Ion Channels Targeted Library

Medicinal and Computational Chemistry Dept., ChemDiv, Inc., 6605 Nancy Ridge Drive, San Diego, CA 92121 USA, Service: +1 877 ChemDiv, Tel: +1 858-794-4860, Fax: +1 858-794-4931, Email: ChemDiv@chemdiv.com

Preamble

Due to poor efficiency of the mass random bioscreening concept in drug discovery, the current paradigm holds that target-specific properties of small-molecule compound libraries must be addressed as early as possible. In general, the existing target and ligand structure-based technologies cannot adequately address all the problems of rational drug design, particularly those connected with virtual screening of large compound databases for novel active chemotypes. An alternative design for target-specific libraries is based on the similarity of molecular physicochemical properties of active compounds for certain protein families. We applied this approach in the design of our ion-channel (IC) focused library using several neural network (NN) QSAR methods, particularly Kohonen and Sammon maps for data analysis and visualization.

1. Ion channels as promising drug targets

Modulation of ion transmembrane channels is the basis of therapy for a variety of illnesses. The development of drugs that modulate the entry of ions into cells may provide clinically significant benefits in the treatment of cardiovascular diseases, cerebral and peripheral vascular disorders, male and female sexual dysfunctions, diabetes, asthma, drug-induced ulcers of the gastrointestinal tract, epilepsy, and several types of neuropathic pain.

Voltage-gated ion channels (VGICs) play an important role in numerous cell types and occur as large families of related genes with cell-specific expression patterns. VGICs are transmembrane proteins that mediate the influx of ions (Ca^{2+}, Na^{+}, K^{+}) in response to membrane depolarization and thereby initiate multiple cellular activities [1]. The phylogenetic trees of the VGIC families and subfamilies can be identified from the several databases [2]. Several drugs already marketed have generated substantial revenues.

Considering the publication of the human genome and the progress in transcription profiling revealing the tissue specific distributions of ion channels, they will play a much more important role

as therapeutic drug targets in the future. These advancements combined with cell-based assays providing biologically relevant information to genomic and proteomic information will make ion channels a favorable class of selective, tissue specific, drug targets.

2. Potassium channels openers

Potassium channels openers (PCOs) are encoded by a substantial multi-gene family in higher organisms. PCOs are a structurally heterogeneous group of compounds that relax vascular smooth muscle and reduce cardiac muscle contractivity by increasing membrane conductance to potassium1. PCOs have therapeutic potential in a number of disease states: hypertension, irritable bladder syndrome, male and female sexual dysfunctions, diabetes, asthma, drug-induced ulcers of gastrointestinal tract, and cardioprotectants in ischemic heart disease [3]. Blockage of the HERG potassium channel can lead in rare cases to drug-induced arrhythmias, which has led to a number of drugs being withdrawn from the market.

PCOs share a common structure of a tetramer of four alpha-subunits, each contributing one P-domain to an ion selective pore (Fig. 1). The properties of the channel can be modified by auxiliary subunits. The human genome contains 77 genes encoding pore-forming subunits. The genome of the fruit-fly *Drosophila melanogaster* contains 21, but surprisingly the simple nematode worm *Caenorhabditis elegans* has 65.

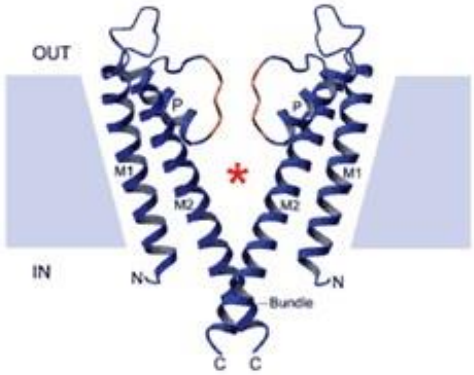


Fig. 1. The transmembrane pore of K^+ channels is composed of four identical subunits, of which two are shown. The ion pathway contains a narrow selectivity filter (*yellow*) and a wide central cavity (*asterisk*). Three helical elements include the outer helix (M1), pore helix (P), and inner helix (M2). The gate is formed by the inner helix bundle.

Analysis of the X-ray crystallography data showed that the potassium channel from *S. lividans* is shaped like a cone or "inverted teepee." According to the published data, the structure helps explain one of the great biophysical mysteries - the chemical nature of the pore's main ion conduction

pathway. Potassium ions are normally surrounded by water. When they slip into the channel, the potassium ions shed water. In order for this to happen, the pore must offer a surrogate for water. Ion discrimination takes place in a region of the pore called the selectivity filter. This area is called a filter because it is narrower than the rest of the channel. When a potassium ion enters the channel, water gets transported out. Oxygen atoms from the protein then surround the ion, making it more stable. Scientists have also wondered why the sodium ion, which is smaller than the potassium ion, does not jump into the potassium channel. Again, the structure may provide insight. There is a suggestion that the selectivity filter, which is held in a very precise conformation, is more tuned for the larger potassium ion.

Phylogenomics uses phylogenetic trees to visualize the relationships between gene family members in genomes. Using the trees it is possible to predict the properties of uncharacterised family members and see how the families might have evolved in the different organisms. Thus, phylogenetic trees constructed for K^+ transporters are depicted in Fig. 2 and Fig. 3. The first figure shows a tree of all K^+ transporters obtained from *A. thaliana*. They has five major branches: a) KUP/HAK/KT transporters (13 genes), b) Trk/HKT transporters (1 gene), c) KCO (2P/4TM) K^+ channels (6 genes), d) Shaker-type (1P/6TM) K^+ channels (9 genes), and e) K^+/H^+ antiporter homologues (6 genes). Predicted membrane topologies for each branch are shown in the figure. The apparent absence of K^+ channels of the 2P/8TM family is remarkable as is the diversity in the AtKUP/HAK/KT transporters. Proteins for which a complete cDNA sequence is available are indicated by bold letters and lines. AGI genome codes are given except for AtKUP3, AtKUP4, AtHAK5, AtHKT1, GORK, KAT2 and AKT2 (*GenBank accessions*) because of errors in the sequences predicted by AGI. Programs used were HMMTOP [4] for topology predictions of the KEA and AtKUP/HAK/KT families, ClustalX [5] for alignments, and tree-view [6] for graphical output.

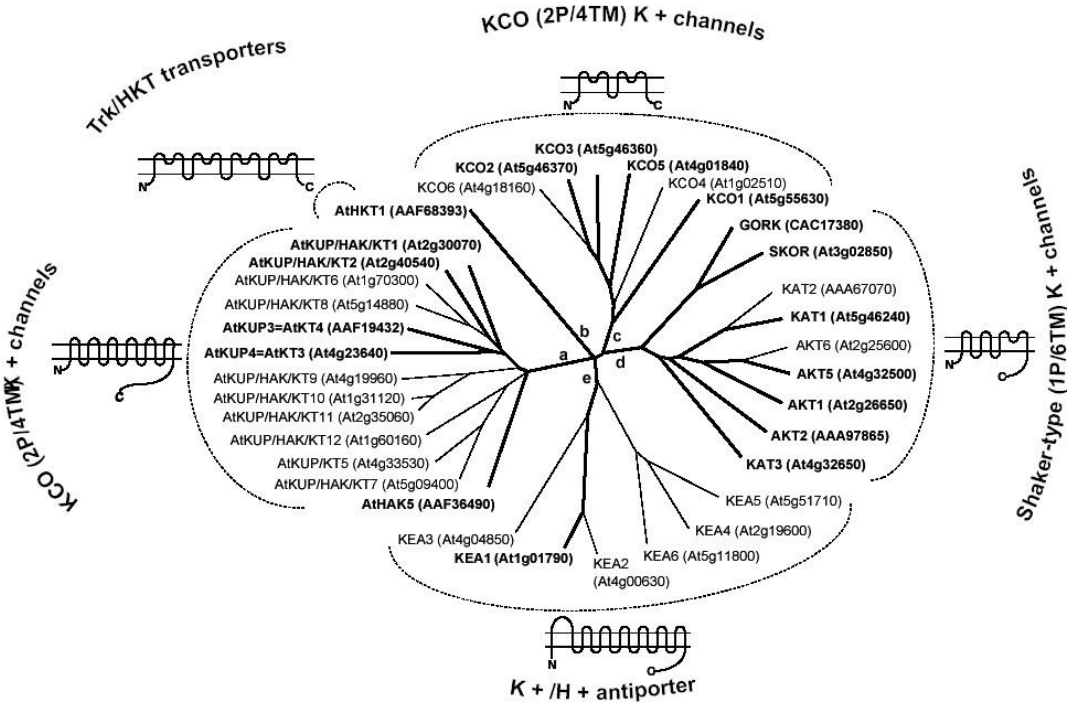


Fig. 2. A tree of all K^+ transporters from *A. thaliana*.

A non-rooted tree (Fig. 3) reflects the structural and functional properties of *A. thaliana* K^+ channels. The two major branches are the 2P/4TM-type and the 1P/6TM (Shaker)-type channels, as depicted by the sketches. For KAT1 the proposed topology has been confirmed experimentally [7]. The 1P/6TM (Shaker-type) channels are further subdivided into the depolarization-activated GORK and SKOR and the KATs and AKTs. All 1P/6TM channels possess a putative cyclic nucleotide-binding site (CNB), and AKT channels also have an ankyrin repeat consensus site (AR, see sketches). P-loops are labeled with asterisks. Proteins for which a complete cDNA sequence is available are indicated by bold letters and lines.

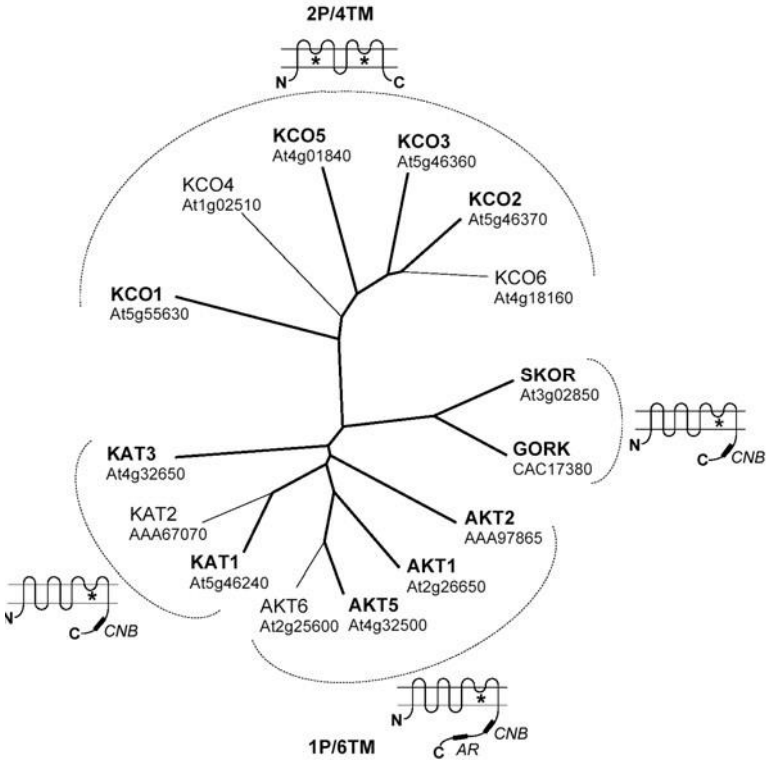


Fig. 3. A non-rooted tree reflects the structural and functional properties of *A. thaliana* K^+ channels.

More than 1K known potassium channel openers (PCOs) were used as a reference database. All compounds were selected from the Ensemble database [8] which is a licensed, and contains known pharmaceutical agents compiled from the patent and scientific literature. This database was used as a source of structural information in the stage of morphing active molecules. Several representative examples of active PCOs were entered into preclinical trials and used as compound prototypes (Fig. 4).

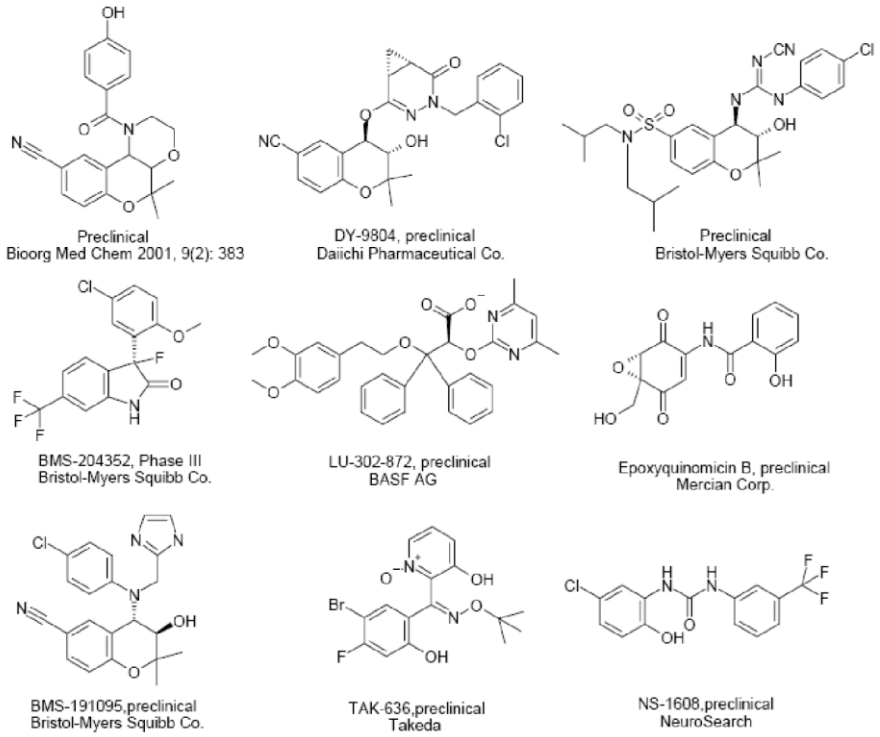


Fig. 4. Reported drugs targeted against PCOs and compounds entered into different clinical trials.

3. Sodium channels

Sodium channels (SCs) play an important role in the neural network by transmitting electrical impulses rapidly throughout cells and cell networks, thereby coordinating higher processes ranging from locomotion to cognition [9]. These channels are large transmembrane proteins, which are able to switch between different states to enable selective permeability for sodium ions. For this process, a potential action is needed to depolarize the membrane, and hence these channels are voltage-gated. The voltage-gated sodium channels could be targeted, either selectively or in combination with other cellular processes, for the treatment of stroke, epilepsy, and several types of neuropathic pain.

More than 190 known sodium channel inhibitors (SCIs) were used as a reference database. All compounds were selected from the Prous Science Integrity Database (*see ref above*) which is a licensed database of known pharmaceutical agents compiled from the patent and scientific literature. This database was used as a source of structural information in the stage of morphing active molecules. Representative examples of active SCIs entered into clinical trials and used as compound-prototypes are shown in Fig. 5.

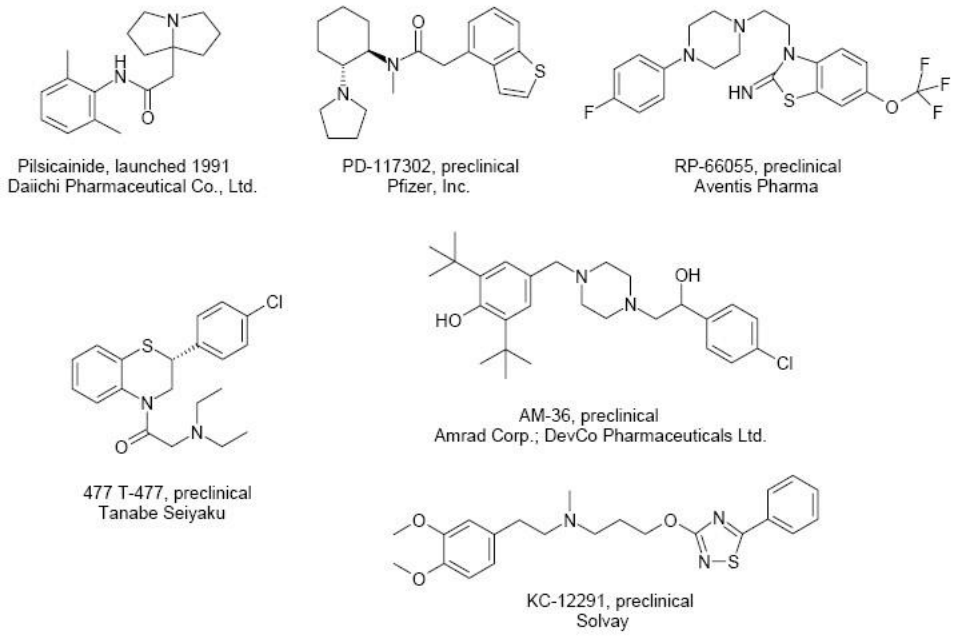


Fig. 5. Examples of clinically validated SCIs.

4. Voltage-dependent calcium channels

Voltage-dependent calcium channels (VDCCs) were first identified in crustacean muscle by Paul Fatt and Bernard Katz (1953). These muscles showed action potentials in the absence of external Na^+ that were dependent on calcium (Ca^{2+}) entry. The first VDCC to be cloned was α1S (skeletal), following purification of the DHP receptor from skeletal muscle; it is concentrated in the T tubules, providing a rich source of starting material [10]. The purified oligomeric complex from muscle consisted of five proteins: α1 (~200kD), α2 (~140 kD), β (~50 kD), δ (~20 kD) and γ (~30 kD). cDNA clones were gathered using primers obtained from the amino acid sequence of proteolytic fragments of the individual proteins. The skeletal muscle α1S subunit is unique in its properties; it activates slowly with relatively large gating charge movements. In addition to the skeletal muscle α1S subunit, three further subtypes of L type calcium channel α1 subunit have been identified: C, D and F, which form the SCDF family of calcium channels. Another VDCC gene, termed α1F, has been identified whose expression is restricted to the retina. On the basis of homology, it is also thought to encode an L type channel and a mutation in this gene has been identified to be responsible for one type of congenital night blindness [11]. In addition, there are also the ABE family of calcium channels composed of several subtypes, including the neuron-specific B type clone (α1B), α1A and α1E, and the GHI family [12]. For example, voltage-dependent presynaptic inhibition can be reconstituted with cloned and expressed calcium channels; it is shown by all three of the first subfamily of channels,

with α1B showing the greatest ability to be modulated by G proteins and α1E the least. There are many different ion channels that fall under the general umbrella of non-selective cation channels. This simply means that they show little selectivity for Ca^{2+}, K^{+} or Na^{+}. The ion flux that occurs depends on the membrane potential and the concentration of each ion on either side of the membrane.

Ca^{2+} is generally present at a concentration of a few mM in the extracellular space, but inside the cell, the cytoplasmic concentration is about 0.1 μM. This is kept low by a number of different pumps and buffering systems, as well as the general impermeability of the plasma membrane to the entry of Ca^{2+}. VDCCs have subsequently been found in all types of excitable cell: vertebrates, invertebrates, and even plants. They fulfill numerous functions depending on the tissue and is thus, not surprising that a number of subclasses of VDCC have been identified. Examination of the biophysical properties of VDCCs required the advent of voltage-clamp and subsequently patch clamp technology. VDCCs are normally closed at resting membrane potentials and open upon depolarization, due partly of the channel structure sensing the change in transmembrane voltage. The resultant current through the cell membrane can be characterized by a number of properties, including the membrane potential range over which the channel opens and the kinetics or time-dependent properties of the current. Different single channel currents can also be identified with varying properties; the task of matching these single channel types with the currents observed in entire cells is a difficult one, but has been made easier by the cloning of the cDNAs for a number of VDCCs and the use of selective drugs and toxins to identify specific current components that correspond to particular channel types.

Low threshold and high threshold voltage-gated calcium channels

In a number of tissues, including certain cardiac muscle cells, neurons, and other excitable cells, it became apparent that there are two types of calcium current. One is activated by small depolarizations and shows rapid voltage-dependent inactivation; this is termed low voltage-activated (LVA), or T for transient. This type is sensitive to dihydropyridines. The second is activated by large depolarizations and is termed high voltage-activated (HVA). They respond for the contraction of smooth, skeletal, cardiac muscle and mediate hormone release. The single calcium channels underlying these currents are also clearly distinct, T type channels being of small conductance (5-9 pS in 110 mM Ba^{2+}) and show rapid inactivation during a voltage step, whereas HVA channels are of larger conductance (13-24 pS) [13]. HVA currents have been further subdivided; in skeletal and cardiac muscle, the HVA current was termed L for long lasting, and was found to be sensitive to a

number of calcium channel antagonist drugs including the 1,4dihydropyridines (DHPs) such as nifedipine, phenylalkylamines, and benzothiazepines. Furthermore, L type current could be enhanced by another drug in the DHP class, called BayK8644, which has proved very useful as a diagnostic tool for the presence of L type channels. Subsequent studies by Tsien and colleagues [14] in sensory neurons showed the presence not only of L-type currents, but also of a second HVA component of current that was termed N (for neuronal). This was found to have an intermediate single channel conductance (13-18 pS) and was not sensitive to DHPs but was irreversibly inhibited by ω-conotoxin GVIA (ω-CTX GVIA), a peptide toxin from the cone shell mollusc *Conus geographus*. These channels mediate neurotransmitter release at some synapses and represent a specific target for various neurotransmitters and hormones. Another subgroup of calcium currents, insensitive to both ω-CTX GVIA and DHPs, has now been reported in many tissues, indicating the presence of further current components. An extreme example is the cerebellar Purkinje cell, where only a small proportion of the calcium current corresponds to N and L current, and the major calcium current in these cells has been termed P type. A selective blocker for the Purkinje cell calcium current has been found in a peptide toxin from the venom of the American funnel web spider *Agelenopsis aperta*, called ω-Agatoxin IVA (ω-Aga IVA). At higher concentrations it also blocks a current component that has been termed Q (the letter after P), although the distinction between P and Q current is not always clear. In many neurons, despite the application of all three blockers, there often remains a substantial proportion that cannot be classified as L, N or P/Q; this residual current has been termed R (for resistant). Thus in native neurons and other cell types, biophysical properties and selective drugs and toxins allow the identification of 5 distinct current components: T, L, N, P/Q and R. The more recent challenge has been to marry these components with the recently cloned VDCC classes.

Receptor channels

Channels that can be classed as non-selective cation channels encompass a number of receptor channels, including nicotinic, $5HT_3$ receptors, the glutamate receptor subclasses termed NMDA and AMPA receptors. These receptors are all thought to form pentamers of subunits, each of which has 4 transmembrane segments and an extracellular N and C terminal (Fig. 6). While this structure is well accepted for the nicotinic acetylcholine receptor, it is not proven for all members of the group. For example, the ATP (P2X) receptor channels appear to have two transmembrane segments and a P region. Of interest, certain subtypes of receptor channels are more Ca^{2+}-permeable than others; their temporal or tissue-specific expression may play a role in a number of functional switches, for example

during development [15]. Other non-selective cation channels include *trp* channels, which have the putative topology of 6 transmembrane α helices and a P loop between S5 and S6 [16], similar to the voltage-gated K^+ channels. They may form a tetrameric channel; it has been suggested that some are involved in the refilling of intracellular Ca^{2+} stores. The capsaicin receptor VR1, which is involved in pain sensation in sensory neurons, is structurally related to these channels [17].

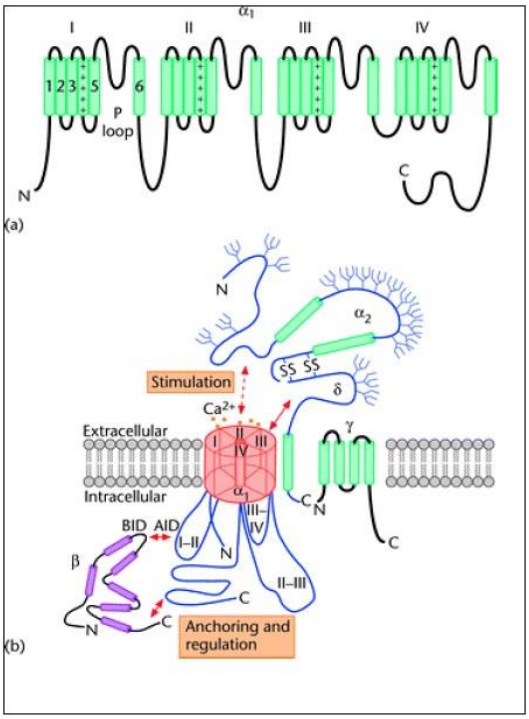


Fig. 6. The simplified model of the Ca^{2+} channel oligomeric complex. a) the topology of the α1 subunit, showing the S4 transmembrane segments, containing a motif of positively charged amino acid residues, and the P regions between S5 and S6; b) the putative structure of the oligomeric VDCC complex.

The inositol trisphosphate and ryanodine receptors are present in the membranes of the endoplasmic and sarcoplasmic reticulum, and are involved in the release of Ca^{2+} into the cytoplasm from these intracellular stores [18]. They have a very similar structure, each consisting of 4 subunits, with an estimated 12 transmembrane segments at the C terminal end and a very large 8 cytoplasmic N terminal domain. This forms a vestibule for drug binding and allosteric effects associated with Ca^{2+}-dependent Ca^{2+} release.

The trigger that opens this channel is IP3, generated by the activation of receptors that stimulate phospholipase C (PLC). The IP3 receptor is thus an integral part of a number of pathways involving G protein coupled receptors, linked to the Gq/11 subclass of GTP binding protein and stimulating PLCβ, or growth factor receptors coupled to PLCγ. Their effect is to increase cytoplasmic

Ca^{2+} from intracellular stores via elevation of IP3, rather than by direct entry across the plasma membrane. At least three IP3 receptor isoforms are known.

The skeletal muscle ryanodine receptor (RyR1) is one of the largest cloned proteins. Each monomer has over 5000 amino acids; thus, the tetrameric channel has a molecular weight of over 2 million. There are also two other ryanodine receptor isoforms (RyR2 and 3) in cardiac muscle, brain, and other tissues. In skeletal muscle, ryanodine receptors are activated by direct mechanical coupling to skeletal muscle L type Ca^{2+} channels brought about by juxtaposition of the T tubules and the sarcoplasmic reticulum; this causes Ca^{2+} release from the sarcoplasmic reticulum without prior Ca^{2+} entry through the L type channels.

5. The core approach and computational methods for IC-targeted library design

In the present study we have effectively applied several advanced methods for in silico evaluation of the specific activity of compounds against particular voltage-gated ion channels (calcium, potassium, and sodium channels).

In the new millennium, pharmaceutical drug discovery is undergoing tremendous changes due to progress in genome research, massive advent of combinatorial synthesis, and high-throughput biological screening. Although these important modern technologies now provide incredible opportunities to pharmaceutical researchers, there are some serious problems associated with the effect of combinatorial explosion. The costs of high-throughput screening or parallel synthesis per one sample may be very low, but they become fairly expensive when multiplied by millions of compounds. Moreover, several papers report that the large number of compounds synthesized and screened did not result in the increase in viable drug candidates [19]; therefore, there is vital need for development of special technologies for making combinatorial synthesis and library design cost-effective.

The main objective of a rational library design is selection of synthetic candidates that possess desirable properties. The “corner stones” of this process are depicted in Fig. 7. Initially, efforts were focused on maximizing diversity [20], sometimes with the introduction of biased pharmacophoric structural motifs. A medicinal chemistry component has subsequently been introduced, resulting in drug and lead-like libraries reflecting the need for soluble molecules with the optimized *in vitro* pharmacokinetic profile. Further interest in a concise screening campaign yielded biased libraries that are focused against a single biological target or a family of related targets (ICs, kinases, GPCRs, NHRs and so forth).

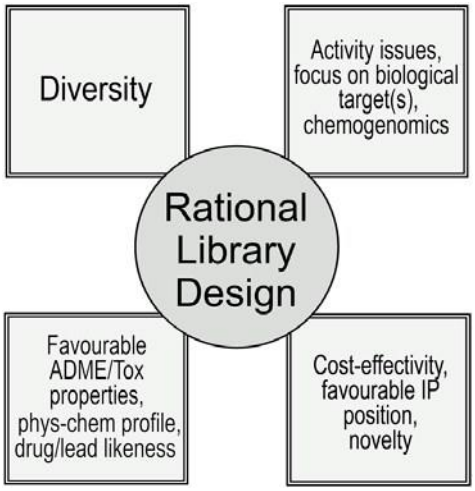


Fig. 7. The "corner stones" in rational library design.

Various ligand and target structure-based design strategies can be implemented in focused library design when a set of known active ligands or 3D structure of the target are available. Additional design elements include cost, synthetic feasibility, physicochemical, PKPD and toxicity properties. These parameters are taken into account by the knowledge-based approaches when relevant experimental and calculated information empowers knowledge-oriented process of rational library design. Moreover, modern computational approaches allow for a simultaneous optimization of several variables. These allow a library designer to 1) control the relative significance of various objectives and 2) intelligently select compounds for the synthesis.

Currently, several advanced computational approaches are widely used to compose rational selecting molecule libraries for synthesis and for further biological evaluation. Specifically, the following conceptually and algorithmically diverse methods can be effectively applied:

(1) ligand structure-based design;
(2) target structure-based approaches;
(3) chemogenomics approaches;
(4) design based on special data mining algorithms;
(5) optimization of ADME/Tox properties.

The current study rationally implicates several approaches within the title multi-step design conception.

5.1. Ligand structure-based design

Historically, ligand structure-based design is the most widely used approach to the design of target-directed chemical libraries. Methods that start from hits or leads are among the most diverse, ranging from 2D substructure search and similarity-based techniques to analysis of 3D pharmacophores and molecular interaction fields (Fig. 8).

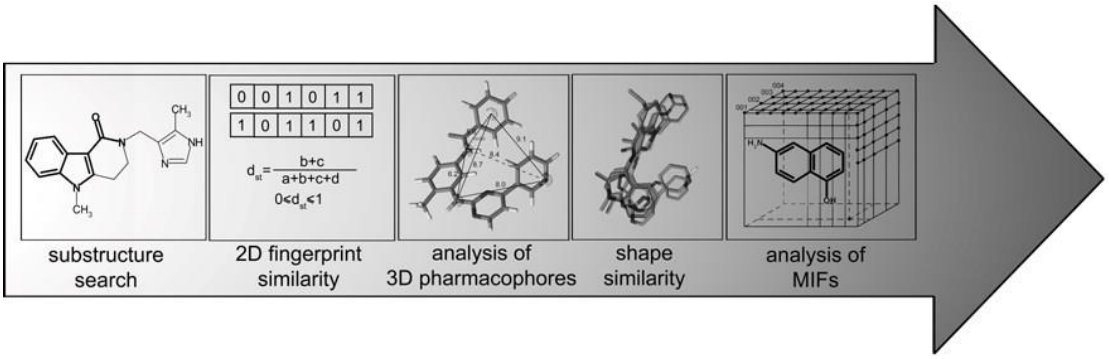


Fig. 8. Computational tools widely applied in ligand structure-based design.

Specific structural fragments of biologically active molecules can be used as the core elements for generating targeted libraries. The most straightforward approach is related to 2D substructure search for analogs of known ligands [21]. These "privileged" substructures [22] have been applied successfully in the framework of ligand-based strategy. Target-directed libraries based on privileged substructures can be effectively designed without any prior knowledge on the structure of endogenous ligand, which in turn means that even orphan receptors can be addressed as potential drug targets [23]. Limitations of this approach include rather restricted availability of privileged substructures for known target families and related IP issues.

Another group of methods address molecular similarity [24]. Similarity methods include two independent aspects: representation of molecules and assessment of their similarity. For example, calculation of 2D molecular fingerprints similarity represents relatively simple yet practical library design principle; it is frequently used to select molecules that have diverse structures but similar activity [25]. Alternatively, individual library compounds are represented by Kier-Hall topological descriptors and molecular similarities between compounds. These are evaluated quantitatively by modified pair wise Euclidean distances in multidimensional descriptor space [26]. This method, called Focus-2D, represents a useful approach to rational design of targeted combinatorial libraries.

Going beyond analysis of 2D structural representation—virtual libraries can be searched using 3D molecular queries [27]. 3D Pharmacophore fingerprints detect the presence of pre-defined pharmacophores in a molecule using a systematic conformational search [28]. Researchers at Tripos

have developed topomer-shape similarity searching, an algorithm that identifies similar compounds by comparing steric interactions between a given query and molecules in a virtual library [29]. This patented technology can effectively generate target-specific libraries around the known ligands used as input queries.

The computational ligand-based strategies are currently progressing to advanced field-fit based methods. In general, such methods remain indispensable in those cases where the structure of binding site of the target protein is unknown.

5.2. Target structure-based approaches

Due to the rapidly increasing availability of structures of target proteins which can be used as templates for virtual screening, combinatorial synthesis and target structure-based design have begun to converge in the process of drug discovery. Many lead generation programs include analysis of X-ray structures of therapeutic biotargets to prioritize compounds for high-throughput screening or to establish a tractable collection for lower throughput assays [30]. A natural trend recognized in the past few years is the application of similar techniques for increasing the likelihood of including active compounds in a focused combinatorial library. There are many examples from literature in which combinatorial library synthesis successfully complemented structure-based design techniques in drug discovery [31].

In the past few years, we witnessed a rapid progress in development of powerful computational technologies, which combine elements of structure-based design and combinatorial chemistry [32]. Computational programs developed on the basis of these approaches generally start from a synthetically accessible combinatorial template that is complimentary to a target binding site. A database of available building blocks for each point of randomization is then considered. The substituents are selected on the basis of their ability to 1) interact with a specific residue(s) in the active site and 2) couple with the template through accessible synthetic reactions compatible with the combinatorial protocol (synthetic feasibility). The generated list of accessible virtual ligands is then computationally screened against the active site and ranked on the basis of the scoring function available. For example, starting with a combinatorial template positioned in the active site of the target protein, the SurflexDock program (Tripos) uses a special scoring function to rank potential substituents at each position on the template. Based on the calculated score, a target-specific library of synthetically accessible molecules is then generated, which may then be prioritized for synthesis and assay.

Alternatively, knowledge of the active site parameters can be used for the generation of pharmacophore hypotheses which are then applied for library design. The pharmacophores define a design space that can be used to select compounds using an informative library design tool. The method was used in prioritizing molecules biased against a cyclin-dependent kinase target, CDK-2. Researchers at Vernalis developed sets of strategies to address receptor flexibility (CDK-2 and HSP90) in virtual screening experiments using multiple crystallographic structures [33]. Based on their assessment, combination of flexible receptor docking algorithm and a robust scoring scheme for hits resulted in a significant improvement of binding affinities.

Customized algorithms, which combine combinatorial library design tools with structure-based design techniques, are viewed by both scientific and business communities as a serious competitive advantage. Despite this fact, there are several key questions about these product:

- What are the performance and limitations of the approach?
- Is the method properly validated? Is the user interface convenient?
- Are the programs compatible with other industry-standard chemoinformatics platforms?

Questions such as these will be taken into consideration should one implement these programs for target directed research. It should also be noted that most of these technologies are still in their infancy, and future practical works will highlight their role in contemporary drug discovery. Practical utility of the target-structure-based approach in the design of chemical libraries is still limited because of the requirement of quality crystallographic data, detailed knowledge of the ligand binding mode and inherent issues concerning scoring functions. The stepwise procedure of selection and filtering using simpler ligand-based technologies can reduce the virtual databases to a manageable size. Such pre-screening procedure leaves the high-ranking molecules for further analysis by biostructure-based docking and scoring, and thus provides both activity enrichment and structural novelty.

5.3. Library design based on special data mining algorithms

Pharmaceutical lead discovery and optimization have historically followed a sequential process in which relatively small sets of individual compounds are synthesized and tested for bioactivity. The information obtained from such experiments is then used for the selection of further molecules. With the advent of high-throughput synthesis and screening technologies, relatively simple statistical techniques of data analysis have been largely replaced by a massive parallel mode of

processing, in which many thousands of molecules are synthesized and tested. As a result, the complete analysis of large sets of diverse molecules and their structural activity patterns have become an emerging problem. Hence, there is considerable interest in novel computational approaches that may be applied to extraction and utilization of useful information from such data sets. Among such 2D and 3D-clustering approaches, top computational dimensionality reduction techniques include Sammon maps, various neural-net-based (NN) methods, back-propagation (BPNNs), feed-forward neural networks (FFNNs), self-organizing maps (SOMs), support-vector machines (SVM), genetic algorithms (GAs), principal component analysis (PCA), and factor analysis (FA). In the current study we have successfully applied BPNN, FFNN, Kohonen-based SOM and Sammon reduction technique to the IC-focused library design.

Visualization techniques

Visual analysis of multivariate data sets have established itself as a powerful means in data mining to detect non-obvious and relevant information for further exploitation, in particular, topology and distance preserving mappings. Using the Kohonen-based SOMs [34] or distance preserving Sammon-based NLM [35], for example, are well suited for data visualization and data mining purposes.

The general idea of Kohonen-based SOMs is to map a set of vectorial samples onto a two-dimensional lattice in a way that preserves the topology of the original space. Kohonen maps were actively used for analysis and visualization of large datasets originated from screening campaigns. In particular, they appeared to be effective in the analysis of large databases created and hosted by the National Cancer Institute (NCI) [36]. Kohonen maps were used by Gasteiger *et al.* for the analysis and visualization of HTS data; the developed structure-activity model was further utilized to design candidates for new sweeteners [37]. The same group of researchers used SOMs for analysis of structure-activity relationships for 5,513 compounds from a combinatorial library [38]. Based on the results of these studies, the authors suggest that the self-organizing maps can serve not only as an indicator of structure-activity relationships, but as the basis of a classification system allowing for the predictive modeling of combinatorial libraries.

By contrast to SOM, NLMs represent relative distances between all pairs of compounds in the descriptor space of a 2D map. The distance between two points on the map directly reflects the similarity of the compounds [39]. NLMs have previously been used for the visualization of protein sequence relationships in two dimensions and comparisons between large compound collections,

which are represented by a set of molecular descriptors [40]. However, for large data sets, NLM computation is becoming more and more intractable. In addition, the approach may generate 2D mapping that poorly approximates the original distances when the number of compounds is large. Several heuristic variants were introduced to alleviate the NLM complexity problem and make it useful for mapping large data sets [41]. Usually, a significant speed gain can be achieved by these modified approaches as compared to NLM. At the same time, they provide better distance and topology preservation as compared to Kohonen maps.

The described computing tools provide interactive, fast, and flexible data visualizations of chemical data that help and even enhance human thought processes. However, visualization alone is often inadequate when multiple data points need to be considered. A number of data mining methods, which seek to identify significant relationships in large multidimensional databases, are now being used for library design.

Classification Methods

Partitioning methods occasionally struggle to provide the accuracy associated with more powerful, albeit less informative techniques such as machine learning and statistical approaches. Due to these reasons, there is a continuing need for the application of more accurate and informative classification techniques to quantitative structure-activity relationship (QSAR) analysis. The goal of a classifier is to produce a model that can separate new untested compounds into classes using a training set of already classified compounds.

It is important that QSAR methods are quick, give unambiguous models, do not rely on any subjective decisions about the functional relationships between structure and activity, and are easy to validate. In the past 10-15 years, methods based on artificial neural networks have been shown to overcome some of these problems. For example, these can manage both linear and nonlinear SARs observed in real practice. There are reports that describe successful application of neural network algorithms to cluster compounds in large datasets with low signal-to-noise values. A recent review [42] on the concepts behind neural networks applied to QSAR analysis, points out problems that may be encountered, suggests ways of avoiding the pitfalls, and introduces several exciting new neural network methods discovered during the last decade. Besides ANNs, there are a number of unique classification methods with high prediction accuracy, including the support vector machine (SVM) and Genetic algorithm.

5.4. Prediction and optimization of ADME/Tox properties based in *in silico* methods

Poor pharmacokinetics and toxicity are important causes of costly late-stage failures in drug development. It is generally recognized that in addition to optimized potency and specificity, chemical libraries should also possess favorable ADME/Tox and drug-like properties [43]. Assessment of a drug-like character is an attempt to decipher molecular features that are likely to lead to a successful *in vivo* and, ultimately clinical candidate [44]. Many of these properties can be predicted before molecules are synthesized, purchased or even tested in order to improve overall lead quality.

Considerable research efforts were focused on novel machine learning algorithms that predict ADME/Tox properties of new chemical entities. Computer-aided techniques now permit enhancement of the latter strategy with an additional set of *in silico* filters (Fig. 9).

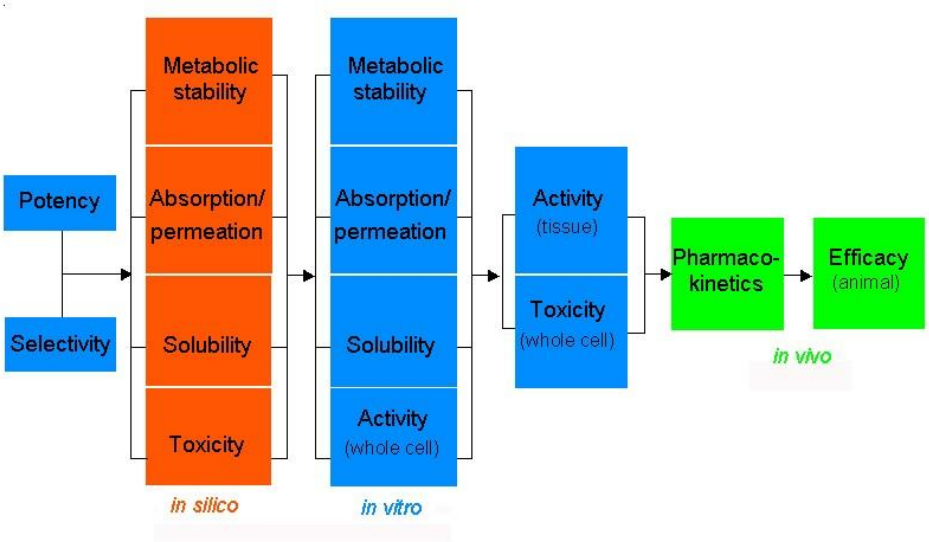


Fig. 9. A common strategy for lead optimization enhanced with parallel *in silico* filtering procedures.

These calculations can be performed with very large numbers of molecules and act as a form of multidimensional selection filter. For example, comparative molecular fields analysis (CoMFA) and pharmacophore approaches (for review, see [45]) have been used to model binding modes of metabolizing cytochrome P450 (CYP) enzymes, transporters such as P-glycoprotein [46], nuclear hormone receptors [47], and ion channels [48], important for drug-drug interactions. Recursive partitioning methods have been used extensively with these large sets of molecules and either continuous or binary data [49]. Kohonen self-organizing and Sammon maps have successfully been applied to model various ADME/Tox properties, including cytochrome P450 mediated drug metabolism [50], blod-brain barrier permeability (BBB), human intestinal absorption (HIA), plasmaprotein binding affinity (PPB), volume of distribution (V_d), plasma half-life time ($T_{1/2}$), and

specific cell toxicity. Many of the reported to-date ADME/Tox models are rule-based. For example, some research groups have used relatively simple filters like the rule of 5 [51] and others [52] to limit the types of molecules evaluated with *in silico* methods and to focus libraries for high throughput screening. However, being designed as rapid "computational alert" tools aimed at single property of interest, they cannot offer a comprehensive picture when it comes to understanding ADME models.

Multivariate data mining techniques can serve as the basis for advanced ADME filters. Thus, we have developed a method for early evaluation of several important pharmacokinetic parameters, including V_d and $T_{1/2}$ [53]. These two parameters determine the dose regimen of a drug; the early prediction of both properties would be of a great benefit. It was demonstrated that such complex properties can be effectively modeled using the non-linear mapping algorithms based on a pre-selected set of electronic, topological, spatial, and structural descriptors. Generated models demonstrated good predictive power in the internal and external validation experiments with up to 80-90% compounds classified accurately. The achieved accuracy level can be used as a guide in modifying and optimizing these pharmacokinetic properties in chemical libraries.

Collection of algorithms for prediction of a number of ADME/Tox related properties is now integrated on the basis of our SmartMining/ADMET software suite available from ChemDiv. To date, they were initially validated on human intestinal absorption, blood–brain-barrier, plasma half-life, volume of distribution, plasma protein binding [54], CYP450 substrate/non-substrate potential [55], and binding affinity [56] models. These algorithms were further extended to evaluation of important physico-chemical properties such as DMSO solubility [57] and target-specific activity [58]. While other software tools for ADME modeling are available (for example [59]), the SmartMiningbased collection of predictive classification tools is both extensive and well validated in multiple library design projects. These methods are particularly suited for rapid evaluation of both large and medium-sized compound libraries in connection with early ADME/Tox profiling.

6. Concept and applications

IC-targeted library design at CDL involves:

• *A combined profiling methodology that provides a consensus score and decision based on various advanced computational tools*:

1. Unique bioisosteric morphing and funneling procedures in designing novel potential ICs ligands with high IP value. 2D/3D-structure similarity and compound diversity. We

apply CDL's proprietary Chemosoft™ software and commercially available solutions from Accelrys, MOE, Daylight and other platforms.

2. Neural Network tools for target-library profiling, in particular Self-organizing Kohonen maps, performed in SmartMining software. We have also used the Sammon mapping and Support vector machine (SVM) methodology as more accurate computational tools to create our IC-focused library.

3. 3D-pharmacophore modeling/searching as well as 3D-molecular docking study for the individual classes of ICs agents.

4. "Rapid Elimination of Swill" (REOS) filters. Computational-based `*in silico*` ADME/Tox assessment for novel compounds includes prediction of human CYP P450-mediated metabolism and toxicity as well as many pharmacokinetic parameters, such as Brain-Blood Barrier (BBB) permeability, Human Intestinal Absorption (HIA), Plasma Protein binding (PPB), Plasma half-life time ($T_{1/2}$), Volume of distribution in human plasma (V_d), etc.

The fundamentals for these applications are described in a series of our recent articles on the design of exploratory small molecule chemistry for bioscreening [for related data visit ChemDiv. Inc. online source: www.chemdiv.com]. Our multi-step *in silico* approach to IC-focused library design is schematically illustrated in Fig. 10.

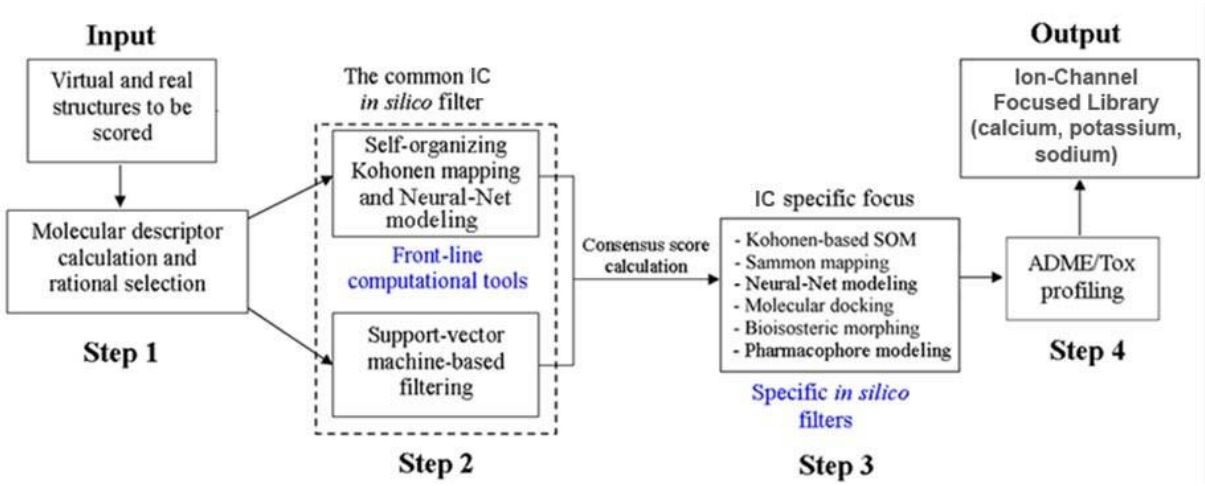


Fig. 10. Multi-step computational approach to IC-targeted library design.

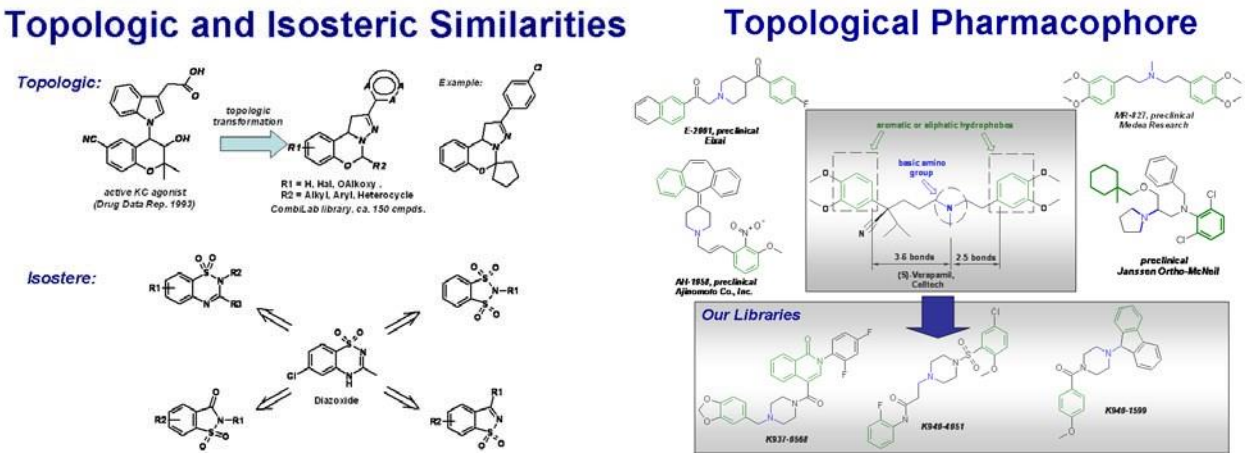
This common approach was effectively applied for the developing of our IC-focused, in particular for calcium, potassium and sodium channels.

• *Synthesis, biological evaluation and SAR study for the selected structures*:

1. High-throughput synthesis with multiple parallel library validation. Synthetic protocols, building blocks and chemical strategies are available.
2. Library activity validation via bioscreening; SAR is implemented in the next library generation.

6.1. Bioisosteric transformations, 2D-structure similarity/diversity and topological pharmacophore

The entitled methods are crucial in drug design and development [60]. For example, the bioisosteric morphing refers to the compounds or substructures that share similar shapes, volumes, electronic distributions, physicochemical properties, therefore having similar biological activity [61]. Bioisosteric approach is useful for morphing the marginal chemotypes; several key bioisosteric transformations, topological similarities, and pharmacophore within ICagents are illustrated in Fig. 11(a,b).



(**a**) (**b**)

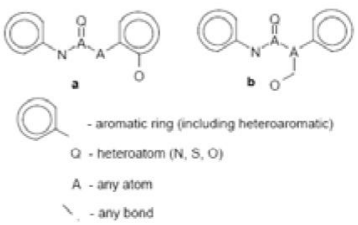
Fig. 11. Implementation of key structure-based *in silico* approaches to our IC-targeted library design: examples of topologic and isosteric similarities (**a**) as well as topological pharmacophore (**b**) within K^+ channels-targeted agents.

For example, the total `first-round` target-specific library of potential potassium channel openers is represented by four main structural types below that contain topological motifs typical of the reported PCOs. These types are divided into small subtypes corresponding to separate combinatorial sublibraries obtained by parallel solution-phase synthesis.

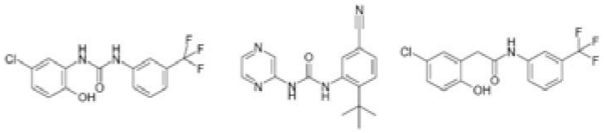
Type 1

General Formula:

Type 1



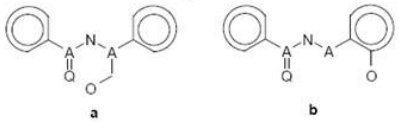
Type 1: Examples of Active Agents



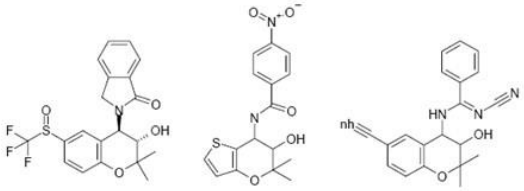
Type 2

General Formula:

Type 2



Type 2: Examples of Active Agents

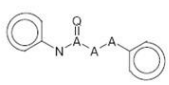


ChemDiv, Inc. 6605 Nancy Ridge Drive, San Diego, CA 92121 USA
PHONE: +1 858-794-4860 FAX: +1 858-794-4931
EMAIL: chemdiv@chemdiv.com, www.chemdiv.com

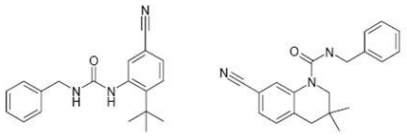
Type 3

General Formula:

Type 3



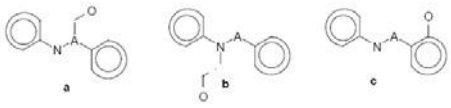
Type 3: Examples of Active Agents



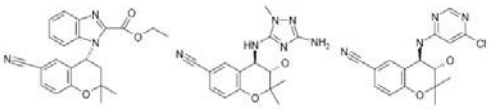
Type 4

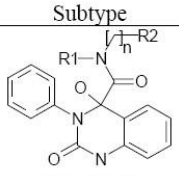
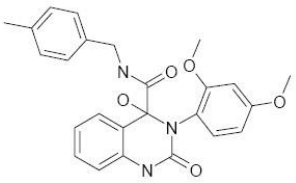
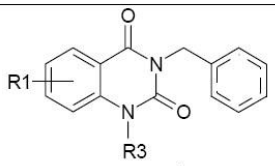
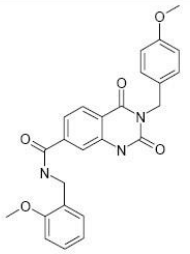
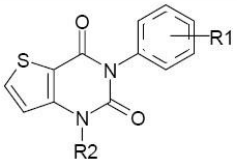
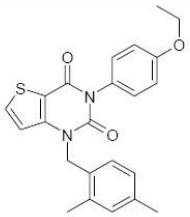
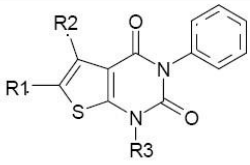
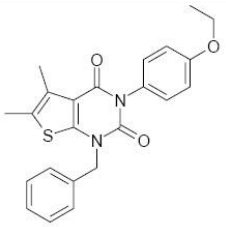
General Formula:

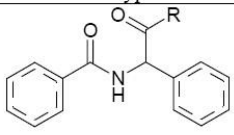
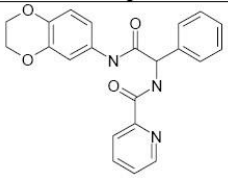
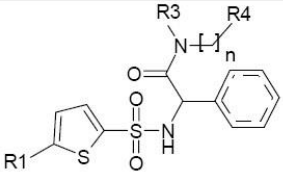
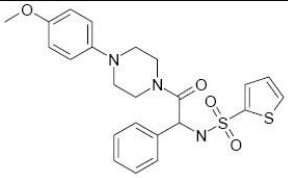
Type 4

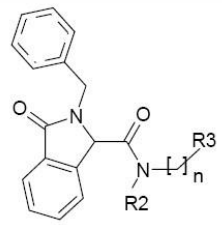
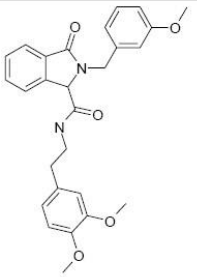
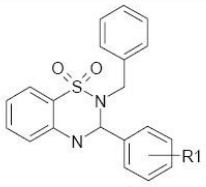
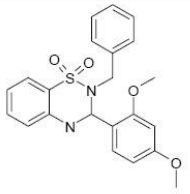
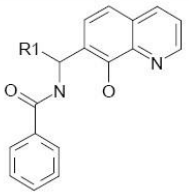
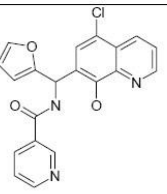


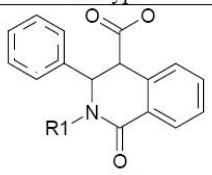
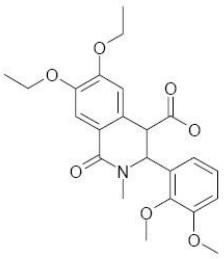
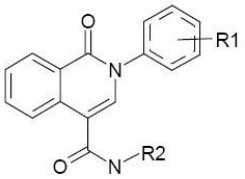
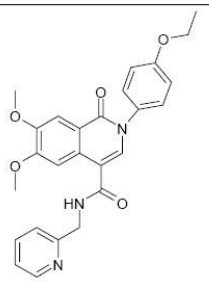
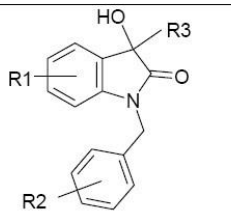
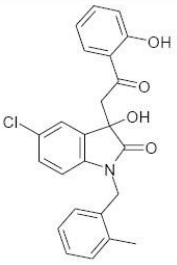
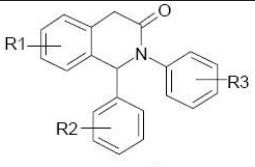
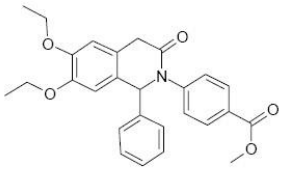
Type 4: Examples of Active Agents

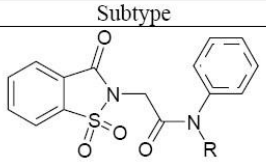
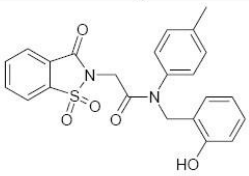
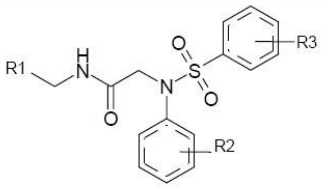
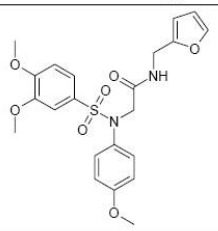
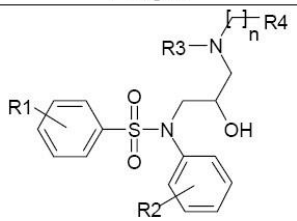
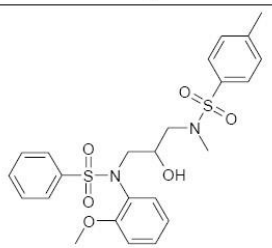
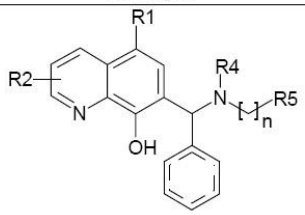
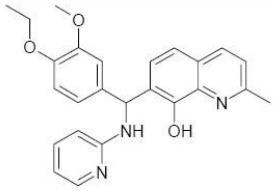
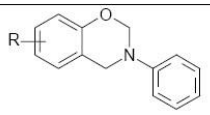
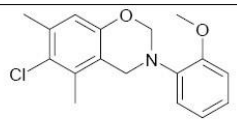


Type 1	
Subtype	Example
 118 cmpds	
 877 cmpds	
 283 cmpds	
 571 cmpds	

Type 2	
Subtype	Example
 5 cmpds	
 12 cmpds	

 39 cmpds	
 7 cmpds	
 8 cmpds	

Type 3	
Subtype	Example
 47 cmpds	
 25 cmpds	
 41 cmpds	
 7 cmpds	

Type 4	
Subtype	Example
 2 cmpds	
 9 cmpds	
 36 cmpds	
 36 cmpds	
 27 cmpds	

Optimization of structural diversity is a natural way to constraining the size of combinatorial libraries. The procedure for generating a diverse library consists of calculating of structural descriptors for the molecules, weighting them according to pre-defined scheme, followed by calculating a similarity coefficient. The cluster-based, dissimilarity-based, or partition-based compound selection is then carried out. The subject has been reviewed elsewhere. It should be noted that different (dis)similarity metrics proposed to-date are not free of subjective decisions, and an application of a particular diversity optimization algorithm often depends on the task. CDL has developed effective algorithms of diversity optimization of compound libraries designed for primary bioscreening; these algorithms are based on the proprietary chemoinformatics platform, ChemoSoft™.

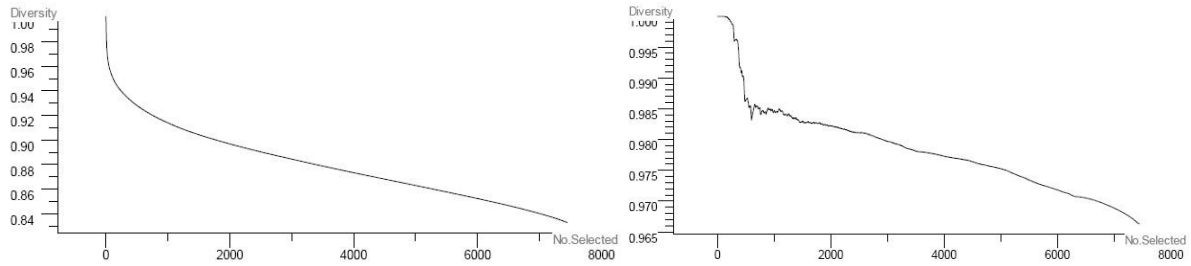
It should be particularly noted that following the original concept of diversity-oriented compound library design we have, sequentially applied three computational methods which are based merely on the structure of known IC-ligands. Initially, we applied 2D-Tanimoto similarity/scoring to rank the compounds from our collection (more than 1140K cmpds) , then among these compounds we selected a representative set of maximal structural diversity (65K cmpds), and lastly, we successfully applied bioisosteric morphing to obtain the optimal `frontline` library (more than 200K cmpds). These have been further evaluated using advanced *in silico* methods (*see the following subsection*).

Some important parameters for the total SC-targeted library are shown in table 1. As evident from the number of screens, number of unique heterocyclic fragments, and diversity coefficient (calculated with the ChemoSoft™ software), the data show that the library has a high level of diversity. Fig. 12a,b shows the substructural and heterocycle diversity profiles for the libraries after diversity sorting procedures.

Table 1. Diversity parameters of the total SC-targeted library.

parameter	value
total number of compounds	7447
number of participating CombiLab libraries	300
number of screens*	6438
number of unique heterocycles	416
substructural diversity	0.8328

* Screens are simple structural fragments, centroids, with the topological distance equal to 1 bond length between the central atom and the atoms maximally remote from it.



(**a**) (**b**)

Fig. 12. (**a**) substructural diversity profile for the total SC-targeted library; (**b**) heterocycle diversity profile for the same library.

6.2. Advanced neural-net approaches for our IC-targeted library design

6.2.1. Kohonen-based SOMs

Choosing structures that are most likely to have a predefined target-specific activity of interest from the vast assortment of structurally dissimilar molecules is a particular challenge in compound selection. This challenge has been tackled with powerful computational methodologies, such as docking available structures into the receptor site and pharmacophore searching for particular geometric relations among elements thought critical for biological activity. The subject is discussed in several comprehensive reviews [62]. Both methodologies focus on conformational flexibility of both target and ligand, which is a complex and computationally intense problem. The latest developments in this field pave the way to wide industrial application of these technologies in drug design and discovery, though the limits of computational power and time still restrict the practical library size selected by these methods. Another popular approach to VS is based on ligand structure and consists of selecting compounds structurally related to hits identified from the initial screening of the existing commercial libraries and active molecules reported in research articles and patents [63]. Although broadly used in the development of SAR profiling libraries, these methods usually perform poorly when it comes to the discovery of structurally novel lead chemotypes. In general, the target and ligand structure-based technologies cannot adequately address all real problems of rational drug design, particularly those connected with VS of large compound databases or discovery of novel lead chemotypes.

An alternative design for target-specific libraries is based on statistical data mining methods, which are able to extract information from knowledge databases of active compounds. Here, we describe a practical approach to limiting the size of virtual combinatorial libraries and selection of molecular subsets with enhanced target-specific informational content. In this work, we used three key methods of QSAR and data visualization based on classical neural-nets (back-propagation network), Kohonen SOMs, and Sammon nonlinear maps. At the initial stage of this work, we collected more than 23,000 compounds for a database of known drugs and compounds entered into preclinical or clinical trials. Each compound in this database is characterized by a defined profile of target-specific activity, focused against 1 of more than 100 different protein targets. The database was filtered based on MW (not more than 800). Molecular features encoding the relevant physicochemical and topological properties of compounds were calculated from 2D molecular representations and selected by PCA. These molecular descriptors encode the most significant molecular features, such

as molecular size, lipophilicity, H-binding capacity, flexibility, and molecular topology. Taken in combination, they define both pharmacokinetic and pharmacodynamic behavior of compounds and are effective for property-based classification of target-specific groups of active agents; however, it should be noted that for each particular target-specific activity group, another,

……more optimal set of descriptors can be found, which provides better classification ability.

A Kohonen SOM of the whole set of pharmaceutical leads and drugs generated as a result of the unsupervised learning procedure is depicted in Fig. 13. It shows that the studied compounds occupy a wide area on the map, which can be characterized as the area of drug-5likeness. Distribution of various target-specific groups of ligands in the Kohonen map demonstrates that most of these groups have distinct locations in specific regions of the map (Figure 14A through Figure 14E). Thus, agents have been shown to be active towards ICs, are located within a separate area.

Therefore, it can be reasonably suggested that these compounds have a common feature profile encoded by a set of descriptor used. A possible explanation of these differences is in that, as a rule, receptors of one type share a structurally conserved ligand-binding site. The structure of this site determines molecular properties that a receptor-selective ligand should possess to properly bind the site. These properties include specific spatial, lipophilic, and H-binding parameters, as well as other features influencing the pharmacodynamic characteristics; therefore, every group of active ligand molecules can be characterized by a unique combination of physicochemical parameters differentiating it from other target-specific groups of ligands. Another explanation of the observed phenomenon can be related to different pharmacokinetic requirements to drugs acting on different biotargets.

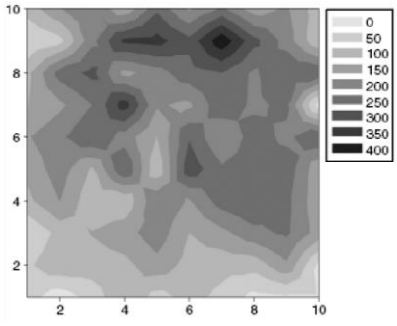
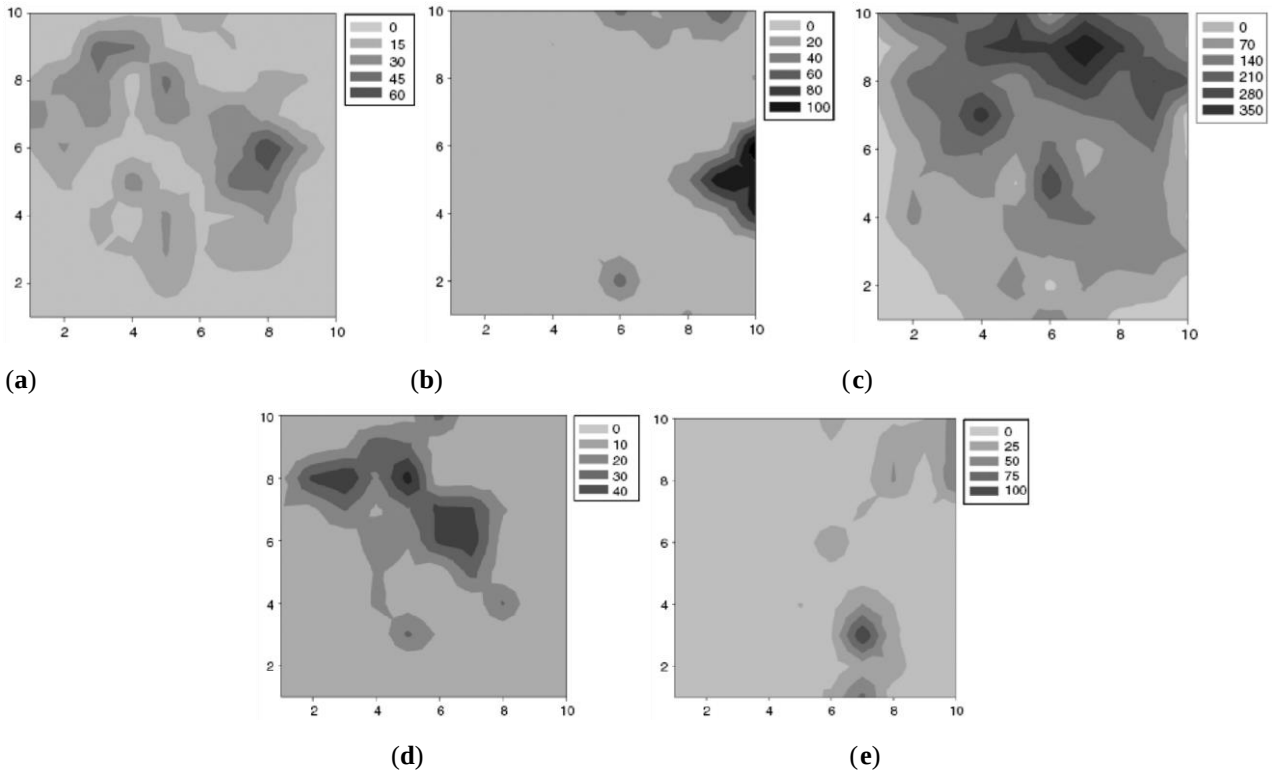


Fig. 13. Property space of 23K pharmaceutical leads and drugs visualized using the Kohonen map (the data have been smoothed).



(**a**) (**b**) (**c**)

(**d**) (**e**)

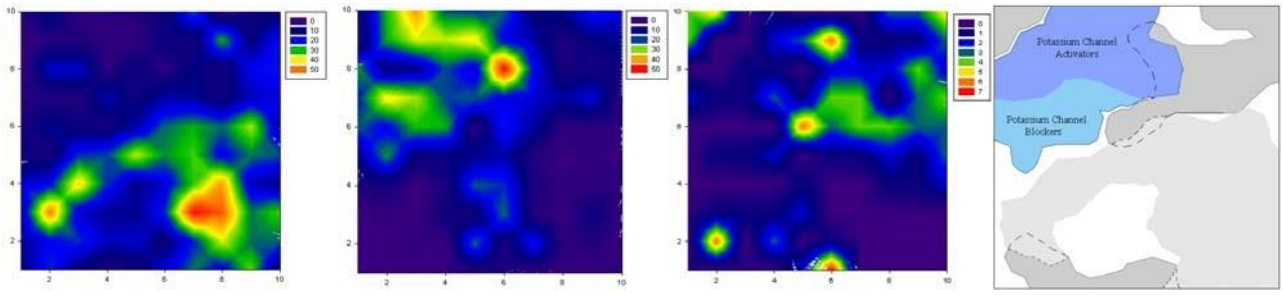
Fig. 14. Distribution of 5 large target-specific groups of pharmaceutical agents on the Kohonen map: (**a**) tyrosine kinase inhibitors (1405 compounds); (**b**) nuclear receptor agonists/antagonists (1021 compounds); (**c**) GPCR agonists/antagonists (12,512 compounds); (**d**) potassium channel activators (1060 compounds); (**e**) calcium channel antagonists (1230 compounds).

The described algorithm represents an effective procedure for selection of target-biased compound subsets compatible with high throughput *in silico* evaluation of large virtual chemical space. Whenever a large enough set of active ligands is available for a particular receptor, the quantitative discrimination function can be generated allowing selection of a series of compounds to be assayed against the target. Once a Kohonen network is trained and specific sites of location of target-activity groups of interest are identified, the model can be used for testing any available chemical databases with the same calculated descriptors. The Kohonen mapping procedure is computationally inexpensive and permits real-time calculations with moderate hardware requirements. Our own experience and literature data demonstrate that Kohonen SOMs are efficient clustering, quantization, classification, and visualization tools useful in the design of chemical libraries. Using the developed model we have predicted the target-specific activity of the selected compounds (200K). As a result, we have estimated that 100K structures have been successfully passed through the model.

Specific Kohonen-based SOM for the development of novel compounds targeted towards individual ICs types

From the initial reference database we have selected a particular set of known agonists and antagonists of the voltage-gated ion channels. Molecular features encoding the relevant physicochemical properties of the compounds were calculated from 2D and 3D molecular representations generated from energy-minimized conformations. Electrostatic, topological, lipophilic, and spatial descriptors were the main contributors to the models represented. We then applied an unsupervised learning procedure to generate the self-organizing Kohonen map of this database.

The distribution of ligands to particular ICs within the Kohonen map yielded interesting results. The IC-specific ligand groups being studied were localized in distinct clusters within specific regions of the map (Fig. 15). These differences in localization illustrate the underlying theory of property-based design; every group of active ligand molecules can be characterized by a unique combination of physico-chemical parameters differentiating it from other target-specific groups of ligands. As described above, the structure of a binding site determines the bundle of properties a receptor-selective ligand should possess in order to properly bind to the site, including specific spatial, lipophilic, topologic, and other features influencing the pharmacodynamic requirements.



(**a**) (**b**) (**c**) (**d**)

Fig. 15. Distributions of three IC-specific ligand groups within the Kohonen map: (**a**) calcium-channel antagonists (1611 cmpds.); (**b**) potassium-channel agonists and antagonists (1060 cmpds.); (**c**) sodium-channel antagonists (190 cmpds.); (**d**) combined map.

Our computational models quantitatively discriminated among different ligand groups, which is consistent with theory. Whenever a large enough set of active ligands is available for a particular receptor, quantitative discrimination functions can be generated that allow selection of a series of compounds to be assayed against the target. Moreover, to be capable of suggesting novel lead

chemotypes, this function must be based on physico-chemical rather than on structural features. As a result, 100,000 compounds selected from the previous step have been assigned to a particular IC-group while 25,000 compounds have been unclassified.

Our experimental work suggests that the neural network approach is a useful *in silico* tool in the combinatorial design of compounds active against families of ion channels and other targets. NN algorithms can be used for limiting the space of virtual combinatorial libraries and enhancing their target-specific informational content. Moreover, NN methods are computationally inexpensive and permit effective real-time calculations with moderate hardware requirements. In addition, we have successfully applied the classical back-propagation neural network to discriminate the particular classes of IC-targeted agents (*these results are not presented here*). We have used these results for calculating the consensus score.

6.2.1. Sammon mapping

Sammon mapping represents a useful alternative or supplement to Kohonen SOM algorithm because this dimensionality reduction technique often provides much greater detail about the individual compounds and their interrelationships. In contrast to Kohonen SOMs, the Sammon mapping preserves distances between the input and projection spaces, which can facilitate the analysis of relationships between the training parameters and output picture. To illustrate the application of Sammon mapping to determine the potential target-specific activity profile of compounds, consider the following example. The sites of distribution of tyrosine kinase inhibitors and potassium channel openers (Figure 14A and Figure 14D, correspondingly) on the Kohonen map are similar and therefore, differentiation using this particular set of molecular descriptors seems to be problematic. To test the classification ability of the alternative nonlinear mapping technique, we processed these two categories of active agents on a 2D Sammon map using the same set of molecular descriptors. The resulting Sammon map is shown in Figure 16.

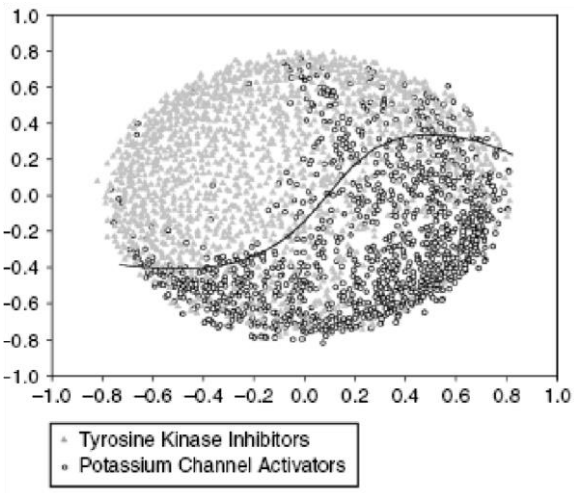


Fig. 16. Sammon map of the combined set of tyrosine kinase inhibitors (1405 agents) and potassium channel activators (1060 agents).

The positioning of the separation line is determined using nonlinear Support Vector Machine algorithm as it implemented in LibSVM-2.4 program [64]. This line provides the largest margin separating the studied classes (margin is defined as a sum of the shortest distances from decision line to the closest points of both classes) and thus, can serve as an optimal discriminator between the two studied compound categories. As Figure 16 and Figure 17 show, the studied categories of active agents—tyrosine kinase inhibitors and potassium channel openers—occupy distinctly different regions on the map. It can be concluded that in this particular case, Sammon mapping algorithm can be used as a complementary to Kohonen map tool, which discriminates the compound categories of interest with greater effectivity.

The analogues maps have been constructed for three classes of IC-targeted ligands. The obtained results have revealed that different IC-agents are located in different areas within the constructed map, and we have used the model for the assessment of compounds (75,000) selected above; thus, we have revealed that 50,000 structures have been successfully passed through the model.

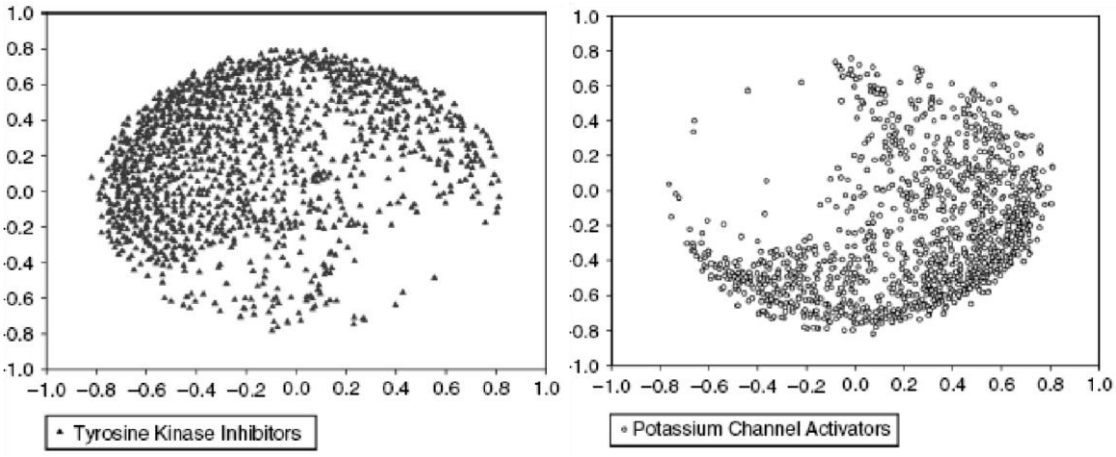


Fig. 17. Areas of distribution of tyrosine kinase inhibitors (**a**) and potassium channel activators; (**b**) shown separately on the same Sammon map.

In a real situation of combinatorial synthesis planning, a virtual combinatorial compound database should be processed on the Sammon map simultaneously with the active compounds. The sites of location of these active agents will define the combinatorial subset, which can be recommended for further synthetic development as a target-biased selection.

Among the other dimensionality reduction techniques that have appeared in statistical literature, Sammon nonlinear mapping is unique for its conceptual simplicity and ability to reproduce the topology and structure of the data space in a faithful and unbiased manner. A major shortcoming of this method is its quadratic dependence on the number of objects scaled, which imposes practical limitations on the size of data sets that can be effectively manipulated.

6.3. 3D-pharmacophore modeling/searching and 3D-molecular docking

In addition to the computational tools described above, we have also applied more accurate computational techniques for our IC-targeted library design. Based on two key 3D-operated *in silico* tools, 3D-pharmacophore modeling/searching, and 3D-molecular docking, we have built a robust model which has been successfully evaluated in ChemDiv.

Following the first computational approach, for example, we initially mined among the known PC-agents (the target-specific library contains topologic and bioisosteric analogs of reported potassium channel openers) using their 3D-conformations. As a result, we have developed the unique 3D-pharmacophore model to predict the target-specific activity of the selected compounds (50,000). Thus, we have found that more than 10,000 compounds have the following 3Dpharmacophore motif (Fig. 18) characteristic of more than 200 known PCOs including those shown in Fig. 4.

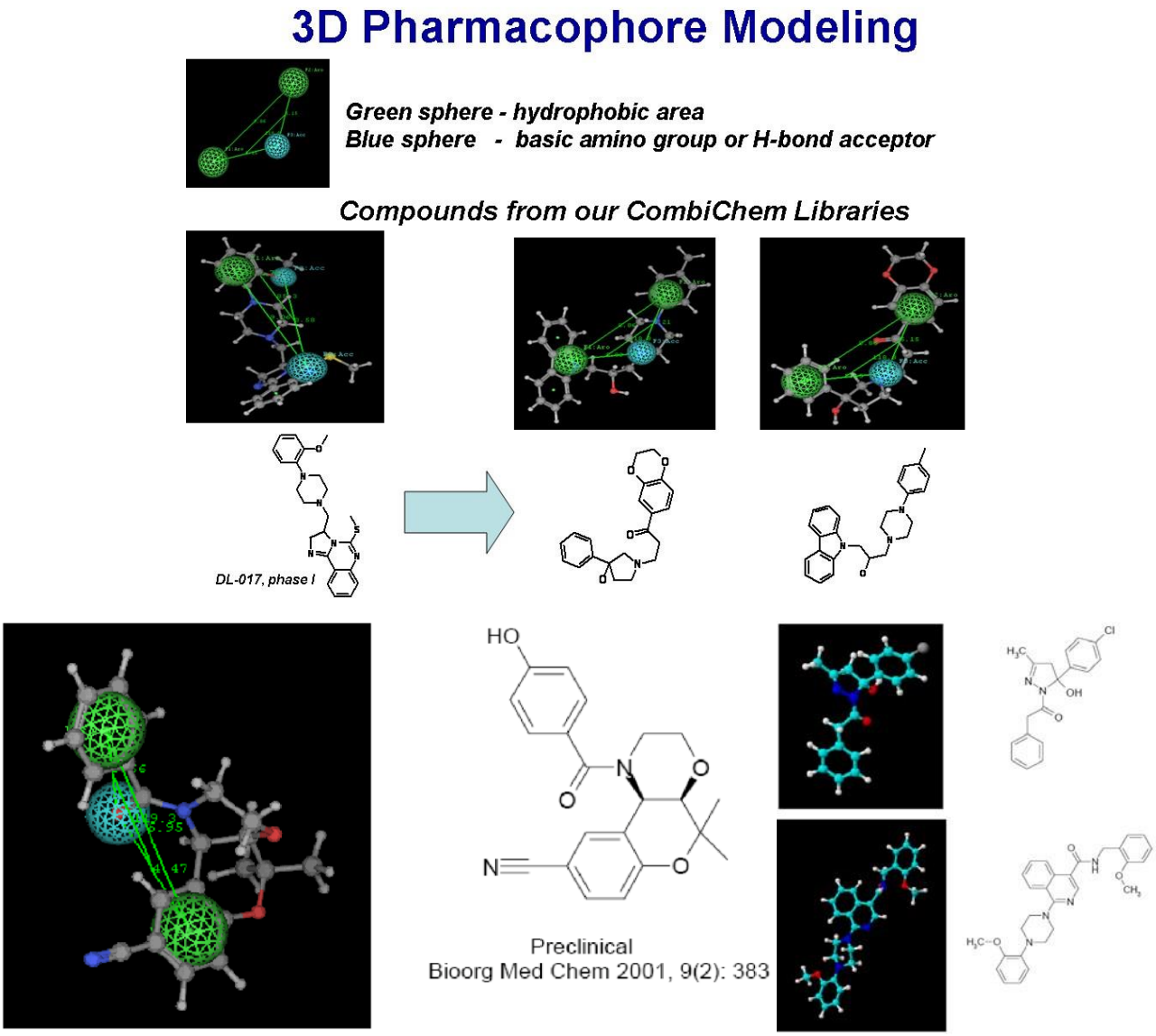


Fig. 18. 3D-pharmacophore model used as a query in the 3D search. Green sphere–centroid of aromatic ring; blue sphere – centroid of H-bond acceptor. The figure also shows the examples of compounds from the target-specific library that contain this pharamacophore motif.

The analogous modeling has been performed for the SC-targeted library using 3Dpharmacophore hypothesis generation and subsequent 3D searching. Fig. 19 below illustrates this approach. In this case, 5,500 structures from 50,000 have been successfully passed through the constructed pharmacophore model. We have also constructed a similar pharmacophore model for the CC-targeted agents. Applying the model to the 50,000 set, we have revealed that more than 15,000 compounds have been passed through the model. The combined set includes PC- (10,000), SC- (5,500), and CC-targeted compounds (15,000), of total – 30,500 unique structures. These structures have been further evaluated using a 3D-molecular docking approach (*these results are not presented*

here). Based on the obtained score we have assigned the compounds selected from the previous step (30,5K) to the individual IC-subclasses. The resulting set included 15,000 structures, which have been finally mined using various ADME/Tox and REOS filters (*see below*).

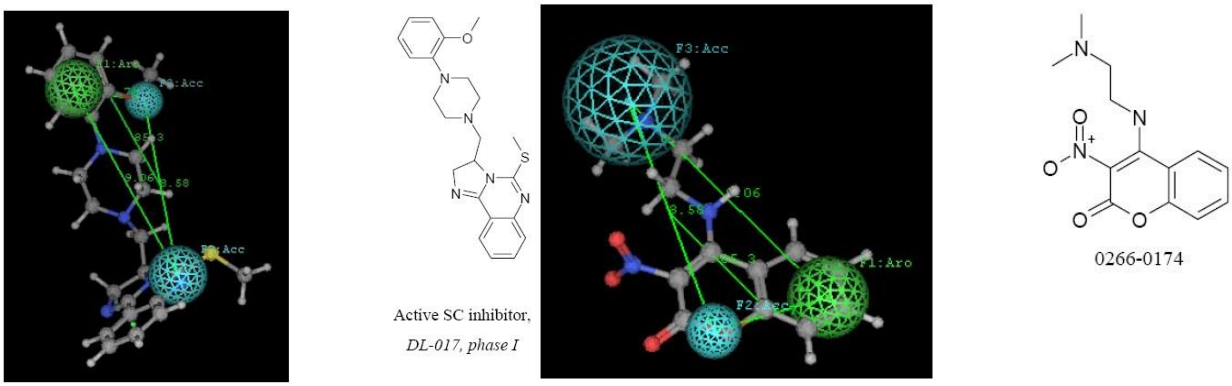
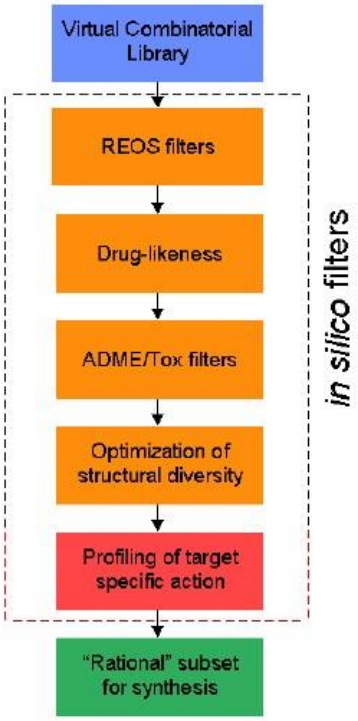


Fig. 19. 2D and 3D structures of an active SC inhibitor and a compound from SC-targeted library.

6.4. "Rapid Elimination of Swill" (REOS) filters

A general approach to limiting the space of virtual libraries consists of implementation of a series of special filtering procedures. The typical filtering stages we apply at CDL are summarized in Scheme 1.



Scheme 1. Selection of a rational subset via application of special filtering procedures.

We use a variety of "Rapid Elimination of Swill" (REOS) filters to "weed out" compounds that do not meet certain criteria. These criteria can include: 1) presence of certain non-desirable functional groups, such as reactive moieties and known toxicophores and 2) molecular size, lipophilicity, the number of H-bond donors/acceptors, and the number of rotatable bonds. For example, a molecule with a molecular weight 1000 and 10 H-bond donors would be eliminated as it is probably possesses very poor pharmacokinetics.

Compounds are thus, not included into the final IC-targeted library if they contain the following molecular fragments:

- alkylators;
- Michael acceptors, including all types of heterocycles containing exo-double C=C bonds, activated by electron-withdrawing group (false positive alarm);
- N-oxides, diazo- (R-N=N-R), and nitroso-compounds;
- thioureas including cyclic ones;
- quinones;
- Schiff bases, hydrazones, amidrazones;
- hydrazines, hydroxylamines and hydroxamic acids;
- more than one NO_2 group per molecule;
- NO_2 group, if NO_2 is the only functional substituent in molecule;
- NO_2 group, if halogen is introduced in the same ring;
- halogen bonded to carbon in the system –N=C-N- or analogous;
- acetals, aminals, thioketals, mixed ketal-like substructures (hidden carbonyls); - more than 3 conjugated rings, if two rings are aromatic.

At the next stage the design is focusing on “drug-likeness” of combinatorial molecules, a convenient starting point to develop a consensus definition(s) of a drug-like molecule is to analyze databases of known pharmaceutical agents [65].

6.5. ADME/Tox profiling

All the compounds/structures successfully passed through the computational filters above have been finally estimated towards their pharmacokinetic properties, such as CYP450-related

metabolism, BBB-permeability, PPB, and cellular toxicity. Several representative filters are described below.

6.5.1. Cytochrome P450-related metabolism

Currently, drug metabolism and toxicity in the human body are primarily the subject of clinical trials. The outcome can be extrapolated based on preclinical experiments, both *in vitro* (hepatocytes, organ slices, etc) and *in vivo* (several animal models). Acute human toxicity can be predicted fairly well, but chronic toxicity only reasonably on average, while nothing can be done for idiosyncratic toxicity.

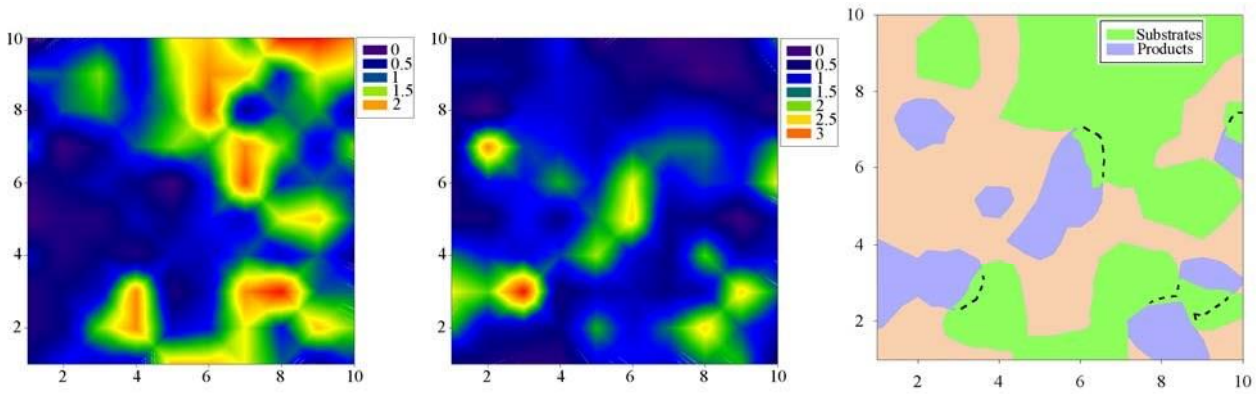
Cytochrome P450 (CYP) enzyme superfamily plays a central role in Phase I processing of xenobiotics. CYP enzymes represent mixed function monooxygenases capable of either inactivating or activating xeno and endobiotics molecules for further processing by Phase II bioconjugation enzymes. The major components of Phase I enzymatic complex are a phospholipid, a flavoprotein, a NADPH-cytochrome P450 oxidoreductase, and the hemoprotein cytochrome P450. CYPs are the terminal binding proteins of monooxygenase electron transport chain, important for catalyzing the oxidation of such endobiotics as fatty acids, steroids, ketones, polycyclic aromatic hydrocarbons, nitrosamines, hydrazines, arylamines. CYPs have been characterized as the most powerful in vivo oxidizing agents. The recent reviews on CYPs detail their chemistry, regulation, membrane topology, molecular biology, and provide the models for substrate binding sites. Out of about 40 human CYP genes cloned and described only three CYP families, a half-dozen subfamilies, and fewer than a dozen isoenzimes have been shown to play any significant role in hepatic processing of drugs. There may be a survival benefit associated with the use of such selected number of CYP isoforms. The active CYP enzymes have broad and overlapping substrate specificity, which poses a serious challenge to prediction of therapeutic or toxic outcomes of xenobiotic metabolism. We believe that experiment-based *in silico* models allow us to consider human metabolism and toxicity at earlier stages of drug discovery; thus, we have developed the unique integral Kohonen-based neural model for the assessment of P450 drug metabolism [66].

A comprehensive set of over 2200 substrate-product reactions for 38 human cytochromes and related expert work—for example, was performed by our partner GeneGo. Based on this database, we have developed a neural network computational algorithm for *in silico* assessment of the probability of cytochrome P450-mediated transformation for any novel drug-like compound. A Kohonen net with a 2D organization of the network nodes (neurons) was used in these experiments.

The smoothed projection of the combined data set of cytochrome substrates onto the 10×10 Kohonen map was conducted using the seven descriptors selected by principal component analysis (Figure 20a). The cytochrome substrates are distributed throughout the map as the irregularly shaped islands, with a clearly defined trend towards the right side of the map. The area occupied by the cytochrome substrates are relatively large, which reflects the broad substrate specificity of the studied set of cytochromes. We suggest that the physico-chemical properties of a molecule falling into the positive regions of the Kohonen map are consistent with the molecule's ability to be a cytochrome substrate.

For the comparison, we also processed the additional data set of 523 products of cytochrome-mediated biotransformations on the same Kohonen map (Figure 20b). This data set occupies distinct areas on the map substantially different from the regions of the substrates localization. The "product" compound category is unified by a combination of physico-chemical properties distinctly different from the cytochrome substrates; therefore, the sites of "products" localization on the Kohonen map can be used for the enhancement of prediction quality.

Based on these distributions, we built the smoothed contour plots of the occurrences of these two compound categories within the Kohonen map (Figure 20c). The area of "substrates" are marked in green, the area of "products" in blue, and the low-populated area in brown. The contours correspond to at least 1.5% of compounds per node, belonging to the particular category; therefore, these areas have a higher concentration of compounds compared to random distribution. Some overlap (5% of the total surface occupied by both substrates and products) is observed between these two distributions, which can be explained by the incompleteness of data for the set of "products." We suggest that a fraction of compounds assigned to the category of "products" from the overlapping regions, indeed, represents the cytochrome substrates. For this reason, the overlapping areas were assigned to "substrates". The model correctly classified 76.7 % of substrates and 62.7 % of products, as defined by their localization in the corresponding areas of the Kohonen map. A part of misclassified compounds (12.6 % and 22.4 % for substrates and products, correspondingly) falls into the area for which no specific assignment could be made. To correctly assess the cytochrome substrate potential for these compounds, one has to apply additional criteria. To summarize, we believe that the developed algorithm permits an effective classification of the compounds based on their ability to be the cytochrome P450 substrates.

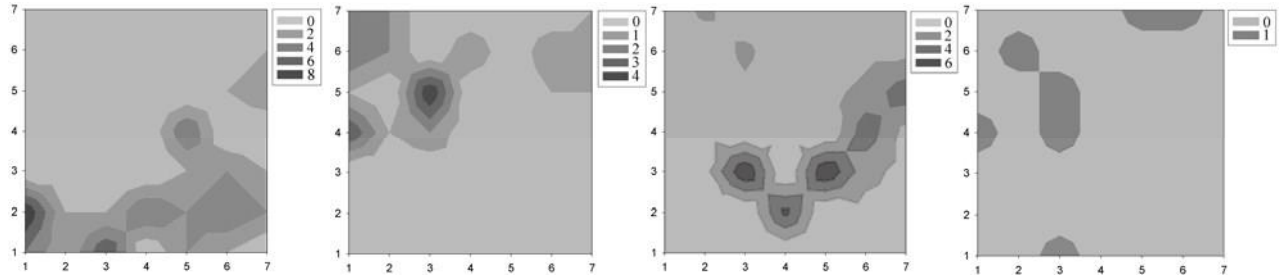


(**a**) (**b**) (**c**)

Fig. 20. The distribution of CYP450 substrates (**a**) and products (**b**) within the same Kohonen map; (**c**) - combined map.

The analogues *in silico* models have been constructed for the prediction of small molecule binding to human cytochrome P450 isoenzymes, particularly for 3A4 and 2D6; thus, a dataset of over 500 literature compounds with experimental apparent Km values for 12 human CYPs was obtained from the commercially available MetaDrugTM database (GeneGo, Inc., St Joseph, MI). Each compound was assigned to at least one enzyme-specific group within which the compounds were conditionally divided into three non-overlapping categories: low K_m (K_m < 10 μM), moderate K_m (K_m =10-100 μM), and high K_m (K_m > 100 μM). Prior to modeling experiments, the molecules were also filtered to ensure they were "drug-like" based on molecular weight (range 150–700) and atom type content (only C, N, O, H, S, P, F, Cl, Br). Some specific compound classes typically not related to drug-like agents, such as polyaromatic compounds and long-chain linear molecules (e. g., leukotrienes, fatty acids), were excluded from this reference set. Ultimately 491 compounds remained from the initial database.

The generation of the Kohonen self-organizing maps was conducted using the SmartMining software. Only one SOM was generated for the entire training set (491 compounds). After the SOM was generated, we studied the distribution of various compound groups (such as strong or poor binders, strong binders to particular isozymes, etc.) as separate maps. For illustration, we have shown the positions of low Km and high Km molecules for both CYP3A4 and CYP2D6 enzymes, which represent the two largest datasets studied (Fig. 21). Compounds with low K_m for both CYP3A4 and CYP2D6 occupied somewhat different sites on the map, though there was some substantial overlap between these enzymes.



(**a**) (**b**) (**c**) (**d**)

Fig. 21. Distribution of low K_m ($K_m < 10$ μM), and high K_m ($K_m > 100$ μM) molecules for two major CYP enzymes on the SOM map: (**a**) CYP3A4 low K_m (38 compounds), (**b**) CYP3A4 high K_m (32 compounds), (**c**) CYP2D6 low K_m (45 compounds), (**b**) CYP2D6 high K_m (7 compounds). The data have been smoothed for presentation purposes.

The distance between nodes on the SOM is a dimensionless parameter; it represents an abstract, discrete distance between the points in a multidimensional property space. For each isozyme-specific group, the areas of strong/poor binders can be identified as the nodes on the map, in which the percentage of strong/poor binders (with respect to their total number equal to 100%) are higher than the percentage of compounds belonging to the opposite category. In the case of CYP3A4, the model correctly classified 91 % high K_m and 97 % low K_m molecules as defined by their localization in the corresponding areas of the SOM. The quality of this discrimination is statistically significant only in the case of CYP3A4 for which a relatively large number of low K_m and high K_m molecules are available. Although the study suggests the method may be able to discriminate between the other CYPs, more data are required for a statistically valid result.

We have also applied the SOM to discriminate between low K_m and high K_m molecules across the whole panel of CYP enzymes. It should be taken into account that a molecule may have a low K_m with one CYP and a high K_m for another CYP. Accordingly, the same compound can be considered either a low Km or high K_m compound, depending on the specific CYP being considered. Such compounds were assigned to the low K_m category—likely due to it being the most important. Within the figure 22, the distribution of low K_m molecules with $K_m < 10$ with respect to at least one CYP isozyme, is shown as the green area, and the high Km molecules with $K_m > 100$ for at least one CYP isozyme (and no low Km values for any other isozyme) is shown as the blue area. These two compound categories occupy distinctly different sites on the map. The classification quality was 65% for low K_m and 82% for high K_m molecules which suggests some utility of this model for predicting global binding to human CYPs, which could clearly be improved upon.

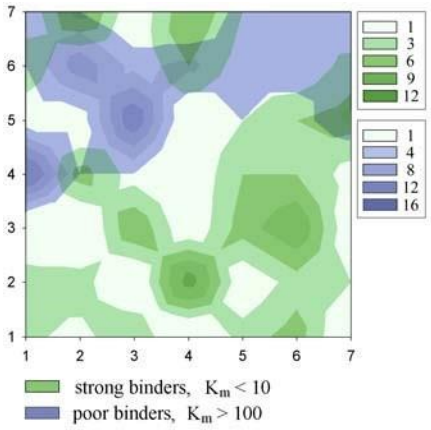
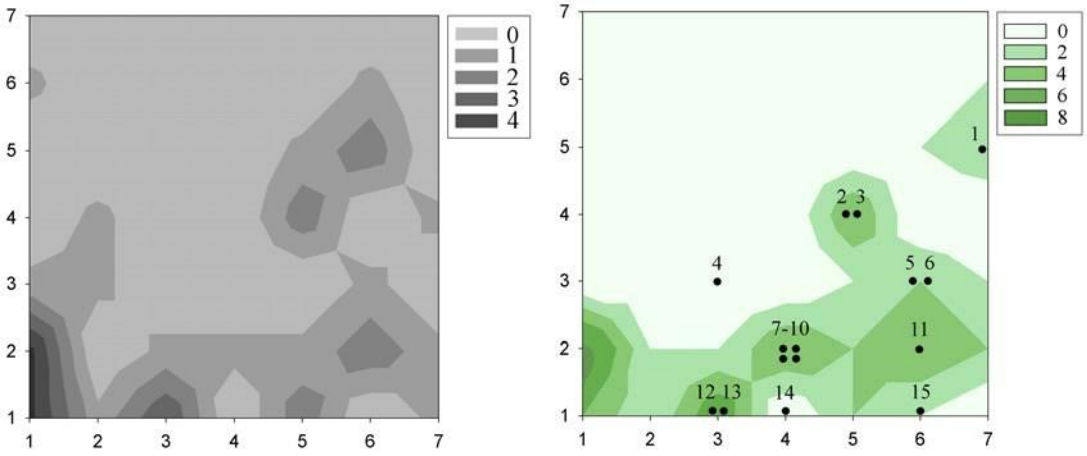


Fig. 22. An SOM map of low K_m (K_m < 10 μM), and high K_m (K_m > 100 μM) molecules for the whole panel of 12 CYP isozymes. The data have been smoothed for presentation purposes.

Two external test sets were used for assessment of a relationship between substrates and inhibitors of human CYPs. One set comprised 33 compounds which were classified in the MetaDrug™ database as reversible competitive CYP3A4 inhibitors. In addition a further 15 CYP3A4 competitive inhibitors were compiled from the literature and selected as an independent test set [67]; thus, 33 compounds overall, were classified in the MetaDrug™ database as reversible competitive CYP3A4 inhibitors and were processed on a SOM (Fig. 23a), of which 94 % were located in the area of low K_m CYP3A4 molecules. The molecular descriptors were calculated for these 15 molecules and then they were positioned on the same SOM (Fig. 23b). In this case 87 % of these molecules were located in the areas of low K_m CYP3A4 molecules. Only two compounds out of the 15—namely **4** and **14** (LY213829 and LY303870)—were misclassified yet are still located close to low K_m CYP3A4 molecules on the SOM.



(**a**) (**b**)

Fig. 23. Distribution of 33 (**a**) and 15 (**b**) competitive CYP3A4 inhibitors from the external test set within the constructed map.

We have also developed a benefit computational model for the prediction of metabolic *N*dealkylation reaction rates [68]; A total of 83 metabolic *N*-dealkylation reactions with experimental log V_{max} values for two major human P450s, CYP3A4, and CYP2D6, were studied in the mentioned work. Together, both P450s are responsible for hepatic metabolism of ca. 80% of drugs in humans [69]. For each reaction, structures of initial substrates and products were obtained from the MetaDrug data base (GeneGo, St. Joseph, MI). These *N*-dealkylation reactions in human enzyme assays generally followed Michaelis-Menten kinetics, allowing calculation of Vmax values—which ranged from 1×10^{-6} to 3.3×10^{3} pmol/min/pmol of enzyme.

Three types of molecular fragments belonging to the initial substrate molecules were considered (Fig. 24); for each initial molecule that the CYP3A4- or CYP2D6-mediated *N*-dealkylation occurred, the whole structure (A), the centroid with topological radius equal to three bonds (B), and the cleaved leaving fragment (C) were studied. Such a dissection strategy arose from our basic theoretical assumption that these elements of a substrate's organization will be crucial for the metabolic *N*-dealkylation rate—such an assumption has a solid theoretical basis.

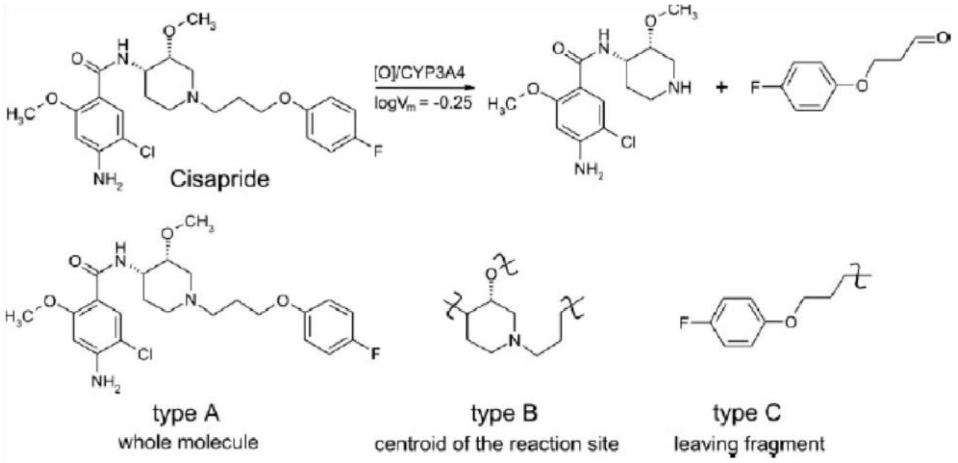
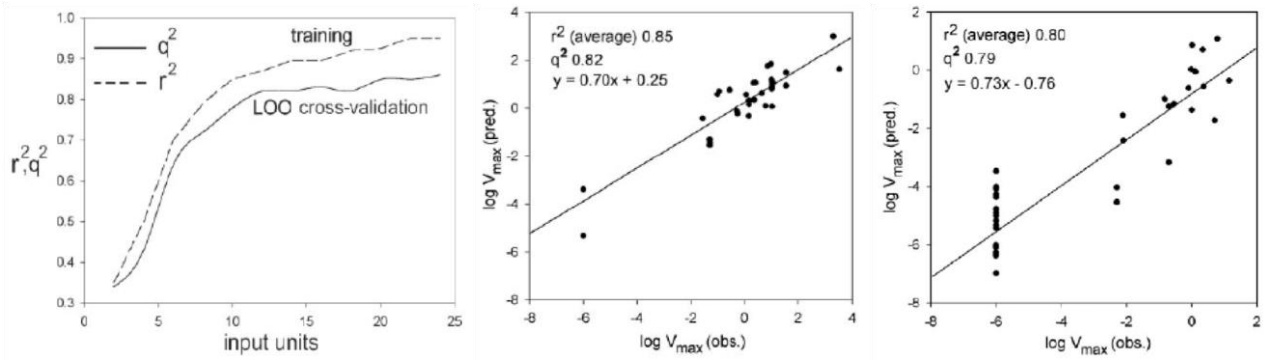


Fig. 24. An example showing the types of molecular fragments of the substrates studied.

A FFNN was generated and trained using the entire training set (31 objects for CYP3A4) and 121 input variables, which included 120 calculated descriptors and one phantom variable. After the neural network had been trained, a sensitivity measure per feature was obtained, and the procedure was repeated three times. We carried out a systematic training-testing experiment based on the crossvalidation leave-one-out (LOO) procedure to further reduce the number of inputs—more accurately select descriptors—and found the optimal architecture of the modular neural network. LOO works by leaving one data point out of the training set and giving the remaining instances (31 in the

case of the P4503A4 reaction set) to the learning algorithms for training. The process was repeated 32 times so that each example is a part of the test set only once. The resulting values for average training (r^2) and cross-validation (q^2) coefficients were calculated. For reasonable regression models, q^2 should be close to r^2, and is usually smaller [70]. Figure 25a shows the dependence of q^2 and average value of r^2 on the number of input variables for LOO cross-validation experiments using the modular neural network with four hidden units. The observed dependences are similar to each other; the change from 5 to 12 input units causes clear increases in r^2 and q^2 values and the predictive ability and the goodness of fit do not change significantly upon a further increase in a variable number.



(**a**) (**b**) (**c**)

Fig. 25. (**a**) Variation of r^2 and q^2 as a function of the number of input nodes used in the LOO cross-validation (CYP3A4 data set); (**b**) plot of the cross-validated log V_{max} (CYP3A4 set, 31 compounds) against the experimental values for the best neural network architecture (modular neural network with 12 input neurons and 2 hidden layers with 4 processing elements); (**c**) a plot of the cross-validated log V_{max} (CYP2D6 set, 36 compounds) against the experimental values for the best neural network architecture (a modular neural network with 12 input neurons and 2 hidden layers with 4 processing elements).

Figure 25b shows the cross-validated versus observed reaction rates for the best LOO crossvalidation experiment. There are no outliers in this model and the overall good conformity between the predicted and observed log V_{max} values resulted in comparable r^2 and q^2 values of 0.85 and 0.82, respectively. For the CYP2D6 training set, we have also performed a LOO procedure, which generated 36 QSMR models. Figure 25c shows the crossvalidated versus observed log V_{max} values for this model. This plot demonstrates good prediction quality with good q^2 and r^2 values (0.79 and 0.80, correspondingly). The general conclusion that emerges from this experiment is that for the CYP2D6 *N*-dealkylation reaction set, the developed QSMR models based on the same 12 molecular descriptors as for CYP3A4 *N*-dealkylation provide reasonable generalization accuracy and predictive power.

There are several ways to evaluate the predictive ability of a computational model; leaving groups out and scrambling the descriptors with the biological activity are perhaps the most widely used. The most valuable test is an external set of molecules that have been excluded from the modelbuilding process. In this study, nine CYP3A4-mediated and five CYP2D6-mediated *N*dealkylation reactions with known V_{max} values were collected from the literature and used to test the respective models. A comparison of the calculated and experimental data for the test set reactions demonstrates a good predictive power of the developed models with R^2 values equal to 0.90 and 0.94 for CYP3A4 and CYP2D6, respectively. Considering an obvious `prediction power` of the constructed models it has been effectively applied for partial ADME/Tox assessment within our IC-targeted library design.

6.5.2. Blood-brain barrier permeability

The BBB is a metabolically active tissue that facilitates and controls the brain uptake of certain solutes while helping to maintain homeostasis within the central nervous system. Optimizing the distribution of therapeutic compounds between brain and blood is one of the key issues in the design of novel CNS-active drugs. Given the problem's importance, reliable new methods of effective presynthetic assessment of BBB permeability are needed for the discovery of CNS-active agents. This problem is especially important in the early stages of the drug discovery process, where high-throughput and efficient estimation of BBB permeability represents a serious bottleneck in the design of CNS-targeted or non-CNS targeted compound libraries. We created a robust qualitative model for the early assessment of the BBB permeability of small and therapeutically relevant molecules. The methodology is based on the extraction of knowledge from a large number of literature sources on BBB permeation of organic compounds using the same unsupervised Kohonen learning approach. A comprehensive set of experimental data on 502 compounds (including 197 and 305 compounds with poor BBB(-) and good BBB permeation BBB(+), respectively) were collected. It was assumed that only passive diffusion mechanisms are involved in the BBB transport of these compounds. Statistical analysis enabled the selection of an optimal set of molecular descriptors for the effective prediction of BBB penetration. The projection of the combined data set of BBB(+) and BBB(-) compounds was generated onto the Kohonen map (Fig. 26). The data set of BBB(+) compounds occupies a distinct area on the map substantially different from that of the BBB(-) compounds; therefore, the location of a compound's site on the Kohonen map can be used for the assessment of its BBB-permeability. The constructed learning model is useful in reducing the size of libraries of potential CNS active agents.

It can also be used as an *in silico* filter to assist in the synthetic design and planning of novel combinatorial libraries.

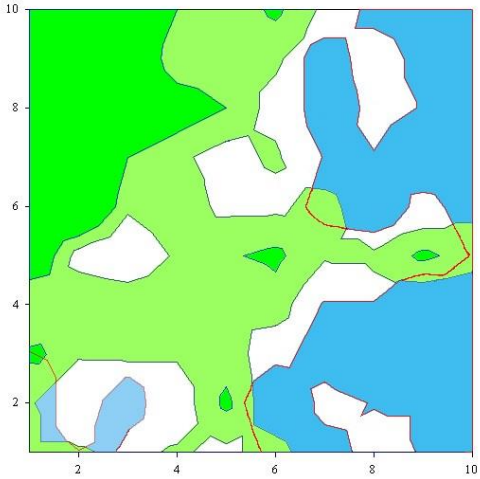
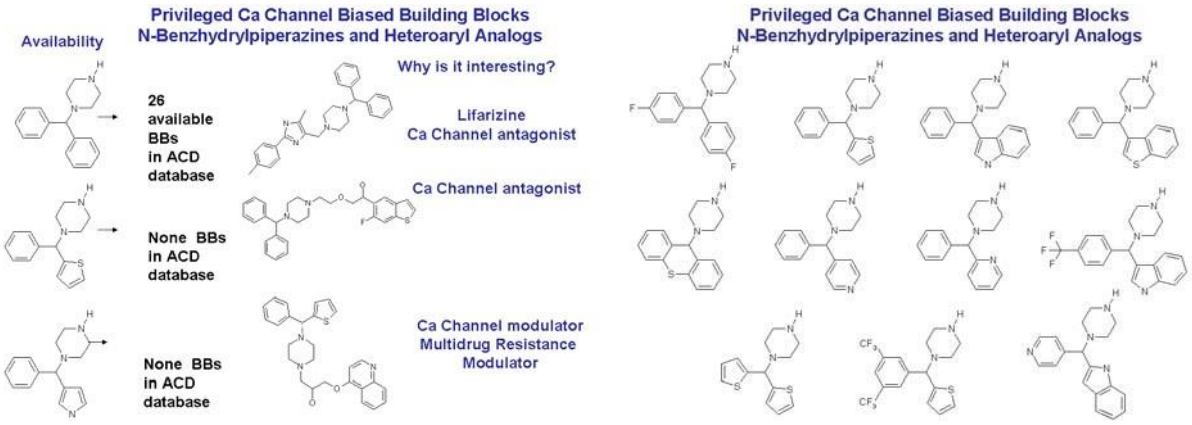
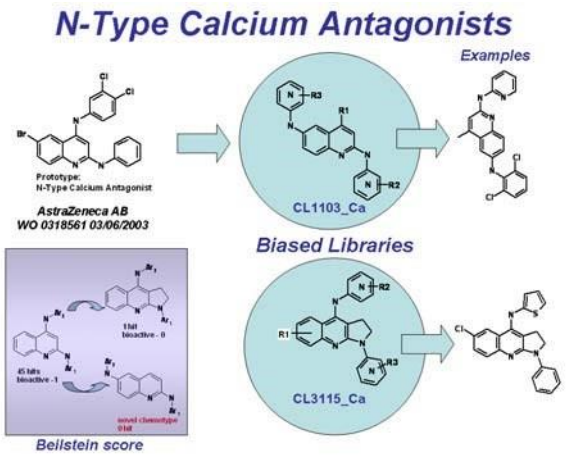


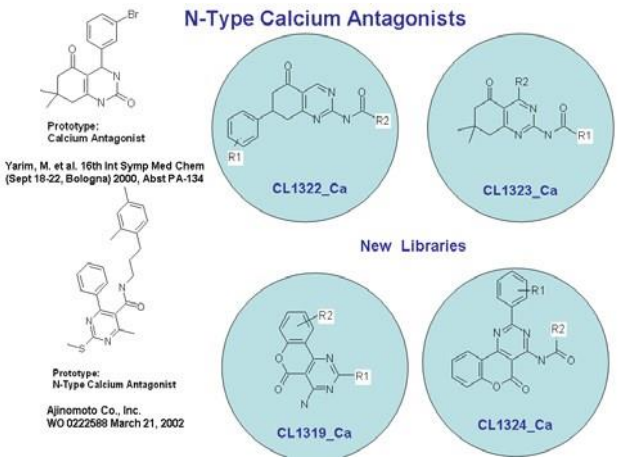
Fig. 26. Smoothed contour plots of the occurrences of BBB(+) (green area) and BBB(-) (blue area) compounds within the Kohonen map. The contours correspond to at least 1.5% of compounds, from a particular category, per node.

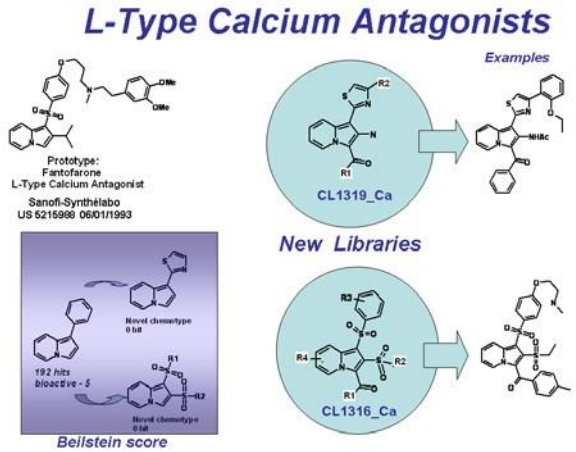
6.6. A brief overview of IC-targeted library

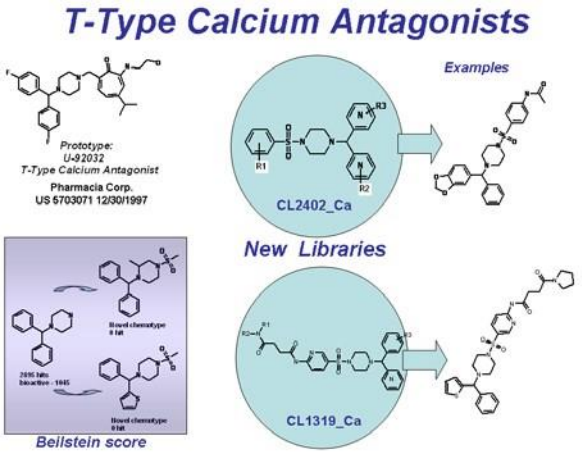
As a final result of our thorough computational exercises, a 7200-Ion Channels focused library has been compared for sale. This library contains compounds with high structural diversity and *in silico* prediction score. A brief overview of our IC-targeted library is presented below.

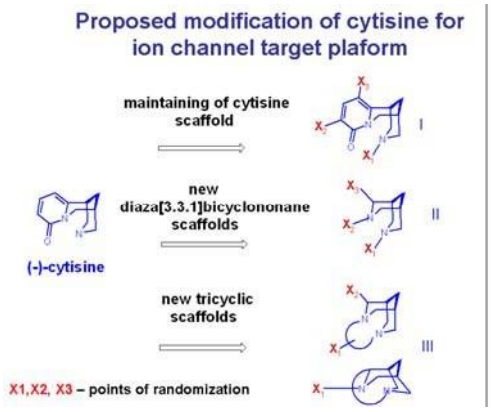


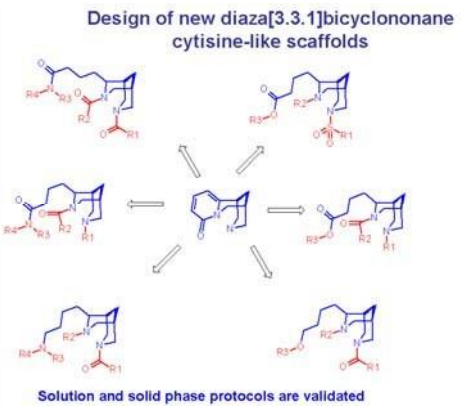

N-Type Calcium Antagonists
Examples
Prototype:
N-Type Calcium Antagonist
AstraZeneca AB
WO 0318561 03/06/2003
CL1103_Ca
Biased Libraries
CL3115_Ca
Beilstein score


N-Type Calcium Antagonists
Prototype:
Calcium Antagonist
Yarim, M. et al. 16th Int Symp Med Chem
(Sept 18-22, Bologna) 2000, Abst PA-134
CL1322_Ca
CL1323_Ca
New Libraries
Prototype:
N-Type Calcium Antagonist
Ajinomoto Co., Inc.
WO 0222588 March 21, 2002
CL1319_Ca
CL1324_Ca


L-Type Calcium Antagonists
Examples
Prototype:
Fantofarone
L-Type Calcium Antagonist
Sanofi-Synthélabo
US 5215988 06/01/1993
CL1319_Ca
New Libraries
CL1316_Ca
Beilstein score


T-Type Calcium Antagonists
Examples
Prototype:
U-92032
T-Type Calcium Antagonist
Pharmacia Corp.
US 5703071 12/30/1997
CL2402_Ca
New Libraries
CL1319_Ca
Beilstein score


Proposed modification of cytisine for
ion channel target plaform
maintaining of cytisine
scaffold
I
new
diaza[3.3.1]bicyclononane
scaffolds
II
(-)-cytisine
new tricyclic
scaffolds
III
X1,X2, X3 – points of randomization


Design of new diaza[3.3.1]bicyclononane
cytisine-like scaffolds
Solution and solid phase protocols are validated

Ca-channels:

L-Type	dihydropyridines, phenylalkylamines, benzothiazepines	***Contraction of smooth, skeletal and cardiac muscle Hormone release***
N-Type	ω-conotoxin GVIA (irreversible), quinolines	***Mediate neurotransmitter release at some synapses Modulated by neurotransmitters and hormones***
P/Q-Type	spider peptide ω-Aga-IVa, ω-conotoxin MVIIC	***Mediate neurotransmitter release at some synapses***
R-Type	Resistant to ω-Aga-IVa, ω-conotoxin MVIIC, DHP	
T-Type	DHP	***Mediate neurotransmitter and hormone release***

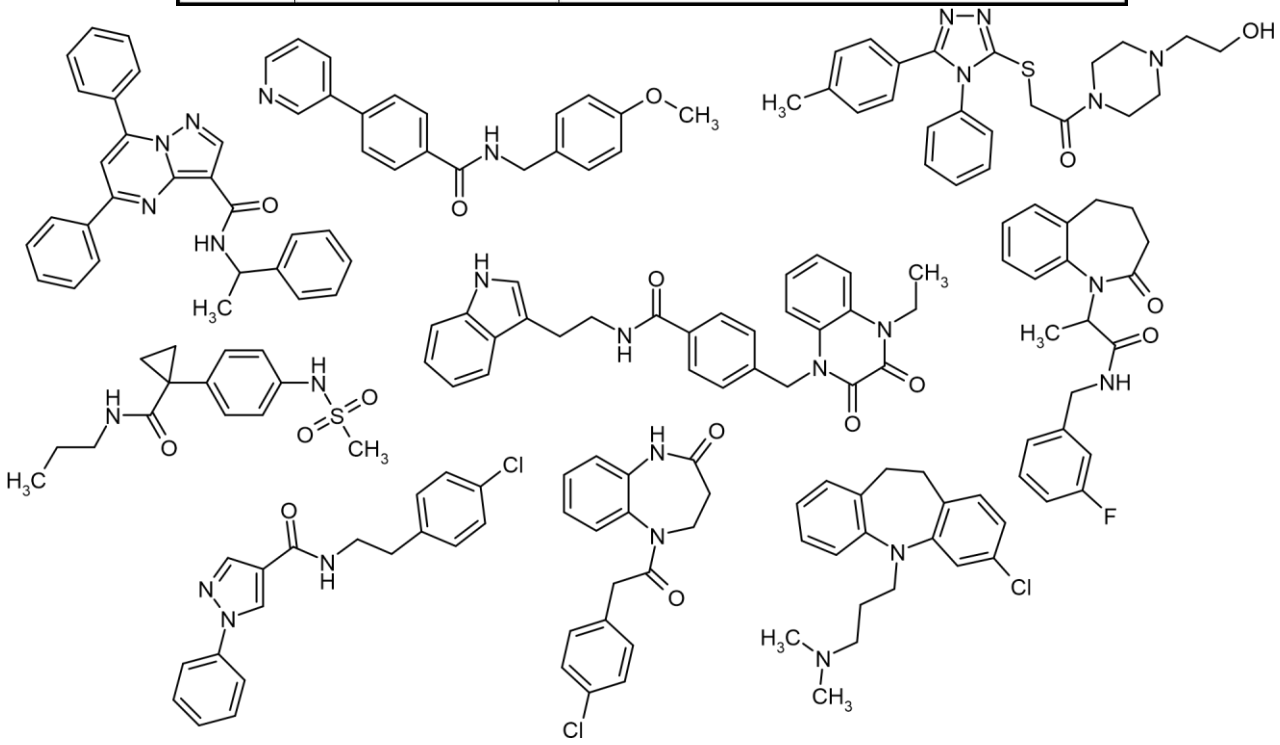


Fig 27. Representative examples of compounds from the IC-targeted library.

IC-targeted library profile:

• About 3K known ligands for known VGICs comprising the knowledge base;

• Over 8,7K compounds from about 240 sub-libraries represented;

• Purity: all the compounds analyzed; >90% average purity;

• The library contains 200 unique heterocycles, over 2,800 “screen” subfragments;

• The library can be expanded in the feasible space of 30K structures.

Property range:

• 218 < MW < 526; 413 on average;

• 1 < H-bond acceptors < 10; 5 on average;

• 0 < H-bond donors < 5; 1 on average;

• 0 < rotatable bonds < 12; 4 on average;

• -0.9 < clogP < 5.8; 3.4 on average;

• -9.9 < clog of solubility in water (pH 7.4) < 9.0; -4.0 on average.

Conclusion

Analysis of information contained within human genome along with innovations in combinatorial synthesis and biological screening provides for the new opportunities in the design of novel effective drugs. However, despite the fact that these high-throughput technologies have become common within the modern drug discovery process, their accuracy and efficiency require further improvement. One of the possible solutions includes development and adoption of special computational technologies for making combinatorial library design cost-effective.

Over the past few years, various computational concepts and methods have been introduced to extract relevant information from the accumulated knowledge of chemists and biologists and to create a robust basis for rational design of chemical libraries. The obvious trend is that molecular diversity alone cannot be considered to be a sufficient library design criterion. We can also observe a clear shift from the ligand-structure-based methods toward more sophisticated docking algorithms. At the same time, rapid, reliable, and conceptually simple ligand-based strategies are still very useful as pre-screening procedures especially for the cases when the structure of a target is unknown. Knowledge-based methods successfully complement the above mentioned strategies to create information-rich compound collections optimized by multiple parameters. Among such *in silico* approaches the mapping techniques presented denote useful approaches to filtering combinatorial libraries for selection of target-specific subsets. They often permit the user to reduce the size of initial chemistry space up to two orders of magnitude and can be recommended as efficient classification and visualization tools for practical combinatorial design. It is important that these property-based methods are complementary to other target and ligand structure-based approaches to VS. Kohonen map-based method is compatible with high throughput VS protocols and that Sammon mapping technique is more relevant to analysis of small-to-medium-sized chemical libraries.

Based on advanced computational algorithms we have developed and effectively applied a multi-step *in silico* approach to design of our IC-targeted library. In particular, we have successfully validated this strategy towards calcium, potassium and sodium channels. The related biological trials have revealed several highly potent inhibitors: we can confidently conclude that described *in silico*

pathway represents an effective method for IC-targeted libraries design. Moreover, we provide rapid and efficient tools for follow-up chemistry on discovered hits, including single isomer chemistry stereoselective synthesis and racemic mixture separation. The developed libraries are updated quarterly based on a "cache" principle. Older scaffolds/compounds are replaced by templates resulting from our in-house development (unique chemistry, literature data, computational approaches) while the overall size of the library remains the same (ca.7-8K compounds). As a result, the libraries are renewed each year, proprietary compounds comprising 50-75% of the entire set. Clients are invited to participate in the template selection process prior to launch of our synthetic effort.

References

1 (a) Anger T. et al. J. Med. Chem. 2001, 44, 115-137; (b) Gilbert A.M. et al. J. Med. Chem. 2000, 43, 1203-1214; (c) Williams M. et al. J. Med. Chem. 1999, 42, 1481-1500; (d) Cox B. and Denyer J.C. Exp. Opin. Ther. Patents 1998, 8, 1237-1250.

2 For example: (a) http://www.iuphar-db.org/iuphar-ic/. 3 (a) Ashcroft S.J.H. and Ashcroft F.M. Cell Signal. 1990, 2, 197-214; (b) Isomoto S. et al. J. Cardiovasc. Electrophysiolog., 1997, 8,
1431-1446; (c) Gilbert A.M. et al. J. Med. Chem. 2000, 43, 1203-1214

4 Tusnady and Simon, 1998

5 Thompson et al., 1997

6 Page, 1996

7 Uozumi et al., 1998

8 http://integrity.prous.com 9 (a) Madge D. J. Annu. Rep. Med. Chem. 1998, 51-60; (b) Taylor C. P. Curr. Pharm. Des. 1996, 2, 375-388; (c) Anger T. et al. J. Med. Chem. 2001, 44, 115-137.

10 Tanabe et al. 1987

11 Bech-Hansen et al. 1998

12 Perez-Reyes et al. 1998

13
Carbone & Lux, 1984

14 Nowycky et al. 1985

15 Aidley & Stanfield, 1996

16 Zhu et al. 1996

17 Caterina et al. 1997

18 Mignery et al. 1989 19 Oprea TI. Chemical space navigation in lead discovery. Curr Opin Chem Biol 2002; 6: 384-9.

20 (a) Dean PM, Lewis RA. Molecular diversity in drug design. Dordrecht: Kluwer, 1999; (b) Agrafiotis D. Diversity of chemical libraries. In: Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR, editors. Encyclopedia of computational chemistry. Chichester: Wiley, 1998. p. 742-61. 21 Merlot C, Domine D, Cleva C, Church DJ. Chemical substructures in drug discovery. Drug Discov Today 2003; 8: 594-602

22 (a) Patchett AA, Nargund RP. Privileged structures – an update. Annu Rep Med Chem 2000; 35: 289-98; (b) Muller G. Medicinal chemistry of target family-directed masterkeys. Drug Discov Today 2003; 8: 681-91.

23 Bondensgaard K, Ankersen M, Thogersen H, Hansen BS, Wulff BS, Bywater RP. Recognition of privileged structures by gprotein coupled receptors. J Med Chem 2004; 47: 888-99

24 (a) Johnson MA, Maggiora GM. Concepts and applications of molecular similarity. New York: John Wiley & Sons, 1990; (b) Downs GM, Willett P. Similarity searches in databases of chemical structures. Rev Comput Chem 1995; 7: 1-66; (c) Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? Drug Disc Today 2002; 7: 903-11.

25 Xue L, Stahura FL, Godden JW, Bajorath J. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. J Chem Inf Comp Sci 2001; 41: 394-401

26 Zheng W, Cho SJ, Tropsha A. Rational combinatorial library design 1 Focus-2D: a new approach to the design of targeted combinatorial chemical libraries. J Chem Inf Comput Sci 1998; 38: 251-8

27 (a) Martin YC. 3D database searching in drug design. J Med Chem 1992; 35: 2145-54; (b) Mason JS, Good AC, Martin EJ. 3D pharmacophores in drug discovery. Curr Pharm Des 2001; 7: 567-97

28 (a) Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Labaudiniere RF. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. J Med Chem 1999; 42: 3251-64; (b) Makara GM. Measuring molecular similarity and diversity: total pharmacophore diversity. J Med Chem 2001; 44: 3563-71.

[29] (a) Andrews KM, Cramer RD. Toward general methods of targeted library design: topomer shape similarity searching with diverse structures as queries. J Med Chem 2000; 43: 1723-40; (b) Cramer RD, Jilek RJ, Andrews KM. Dbtop: topomer similarity searching of conventional structure databases. J Mol Graph Model 2002; 20: 447-62.

[30] Laird ER, Blake JF. Structure-based generation of viable leads from small combinatorial libraries. Curr Opin Drug Discov Devel 2004; 7: 354-9

[31] Tondi D, Costi MP. Enhancing the drug discovery process by integration of structure-based design and combinatorial synthesis. In:

Viswanadhan AK, Ghose VN, editors. Combinatorial library design and evaluation. New York: Marcel Dekker Inc, 2001. p. 563-604

[32] (a) An J, Totrov M, Abagyan R. Comprehensive identification of "druggable" protein Ligand binding sides. Genome Informatics 2004; 15: 31-41; (b) Murray CW, Clark DE, Auton TR, Firth MA, Li J, Sykes RA, Waszkowycz B, Westhead DR, Young SC. PRO_SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. J Comput Aided Mol Des 1997; 11: 193-207; (c) Makino S, Ewing TJ, Kuntz ID. DREAM++: flexible docking program for virtual combinatorial libraries. J Comput Aided Mol Des 1999; 13: 513-32; (d) Rarey M, Lengauer T. A recursive algorithm for efficient combinatorial library docking perspective in drug discovery and design. Perspective in Drug Discovery and Design 2000; 20: 63-81; (e) Sprous DG, Lowis DR, Leonard JM, Heritage T, Burkett SN, Baker DS, Clark RD. OptiDock: virtual HTS of combinatorial libraries by efficient sampling of binding modes in product space. J Comb Chem 2004; 6: 530-9; (f) Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring 1 Method and assessment of docking accuracy. J Med Chem 2004; 47: 1739-49.

[33] Barril X, Morley SD. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. J Med Chem 2005; 48: 4432-43

[34] Kohonen T. Self-Organizing Maps, 3rd edn. New York: Springer Verlag, 2000

[35] Sammon JE. A nonlinear mapping for data structure analysis. IEEE Trans Comput 1969; C-18: 401-9

[36] Rabow AA, Shoemaker RH, Sausville EA, Covell DG. Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. J Med Chem 2002; 45: 818-40

[37] Polanski J, Jarzembek K, Gasteiger J. Self-organizing neural networks for screening and development of novel artificial sweetener candidates. Comb Chem High Throughput Screen 2000; 3: 481-95

[38] Teckentrup A, Briem H, Gasteiger J. Mining high-throughput screening data of combinatorial libraries: development of a filter to distinguish hits from nonhits. J Chem Inf Comput Sci 2004; 44: 626-34

[39] Sammon JE. A nonlinear mapping for data structure analysis. IEEE Trans Comput 1969; C-18: 401-9

[40] Agrafiotis DK, Myslik JC, Salemme FR. Advances in diversity profiling and combinatorial series design . Mol Div 1999; 4: 1-22

[41] (a) Agrafiotis DK, Lobanov VS. Nonlinear mapping networks. J Chem Inf Comput Sci 1997; 40: 1356-62; (b) Xie D, Tropsha A, Schlick T. An efficient projection protocol for chemical databases: singular value decomposition combined with truncated-newton minimization. J Chem Inf Comput Sci 2000; 40: 167-77; (c) Garrido L, Goméz S, Roca J. Improved multidimensional scaling analysis using neural networks with distance-error backpropagation. Neural Comput 1999; 11: 595-600; (d) Pal NR, Eluri VK. Two efficient connectionist schemes for structure preserving dimensionality reduction. IEEE Trans Neural Networks 1998; 9: 1142-54; (e) König A. Interactive visualization and analysis of hierarchical neural projections for data mining. IEEE Trans Neural Networks 2000; 11: 615-24. [42] Winkler DA. Neural networks as robust tools in drug lead discovery and development. Mol Biotechn 2004; 27: 139-68

[43] (a) Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. J Pharm Toxicol Methods 2000; 44: 235-49; (b) Walters WP, Murcko MA. Prediction of 'drug-likeness'. Adv Drug Del Rev 2002; 54: 255-71; (c) Oprea TI. Current trends in lead discovery: are we looking for the appropriate properties? J Comput Aided Mol Des 2002; 16: 325-34; (d) Wenlock MC, Austin RP, Barton P, Davis AM, Leeson PD. A comparison of physiochemical property profiles of development and marketed oral drugs. J Med Chem 2003; 4: 1250-6

[44] (a) Ekins S, Boulanger B, Swaan PW, Hupcey MAZ. Towards a new age of virtual ADME/TOX and multidimensional drug discovery. J Comput Aided Mol Des 2002; 16: 381-401; (b) Ekins S, Rose JP. In silico ADME/Tox: the state of the art. J Mol Graph 2002; 20: 305-9; (c) Van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2003; 2: 192-204

[45] (a) de Groot MJ, Ekins S. Pharmacophore modeling of cytochromes P450. Adv Drug Del Rev 2002; 54: 367-73; (b) Ekins S, de Groot M, Jones JP. Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome P450 active sites. Drug Metab Dispos 2001; 29: 936-44

[46] Ekins S, Kim RB, Leake BF, Dantzig AH, Schuetz E, Lan LB, Yasuda K, Shepard RL, Winter MA, Schuetz JD, Wikel JH, Wrighton SA. Three-dimensional quantitative structure-activity relationships of inhibitors of P-glycoprotein Mol Pharmacol 2002; 61: 964-73

[47] Ekins S, Erickson JA. A pharmacophore for human pregnane X-receptor ligands. Drug Metab Dispos 2002; 30: 96-9

[48] Aronov AM, Goldman BB. A model for identifying HERG K+ channel blockers. Bioorg Med Chem 2004; 12: 2307-37

[49] (a) Ekins S, Berbaum J, Harrison RK. Generation and validation of rapid computational filters for Cyp2D6 and Cyp3A4. Drug Metab Dispos 2003; 31: 1077-80; (b) Young SS, Gombar VK, Emptage MR, Cariello NF, Lambert C. Mixture deconvolution and analysis of Ames mutagenicity data. Chemo Intell Lab Sys 2002; 60: 5-22

[50]

Korolev D, Balakin KV, Nikolsky Y, Kirillov E, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Nikolskaya T. Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach. J Med Chem 2003; 46: 3631-43

51 Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Del Rev 1997; 46: 3-25

52 Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 2002; 45: 2615-23

53 Balakin KV, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Ekins S. Comprehensive computational assessment of ADME properties using mapping techniques. Current Drug Discovery Technologies 2005; 2: 99-113

54 Balakin KV, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Ekins S. Comprehensive computational assessment of ADME properties using mapping techniques. Current Drug Discovery Technologies 2005; 2: 99-113

55 Korolev D, Balakin KV, Nikolsky Y, Kirillov E, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Nikolskaya T. Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach. J Med Chem 2003; 46: 3631-43

56 Balakin KV, Ekins S, Bugrim A, Ivanenkov YA, Korolev D, Nikolsky YV, Skorenko AV, Ivashchenko AA, Savchuk NP, Nikolskaya T. Kohonen maps for prediction of binding to human cytochrome P450 3A4. Drug Metab Dispos 2004; 32: 1183-9

57 Balakin KV, Ivanenkov YA, Skorenko AV, Nikolsky YV, Savchuk NP, Ivashchenko AA. In silico estimation of DMSO solubility of organic compounds for bioscreening. J Biomol Scr 2004; 9: 22-31

58 Savchuk NP, Balakin KV. Data mining approaches for enhancement of knowledge-based content of de novo chemical libraries. In: Alvarez H, Shoichet B, editors. Virtual screening in drug discovery. New York: CRC Press, 2005. p. 121-49

59 (a) Davis AM, Riley RJ. Predictive ADMET studies The challenges and the opportunities. Curr Opin Chem Biol 2004; 8: 378-86;
(b) Van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2003; 2: 192204

60 (a) Burger, A. Prog. Drug Res., 1991, 37, 287; (b) Patani, G.A.; LaVoie, E.J. Chem. Rev., 1996, 96, 3147; (c) Olesen, P.H. Curr. Opin. Drug Discov. Devel., 2001, 4, 471; (d) Chen, X.; Wang, W. Ann. Reports Med. Chem., 2003, 38, 333.

61 (a) Patani, G.A.; LaVoie, E.J. Chem. Rev., 1996, 96, 3147; (b) Chen, X.; Wang, W. Ann. Reports Med. Chem., 2003, 38, 333.

62 (a) J. Krumrine, F. Raubacher, N. Brooijmans, and I. Kuntz, Principles and methods of docking and ligand design, Methods Biochem. Anal. 44:443–476, 2003; (b) D. Xu, Y. Xu, and E.C. Uberbacher, Computational tools for protein modeling, Curr. Protein Pept. Sci. 1:1–21, 2000; (c) R.D. Taylor, P.J. Jewsbury, and J.W. Essex, A review of protein-small molecule docking methods, J. Computer-Aided Mol. Des. 16:151–166, 2002; (d) P.J. Gane and P.M. Dean, Recent advances in structure-based rational drug design, Curr. Opin. Drug Discov. Dev. 10:401–404, 2000

63 (a) P. Willett, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38:983–996, 1998; (b) M.I. Skvortsova, I.I. Baskin, I.V. Stankevich, V.A. Palyulin, and N.S. Zefirov, Molecular similarity. 1. Analytical description of the set of graph similarity measures, J. Chem. Inf. Comput. Sci. 38:785–790,1998; (c) R.P. Sheridan and S.K. Kearsley, Why do we need so many chemical similarity search methods?, Drug Discov. Today 7:903–911, 2002; (d) C. Merlot, D. Domine, C. Cleva, and D.J. Church, Chemical substructures in drug discovery, Drug Discov. Today 8:594–602, 2003.

64 C. Chang and C-J. Lin, LibSVM: a library for support vector machines, 2001. URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

65 Lipinski, 1997 66

(a) Konstantin V. Balakin, Sean Ekins, Andrey Bugrim, Yan A. Ivanenkov, Dmitry Korolev, Yuri V. Nikolsky, Andrey V. Skorenko, Andrey A. Ivashchenko, Nikolay P. Savchuk, and Tatiana Nikolskaya. KOHONEN MAPS FOR PREDICTION OF BINDING TO HUMAN CYTOCHROME P450 3A4. DMD 32:1183–1189, 2004; (b) Konstantin V. Balakin, Sean Ekins, Andrey Bugrim, Yan A. Ivanenkov, Dmitry Korolev, Yuri V. Nikolsky, Andrey A. Ivashchenko, Nikolay P. Savchuk, and Tatiana Nikolskaya. QUANTITATIVE STRUCTURE-METABOLISM RELATIONSHIP MODELING OF METABOLIC N-DEALKYLATION REACTION RATES. DMD 32:1111–1120, 2004; (c) Dmitry Korolev, Konstantin V. Balakin, Yuri Nikolsky, Eugene Kirillov, Yan A. Ivanenkov, Nikolay P. Savchuk, Andrey A. Ivashchenko, and Tatiana Nikolskaya. Modeling of Human Cytochrome P450-Mediated Drug Metabolism Using Unsupervised Machine Learning Approach. J. Med. Chem. 2003, 46, 36313643.

67 Ekins et al., 1999a; Gibbs et al., 1999; He et al., 1998; Iribarne et al., 1998; Katoh et al., 2000; Zhang et al., 2002

68 Konstantin V. Balakin, Sean Ekins, Andrey Bugrim, Yan A. Ivanenkov, Dmitry Korolev, Yuri V. Nikolsky, Andrey A. Ivashchenko, Nikolay P. Savchuk, and Tatiana Nikolskaya. Quantitative Structure-Metabolism Relationship Modeling Of Metabolic N-Dealkylation Reaction Rates. Drug Metabolism And Disposition. 32:1111–1120, 2004.

69 Yan and Caldwell, 2001

70 Wold, 1991