecular target for the development of antiviral agents. In accordance with the scheme given on Figure **9** in section **4.8**, a data set of 42 anti-HPV compounds was divided into three subsets: 9 valence fragments (FI), 8 bridge fragments and 23 valence fragments (FII). FQSAR model was generated for prediction of the antiviral activities.[IV].

5.5 QSAR: Link between cause and effect

Title: "Quantitative structure–activity/property relationships: the ubiquitous links between cause and effect"

The universal applications of the QSAR approach were explored in various research fields. The predictions and modelling of QSAR within the applicability domain can be useful, reliable and cost effective for the whole drug discovery process depending on the dataset. Recent improvements in the QSAR approach have given a vision beyond the classical QSAR paradigm by detailed consideration of the molecular conformers, protein-ligand receptor complexes, and molecular dynamics. The only problem arises when there is a lack in the availability of 3D structures of protein targets to consider the improved methodology. QSAR is also widely used in designing novel compounds with improved activity, evaluating their toxicity in the field of materials science, nanotechnology, agrochemicals, pharmaceuticals and personal care products. The assessment of dimensionality in statistical QSAR conveys that not all models can predict activity for novel compounds and not all model validations are reliable for different applications. QSAR is a scientific method with its own benefits and drawbacks. Nevertheless, it is a powerful technique capable to cover huge chemical space, which is inaccessible with any other methodology. [V]

6. SUMMARY

In this thesis, QSAR methods in combination with pharmacophore assessment and molecular modelling were applied to generate predictive models for biological activities and potential drug candidates. The focus was set on a closer study to the conformational analysis, fragment- and ligand -based methods, global QSAR and molecular docking in prospect of drug design.

QSLR models for relative abundance (Article I) revealed importance of the bond distance b(N-C) in N→N acyl transfer and Balaban index in the chemical ligation. The reproducibility of the model was further proven by experimental validation for the predicted relative abundance by synthesizing 3 more compounds.

New chemotypes for mosquito repellents (Article II) were identified using the hit expansion technique which can lead to the discovery of less toxic and long lasting repellents. Molecular docking and pharmacophore based QSAR modelling was applied to identify new repellents from chemical libraries.

Comparative analysis of dual inhibition studies on T2DM and AD (Article III) showed overlapping of biological activities for 30 compounds. The identification of common pharmacophoric patterns may lead to the design of multi-target drugs in the future.

Finally, FQSAR method (Article IV) was applied for the prediction of novel potential inhibitors against HPV. A new set of techniques on fragmentation method and calculation of fragment based descriptor matrix were introduced in this work. The reported model had interpretable descriptors and better statistical parameters of prediction as compared to those of linear QSAR approach.

In summary, an improved QSAR approach was designed with adjacent computational methods to overcome the limitations of ligand-based methods. This will also provide the researchers with dependable tools for precise model predictions within the applicability domain to elucidate new drug candidates.

7. SUMMARY IN ESTONIAN

Keemiliste, biokeemiliste ja biomeditsiiniliste omaduste arvutuslik modelleerimine

Käesolevas dissertatsioonis kasutati QSAR meetodeid kombinatsioonis farmakofooride ja molekulaarmodelleerimisega ennustusvõimeliste mudelite loomiseks ning uute ravimikandidaatide leidmiseks. Töö eesmärgiks oli uurida konformatsioonianalüüsi ning fragmendi- ja ligandipõhiste meetodite ja molekulaarsildamise meetodite võimalikku kasutamist ravimiarenduses.

Artiklis I arendatud QSLR mudelis suhteliste saagiste jaoks peptiidide sünteesil ilmnes, et sidemete kaugus (*b(N-C) N>N* tsüüli üleminekus) ja Balabani indeks mängivad olulist rolli kirjeldamaks keemilist seostumist (*chemical ligation*). Mudeli pädevust tõestati kolme uue aine sünteesiga ja vastavate eksperimentaalsete mõõtmistega.

Artiklis II leiti uued kemotüübid repellentidele, mis omakorda andis juhtnöörid uute, vähemtoksiliste ja kauakestvate sääsetõrjevahendite leidmiseks. Kasutati ka molekulaarsildamist ja farmakofooripõhist QSAR meetodit sobivate kandidaatide väljaselekteerimiseks kemikaalide andmebaasist.

Artiklis III leiti, et 30 ühendit inhibeerivad samaaegsel nii diabeeti kui ka Alzheimeir tõbe. Taoliste ühiste farmakofooriliste mustrite avastamine võib tuleviks olla kasuks multifunktsionaalsete ravimite väljatöötamiseks.

Artiklis IV arendati FQSAR meetodi abil välja papiloom viiruse inhibiitorite aktiivsust ennustav mudel ning disainiti uued, potentsiaalsed antiviraalsed ühendid.

Kokkuvõtteks võib öelda, et erinevaid arvutumeetodeid kombineerides arendati täiendatud QSAR meetod, et saada üle ligandipõhiste meetoditega kaasnevatest piirangutest.

APPENDIX

A. Multiple Linear Regression Functions and Parameters[123]

BMLR relies on the following assumptions:

a) The relationship between the independent (x) variables and the dependent (y) variable is linear;
b) The residuals between the actual and the estimated values of y follow a normal distribution;
c) The independent variables x_1, x_2 …x_n should be uncorrelated ($R < 0.5$). As the calculation of the regression coefficients is done through matrix inversion if multi-collinearity is present the inversion matrix would be unstable.

The following signs can indicate the presence of multi-collinearity between the descriptors in a given descriptor pool:

a) The F-test of the QSAR equation as a whole is significant while none of the t-ratios of the coefficients are statistically significant;
b) The addition of a new descriptor to the equation radically changes either the size or the sign (plus or minus) of the regression coefficients of the remaining descriptors. Most QSAR operates on large descriptors pools. However, only a few of the descriptors are relevant to the modelled endpoint. Thus, feature selection algorithms able to extract a small subset of descriptors from a larger pool are often used.

A.1. Residual Sum of Squares, *RSS* (error sum of squares). The sum of squared differences between the observed (y) and estimated response:

$$\boldsymbol{RSS} = \sum_{\boldsymbol{i-1}}^{\boldsymbol{n}} (\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}})^{\boldsymbol{2}}$$

being *n* the number of training objects. This quantity is minimized by the least square estimator.

A.2. Model Sum of Squares, *MSS*, defined as the sum of the squared differences between the estimated responses and the average response:

$$\boldsymbol{MSS} = \sum_{\boldsymbol{i-1}}^{\boldsymbol{n}} (\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \bar{\boldsymbol{y}})^{\boldsymbol{2}}$$

This is a part of the total variance explained by the regression model as opposed to the residual sum of squares *RSS*.

A.3. Total Sum of Squares, *TSS*, defined as the sum of the squared differences between the experimental responses and the average response:

$$\boldsymbol{TSS = \sum_{i-1}^{n} (y_i - \bar{y})^2}$$

This is the total variance that a regression model has to explain and is used as a no-model reference quantity to calculate standard quality parameters such as the coefficient of determination.

A.4. Coefficient of determination, R^2. The squared multiple correlation coefficient that is the total variance of the response explained by a regression model. It can be calculated from the model sum of squares *MSS* or from the residual sum of squares RSS:

$$\boldsymbol{R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i-1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i-1}^{n} (y_i - \bar{y})^2}}$$

where *TSS* is the total sum of squares around the mean. A value of one indicates perfect fit, i.e. a model with zero error term.

A.5. Residual Mean Square, *RMS* or s^2 (: mean square error, expected squared error). The estimate s^2 of the error variance σ^2, defined as:

$$\boldsymbol{s^2 = \frac{RSS}{df_E}}$$

where *RSS* is the residual sum of squares and df_E is the error degrees of freedom, i.e. to $n - \mathbf{p}'$, where n is the number of objects (samples), p' the number of model parameters (for example, $n - p - 1$ for a regression model with p variables and the intercept). The standard error of the estimates is the square root of the residual mean square.

A.6. Standard Deviation Error in Calculation, *SDEC* also known as standard error in calculation, *SEC*. A function of the residual sum of squares, defined as:

$$\boldsymbol{SDEC = \sqrt{\frac{\sum_{i-1}^{n} (\hat{y}_i - y_i)^2}{n}} = \sqrt{\frac{RSS}{n}}}$$

A.7. F Fisher function. Among the most known statistical tests, it is defined as the ratio between the model sum of squares *MSS* and the residual sum of squares *RSS*:

$$\boldsymbol{F = \frac{MSS/df_M}{RSS/df_E}}$$

where df_M and df_E refer to the degrees of freedom of the model and error, respectively. The calculated value is compared with the critical value F crit for the corresponding degrees of freedom. It is a comparison between the model explained variance and the residual variance: high values of the F-ratio test indicate reliable models.

A.8. Adjusted $\mathbf{R^2}$, *f*. A fitness parameter adjusted for the degrees of freedom, so that it can be used for comparing models with different numbers of predictor variables:

$$\boldsymbol{R^2_{adj} = 1 - \frac{RSS/df_E}{TSS/df_T} = 1 - \left(1 - R^2\right).\left(\frac{n-1}{n-p'}\right)}$$

where *RSS* and *TSS* are the residual sum of squares and the total sum of squares, respectively; df_T refers to the total degrees of freedom; R^2 is the coefficient of determination.

A.9. Predictive Residual Sum of Squares, *PRESS*. The sum of squared differences between the observed and estimated response by validation techniques:

$$\boldsymbol{PRESS = \sum_{i-1}^{n}\left(y_i - \hat{y}_{i/i}\right)^2}$$

where $y_{i/i}$ denotes the response of the i-th object estimated by using a model obtained without using the i-th object. Using validation techniques minimizes this quantity.

A.10. Cross-validated $\mathbf{R^2}$, R^2cv (or Q^2). The explained variance in prediction:

$$\boldsymbol{R^2_{cv} = Q^2 = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i-1}^{n}\left(y_i - \hat{y}_{i/i}\right)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2}}$$

where *PRESS* is the predictive error sum of squares and *TSS* the total sum of squares.

A.11. External Q^2. The explained variance in prediction:

$$\boldsymbol{Q^2_{EXT} = 1 - \frac{\sum_{i-1}^{n_{ext}} \left(\hat{y}_{i/i} - y_i\right)^2 / n_{EXT}}{\sum_{i-1}^{n_{tr}} (y_i - \bar{y}_{TR})^2 / n_{TR}} = 1 - \frac{PRESS/n_{EXT}}{TSS/n_{TR}}}$$

where the sum runs over the test set objects (next) and $\bar{y}$ is the average value of the training set responses.

A.12. Standard Deviation Error of Prediction, *SDEP* also known as standard error in prediction *SEP* or *PSE*. A function of the predictive residual sum of squares, defined as:

$$\boldsymbol{SDEP = \sqrt{\frac{\sum_{i-1}^{n} \left(y_i - \hat{y}_{i/i}\right)^2}{n}} = \sqrt{\frac{PRESS}{n}}}$$

A.13. Total correlation in the model predictors, K_X:

$$\boldsymbol{K_x = \frac{\sum_{i=1}^{p} \left| \frac{\lambda_i}{\sum \lambda_n} - \frac{1}{p} \right|}{\frac{2.(p-1)}{p}}}$$

where λ are the eigenvalues obtained from the correlation matrix of the data set **X**(*n*, *p*), being *n* the number of objects and *p* the number of variables. Total correlation in the set given by the model predictors **X** plus the response **Y**. K_{XY} is calculated by the above formula (K_X) adding the y response to the set of predictor variables.

Appendix B
B. Genetic Algorithm Parameters

B.1. Random initialization of the population: The model population is built initially by random models with a number of variables between 1 and L. The value of the selected objective function of each model is calculated in a process called evaluation. The models are then ordered with respect to the selected objective function – model quality - (the best model is in first place in the population, the worst at position P);

B.2. Crossover: From the actual population, pairs of models are selected (randomly or with a probability function of their quality). Then, from each pair of selected models (parents), a new model is generated, preserving the common characteristics of the parents (i.e. variables excluded in both

models remain excluded, variables included in both models remain included) and mixing the opposite characteristics according to the crossover probability. If the generated sibling coincides with one of the individuals already present in the actual population, it is rejected; otherwise, it is evaluated. If the objective function value is better than the worst value in the population, the model is included in the population, in the place corresponding to its rank; otherwise, it is no longer considered. This procedure is repeated for several pairs;

B.3. Mutation: After a number of crossover iterations, the population proceeds through the mutation process. This means that for each individual of the population every gene is randomly changed into its opposite or left unchanged. Mutated individuals are evaluated and included in the population if their quality is acceptable. This process is controlled by mutation probability which is commonly set at low values, thus allowing only a few mutations and new individuals not too far away from the generating individual.

B.4. New generation: After a number of iterations, a new generation of the population can be performed killing a defined percent of individuals and randomly recreating them. In MobyDigs the 50% of population individuals is iteratively recreated after a user-defined number of iterations; the killed individuals are the worst ones. This process is useful in better exploring the solution space.

B.5. Population size: maximum number of models in a population (default: 50).

B.2. Maximum allowed variables in a model: maximum number of variables in a model (default: 3).

B.3. Start calculation with all subset models until: sets the maximum size of models searched by the All Subset Model approach (the default 0 indicates that not all the subset model search is performed

B.4. Number of retained models for each size: number of the best models for each size surviving in the population regardless of their quality (default: 3). This option is important to save, in the final population, also the best models of lower complexity e.g., the first best three models with one variable, the first three models with two variables, etc.

B.5. Trade-off between crossover and mutation: user-defined value of the T parameter which sets the values of the crossover and mutation probabilities (default: 0.5; T = 0 only crossover; T = 1 only mutation).

B.6. Selection bias: user-defined value of the B parameter which sets the parent selection operator (default: 0.5 roulette-wheel selection; B = 0 random selection).

B.7. Add noisy variables: addition of normal and uniformly distributed variables, which test chance correlation during the evolution procedure. The user can add up to 200 noisy variables to each population with labels ZZNxx when normally distributed and ZZUxx when uniformly distributed (xx is an ID number associated to the noisy variable).

REFERENCES

(1) Wilson, G. L.; Lill, M. A. *Future Med. Chem.* **2011**, *3*(6):735–50.
(2) Amaravadhi, H.; Baek, K.; Yoon, H. S. *Curr. Top. Med. Chem.* **2014**, *14*, 1890.
(3) Lionta, E.; Spyrou, G.; Vassilatis, D. K.; Cournia, Z. *Curr. Top. Med. Chem.* **2014**, *14*, 1923.
(4) Merz, K.M.; Ringe, D.; Reynolds, C.H. *Drug Design: Structure- and Ligand-Based Approaches*; Cambridge University Press, Cambridge, **2010**; p. 274.
(5) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. *PLoS Comput. Biol.* **2012**, *8*, e1002380.
(6) Hall, L. H.; Hall, L. M. *SAR QSAR Environ. Res.* **2007**, *16*, 13.
(7) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. *Nat. Rev. Drug Discov.* **2013**, *12*, 948.
(8) Wieland, T.; Bokelmann, E.; Bauer, L.; Lang, H. U.; Lau, H. *Justus Liebigs Ann. Chem.* **1953**, *583*, 129.
(9) Dawson, P.; Muir, T.; Clark-Lewis, I.; Kent, S. *Science.* **1994**, *(80)*266, 776.
(10) Lemieux, G. *Trends Biotechnol.* **1998**, *16*, 506.
(11) Dang, B.; Kubota, T.; Mandal, K.; Bezanilla, F.; Kent, S. B. H. *J. Am. Chem. Soc.* **2013**, *135*, 11911.
(12) McGrath, N. A.; Raines, R. T. *Acc. Chem. Res.* **2011**, *44*, 752.
(13) Pentelute, B. L. *New Chemical Methods for the Synthesis of Proteins and Their Application to the Elucidation of Protein Structure by Racemic Protein Crystallography*; ProQuest, 2008; p. 353.
(14) Harris, P. W. R.; Brimble, M. A. *Biopolymers* **2010**, *94*, 542.
(15) Low, D. W.; Hill, M. G.; Carrasco, M. R.; Kent, S. B.; Botti, P. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 6554.
(16) Bosman, A. W.; Janssen, H. M.; Meijer, E. W. *Chem. Rev.* **1999**, *99*, 1665.
(17) Hojo, H.; Ozawa, C.; Katayama, H.; Ueki, A.; Nakahara, Y. Y.; Nakahara, Y. Y. *Angew. Chemie* **2010**, *49*, 5446.
(18) Nilsson, B. L.; Kiessling, L. L.; Raines, R. T. *Org. Lett.* **2000**, *2*, 1939.
(19) Restituyo, J. A.; Comstock, L. R.; Petersen, S. G.; Stringfellow, T.; Rajski, S. R. *Org. Lett.* **2003**, *5*, 4357.
(20) Biswas, S.; Kayaleh, R.; Pillai, G. G.; Seon, C.; Roberts, I.; Popov, V.; Alamry, K. A.; Katritzky, A. R. *Chemistry: A Euro. J.* **2014**, *20*, 8189.
(21) Strickman, D.; Debboun, M.; Frances, S. *Insect Repellents: Principles, Methods, and Uses*; 1st ed.; CRC Press: London, UK. **2006**, p. 495
(22) Suryanarayana, M. V. S.; Pandey, K. S.; Prakash, S.; Raghuveeran, C. D.; Dangi, R. S.; Swamy, R. V.; Rao, K. M. *J. Pharm. Sci.* **1991**, *80*, 1055.
(23) Van der Goes van Naters, W.; Carlson, J. R. *Nature* **2006**, *444*, 302.
(24) Stanczyk, N. M.; Brookfield, J. F. Y.; Ignell, R.; Logan, J. G.; Field, L. M. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 8575.
(25) Belova, O. A.; Burenkova, L. A.; Karganova, G. G. *Ticks Tick. Borne. Dis.* **2012**, *3*, 240.
(26) Corbel, V.; Stankiewicz, M.; Pennetier, C.; Fournier, D.; Stojan, J.; Girard, E.; Dimitrov, M.; Molgó, J.; Hougard, J.-M.; Lapied, B. *BMC Biol.* **2009**, *7*, 47.
(27) Bohbot, J. D.; Dickens, J. C. *PLoS One* **2010**, *5*, e12138.
(28) Paluch, G.; Bartholomay, L.; Coats, J. *Pest Manag. Sci.* **2010**, *66*, 925.
(29) Syed, Z.; Pelletier, J.; Flounders, E.; Chitolina, R. F.; Leal, W. S. *PLoS One* **2011**, *6*, e17705.

(30) Basak, S. C.; Ramanathan, N.; Mills, D. *ICCOMP Proceedings,* **2005**, *41*.
(31) Zhu, Y.-P.; Yin, L.-J.; Cheng, Y.-Q.; Yamaki, K.; Mori, Y.; Su, Y.-C.; Li, L.-T. *Food Chem.* **2008**, *109*, 737.
(32) Du, Z.; Liu, R.; Shao, W.; Mao, X.; Ma, L.; Gu, L.; Huang, Z.; Chan, A. S. C. *Eur. J. Med. Chem.* **2006**, *41*, 213.
(33) Liu, Y.; Zou, L.; Ma, L.; Chen, W.-H.; Wang, B.; Xu, Z.-L. *Bioorg. Med. Chem.* **2006**, *14*, 5683.
(34) McCulloch, D. K.; Kurtz, A. B.; Tattersall, R. B. *Diabetes Care* **1983**, *6*, 483.
(35) Van de Laar, F. A.; Lucassen, P. L. B. J.; Akkermans, R. P.; Van de Lisdonk, E. H.; Rutten, G. E. H. M.; Van Weel, C. *Cochrane database Syst. Rev.* **2005**, CD003639.
(36) Iwamoto, Y.; Kashiwagi, A.; Yamada, N.; Terao, S.; Mimori, N.; Suzuki, M.; Tachibana, H. *Diabetes. Obes. Metab.* **2010**, *12*, 700.
(37) Fukaya, N.; Mochizuki, K.; Tanaka, Y.; Kumazawa, T.; Jiuxin, Z.; Fuchigami, M.; Goda, T. *Eur. J. Pharmacol.* **2009**, *624*, 51.
(38) Greig, N. H.; Utsuki, T.; Ingram, D. K.; Wang, Y.; Pepeu, G.; Scali, C.; Yu, Q.-S.; Mamczarz, J.; Holloway, H. W.; Giordano, T.; Chen, D.; Furukawa, K.; Sambamurti, K.; Brossi, A.; Lahiri, D. K. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17213.
(39) Giacobini, E. *Drugs and Aging*, **2001**, *18*(12), 891–898.
(40) Decker, M.; Kraus, B.; Heilmann, J. *Bioorg. Med. Chem.* **2008**, *16*, 4252.
(41) Han, W.; Li, C. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 6557.
(42) De la Monte, S. M. *BMB Rep.* **2009**, *42*, 475.
(43) Götz, J.; Ittner, L. M.; Lim, Y. *Cell. Mol. Life Sci.* **2009**, *66*, 1321.
(44) Lue, L.-F.; Andrade, C.; Sabbagh, M.; Walker, D. *Int. J. Alzheimers. Dis.* **2012**, *2012*, 918680.
(45) Akter, K.; Lanza, E. A.; Martin, S. A.; Myronyuk, N.; Rua, M.; Raffa, R. B. *Br. J. Clin. Pharmacol.* **2011**, *71*, 365.
(46) Sridhar, G. R.; Thota, H.; Allam, A. R.; Suresh Babu, C.; Siva Prasad, A.; Divakar, C. *Lipids Health Dis.* **2006**, *5*, 28.
(47) Lu, J.-J.; Pan, W.; Hu, Y.-J.; Wang, Y.-T. *PLoS One* **2012**, *7*, e40262.
(48) Faucher, A.-M.; White, P. W.; Brochu, C.; Grand-Maître, C.; Rancourt, J.; Fazal, G. *J. Med. Chem.* **2004**, *47*, 18.
(49) De Villiers, E.-M.; Fauquet, C.; Broker, T. R.; Bernard, H.-U.; zur Hausen, H. *Virology* **2004**, *324*, 17.
(50) White, P. W.; Pelletier, A.; Brault, K.; Titolo, S.; Welchner, E.; Thauvette, L.; Fazekas, M.; Cordingley, M. G.; Archambault, J. *J. Biol. Chem.* **2001**, *276*, 22426.
(51) Dochez, C.; Bogers, J. J.; Verhelst, R.; Rees, H. *Vaccine* **2014**, *32*, 1595.
(52) Kumar, S.; Jena, L.; Galande, S.; Daf, S.; Mohod, K.; Varma, A. K. *Genomics Inform.* **2014**, *12*, 64.
(53) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.
(54) García-Domenech, R.; Galvez, J.; de Julian-Ortiz, J. V; Pogliani, L. *Chem. Rev.* **2008**, *108*, 1127.
(55) Zhan, W.; Liang, Z.; Zhu, A.; Kurtkaya, S.; Shim, H.; Snyder, J. P.; Liotta, D. C. *J. Med. Chem.* **2007**, *50*, 5655.
(56) Radchenko, E. V.; Makhaeva, G. F.; Malygin, V. V.; Sokolov, V. B.; Palyulin, V. A.; Zefirov, N. S. *Dokl. Biochem. Biophys.* **2011**, *418*, 47.

(57) Chupakhin, V. I.; Bobrov, S. V.; Radchenko, E. V.; Palyulin, V. A.; Zefirov, N. S. *Dokl. Chem.* **2008**, *422*, 227.
(58) Morris, G. M.; Lim-Wilby, M. *Methods Mol. Biol.* **2008**, *443*, 365.
(59) Huang, N.; Shoichet, B. K.; Irwin, J. J. *J. Med. Chem.* **2006**, *49*, 6789.
(60) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. *J. Med. Chem.* **2006**, *49*, 5912.
(61) Ojha, P.; Roy, K. *Comb. Chem. High Throughput Screen.* **2013**, *16*, 739.
(62) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. *J. Comput. Chem.* **2009**, *30*, 2785.
(63) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739.
(64) Johnson, A. M.; Maggiora, G. M.; A. M. Johnson, G. M. M. *Concepts and Applications of Molecular Similarity*; John Willey & Sons: New York, USA, **1990**. p. 393
(65) Willett, P. *Methods Mol. Biol.* **2011**, *672*, 133.
(66) Bender, A.; Glen, R. C. *Org. Biomol. Chem.* **2004**, *2*, 3204.
(67) Kubinyi, H. *Perspect. Drug Discov. Des.* **1998**, *9–11*, 225.
(68) Instant JChem 6.0, Chemaxon Kft, Hungary, **2013**.
(69) Discovery Studio Visualizer 4.0, Accelrys (Biovia) Inc., USA, **2014**.
(70) Katritzky, A. R.; Oliferenko, A.; Lomaka, A.; Karelson, M. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 3453.
(71) Zanni, R.; Galvez-Llompart, M.; Galvez, J.; García-Domenech, R. *Curr. Comput. Aided-Drug Des.* **2014**, *10*, 129.
(72) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. *J. Med. Chem.* **2014**, *57*, 4977.
(73) Tropsha, A. *Mol. Inform.* **2010**, *29*, 476.
(74) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)*; John Wiley & Sons, 2009; p. 1257.
(75) Mamy, L.; Patureau, D.; Barriuso, E.; Bedos, C.; Bessac, F.; Louchart, X.; Martin-Laurent, F.; Miege, C.; Benoit, P. *Crit. Rev. Environ. Sci. Technol.* **2015**, *45*, 1277.
(76) Tetko, I. V; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. *J. Comput. Aided. Mol. Des.* **2005**, *19*, 453.
(77) Todeschini, R.; Consonni, V.; Gramatica, P. *Comprehensive Chemometrics*; Elsevier, **2009**; pp. 129–172.
(78) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(79) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. *Curr. Pharm. Des.* **2006**, *12*, 2111.
(80) Vařeková, R.S.; Jirouskova, Z.; Vanek, J.; Suchomel, Š.; Koca, J. *Int. J. Mol. Sci.*, **2007**, *8*(7), 572.
(81) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.
(82) No, K. T.; Grant, J. A.; Jhon, M. S.; Scheraga, H. A. *J. Phys. Chem.* **1990**, *94*, 4740.

(83) Karelson, M. *Molecular Descriptors in QSAR/QSPR,* **2000**, John Wiley & Sons Inc, New York, USA. p. 448
(84) Shim, J.; Mackerell, A. D. *Medchemcomm* **2011**, *2*, 356.
(85) Fourches, D.; Muratov, E.; Tropsha, A. *J. Chem. Inf. Model.* **2010**, *50*, 1189.
(86) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. *SAR QSAR Environ. Res.* **2009**, *20*, 241.
(87) Johnson, S. R. *J. Chem. Inf. Model.* **2007**, *48*, 25.
(88) Varmuza, K.; Filzmoser, P.; Dehmer, M. *Comput. Struct. Biotechnol. J.* **2013**, *5*, e201302007.
(89) Darlington, R.B. *Regression and Linear Models,* **1990**, McGraw-Hill Inc.,US. p. 542.
(90) Katritzky, A.; Pacureanu, L.; Slavov, S.; Dobchev, D.; Shah, D.; Karelson, M. *Comput. Chem. Eng.* **2009**, *33*, 321.
(91) Cartwright, H. M. *Applications of Artificial Intelligence in Chemistry (Oxford Chemistry Primers)*; Oxford University Press, USA, **1994**; p. 96.
(92) Leardi, R.; Boggia, R.; Terrile, M. *J. Chemom.* **1992**, *6*, 267.
(93) Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. *J. Comput. Chem.* **2013**, *34*, 2121.
(94) Pavan, M.; Netzeva, T. I.; Worth, A. P. *SAR QSAR Environ. Res.* **2006**, *17*, 147.
(95) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Mannhold, R.; Kubinyi, H.; Folkers, G., Eds.; 2nd ed.; Wiley-VCH, **2009**; p. 1257.
(96) Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579.
(97) Pratim Roy, P.; Paul, S.; Mitra, I.; Roy, K. *Molecules* **2009**, *14*, 1660.
(98) Mitra, I.; Saha, A.; Roy, K. *J. Mol. Model.* **2012**, *18*, 1819.
(99) Roy, K.; Mitra, I.; Kar, S.; Ojha, P. K.; Das, R. N.; Kabir, H. *J. Chem. Inf. Model.* **2012**, *52*, 396.
(100) Rocke, D. M.; Woodruff, D. L. *J. Am. Stat. Assoc.* **1996**. *91*(435), 1047–1061.
(101) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Öberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V; Tetko, I. V. *J. Chem. Inf. Model.* **2010**, *50*, 2094.
(102) Tropsha, A.; Golbraikh, A. *Curr. Pharm. Des.* **2007**, *13*, 3494.
(103) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. *Chem. Rev.* **2010**, *110*, 5714.
(104) Crawley M.J. *Statistics: An Introduction using R,* **2011**, John Wiley & Sons Ltd., p. 327.
(105) Masuda, T.; Jikihara, T.; Nakamura, K.; Kimura, A.; Takagi, T.; Fujiwara, H. *J. Pharm. Sci.* **1997**, *86*, 57.
(106) Thorburn, W. M. *Mind,* **1918**, *27*(107), 345–353.
(107) Livingstone, D. J. *J. Chem. Inf. Model.* **2000**, *40*, 195.
(108) Guha, R.; Van Drie, J. H. *J. Chem. Inf. Model.* **2008**, *48*, 646.
(109) Hajduk, P. J.; Greer, J. *Nat. Rev. Drug Discov.* **2007**, *6*, 211.
(110) Marcou, G.; Rognan, D. *J. Chem. Inf. Model.* **2006**, *47*, 195.
(111) Kumar, A.; Voet, A.; Zhang, K. Y. J. *Curr. Med. Chem.* **2012**, *19*, 5128.

(112) Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Hür, E.; Tämm, K.; Kurunczi, L.; Karelson, M.; Varnek, A.; Solov'ev, V. P. *J. Med. Chem.* **2006**, *49*, 3305.
(113) Myint, K. Z.; Xie, X.-Q. *Int. J. Mol. Sci.* **2010**, *11*, 3846.
(114) Yang, S.-Y. *Drug Discov. Today* **2010**, *15*, 444.
(115) Spark v10, Cresset-UK, **2014**.
(116) Khanna, V.; Ranganathan, S. *J. Cheminform.* **2011**, *3*, 30.
(117) Low, C. M. R.; Buck, I. M.; Cooke, T.; Cushnir, J. R.; Kalindjian, S. B.; Kotecha, A.; Pether, M. J.; Shankley, N. P.; Vinter, J. G.; Wright, L. *J. Med. Chem.* **2005**, *48*, 6790.
(118) Gilbert, K. E. PC Model v9.3, **2011**.
(119) Katritzky, A. R.; Wang, Z.; Slavov, S.; Dobchev, D. A.; Hall, C. D.; Tsikolia, M.; Bernier, U. R.; Elejalde, N. M.; Clark, G. G.; Linthicum, K. J. *J. Med. Entomol.* **2010**, *47*, 924.
(120) Katritzky, A. R.; Wang, Z.; Slavov, S.; Tsikolia, M.; Dobchev, D.; Akhmedov, N. G.; Hall, C. D.; Bernier, U. R.; Clark, G. G.; Linthicum, K. J. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 7359.
(121) Paluch, G.; Grodnitzky, J.; Bartholomay, L.; Coats, J. *J. Agric. Food Chem.* **2009**, *57*, 7618.
(122) D'Abramo, C. M.; Archambault, J. *Open Virol. J.* **2011**, *5*, 80.
(123) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics, **2009**. Springer New York: New York, NY. pp. 43–99.

ACKNOWLEDGEMENTS

First and foremost I offer my sincere gratitude to my supervisors, Prof. Mati Karelson, The Late Prof. Alan Roy Katritzky and Dr. Kaido Tämm who have supported me throughout my thesis with patience and knowledge. I attribute the level of my Doctoral degree to their encouragement and effort considering me to the research group. I also thank my colleagues, collaborators and the co-authors of my research publications for their contribution to research studies.

In the Florida Center for Heterocyclic Compounds, University of Florida, USA, I have been aided for three years in running the theoretical experiments and calculations with the guidance of the Late Prof. Alan R. Katritzky (1928-2014) and Dr. Alexander A. Oliferenko, a fine researcher who guided throughout my stay at University of Florida. I extend my thanks to Dr. Hall, Dr. Biswas, Dr. Polina, Dr. Jabeen, Dr. Panda, Dr. Khanh Ha and Dr. Lebedeva who helped me with laboratory skills in organic chemistry, medicinal chemistry approaches in my research and to analyse the data. This work was supported by US Department of Agriculture – ARS and Kenan Foundation, University of Florida to carry out my research in the USA.

In the Institute of Chemistry, University of Tartu I have been aided with the support and guidance of the Chair Molecular Technology Researchers. I thank Dr. Maran, Dr. Sild, Dr. Tulp, Dr. Garcia-Sosa, Birgit, Geven, Mare, Laura, Maikki and Kalev for creating excellent office atmosphere and helping me throughout my stay at Estonia. The institute has provided with the research support and facility that I have needed to produce and complete my thesis. I am thankful to ChemAxon, SemiChem, QSARINS, Cresset-UK, BioSolveIT and COSMOlogic for providing me with personal software licenses. I extend my gratitude to Dr. Tarmo Tamm and Dr. Dennis Hall for their critical review and linguistics support for the thesis. This work was partially supported by the graduate school "Functional materials and technologies," for funding from the European Social Fund under project 1.2.0401.09-0079, the EU European Regional Development Fund through the Center of Excellence in Chemical Biology, Estonia, and by targeted financing from the Estonian Ministry of Education & Research (SF0140031As09).

Finally, I am thankful to my dearest wife Archana and daughter Gangaa for their patience and support throughout the efforts for the completion of my doctoral studies. Last but not least, I want to thank my parents who have always believed in me and been very supportive and inspiration.

ORIGINAL PUBLICATIONS

CURRICULUM VITAE

Name:	Girinath Gopinathan Pillai
Data of Birth:	October 12
Citizenship:	Indian
Address:	#63, Nyro Puthen Veedu, Sarayu Nagar, Asramam, Kollam, Kerala, India
Phone number:	+1-352-316-7801, +91-9995981141
E-mail:	giribio@mail.com

Education:

2012–Present	PhD in Chemistry (*Computational*), Institute of Chemistry, University of Tartu, Estonia
2004–2006	M.Sc. in Bioinformatics, Bharathidasan University, Trichy, India
2001–2004	B.Sc. in Biochemistry, Kerala University, Trivandrum, India

Professional career:

2014–Present	Marie Curie Actions Research Fellow, Molcode Ltd., Tallinn, Estonia
2011–2015	Research Scholar, Florida Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, USA
2014–2014	Chemist, Chair of Molecular Technology, Chemicum, University of Tartu, Tartu, Estonia
2012–2012	Research Scholar, Chair of Molecular Technology, Chemicum, University of Tartu, Tartu, Estonia
2010–2011	Project Fellow, Center for Bioinformatics (IGNOU), University of Kerala, Thiruvananthapuram, India
2009–2010	Technical Consultant, Partner of QChem, Chemaxon, & Molcode, Mumbai, India
2007–2009	Scientific Support Manager, Partner of Accelrys, BioSolveIT & Cosmologic, Apsara Innovations Pvt. Ltd., Bangalore, India
2006–2007	Lecturer, Vinayaka Missions Kirupananda Variyar Engineering College, Vinayaka Missions University, Salem, India

Main field of research:
Computational chemistry, Molecular modelling, Cheminformatics, QSAR and Drug design.

Scholarships and Awards:

- Marie Curie Actions Research Fellow (Recognised Researcher – R2), Molcode Ltd. EU FP7/People Marie Curie Actions Programme, 612275, 2014.
- Graduate School Research Scholarship, FMTDK, European Social Fund (1.2.0401.09-0079), 2013–2014.
- Research Project Funding, US Department of Agriculture – ARS, Florida. USA. 2011–2014.
- University First Rank, MSc. Bioinformatics, Bharathidasan University, India. 2006.
- Project Studentship on Malaria, Tamilnadu State Council for Science & Technology, India. 2005.
- Best Poster Award – FloHet Conference, USA and BioDesign India.

List of Patents:

Oliferenko, P.; Oliferenko, A.; Ibrahim, Mohamed, A.; Panda, Siva, S.; Katritzky, Alan, R.; **Pillai, G.G.**; Panda, Chandramukhi, S. Macrocyclic Peptidomimetics with Nanomolar Antifungal And Antimicrobial Activity. WO/2014/182839, November 14, 2014.

List of Publications:

1. Dearden, J.; Hewitt, M.; Roberts, D.; Enoch, S.; Rowe, P.; Przybylak, K.; Vaughan-Williams, D.; Smith, M.; **Pillai, G.G**; Katritzky, A. Mechanism-based QSAR modeling of skin sensitization. *ACS Chem. Res. Tox.* **2015**, (In Press).
2. Wang, Z.; Zhang, H.; Killian, B.J.; Jabeen, F.; **Pillai, G.G**.; Berman, H.M.; Mathelier, M.; Sibble, A.J.; Yeung, J.; Zhou, W.; Steel, P.J.; Hall, C.D.; Katritzky, A. R. Synthesis, Characterization and Energetic Properties of 1,3,4-Oxadiazoles. *Eur. J Org. Chem.* **2015**, DOI: 10.1002/ejoc.201500583.
3. Girgis, A. S.; Panda, S. S.; Farag, I. S. A.; El-Shabiny, A. M.; Moustafa, A. M.; Ismail, N. S. M.; **Pillai, G. G.**; Panda, C. S.; Hall, C. D.; Katritzky, A. R. Synthesis, and QSAR Analysis of Anti-Oncological Active Spiro-Alkaloids. *Org. Biomol. Chem.* **2015**, *13* (6), 1741–1753.
4. **Pillai, G. G**.; Sikk, L.; Tamm, T.; Karelson, M.; Burk, P.; Tämm, K.; Tamm, K. Theoretical Modeling of HPV: QSAR and Novodesign with Fragment Approach. *Curr. Comput. Aided. Drug Des.* **2014**, *10* (4), 303–314.
5. Ha, K.; Lebedyeva, I.; Hamedzadeh, S.; Li, Z.; Quiñones, R.; **Pillai, G. G**.; Williams, B.; Nasajpour, A.; Martin, K.; Asiri, A. M.; Katritzky, A. R. Tandem Deprotection-Dimerization-Macrocyclization Route to C(2) Symmetric Cyclo-Tetrapeptides. *Chem. Eur. J.* **2014**, *20* (17), 4874–4879. *(Frontispiece)*
6. Girgis, A. S.; Saleh, D. O.; George, R. F.; Srour, A. M.; **Pillai, G. G**.; Panda, C. S.; Katritzky, A. R. Synthesis, Bioassay, and QSAR Study of

Bronchodilatory Active 4H-pyrano[3,2-C]pyridine-3-Carbonitriles. *Eur. J. Med. Chem.* **2014**, *89*, 835–843.
7. Jabeen, F.; Oliferenko, P. V.; Oliferenko, A. A.; **Pillai, G. G**.; Ansari, F. L.; Hall, C. D.; Katritzky, A. R. Dual Inhibition of the A-Glucosidase and Butyrylcholinesterase Studied by Molecular Field Topology Analysis. *Eur. J. Med. Chem.* **2014**, *80*, 228–242.
8. Oliferenko, P. V; Oliferenko, A. A.; Girgis, A. S.; Saleh, D. O.; Srour, A. M.; George, R. F.; **Pillai, G. G**.; Panda, C. S.; Katritzky, A. R.; Hall, C. D. Synthesis, Bioassay, and Molecular Field Topology Analysis of Diverse Vasodilatory Heterocycles. *J. Chem. Inf. Model.* **2014**, *54* (4), 1103–1116.
9. Biswas, S.; Kayaleh, R.; **Pillai, G. G**.; Seon, C.; Roberts, I.; Popov, V.; Alamry, K. A.; Katritzky, A. R. Long-Range Chemical Ligation from N→N Acyl Migrations in Tryptophan Peptides via Cyclic Transition States of 10- to 18-Members. *Chemistry* **2014**, *20* (26), 8189–8198.
10. Li, Z.; Lebedyeva, I.; Zhao, D.; Myers, L.; **Pillai, G. G**.; Hall, C. D.; Katritzky, A. R. Synthesis of L-Lys-Aminoxy-Goralatide. *J. Pept. Sci.* **2014**, *20* (12), 923–927.
11. Dobchev, D. A.; **Pillai, G. G**.; Karelson, M. *In Silico* Machine Learning Methods in Drug Development. *Curr. Top. Med. Chem.* **2014**, *14* (16), 1913–1922.
12. Oliferenko, P. V.; Oliferenko, A. A.; Poda, G. I.; Osolodkin, D. I.; **Pillai, G. G**.; Bernier, U. R.; Tsikolia, M.; Agramonte, N. M.; Clark, G. G.; Linthicum, K. J.; Katritzky, A. R.; Polina V. Oliferenko Alexander A. Oliferenko, A. R. K. Promising *Aedes Aegypti* Repellent Chemotypes Identified through Integrated QSAR, Virtual Screening, Synthesis, and Bioassay. *PLoS One* **2013**, *8* (9), e64547.
13. El-Gendy, B. E. D. M.; Ghazvini Zadeh, E. H.; Sotuyo, A. C.; **Pillai, G. G**.; Katritzky, A. R. α-Substitution Effects on the Ease of S→ N-Acyl Transfer in Aminothioesters. *Chem. Biol. Drug Des.* **2013**, *81* (5), 577–582.
14. Ha, K.; Lebedyeva, I.; Li, Z.; Martin, K.; Williams, B.; Faby, E.; Nasajpour, A.; **Pillai, G. G**.; Al-Youbi, A. O.; Katritzky, A. R. Conformationally Assisted Lactamizations for the Synthesis of Symmetrical and Unsymmetrical Bis-2,5-Diketopiperazines. *J. Org. Chem.* **2013**, *78* (17), 8510–8523.
15. Ha, K.; Monbaliu, J.-C. M.; Williams, B. C.; **Pillai, G. G**.; Ocampo, C. E.; Zeller, M.; Stevens, C. V; Katritzky, A. R. A Convenient Synthesis of Difficult Medium-Sized Cyclic Peptides by Staudinger Mediated Ring-Closure. *Organic & Biomolecular Chemistry*, **2012**, *10*, 8055.
16. Berhanu, W. M.; **Pillai, G. G**.; Oliferenko, A. A.; Katritzky, A. R. Quantitative Structure-Activity/property Relationships: The Ubiquitous Links between Cause and Effect. *ChemPlusChem*, **2012**, *77*, 507–517.
17. Singh, N. K.; Pakkkianathan, B. C.; Kumar, M.; Daddam, J. R.; Jayavel, S.; Kannan, M.; **Pillai, G. G**.; Krishnan, M. Computational Studies on Molecular Interactions of 6-Thioguanosine Analogs with Anthrax Toxin Receptor 1. *Interdiscip. Sci.* **2012**, *4* (3), 183–189.

18. Berhanu, W. M.; Ibrahim, M. A.; **Pillai, G. G**.; Oliferenko, A. A.; Khelashvili, L.; Jabeen, F.; Mirza, B.; Ansari, F. L.; Ul-Haq, I.; El-Feky, S. A.; Katritzky, A. R. Supporting Information for Similarity Analysis, Synthesis, and Bioassay of Antibacterial Cyclic Peptidomimetics. *Beilstein J. Org. Chem.* **2012**, *8* (1), 1146–1160.
19. Berhanu, W. M.; Ibrahim, M. A.; **Pillai, G. G**.; Oliferenko, A. A.; Khelashvili, L.; Jabeen, F.; Mirza, B.; Ansari, F. L.; Ul-Haq, I.; El-Feky, S. A.; Katritzky, A. R. Similarity Analysis, Synthesis, and Bioassay of Antibacterial Cyclic Peptidomimetics. *Beilstein J. Org. Chem.* **2012**, *8*, 1146–1160. *(Frontispiece)*
20. Zhong, M.; Jang, M.; Oliferenko, A.; **Pillai, G. G**.; Katritzky, A. R. The SOA Formation Model Combined with Semiempirical Quantum Chemistry for Predicting UV-Vis Absorption of Secondary Organic Aerosols. *Phys. Chem. Chem. Phys.* **2012**, *14* (25), 9058–9066.

ELULOOKIRJELDUS

Nimi:	Girinath Gopinathan Pillai
Sünniaeg:	Oktoober 12
Kodakondsus:	India
Aadress:	63, Nyro Puthen Veedu, Sarayu Nagar, Asramam, Kollam, Kerala, India
Telefon:	+1-352-316-7801, +91-9995981141
E-post:	giribio@mail.com

Haridus:

2012–Present	Doktorant – Keemia, Keemia Instituut, Tartu Ülikool, Eesti
2004–2004	MSc, Bioinformaatika Bharathidasan Ülikool, Trichy, India
2001–2004	BSc Biokeemia Kerala Ülikool, Trivandrum, India

Teenistuskäik:

2014–Present	Marie Curie Actions meetme stipendiaat, Molcode Ltd., Tallinn, Estonia
2011–2015	Teadur, Florida Center for Heterocyclic Compounds, Keemia osakond, Florida Ülikool, USA
2014–2014	Keemik, Molekulaartehnoloogia õppetool, Chemicum, Tartu Ülikool, Tartu, Estonia
2012–2012	Teadur, Molekulaartehnoloogia õppetool, Chemicum, Tartu Ülikool, Tartu, Estonia
2010–2011	Projekti täitja, Bioinformaatika (IGNOU), Kerala Ülikool, Thiruvananthapuram, India
2009–2010	Tehniline konsultant, Qchem, Chemaxon, & Molcode, Mumbai, India
2007–2009	Tehniline tugi, Accelrys, BioSolveIT & Cosmologic, Apsara Innovations Pvt. Ltd., Bangalore, India
2006–2007	Lektor, VMKV Engineering College, Vinayaka Missions Ülikool, Salem, India

Peamised uurimisvaldkonnad:
Molekulaarne modelleerimine, Keemia-informaatika molekulaardisain, Kemo-informaatika, QSPR/QSAR ja ravimite väljatöötamine

Teenistuskäik:

- Marie Curie Actions meetme stipendiaat, Molcode Ltd. EU FP7 programm, 2014.
- FMTDK (Funktionaalsed materjalid ja tehnoloogiad) stipendiaat, Euroopa Sotsiaalfond (1.2.0401.09-0079), 2013-2014.
- Uurimistöö projekti rahastus, USA põllumajandusosakond – ARS, Florida. USA. 2011-2014.
- University First Rank, MSc. Bioinformaatika, Bharathidasan Ülikool, India. 2006.
- Malaaria projekti stipendium, Tamilnadu State Council for Science & Technology, India. 2005.
- Parima postri auhind – FloHet konverents, USA and BioDesign India.

Patendid

Oliferenko, P.; Oliferenko, A.; Ibrahim, Mohamed, A.; Panda, Siva, S.; Katritzky, Alan, R.; **Pillai, G.G.**; Panda, Chandramukhi, S. Macrocyclic Peptidomimetics with Nanomolar Antifungal And Antimicrobial Activity. WO/2014/182839, November 14, 2014.

Publikatsioonide loetelu:

1. Dearden, J.; Hewitt, M.; Roberts, D.; Enoch, S.; Rowe, P.; Przybylak, K.; Vaughan-Williams, D.; Smith, M.; **Pillai, G.G**; Katritzky, A. Mechanism-based QSAR modeling of skin sensitization. *ACS Chem. Res. Tox.* **2015**, (In Press).
2. Wang, Z.; Zhang, H.; Killian, B.J.; Jabeen, F.; **Pillai, G.G**.; Berman, H.M.; Mathelier, M.; Sibble, A.J.; Yeung, J.; Zhou, W.; Steel, P.J.; Hall, C.D.; Katritzky, A. R. Synthesis, Characterization and Energetic Properties of 1,3,4-Oxadiazoles. *Eur. J Org. Chem.* **2015**, DOI: 10.1002/ejoc.201500583.
3. Girgis, A. S.; Panda, S. S.; Farag, I. S. A.; El-Shabiny, A. M.; Moustafa, A. M.; Ismail, N. S. M.; **Pillai, G. G.**; Panda, C. S.; Hall, C. D.; Katritzky, A. R. Synthesis, and QSAR Analysis of Anti-Oncological Active Spiro-Alkaloids. *Org. Biomol. Chem.* **2015**, *13* (6), 1741–1753.
4. **Pillai, G. G**.; Sikk, L.; Tamm, T.; Karelson, M.; Burk, P.; Tämm, K.; Tamm, K. Theoretical Modeling of HPV: QSAR and Novodesign with Fragment Approach. *Curr. Comput. Aided. Drug Des.* **2014**, *10* (4), 303–314.
5. Ha, K.; Lebedyeva, I.; Hamedzadeh, S.; Li, Z.; Quiñones, R.; **Pillai, G. G**.; Williams, B.; Nasajpour, A.; Martin, K.; Asiri, A. M.; Katritzky, A. R. Tandem Deprotection-Dimerization-Macrocyclization Route to C(2) Symmetric Cyclo-Tetrapeptides. *Chem. - A Eur. J.* **2014**, *20* (17), 4874–4879. *(Frontispiece)*
6. Girgis, A. S.; Saleh, D. O.; George, R. F.; Srour, A. M.; **Pillai, G. G**.; Panda, C. S.; Katritzky, A. R. Synthesis, Bioassay, and QSAR Study of

Bronchodilatory Active 4H-pyrano[3,2-C]pyridine-3-Carbonitriles. *Eur. J. Med. Chem.* **2014**, *89*, 835–843.
7. Jabeen, F.; Oliferenko, P. V.; Oliferenko, A. A.; **Pillai, G. G**.; Ansari, F. L.; Hall, C. D.; Katritzky, A. R. Dual Inhibition of the A-Glucosidase and Butyrylcholinesterase Studied by Molecular Field Topology Analysis. *Eur. J. Med. Chem.* **2014**, *80*, 228–242.
8. Oliferenko, P. V; Oliferenko, A. A.; Girgis, A. S.; Saleh, D. O.; Srour, A. M.; George, R. F.; **Pillai, G. G**.; Panda, C. S.; Katritzky, A. R.; Hall, C. D. Synthesis, Bioassay, and Molecular Field Topology Analysis of Diverse Vasodilatory Heterocycles. *J. Chem. Inf. Model.* **2014**, *54* (4), 1103–1116.
9. Biswas, S.; Kayaleh, R.; **Pillai, G. G**.; Seon, C.; Roberts, I.; Popov, V.; Alamry, K. A.; Katritzky, A. R. Long-Range Chemical Ligation from N→N Acyl Migrations in Tryptophan Peptides via Cyclic Transition States of 10- to 18-Members. *Chemistry* **2014**, *20* (26), 8189–8198.
10. Li, Z.; Lebedyeva, I.; Zhao, D.; Myers, L.; **Pillai, G. G**.; Hall, C. D.; Katritzky, A. R. Synthesis of L-Lys-Aminoxy-Goralatide. *J. Pept. Sci.* **2014**, *20* (12), 923–927.
11. Dobchev, D. A.; **Pillai, G. G**.; Karelson, M. In Silico Machine Learning Methods in Drug Development. *Curr. Top. Med. Chem.* **2014**, *14* (16), 1913–1922.
12. Oliferenko, P. V.; Oliferenko, A. A.; Poda, G. I.; Osolodkin, D. I.; **Pillai, G. G**.; Bernier, U. R.; Tsikolia, M.; Agramonte, N. M.; Clark, G. G.; Linthicum, K. J.; Katritzky, A. R.; Polina V. Oliferenko Alexander A. Oliferenko, A. R. K. Promising Aedes Aegypti Repellent Chemotypes Identified through Integrated QSAR, Virtual Screening, Synthesis, and Bioassay. *PLoS One* **2013**, *8* (9), e64547.
13. El-Gendy, B. E. D. M.; Ghazvini Zadeh, E. H.; Sotuyo, A. C.; **Pillai, G. G**.; Katritzky, A. R. α-Substitution Effects on the Ease of S→ N-Acyl Transfer in Aminothioesters. *Chem. Biol. Drug Des.* **2013**, *81* (5), 577–582.
14. Ha, K.; Lebedyeva, I.; Li, Z.; Martin, K.; Williams, B.; Faby, E.; Nasajpour, A.; **Pillai, G. G**.; Al-Youbi, A. O.; Katritzky, A. R. Conformationally Assisted Lactamizations for the Synthesis of Symmetrical and Unsymmetrical Bis-2,5-Diketopiperazines. *J. Org. Chem.* **2013**, *78* (17), 8510–8523.
15. Ha, K.; Monbaliu, J.-C. M.; Williams, B. C.; **Pillai, G. G**.; Ocampo, C. E.; Zeller, M.; Stevens, C. V; Katritzky, A. R. A Convenient Synthesis of Difficult Medium-Sized Cyclic Peptides by Staudinger Mediated Ring-Closure. *Organic & Biomolecular Chemistry*, **2012**, *10*, 8055.
16. Berhanu, W. M.; **Pillai, G. G**.; Oliferenko, A. A.; Katritzky, A. R. Quantitative Structure-Activity/property Relationships: The Ubiquitous Links between Cause and Effect. *ChemPlusChem*, **2012**, *77*, 507–517.
17. Singh, N. K.; Pakkkianathan, B. C.; Kumar, M.; Daddam, J. R.; Jayavel, S.; Kannan, M.; **Pillai, G. G**.; Krishnan, M. Computational Studies on Molecular Interactions of 6-Thioguanosine Analogs with Anthrax Toxin Receptor 1. *Interdiscip. Sci.* **2012**, *4* (3), 183–189.

18. Berhanu, W. M.; Ibrahim, M. A.; **Pillai, G. G**.; Oliferenko, A. A.; Khelashvili, L.; Jabeen, F.; Mirza, B.; Ansari, F. L.; Ul-Haq, I.; El-Feky, S. A.; Katritzky, A. R. Supporting Information for Similarity Analysis, Synthesis, and Bioassay of Antibacterial Cyclic Peptidomimetics. *Beilstein J. Org. Chem.* **2012**, *8* (1), 1146–1160.
19. Berhanu, W. M.; Ibrahim, M. A.; **Pillai, G. G**.; Oliferenko, A. A.; Khelashvili, L.; Jabeen, F.; Mirza, B.; Ansari, F. L.; Ul-Haq, I.; El-Feky, S. A.; Katritzky, A. R. Similarity Analysis, Synthesis, and Bioassay of Antibacterial Cyclic Peptidomimetics. *Beilstein J. Org. Chem.* **2012**, *8*, 1146–1160. *(Frontispiece)*
20. Zhong, M.; Jang, M.; Oliferenko, A.; **Pillai, G. G**.; Katritzky, A. R. The SOA Formation Model Combined with Semiempirical Quantum Chemistry for Predicting UV-Vis Absorption of Secondary Organic Aerosols. *Phys. Chem. Chem. Phys.* **2012**, *14* (25), 9058–9066.

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS

1. **Toomas Tamm.** Quantum-chemical simulation of solvent effects. Tartu, 1993, 110 p.
2. **Peeter Burk.** Theoretical study of gas-phase acid-base equilibria. Tartu, 1994, 96 p.
3. **Victor Lobanov.** Quantitative structure-property relationships in large descriptor spaces. Tartu, 1995, 135 p.
4. **Vahur Mäemets.** The ^{17}O and ^{1}H nuclear magnetic resonance study of H_2O in individual solvents and its charged clusters in aqueous solutions of electrolytes. Tartu, 1997, 140 p.
5. **Andrus Metsala.** Microcanonical rate constant in nonequilibrium distribution of vibrational energy and in restricted intramolecular vibrational energy redistribution on the basis of slater's theory of unimolecular reactions. Tartu, 1997, 150 p.
6. **Uko Maran.** Quantum-mechanical study of potential energy surfaces in different environments. Tartu, 1997, 137 p.
7. **Alar Jänes.** Adsorption of organic compounds on antimony, bismuth and cadmium electrodes. Tartu, 1998, 219 p.
8. **Kaido Tammeveski.** Oxygen electroreduction on thin platinum films and the electrochemical detection of superoxide anion. Tartu, 1998, 139 p.
9. **Ivo Leito.** Studies of Brønsted acid-base equilibria in water and non-aqueous media. Tartu, 1998, 101 p.
10. **Jaan Leis.** Conformational dynamics and equilibria in amides. Tartu, 1998, 131 p.
11. **Toonika Rinken.** The modelling of amperometric biosensors based on oxidoreductases. Tartu, 2000, 108 p.
12. **Dmitri Panov.** Partially solvated Grignard reagents. Tartu, 2000, 64 p.
13. **Kaja Orupõld.** Treatment and analysis of phenolic wastewater with microorganisms. Tartu, 2000, 123 p.
14. **Jüri Ivask.** Ion Chromatographic determination of major anions and cations in polar ice core. Tartu, 2000, 85 p.
15. **Lauri Vares.** Stereoselective Synthesis of Tetrahydrofuran and Tetrahydropyran Derivatives by Use of Asymmetric Horner-Wadsworth-Emmons and Ring Closure Reactions. Tartu, 2000, 184 p.
16. **Martin Lepiku.** Kinetic aspects of dopamine D_2 receptor interactions with specific ligands. Tartu, 2000, 81 p.
17. **Katrin Sak.** Some aspects of ligand specificity of P2Y receptors. Tartu, 2000, 106 p.
18. **Vello Pällin.** The role of solvation in the formation of iotsitch complexes. Tartu, 2001, 95 p.

19. **Katrin Kollist.** Interactions between polycyclic aromatic compounds and humic substances. Tartu, 2001, 93 p.
20. **Ivar Koppel.** Quantum chemical study of acidity of strong and superstrong Brønsted acids. Tartu, 2001, 104 p.
21. **Viljar Pihl.** The study of the substituent and solvent effects on the acidity of OH and CH acids. Tartu, 2001, 132 p.
22. **Natalia Palm.** Specification of the minimum, sufficient and significant set of descriptors for general description of solvent effects. Tartu, 2001, 134 p.
23. **Sulev Sild.** QSPR/QSAR approaches for complex molecular systems. Tartu, 2001, 134 p.
24. **Ruslan Petrukhin.** Industrial applications of the quantitative structure-property relationships. Tartu, 2001, 162 p.
25. **Boris V. Rogovoy.** Synthesis of (benzotriazolyl)carboximidamides and their application in relations with *N*- and *S*-nucleophyles. Tartu, 2002, 84 p.
26. **Koit Herodes.** Solvent effects on UV-vis absorption spectra of some solvatochromic substances in binary solvent mixtures: the preferential solvation model. Tartu, 2002, 102 p.
27. **Anti Perkson.** Synthesis and characterisation of nanostructured carbon. Tartu, 2002, 152 p.
28. **Ivari Kaljurand.** Self-consistent acidity scales of neutral and cationic Brønsted acids in acetonitrile and tetrahydrofuran. Tartu, 2003, 108 p.
29. **Karmen Lust.** Adsorption of anions on bismuth single crystal electrodes. Tartu, 2003, 128 p.
30. **Mare Piirsalu.** Substituent, temperature and solvent effects on the alkaline hydrolysis of substituted phenyl and alkyl esters of benzoic acid. Tartu, 2003, 156 p.
31. **Meeri Sassian.** Reactions of partially solvated Grignard reagents. Tartu, 2003, 78 p.
32. **Tarmo Tamm.** Quantum chemical modelling of polypyrrole. Tartu, 2003. 100 p.
33. **Erik Teinemaa.** The environmental fate of the particulate matter and organic pollutants from an oil shale power plant. Tartu, 2003. 102 p.
34. **Jaana Tammiku-Taul.** Quantum chemical study of the properties of Grignard reagents. Tartu, 2003. 120 p.
35. **Andre Lomaka.** Biomedical applications of predictive computational chemistry. Tartu, 2003. 132 p.
36. **Kostyantyn Kirichenko.** Benzotriazole – Mediated Carbon–Carbon Bond Formation. Tartu, 2003. 132 p.
37. **Gunnar Nurk.** Adsorption kinetics of some organic compounds on bismuth single crystal electrodes. Tartu, 2003, 170 p.
38. **Mati Arulepp.** Electrochemical characteristics of porous carbon materials and electrical double layer capacitors. Tartu, 2003, 196 p.

39. **Dan Cornel Fara.** QSPR modeling of complexation and distribution of organic compounds. Tartu, 2004, 126 p.
40. **Riina Mahlapuu.** Signalling of galanin and amyloid precursor protein through adenylate cyclase. Tartu, 2004, 124 p.
41. **Mihkel Kerikmäe.** Some luminescent materials for dosimetric applications and physical research. Tartu, 2004, 143 p.
42. **Jaanus Kruusma.** Determination of some important trace metal ions in human blood. Tartu, 2004, 115 p.
43. **Urmas Johanson.** Investigations of the electrochemical properties of polypyrrole modified electrodes. Tartu, 2004, 91 p.
44. **Kaido Sillar.** Computational study of the acid sites in zeolite ZSM-5. Tartu, 2004, 80 p.
45. **Aldo Oras.** Kinetic aspects of dATPαS interaction with $P2Y_1$ receptor. Tartu, 2004, 75 p.
46. **Erik Mölder.** Measurement of the oxygen mass transfer through the air-water interface. Tartu, 2005, 73 p.
47. **Thomas Thomberg.** The kinetics of electroreduction of peroxodisulfate anion on cadmium (0001) single crystal electrode. Tartu, 2005, 95 p.
48. **Olavi Loog.** Aspects of condensations of carbonyl compounds and their imine analogues. Tartu, 2005, 83 p.
49. **Siim Salmar.** Effect of ultrasound on ester hydrolysis in aqueous ethanol. Tartu, 2006, 73 p.
50. **Ain Uustare.** Modulation of signal transduction of heptahelical receptors by other receptors and G proteins. Tartu, 2006, 121 p.
51. **Sergei Yurchenko.** Determination of some carcinogenic contaminants in food. Tartu, 2006, 143 p.
52. **Kaido Tämm.** QSPR modeling of some properties of organic compounds. Tartu, 2006, 67 p.
53. **Olga Tšubrik.** New methods in the synthesis of multisubstituted hydrazines. Tartu. 2006, 183 p.
54. **Lilli Sooväli.** Spectrophotometric measurements and their uncertainty in chemical analysis and dissociation constant measurements. Tartu, 2006, 125 p.
55. **Eve Koort.** Uncertainty estimation of potentiometrically measured ph and pK_a values. Tartu, 2006, 139 p.
56. **Sergei Kopanchuk.** Regulation of ligand binding to melanocortin receptor subtypes. Tartu, 2006, 119 p.
57. **Silvar Kallip.** Surface structure of some bismuth and antimony single crystal electrodes. Tartu, 2006, 107 p.
58. **Kristjan Saal.** Surface silanization and its application in biomolecule coupling. Tartu, 2006, 77 p.
59. **Tanel Tätte.** High viscosity $Sn(OBu)_4$ oligomeric concentrates and their applications in technology. Tartu, 2006, 91 p.

60. **Dimitar Atanasov Dobchev**. Robust QSAR methods for the prediction of properties from molecular structure. Tartu, 2006, 118 p.
61. **Hannes Hagu**. Impact of ultrasound on hydrophobic interactions in solutions. Tartu, 2007, 81 p.
62. **Rutha Jäger.** Electroreduction of peroxodisulfate anion on bismuth electrodes. Tartu, 2007, 142 p.
63. **Kaido Viht.** Immobilizable bisubstrate-analogue inhibitors of basophilic protein kinases: development and application in biosensors. Tartu, 2007, 88 p.
64. **Eva-Ingrid Rõõm.** Acid-base equilibria in nonpolar media. Tartu, 2007, 156 p.
65. **Sven Tamp.** DFT study of the cesium cation containing complexes relevant to the cesium cation binding by the humic acids. Tartu, 2007, 102 p.
66. **Jaak Nerut.** Electroreduction of hexacyanoferrate(III) anion on Cadmium (0001) single crystal electrode. Tartu, 2007, 180 p.
67. **Lauri Jalukse.** Measurement uncertainty estimation in amperometric dissolved oxygen concentration measurement. Tartu, 2007, 112 p.
68. **Aime Lust.** Charge state of dopants and ordered clusters formation in CaF_2:Mn and CaF_2:Eu luminophors. Tartu, 2007, 100 p.
69. **Iiris Kahn**. Quantitative Structure-Activity Relationships of environmentally relevant properties. Tartu, 2007, 98 p.
70. **Mari Reinik.** Nitrates, nitrites, N-nitrosamines and polycyclic aromatic hydrocarbons in food: analytical methods, occurrence and dietary intake. Tartu, 2007, 172 p.
71. **Heili Kasuk.** Thermodynamic parameters and adsorption kinetics of organic compounds forming the compact adsorption layer at Bi single crystal electrodes. Tartu, 2007, 212 p.
72. **Erki Enkvist.** Synthesis of adenosine-peptide conjugates for biological applications. Tartu, 2007, 114 p.
73. **Svetoslav Hristov Slavov**. Biomedical applications of the QSAR approach. Tartu, 2007, 146 p.
74. **Eneli Härk.** Electroreduction of complex cations on electrochemically polished Bi(*hkl*) single crystal electrodes. Tartu, 2008, 158 p.
75. **Priit Möller.** Electrochemical characteristics of some cathodes for medium temperature solid oxide fuel cells, synthesized by solid state reaction technique. Tartu, 2008, 90 p.
76. **Signe Viggor.** Impact of biochemical parameters of genetically different pseudomonads at the degradation of phenolic compounds. Tartu, 2008, 122 p.
77. **Ave Sarapuu.** Electrochemical reduction of oxygen on quinone-modified carbon electrodes and on thin films of platinum and gold. Tartu, 2008, 134 p.
78. **Agnes Kütt.** Studies of acid-base equilibria in non-aqueous media. Tartu, 2008, 198 p.

79. **Rouvim Kadis.** Evaluation of measurement uncertainty in analytical chemistry: related concepts and some points of misinterpretation. Tartu, 2008, 118 p.
80. **Valter Reedo.** Elaboration of IVB group metal oxide structures and their possible applications. Tartu, 2008, 98 p.
81. **Aleksei Kuznetsov.** Allosteric effects in reactions catalyzed by the cAMP-dependent protein kinase catalytic subunit. Tartu, 2009, 133 p.
82. **Aleksei Bredihhin.** Use of mono- and polyanions in the synthesis of multisubstituted hydrazine derivatives. Tartu, 2009, 105 p.
83. **Anu Ploom.** Quantitative structure-reactivity analysis in organosilicon chemistry. Tartu, 2009, 99 p.
84. **Argo Vonk.** Determination of adenosine A_{2A}- and dopamine D_1 receptor-specific modulation of adenylate cyclase activity in rat striatum. Tartu, 2009, 129 p.
85. **Indrek Kivi.** Synthesis and electrochemical characterization of porous cathode materials for intermediate temperature solid oxide fuel cells. Tartu, 2009, 177 p.
86. **Jaanus Eskusson.** Synthesis and characterisation of diamond-like carbon thin films prepared by pulsed laser deposition method. Tartu, 2009, 117 p.
87. **Marko Lätt.** Carbide derived microporous carbon and electrical double layer capacitors. Tartu, 2009, 107 p.
88. **Vladimir Stepanov.** Slow conformational changes in dopamine transporter interaction with its ligands. Tartu, 2009, 103 p.
89. **Aleksander Trummal.** Computational Study of Structural and Solvent Effects on Acidities of Some Brønsted Acids. Tartu, 2009, 103 p.
90. **Eerold Vellemäe.** Applications of mischmetal in organic synthesis. Tartu, 2009, 93 p.
91. **Sven Parkel.** Ligand binding to 5-HT_{1A} receptors and its regulation by Mg^{2+} and Mn^{2+}. Tartu, 2010, 99 p.
92. **Signe Vahur.** Expanding the possibilities of ATR-FT-IR spectroscopy in determination of inorganic pigments. Tartu, 2010, 184 p.
93. **Tavo Romann**. Preparation and surface modification of bismuth thin film, porous, and microelectrodes. Tartu, 2010, 155 p.
94. **Nadežda Aleksejeva.** Electrocatalytic reduction of oxygen on carbon nanotube-based nanocomposite materials. Tartu, 2010, 147 p.
95. **Marko Kullapere.** Electrochemical properties of glassy carbon, nickel and gold electrodes modified with aryl groups. Tartu, 2010, 233 p.
96. **Liis Siinor.** Adsorption kinetics of ions at Bi single crystal planes from aqueous electrolyte solutions and room-temperature ionic liquids. Tartu, 2010, 101 p.
97. **Angela Vaasa.** Development of fluorescence-based kinetic and binding assays for characterization of protein kinases and their inhibitors. Tartu 2010, 101 p.

98. **Indrek Tulp.** Multivariate analysis of chemical and biological properties. Tartu 2010, 105 p.
99. **Aare Selberg.** Evaluation of environmental quality in Northern Estonia by the analysis of leachate. Tartu 2010, 117 p.
100. **Darja Lavõgina.** Development of protein kinase inhibitors based on adenosine analogue-oligoarginine conjugates. Tartu 2010, 248 p.
101. **Laura Herm.** Biochemistry of dopamine D_2 receptors and its association with motivated behaviour. Tartu 2010, 156 p.
102. **Terje Raudsepp.** Influence of dopant anions on the electrochemical properties of polypyrrole films. Tartu 2010, 112 p.
103. **Margus Marandi.** Electroformation of Polypyrrole Films: *In-situ* AFM and STM Study. Tartu 2011, 116 p.
104. **Kairi Kivirand.** Diamine oxidase-based biosensors: construction and working principles. Tartu, 2011, 140 p.
105. **Anneli Kruve.** Matrix effects in liquid-chromatography electrospray mass-spectrometry. Tartu, 2011, 156 p.
106. **Gary Urb.** Assessment of environmental impact of oil shale fly ash from PF and CFB combustion. Tartu, 2011, 108 p.
107. **Nikita Oskolkov.** A novel strategy for peptide-mediated cellular delivery and induction of endosomal escape. Tartu, 2011, 106 p.
108. **Dana Martin.** The QSPR/QSAR approach for the prediction of properties of fullerene derivatives. Tartu, 2011, 98 p.
109. **Säde Viirlaid.** Novel glutathione analogues and their antioxidant activity. Tartu, 2011, 106 p.
110. **Ülis Sõukand.** Simultaneous adsorption of Cd^{2+}, Ni^{2+}, and Pb^{2+} on peat. Tartu, 2011, 124 p.
111. **Lauri Lipping.** The acidity of strong and superstrong Brønsted acids, an outreach for the "limits of growth": a quantum chemical study. Tartu, 2011, 124 p.
112. **Heisi Kurig.** Electrical double-layer capacitors based on ionic liquids as electrolytes. Tartu, 2011, 146 p.
113. **Marje Kasari.** Bisubstrate luminescent probes, optical sensors and affinity adsorbents for measurement of active protein kinases in biological samples. Tartu, 2012, 126 p.
114. **Kalev Takkis.** Virtual screening of chemical databases for bioactive molecules. Tartu, 2012, 122 p.
115. **Ksenija Kisseljova**. Synthesis of aza-β^3-amino acid containing peptides and kinetic study of their phosphorylation by protein kinase A. Tartu, 2012, 104 p.
116. **Riin Rebane.** Advanced method development strategy for derivatization LC/ESI/MS. Tartu, 2012, 184 p.

117. **Vladislav Ivaništšev.** Double layer structure and adsorption kinetics of ions at metal electrodes in room temperature ionic liquids. Tartu, 2012, 128 p.
118. **Irja Helm.** High accuracy gravimetric Winkler method for determination of dissolved oxygen. Tartu, 2012, 139 p.
119. **Karin Kipper.** Fluoroalcohols as Components of LC-ESI-MS Eluents: Usage and Applications. Tartu, 2012, 164 p.
120. **Arno Ratas.** Energy storage and transfer in dosimetric luminescent materials. Tartu, 2012, 163 p.
121. **Reet Reinart-Okugbeni**. Assay systems for characterisation of subtype-selective binding and functional activity of ligands on dopamine receptors. Tartu, 2012, 159 p.
122. **Lauri Sikk.** Computational study of the Sonogashira cross-coupling reaction. Tartu, 2012, 81 p.
123. **Karita Raudkivi.** Neurochemical studies on inter-individual differences in affect-related behaviour of the laboratory rat. Tartu, 2012, 161 p.
124. **Indrek Saar.** Design of GalR2 subtype specific ligands: their role in depression-like behavior and feeding regulation. Tartu, 2013, 126 p.
125. **Ann Laheäär.** Electrochemical characterization of alkali metal salt based non-aqueous electrolytes for supercapacitors. Tartu, 2013, 127 p.
126. **Kerli Tõnurist.** Influence of electrospun separator materials properties on electrochemical performance of electrical double-layer capacitors. Tartu, 2013, 147 p.
127. **Kaija Põhako-Esko.** Novel organic and inorganic ionogels: preparation and characterization. Tartu, 2013, 124 p.
128. **Ivar Kruusenberg.** Electroreduction of oxygen on carbon nanomaterial-based catalysts. Tartu, 2013, 191 p.
129. **Sander Piiskop.** Kinetic effects of ultrasound in aqueous acetonitrile solutions. Tartu, 2013, 95 p.
130. **Ilona Faustova**. Regulatory role of L-type pyruvate kinase N-terminal domain. Tartu, 2013, 109 p.
131. **Kadi Tamm.** Synthesis and characterization of the micro-mesoporous anode materials and testing of the medium temperature solid oxide fuel cell single cells. Tartu, 2013, 138 p.
132. **Iva Bozhidarova Stoyanova-Slavova.** Validation of QSAR/QSPR for regulatory purposes. Tartu, 2013, 109 p.
133. **Vitali Grozovski**. Adsorption of organic molecules at single crystal electrodes studied by *in situ* STM method. Tartu, 2014, 146 p.
134. **Santa Veikšina**. Development of assay systems for characterisation of ligand binding properties to melanocortin 4 receptors. Tartu, 2014, 151 p.
135. **Jüri Liiv.** PVDF (polyvinylidene difluoride) as material for active element of twisting-ball displays. Tartu, 2014, 111 p.

136. **Kersti Vaarmets.** Electrochemical and physical characterization of pristine and activated molybdenum carbide-derived carbon electrodes for the oxygen electroreduction reaction. Tartu, 2014, 131 p.
137. **Lauri Tõntson**. Regulation of G-protein subtypes by receptors, guanine nucleotides and Mn^{2+}. Tartu, 2014, 105 p.
138. **Aiko Adamson.** Properties of amine-boranes and phosphorus analogues in the gas phase. Tartu, 2014, 78 p.
139. **Elo Kibena**. Electrochemical grafting of glassy carbon, gold, highly oriented pyrolytic graphite and chemical vapour deposition-grown graphene electrodes by diazonium reduction method. Tartu, 2014, 184 p.
140. **Teemu Näykki.** Novel Tools for Water Quality Monitoring – From Field to Laboratory. Tartu, 2014, 202 p.
141. **Karl Kaupmees.** Acidity and basicity in non-aqueous media: importance of solvent properties and purity. Tartu, 2014, 128 p.
142. **Oleg Lebedev**. Hydrazine polyanions: different strategies in the synthesis of heterocycles. Tartu, 2015, 118 p.
143. **Geven Piir.** Environmental risk assessment of chemicals using QSAR methods. Tartu, 2015, 123 p.
144. **Olga Mazina.** Development and application of the biosensor assay for measurements of cyclic adenosine monophosphate in studies of G protein-coupled receptor signalinga. Tartu, 2015, 116 p.
145. **Sandip Ashokrao Kadam.** Anion receptors: synthesis and accurate binding measurements. Tartu, 2015, 116 p.
146. **Indrek Tallo.** Synthesis and characterization of new micro-mesoporous carbide derived carbon materials for high energy and power density electrical double layer capacitors. Tartu, 2015, 148 p.
147. **Heiki Erikson.** Electrochemical reduction of oxygen on nanostructured palladium and gold catalysts. Tartu, 2015, 204 p.
148. **Erik Anderson.** *In situ* Scanning Tunnelling Microscopy studies of the interfacial structure between Bi(111) electrode and a room temperature ionic liquid. Tartu, 2015, 118 p.