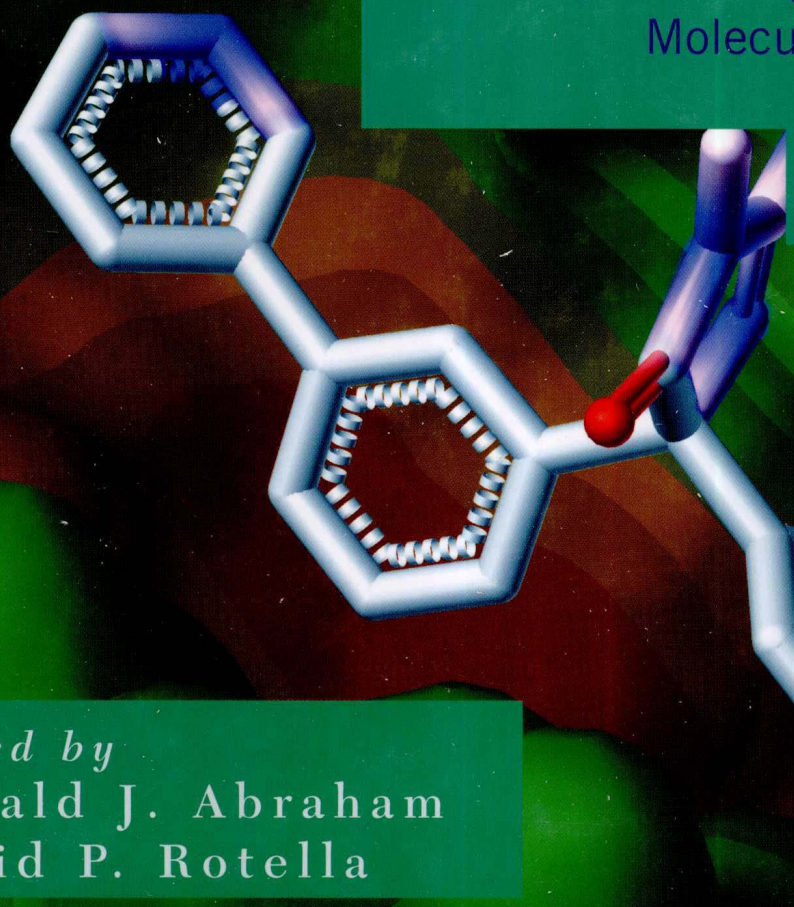
SEVENTH EDITION

BURGER'S MEDICINAL CHEMISTRY, DRUG DISCOVERY, AND DEVELOPMENT

Discovering Lead Molecules

Volume 2



Edited by
Donald J. Abraham
David P. Rotella

WILEY

BURGER'S MEDICINAL CHEMISTRY, DRUG DISCOVERY AND DEVELOPMENT

Seventh Edition

Volume 2: Discovery Lead Molecules

Edited by

Donald J. Abraham
Virginia Commonwealth University

David P. Rotella
Wyeth Research

Burger's Medicinal Chemistry, Drug Discovery and Development
is available Online in full color at
http://mrw.interscience.wiley.com/emrw/9780471266945/home/

WILEY
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Abraham, Donald J., 1936-
Burger's medicinal chemistry, drug discovery, and development/Donald J. Abraham, David P. Rotella. – 7th ed.
p. ; cm.
Other title: Medicinal chemistry, drug discovery, and development
Rev. ed. of: Burger's medicinal chemistry and drug discovery. 6th ed. / edited by Donald J. Abraham. c2003.
Includes bibliographical references and index.
ISBN 978-0-470-27815-4 (cloth)
1. Pharmaceutical chemistry. 2. Drug development. I. Rotella, David P. II. Burger, Alfred, 1905-2000. III. Burger's medicinal chemistry and drug discovery. IV. Title. V. Title: Medicinal chemistry, drug discovery, and development.
[DNLM: 1. Chemistry, Pharmaceutical–methods. 2. Biopharmaceutics–methods. 3. Drug Compounding–methods. QV 744 A105b 2010]
RS403.B8 2010
615'.19–dc22 2010010779

Printed in Singapore

10 9 8 7 6 5 4 3 2 1

PREFACE

The seventh edition of Burger's Medicinal Chemistry resulted from a collaboration established between John Wiley & Sons, the editorial board, authors, and coeditors over the last 3 years. The editorial board for the seventh edition provided important advice to the editors on topics and contributors. Wiley staff effectively handled the complex tasks of manuscript production and editing and effectively tracked the process from beginning to end. Authors provided well-written, comprehensive summaries of their topics and responded to editorial requests in a timely manner. This edition, with 8 volumes and 116 chapters, like the previous editions, is a reflection of the expanding complexity of medicinal chemistry and associated disciplines. Separate volumes have been added on anti-infectives, cancer, and the process of drug development. In addition, the coeditors elected to expand coverage of cardiovascular and metabolic disorders, aspects of CNS-related medicinal chemistry, and computational drug discovery. This provided the opportunity to delve into many subjects in greater detail and resulted in specific chapters on important subjects such as biologics and protein drug discovery, HIV, new diabetes drug targets, amyloid-based targets for treatment of Alzheimer's disease, high-throughput and other screening methods, and the key role played by metabolism and other pharmacokinetic properties in drug development.

The following individuals merit special thanks for their contributions to this complex endeavor: Surlan Alexander of John Wiley & Sons for her organizational skills and attention to detail, Sanchari Sil of Thomson Digital for processing the galley proofs, Jonathan Mason of Lundbeck, Andrea Mozzarelli of the University of Parma, Alex Tropsha of the University of North Carolina, John Block of Oregon State University, Paul Reider of Princeton University, William (Rick) Ewing of Bristol-Myers Squibb, William Hagmann of Merck, John Primeau and Rob Bradbury of AstraZeneca, Bryan Norman of Eli Lilly, Al Robichaud of Wyeth, and John Lowe for their input on topics and potential authors. The many reviewers for these chapters deserve special thanks for the constructive comments they provided to authors. Finally, we must express gratitude to our lovely, devoted wives, Nancy and Mary Beth, for their tolerance as we spent time with this task, rather than with them.

As coeditors, we sincerely hope that this edition meets the high expectations of the scientific community. We assembled this edition with the guiding vision of its namesake in mind and would like to dedicate it to Professor H.C. Brown and Professor Donald T. Witiak. Don collaborated with Dr. Witiak in the early days of his research in sickle cell drug discovery. Professor Witiak was Dave's doctoral advisor at Ohio State University and provided essential guidance to a young

scientist. Professor Brown, whose love for chemistry infected all organic graduate students at Purdue University, arranged for Don to become a medicinal chemist by securing a postdoctoral position for him with Professor Alfred Burger.

It has been a real pleasure to work with all concerned to assemble an outstanding and up-to-date edition in this series.

DONALD J. ABRAHAM
DAVID P. ROTELLA

March 2010

CONTRIBUTORS

Adriano D. Andricopulo, Universidade de São Paulo, São Carlos-SP, Brazil
Valerie S. Bernan, Wyeth Research, Pearl River, NY
Julian Blagg, Haddow Laboratories, Sutton, Surrey, UK
Frank E. Blaney, GlaxoSmithKline plc, Harlow, Essex, UK
John H. Block, Oregon State University, Corvallis, OR
Nicholas Bodor, University of Florida, Gainesville, FL
Peter Buchwald, University of Miami, Miami, FL
Guy T. Carter, Wyeth Research, Princeton, NJ; Wyeth Research, Pearl River, NY
Richard J. Clark, The University of Queensland, Brisbane, Australia
Pietro Cozzini, University of Parma, Parma, Italy; Italian National Institute of Biostructures and Biosystems, Rome, Italy
Gordon M. Cragg, NCI-Frederick, Frederick, MD
David J. Craik, The University of Queensland, Brisbane, Australia
Li Di, Wyeth Research, Princeton, NJ
Alexander Dömling, University of Pittsburgh, Pittsburgh, PA
Emir Duzic, Discovery Research Cephalon, Inc., West Chester, PA
Sean Ekins, Collaborations in Chemistry, Jenkintown, PA; University of Maryland, Baltimore, MD; University of Medicine and Dentistry, Piscataway, NJ; Collaborative Drug Discovery, Burlingame, CA
Steven L. Garland, GlaxoSmithKline plc, Harlow, Essex, UK
Holger Gohlke, Heinrich-Heine-University, Düsseldorf, Germany
Rafael V.C. Guido, Universidade de São Paulo, São Carlos-SP, Brazil
J. Dana Honeycutt, Accelrys, San Diego, CA
Edgar Jacoby, Novartis Institutes for BioMedical Research, Basel, Switzerland
Glen Eugene Kellogg, Virginia Commonwealth University, Richmond, VA
Edward H. Kerns, Wyeth Research, Princeton, NJ
Frank E. Koehn, Wyeth Research, Pearl River, NY
Hannes Kopitz, Heinrich-Heine-University, Düsseldorf, Germany
Andrei Leitão, University of New Mexico, Albuquerque, NM
Michael J. Marino, Discovery Research Cephalon, Inc., West Chester, PA
James T. Metz, Abbott Laboratories, Abbott Park, IL
Carlos Montanari, Universidade de São Paulo, São Carlos-SP, Brazil
Ingo Muegge, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT
Adrian J. Mulholland, University of Bristol, Bristol, UK
David J. Newman, NCI-Frederick, Frederick, MD
Scott Oloff, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT
Igor Polikarpov, Universidade de São Paulo, São Carlos-SP, Brazil
Iosif-Daniel Rosca, University of Bucharest, Bucharest, Romania; Concordia University, Montreal, Canada
Katherine E. Shaw, University of Bristol, Bristol, UK
Phillippa A. Smith, The University of Queensland, Brisbane, Australia
Francesca Spyrakis, University of Parma, Parma, Italy; Italian National Institute of Biostructures and Biosystems, Rome, Italy
Peter W. Swaan, University of Maryland, Baltimore, MD
Daniela B.B. Trivella, Universidade de São Paulo, São Carlos-SP, Brazil
Jean-Maurice Vergnaud, University of Saint-Etienne, Saint-Chamond, France
Michael Williams, Discovery Research Cephalon, Inc., West Chester, PA
Christopher J. Woods, University of Bristol, Bristol, UK

CONTENTS

VIRTUAL SCREENING

INGO MUEGGE
SCOTT OLOFF
Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT

1. INTRODUCTION

Since the 6th edition of *Burger's Medicinal Chemistry and Drug Discovery* went to print six years ago, the field of virtual screening (VS) has expanded dramatically in scope as well as in breadth of application. Figure 1 attests eloquently to the increased use of VS in medicinal chemistry-related research and drug discovery. In 2008 alone, there are more than 404 articles in journals published by the American Chemical Society that contain the phrase "virtual screening"; 50 of them maintain this phrase in its title (compared to only 6 in 2002). These numbers amount to a >2.5-fold increase in yearly publications related to VS over these 6 years. New areas of interest surfaced during the past few years including a resurgence of fragment-based screening techniques often guided by VS, the virtual screening of trillions of combinatorial compounds, a growing interest in the *in silico* screening of synthetically accessible compounds (also often know as *de novo* design), the increased use of machine learning techniques, the emphasis on scaffold hopping, and combining ligand-based and structure-based approaches in synergy to name a few. In addition to highlighting some of these new developments in VS, we have revised the topics presented in the last edition and include more recent examples for VS successes.

Virtual screening, sometimes also called *in silico* screening, is an established branch of medicinal chemistry that represents a fast and cost-effective tool for computationally screening virtual compound databases in search for novel drug leads. The roots for virtual screening go back to structure-based drug design and molecular modeling. In the 1970s, researchers hoped to find novel drugs designed rationally using a fast growing number of diverse protein structures being solved by X-ray crystallography [1,2] or nuclear magnetic resonance spectroscopy (NMR) [3]. However, only very few drugs resulted from those early efforts. Examples include captopril as an angiotensin-converting enzyme inhibitor [4] and metothrexate as a dihydrofolate reductase inhibitor [5]. The reasons for this somewhat disappointing drug yield lie in the low resolution of the protein structures as well as limitations in compute power and methods. Researchers have often tried to *de novo* design the final drug candidate on a computer screen. The compounds suggested have often been difficult to synthesize; initial failure in exhibiting potency has often resulted in the termination of structure-based projects. At the end of the 1980s, rational drug design techniques became somewhat discredited due to the high failure rate in drug discovery projects.

In the 1990s, drastic changes occurred in the way drugs are discovered in the pharmaceutical industry. High-throughput synthesis [6,7] and screening techniques [8] changed the lead identification process that is now governed not only by large numbers of compounds processed but also by fast prosecution of many putative drug targets in parallel. The characterization of the human genome has resulted in a large number of novel putative drug targets. Improved screening techniques make it also possible to look at the entire gene families, at orphan targets, or at otherwise uncharacterized putative drug targets. In this environment of data explosion, rational design techniques experienced a comeback [9]. While the growing number of solved protein structures at high resolution makes it possible to embark on structure-based design for many drug targets, virtual screening—the computational counterpart to high-throughput screening—has become a particularly successful computational tool for lead finding in drug discovery.

Proprietary screening collections typically hold approximately 10^6 compounds. This is only a tiny fraction of the conceivable drug-like chemical space that is estimated to hold more than 10^{60} compounds [10,11]. Even if this space is reduced to compounds that are comparably easy to access synthetically it is still an astronomical number. Virtual screening attempts to suggest which compounds

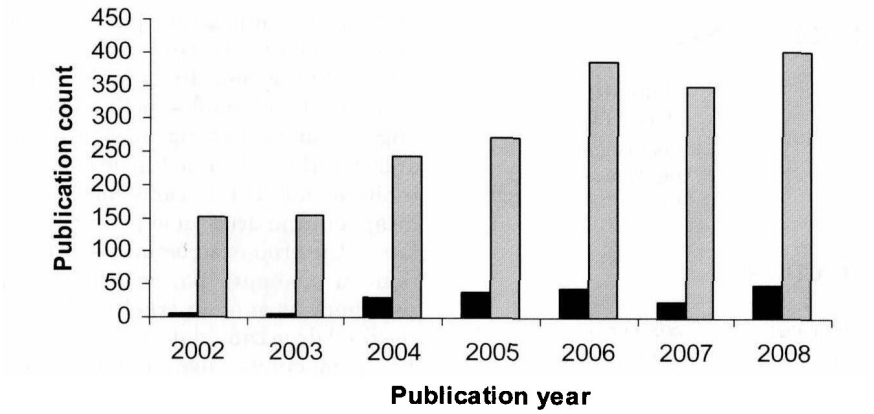


Figure 1. Number of ACS publications with the phrase "virtual screening" in the text (gray patterned bars) or in the title (black bars).

should actually be synthesized or tested against a drug target of interest. Large virtual libraries of up to 10^{12} individual compounds or 10^{14} combinatorial chemistry compounds or 10^{18}*de novo* compounds reassembled from fragments [12] can be screened today using a cascade of various screening tools to reduce the chemical space.

This chapter describes the different concepts and tools used today for virtual screening. They reach from the assessment of the overall "drug-likeness" of a small organic molecule to its ability to specifically bind to a given drug target. Ligand-based as well as structure-based approaches are described and illustrated with specific examples. The ability of virtual screening to assist in scaffold hopping (identifying isofunctional molecular structures with significantly different molecular backbones [13]) is also discussed. Additional topics include the triaging of VS hit sets, the use of machine learning techniques, and finding synergies between different VS approaches. For additional information, the interested reader is referred to a selection of recent books and reviews on the subject of virtual screening [14–25].

2. LIGAND-BASED VIRTUAL SCREENING

Casting a wide net, ligand-based virtual screening (LBVS) encompasses all techniques to computationally screen for novel compounds using ligand information only without including any information of the target structure. LBVS ranges from compound similarity methods for which only a single template molecule needs to be known to very sophisticated pharmacophore elucidation and machine learning methods using a multitude of information on multiple compounds. The interested reader is referred to recent reviews on the topic of similarity searching [21,26], pharmacophore elucidation methods [27,28], and machine learning [29,30] for virtual screening.

2.1. Compound Similarity

Compound similarity as basis for virtual screening relies on the assumption that similar compounds exhibit similar biological activities. The degree to which this is true varies substantially, however. Martin et al. showed from the analysis of a series of HTS experiments at Abbott Laboratories that only approximately 30% of compounds with Daylight fingerprint Tanimoto similarities >0.85 show activity against the same drug target [31]. Compound similarity is generally measured in either structural terms or property terms. Topological fingerprints such as Daylight fingerprints [32] or atom pair descriptors [33], circular fingerprints such as SciTegic fingerprints [34], and structural keys such as ISIS keys [35] encode structural features of the molecules. MolconnZ descriptors [36] present a mixture of molecular connectivity and prop-

erty description, graph theoretical description, and topological description. Properties such as lipophilicity (calculated log P), molecular weight, the number of rotatable bonds, hydrogen-bond acceptors and donors, pK_a values etc. form a group of property descriptors that can link compounds that are structurally quite different. Using such molecular descriptors is quite popular for virtual screening purposes. Especially 2D fingerprints have performed very well in virtual screening [37–41]. However, 3D descriptors have gained in popularity [42–44]. Jenkins et al. showed that 3D pharmacophore feature descriptors called FEPOPS can outperform standard 2D descriptors in a virtual screening protocol [43]. Similarly, Good et al. have demonstrated applying a variety of 2D and 3D descriptors that 3D pharmacophore fingerprints are better suited than 2D descriptors for scaffold hopping [45]. 3D descriptors used in the context of similarity searching often capture pharmacophore features such as hydrophobic or aromatic moieties, hydrogen-bond (HB) acceptors or donors, and both negative and positive ionizable groups. 3D pharmacophore fingerprints store the information of pairs, triplets, or quartets of these pharmacophore features for multiple conformations in the form of binned distance ranges [46–53]. 3D pharmacophore fingerprints are used more often now as descriptors for QSAR as well as for the design of compound libraries [52]. The usefulness of 3D pharmacophore fingerprints in finding new chemotypes through virtual screening has recently been demonstrated [54]. The use of fingerprints to describe chemotypes and biological effects are described in Chapter 12 of Volume 1.

2.1.1. Example 1: Virtual Screening for 15-Lipoxygenase Inhibitors Weinstein et al. reported the identification of a new 15-lipoxygenase inhibitor from the virtual screen of a corporate database of compounds [55]. Starting from the literature compound PD-146176 as a template, 4-point pharmacophore fingerprints [48] were generated for the template compound as well as for all database compounds and then compared. Figure 2 shows the template compound as well as the resulting virtual screening hit. The figure illustrates how structurally different the virtual screening hit is from the template molecule. Some resemblance of the indole moiety is noticeable. This example illustrates the scaffold hopping abilities of 3D pharmacophore fingerprints.

2.2. Pharmacophore Screening

Pharmacophore models are built more often based on small-molecule information. They are a preferred filter tool for virtual screening [56]. Feature-based pharmacophore eluci-

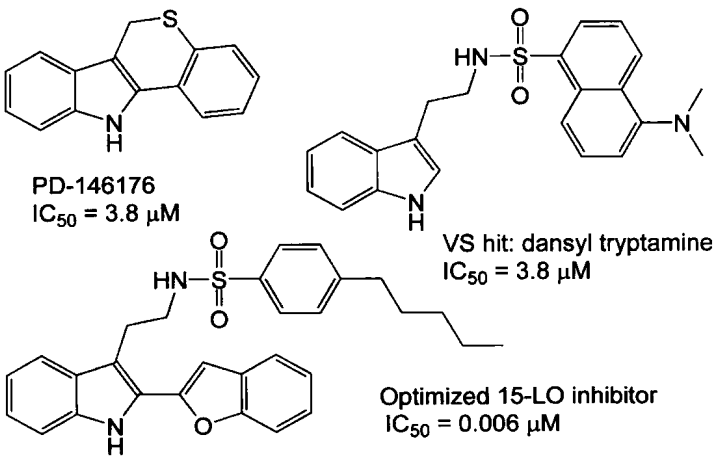


Figure 2. Discovery of a new class of 15-lipoxygenase inhibitors through virtual screening using a 3D pharmacophore fingerprint similarity technique.

dation algorithms such as HipHop [57] and other automated pharmacophore query builders [58] allow for the fast generation of pharmacophore hypotheses. Today sophisticated tool such as LigandScout [59] are available that extract pharmacophore models from one or more protein–ligand crystal structures making pharmacophore modeling a hybrid between ligand-based and structure-based modelings. Traditionally, however, pharmacophore modeling has been most often applied to the target class of GPCRs [60] where until very recently structural information [61–65] of the target receptors was unavailable. Pharmacophore filters are generally much faster than molecular docking approaches and can, therefore, greatly reduce the number of compounds screened further using more expensive docking applications. Today, pharmacophore features are guiding most 3D virtual screening approaches including docking in the form of constraints or prefilters. Most time consuming is still the generation of 3D conformations for small molecules—a task that can only be solved for a small number of molecules when precalculation and storage is necessary.

An interesting aspect of pharmacophores in virtual screening is 3D pharmacophore diversity. While the diversity concept for virtual compounds in general is not applicable because of the enormity of the chemical space, diversity in pharmacophore space is a feasible concept. Virtual libraries can therefore be optimized to cover a wide pharmacophore space [66].

2.2.1. Introduction to Pharmacophores In 1894, Emil Fischer proposed the "lock and key" hypothesis to characterize the binding of compounds to proteins [67]. This can be considered the first attempt to explain binding of small molecules to a biological target. Proteins recognize substrates through specific interactions, forming the pharmacophore. Inhibitors intended to block substrate binding to the protein, should also capture these interactions. The first definition of the pharmacophore formulated by Paul Ehrlich was "a molecular framework that carries (phoros) the essential features responsible for a drug's (pharmacon) biological activity" [68]. This definition was slightly modified by Peter Gund to "a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity" [69]. An example is shown in Fig. 3. An X-ray structure of CDK2 complexed with the adenine-derived inhibitor H717 [70–72] was solved. Interactions that are essential to substrate and inhibitor binding to the enzyme will form the pharmacophore that should be captured by inhibitors binding the same way H717 does. As shown in Fig. 3, the inhibitor binds to the hinge region (Phe82 and Leu83) through two hydrogen bonds, to a hydrophobic region through the cyclopentyl group, and to Asp145 and Asn132 through hydrogen bonds. The pharmacophore that reflects these interactions has a hydrogen-bond donor and a hydrogen-bond acceptor pair that ensures binding to the hinge region, a hydrophobic group that corresponds to the cyclopentyl binding site, and a hydrogen-bond donor that ensures binding to Asp145 and/or Asn132. Pharmacophore hypotheses can be generated using structural information from ligands or from the protein active site itself. Pharmacophore-based drug design is further elaborated in Chapter 11 of Volume 1.

2.2.2. Databases of Organic Compounds It is much more practical to virtually screen databases of compounds that have already been synthesized than to operate in a completely virtual chemical space. If no information about possible lead structures is available, if in-house efforts such as HTS were not successful, and if cost and time are an issue, virtual screening of vendor databases and the subsequent purchase of hits is the method of choice for the generation of novel lead chemical matter.

There is a wealth of databases of available chemical matter including academic sources such as the ZINC database [73] and the National Cancer Institute Database [74,75] as well as commercial vendor databases such as eMolecules [76] that holds approximately 10 million commercially available compounds, Chemspider [77] that is publicly available holding approximately 20 million compounds, and the Available Chemicals Directory screening collection [78]. Some vendors maintain

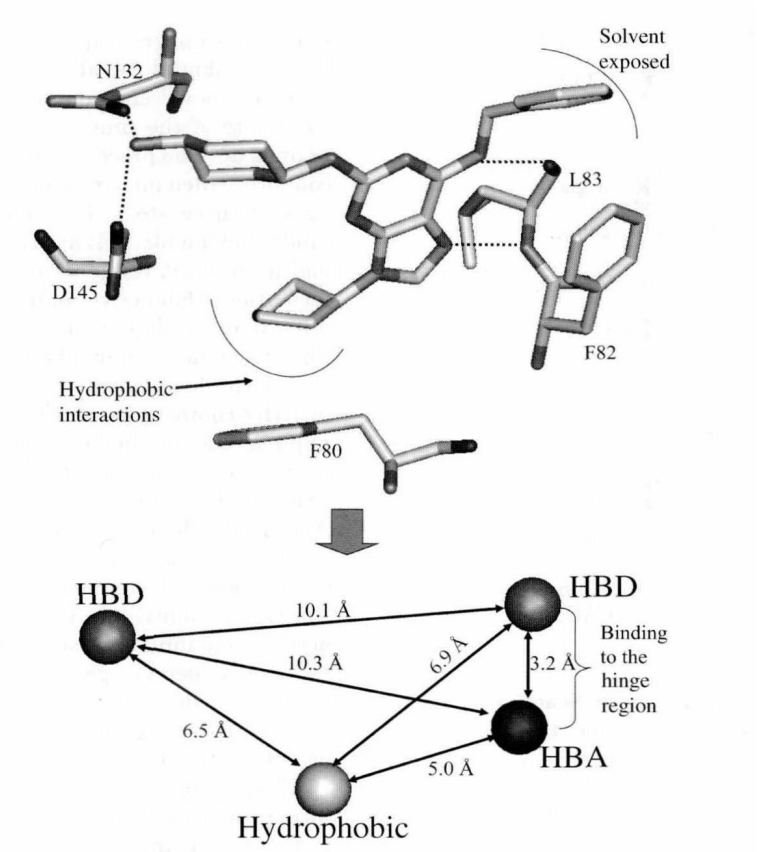


Figure 3. Pharmacophore derived based on the interactions between human cyclin-dependent kinase 2 and the adenine-derived inhibitor H717 as observed in the X-ray structure of the complex (PDB ID 1g5s). Dashed lines highlight hydrogen-bonding interactions. HBD: hydrogen-bond donor, HBA: hydrogen-bond acceptor, and the hinge region is linking the N- and C-terminal domains of a kinase.

databases of virtual compounds that can be custom synthesized. Enamine's REAL database of close to 10 million compounds is an example [79]. In addition, many vendors of chemicals also provide searchable databases with 2D-structure and property information of their compounds. Sometimes, compounds are coded in linear representations such as the SMILES [80,81] notation. The SMILES codes obtained using CACTVS and Daylight programs for 4-benzyl pyridine and R-cocaine are shown in Fig. 4.

Managing compound collections from different sources is not trivial. Chemical databases frequently contain incorrect structures. Careful treatment and curation of compound collections is therefore important to assure optimal virtual screening outcomes. This includes the appropriate assignment of charges, chirality, and tautomerization states, filtering for duplicates, and unwanted structures.

The primary source of 3D experimental structures of organic molecules is the Cambridge Structural Database [82]. Alternatively, 2D databases of organic compounds can be converted into 3D databases using several software programs [83]. Each program starts with generating a crude structure that is subsequently optimized using a force field. To mention the most commonly used programs, CONCORD [84] applies rules derived from experimental structures and a univariate strain function for building an initial structure; CORINA [85] generates an initial structure using a standard set of bond lengths, angles, and dihedrals and rules for cyclic systems.

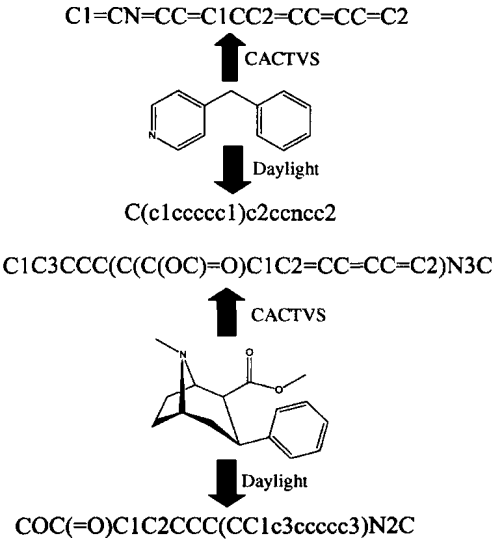


Figure 4. Examples of SMILES notations for two compounds obtained using CACTVS and Daylight.

2.2.3. 2D Pharmacophore Searching Different strategies are pursued to search a 2D database. An exact structure search is applied to find out if a compound is present in the database. Substructure searches are used to find larger molecules with the query imbedded. Superstructure searches are used to find smaller molecules that are embedded in the query. Similarity searches are used to find compounds that are structurally related to the query. Some searching software combines similarity with substructure or super structure searches. Flexible match searches are used for identifying compounds that differ from the query structure in user specified ways. In addition, isomer, tautomer, parent molecule searches may be done to find in a database isomers, tautomers or parent molecules of the query.

2.2.4. Ligand-Based 3D Pharmacophore Generation Ligand-based pharmacophores are typically used when the crystallographic, solution structure, or modeled structure of a protein cannot be obtained. When a set of active compounds is known and it is hypothesized that all compounds bind in a similar way to the protein, then common groups should interact with the same protein residues. Thus, a pharmacophore capturing these common features should be able to identify from a database novel compounds that bind to the same site of the protein as the known compounds do. The process of deriving a pharmacophore, called pharmacophore mapping, consists of three steps: (1) identifying common binding elements that are responsible for biological activity, (2) generating potential conformations that active compounds may adopt, and (3) measuring distances between binding pharmacophore elements in each conformation. To build a pharmacophore based on a set of active compounds, two methods are usually applied. One method generates a set of minimum energy conformations for each ligand and searches for common structural features. Another method considers all possible conformations of each ligand to evaluate shared orientations of common functional groups. Analyzing many low energy conformers of active compounds can suggest a range of distances between key groups that will take into account the flexibility of the ligands and of the protein. This task can be performed manually or automatically.

Pharmacophore generation through conformational analysis and manual alignment is a time-consuming task especially when the list of active ligands is large and the elements of the pharmacophore model are not obvious. There are several software products and algorithms such as HipHop [86], HypoGen [87], PHASE [88], MOE [89], and older tools such as Disco [90], Gasp [91], Flo [92], APEX [93] that can automatically generate potential pharmacophores from a list of known inhibitors. A collection of views on pharmacophore elucidation can be found in Chapter 11 of Volume 1 as well as in the book by Güner [94]. The programs use algorithms that identify common pharmacophore features in the training set molecules that are ranked with a scoring function. Common pharmacophore features include: hydrogen-bond donors, acceptors, negative and positive charges or ionizable centers, and surface accessible hydrophobic regions that can be aliphatic, aromatic, or nonspecific. Most of the programs incorporate ligand flexibility when generating pharmacophores since compounds may not bind to the protein in the minimum energy conformation.

2.2.5. Example 2: Virtual Screening for Dopamine Transporter Inhibitors The dopamine transporter (DAT) is a 12 transmembrane helix protein that plays a critical role in terminating dopamine neurotransmission by taking up dopamine released into the synapse. DAT is involved in several diseases such as drug addiction and attention deficit disorder [95]. DAT inhibitors share one or more common 3D pharmacophore models [96–98]. A pharmacophore model was derived based on two known potent DAT inhibitors R-cocaine and WIN-35065-2 (Fig. 5) [98]. The common binding elements of these compounds are a ring nitrogen that may be substituted, a carbonyl oxygen, and an aromatic ring that can be defined by the position of its center. A systematic conformational search followed by conformer clustering was performed to obtain all possible conformations these compounds could assume when bound to DAT. The resulting 3D pharmacophore model (Fig. 5) was used to search the NCI 3D-database [75] of 206,876 compounds using the program Chem-X [99]. A total of 4094 compounds, 2% of the database, were identified as hits. After further reduction using molecular weight, structural novelty, simplicity, diversity, and hydrogen-bond acceptor nitrogen filters, 70 compounds were selected for testing in biochemical assays.

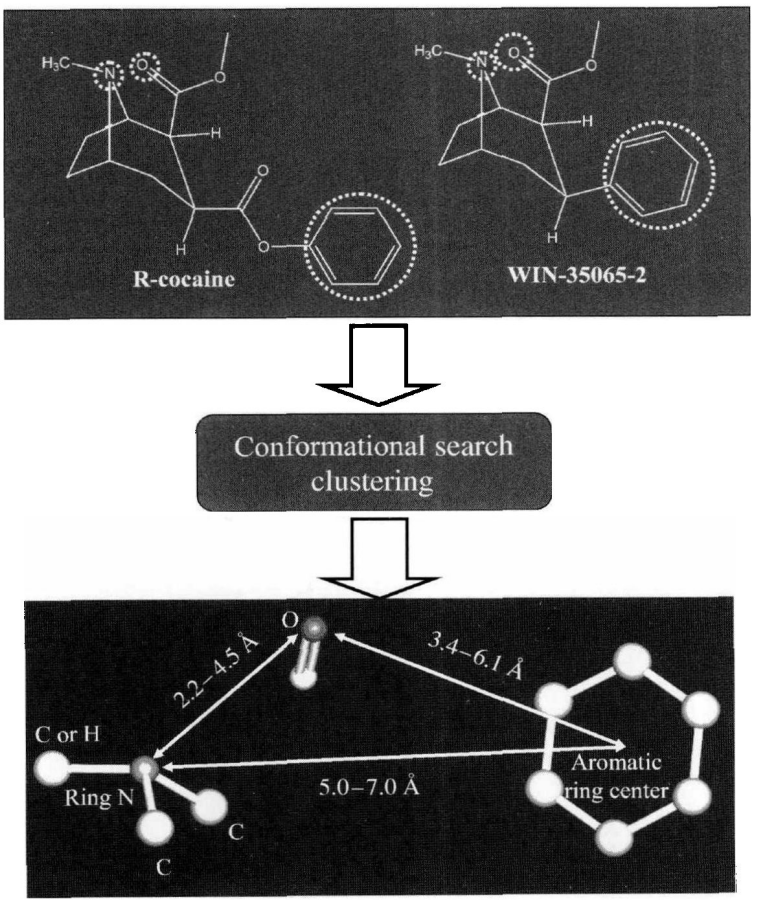


Figure 5. Pharmacophore proposed for identifying DAT inhibitors. The pharmacophore was obtained based on two known DAT inhibitors, R-cocaine and WIN-35065-2. Distance ranges between pharmacophore points were obtained through systematic search of all possible conformations that the two compounds may adopt when bound to DAT. (This figure is available in full color at http://mrw.interscience.wiley.com/emrw/9780471266945/home.)

Forty-four compounds displayed more than 20% inhibition at 10 μM in a [^{3}H]mazindol binding assay. Figure 6 shows selected DAT inhibitors found in the virtual screen.

2.2.6. Receptor-Based 3D Pharmacophore Generation If the 3D structure of a receptor is known a pharmacophore model can be derived directly from the receptor active site. The presence of a cocrystallized ligand in the active site can greatly help with the identification of the pharmacophore. The program LigandScout generates pharmacophore hypotheses in a fully automated way from protein–ligand complexes [59]. In the absence of a ligand costructure, biochemical data can be used for identifying key residues that are important for substrate and/or inhibitor binding. Most ligands bind to proteins through nonbonded interactions such as hydrogen-bond and hydrophobic interactions. Programs such as LUDI [100] or GRID [101] can use the structure of the protein to generate interaction sites or grids to characterize favorable positions that ligand atoms should occupy. Since proteins are not rigid, Carlson et al. [102] proposed using molecular dynamics simulation for generating a set of diverse protein conformations to include protein flexibility in the pharmacophore development. In this case,

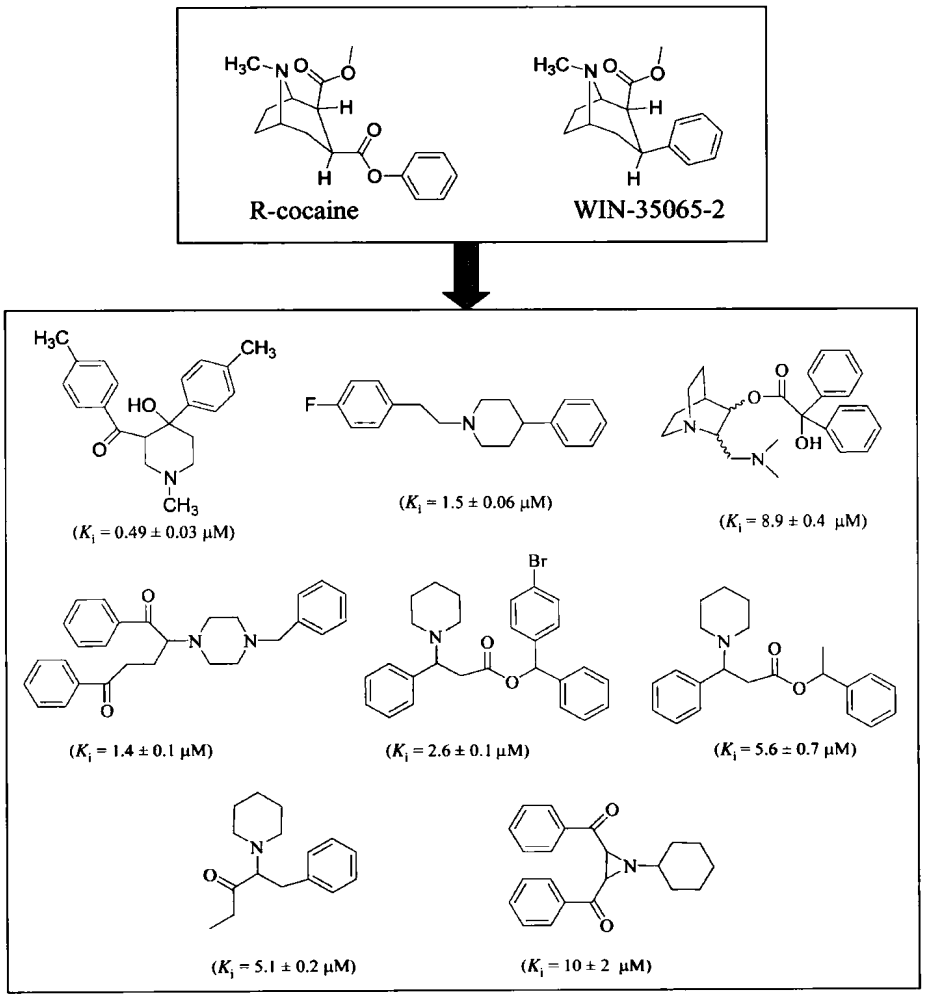


Figure 6. Selected DAT inhibitors identified from the NCI database through virtual screening.

distance ranges between pharmacophores are obtained by examining several conformations of the protein. This technique is similar to the one used for the generation of flexible pharmacophores based on active compounds when several conformations of compound(s) are considered for pharmacophore mapping.

2.2.7. Example 3: Virtual Screening for Novel PPAR Ligands To illustrate how receptor-derived pharmacophore models are used in virtual screening we discuss here a recent example of identifying novel PPAR ligands [103]. A series of LigandScout models were generated based on agonist-PPAR complex structures from the PDB. In addition, ligand-based HipHop models were generated. The pharmacophore models were evaluated using 357 PPAR ligands and 12,775 PPAR decoys. The most selective PPARα agonist model was obtained from LigandScout applied to the PPARα complex structure with PDB ID 1k7l. The best PPARδ agonist model was generated likewise using the PDB complex 1gwx. The best PPARγ agonist model was obtained from a HipHop model. More than one million compounds were screened in Catalyst [104] using all three PPAR-agonist models. A total of 14,311 virtual screening hits were obtained. Filtering by physicochemical properties as found in 321 PPAR agonists retained 5898 hits. 3D shape screening, electrostatic similarity to known PPAR agonists, diversity clustering, visual inspection, and elimination of compounds with reported PPAR activity led to 21 compounds of which 10 were purchased and tested. Several novel PPAR agonists with varying PPAR selectivity were discovered (Fig. 7).

2.2.8. 3D Pharmacophore-Based Virtual Screening Techniques Pharmacophore-based virtual screening is the process of matching atoms and/or functional groups and the geometric relations between them to the pharmacophoric query. Once a pharmacophore model is generated, virtual screening of a database against such a query model becomes straightforward. Programs that perform pharmacophore-based searches include Catalyst [104], UNITY [105], ROCS [106], PHASE, and MOE. There are also some Web-based applications that can perform pharmacophore searches [107,108]. Usually pharmacophore-based searches are done in two steps. (1) The software checks if the compound contains the atom types and/or functional groups prescribed by the pharmacophore. (2) It checks if the spatial arrangement of these elements matches the query. Since most small molecules are flexible, multiple conformations have to be checked. These conformations can be either precomputed and stored (such as in MOE and Catalyst) or calculated on the fly. Precalculating conformations will speed up the virtual screen tremendously. Therefore, this is the method of choice for databases of existing compounds that are screened repeatedly. On the other hand, for larger virtual libraries the data handling requirements become too complex. Therefore, generating conformations on the fly and discarding them after the pharmacophore query has been checked remains an important option for pharmacophore screening programs. Generating a representative ensemble of conformations is essential for the success of a virtual screen using a 3D pharmacophore query.

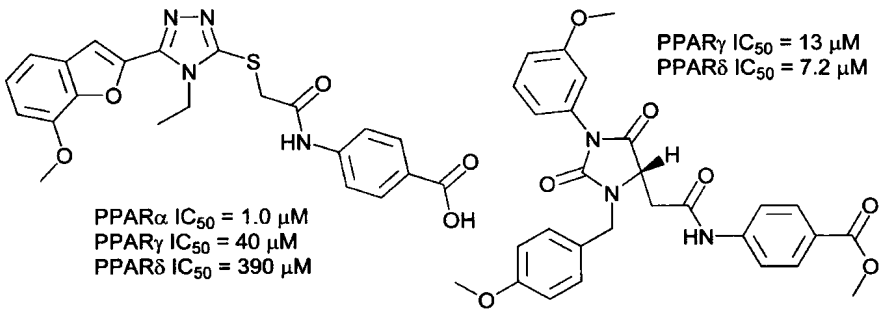


Figure 7. Novel PPAR agonists identified through structure-based pharmacophore screening.

Many ligands do not adopt a minimum energy conformation as the binding pose [109–111]. Therefore, continued research is necessary to represent a conformational ensemble of a ligand including higher energy conformations with a limited number of conformations.

2.3. Application of Machine Learning for Virtual Screening

2.3.1. QSAR Model Generation and Validation

Machine learning is a form of ligand-based VS that builds predictive quantitative structure–activity relationship (QSAR) models that are based on available assay data. Compounds are described with various molecular descriptors to provide numerical representations of a compound's properties. There are numerous types of 2D and 3D descriptors that are commonly used but will not be discussed in detail here [112–117]. Particular descriptors and machine learning methods are chosen because they are believed to be linked or directly correlated to the property being modeled. The process for model generation, validation, and application to virtual screening is shown in Fig. 8 (figure adapted from a recent review by Tropsha and Golbraikh [118]). Many different algorithms and computer software are available for the purpose of predictive modeling; they are based on linear as well as nonlinear methods. In all approaches, descriptors serve as independent variables, and biological activities as dependent variables.

To establish the ability of a QSAR model to predict biological activities the model needs to be thoroughly validated. This is arguably the most important part of QSAR model development [119,120]. Most QSAR modeling methods implement a leave-one-out (LOO) (or leave-some-out) cross-validation procedure. The outcome of this procedure is a cross-validated R^2 (q^2), which does not always guarantee the predictive ability of the model [119,121] illustrating the necessity for thorough model validation. A widely used approach to establish the model's robustness is called y-randomization (randomization of the dependent variable, that is, biological activities). Y-randomization consists of repeating the calculation procedure with randomized activities and subsequent probability assessment of the results. Often, it is used along with the LOO cross-validation to ensure the model is not based on a chance correlation.

2.3.2. Machine Learning Algorithms Commonly Used in QSAR Models

Collections of validated QSAR models can be useful tools for VS

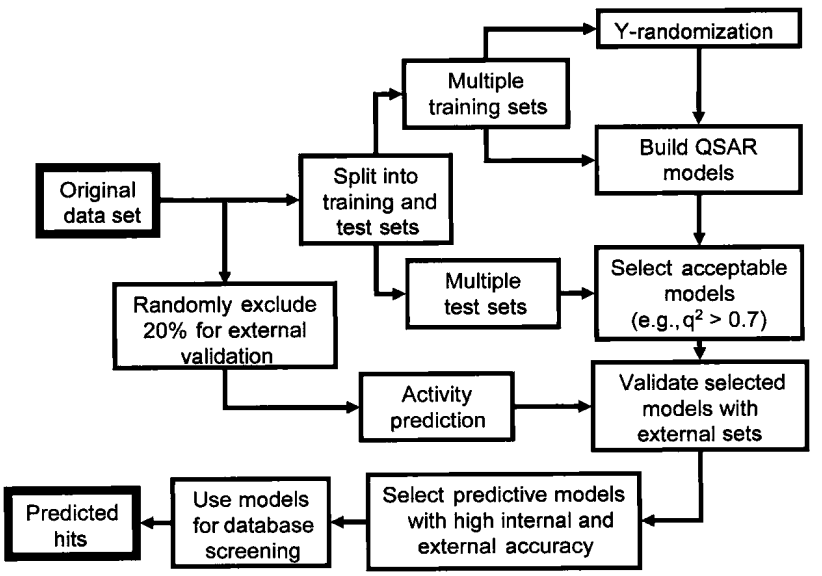


Figure 8. Workflow for generating and validating QSAR models for virtual screening