JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

PENG: A Neural Gas-Based Approach for Pharmacophore Elucidation. Method Design, Validation, and Virtual Screening for Novel Ligands of LTA4H

Daniel Moser,[†,‡,§,#] Sandra K. Wittmann,[†,#] Jan Kramer,[†] René Blöcher,[†] Janosch Achenbach,[†,∥] Denys Pogoryelov,[⊥] and Ewgenij Proschak*[,†,‡,§]

[†]Institute of Pharmaceutical Chemistry, Goethe University, 60438 Frankfurt, Germany

[‡]German Cancer Consortium (DKTK), 60590 Frankfurt, Germany

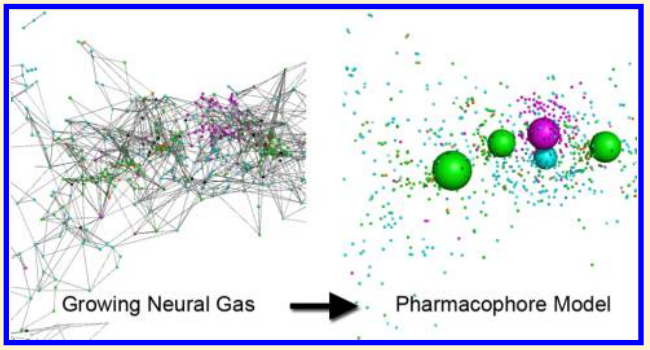
[§]German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

[∥]BASF SE, 67056 Ludwigshafen, Germany

[⊥]Institute of Biochemistry, Goethe University, 60438 Frankfurt, Germany

Supporting Information

ABSTRACT: The pharmacophore concept is commonly employed in virtual screening for hit identification. A pharmacophore is generally defined as the three-dimensional arrangement of the structural and physicochemical features of a compound responsible for its affinity to a pharmacological target. Given a number of active ligands binding to a particular target in the same manner, it can reasonably be assumed that they have some shared features, a common pharmacophore. We present a growing neural gas (GNG)-based approach for the extraction of the relevant features which we called PENG (pharmacophore elucidation by neural gas). Results of retrospective validation indicate an acceptable quality of the generated models. Additionally a prospective virtual screening for leukotriene A4 hydrolase (LTA4H) inhibitors was performed. LTA4H is a bifunctional zinc metalloprotease which displays both epoxide hydrolase and aminopeptidase activity. We could show that the PENG approach is able to predict the binding mode of the ligand by X-ray crystallography. Furthermore, we identified a novel chemotype of LTA4H inhibitors.



INTRODUCTION

In general, methods used to search for novel ligands of a given target or to explain ligand−target interactions can be divided into structure-based design (SBD) and ligand-based design (LBD).[1] While SBD depends on the availability of a three-dimensional (3D) structure of the protein, LBD focuses solely on known active (and inactive) ligands and can therefore be used when a 3D structure of the protein is not available. A pharmacophore[2,3] is a central concept in medicinal chemistry and is defined by IUPAC as an "ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response".[4] A pharmacophore of a given target thus defines the structural requirements to a chemical structure which are necessary to ensure the binding properties. Pharmacophore models are assumptions which describe the pharmacophore of a target and are—while also applicable in SBD—a widely used concept in LBD to explain relevant structural and/or physicochemical features of known active molecules. Usually it is assumed that active ligands form similar interactions (i.e., have the same binding mode) with the protein and therefore should have some shared features, i.e., a common pharmacophore (CP). A common pharmacophore induces an overlay of the ligands which hopefully resembles the conformations when bound to the target.[5] Vice versa it is possible to deduce a CP from a given overlay of ligands. Numerous—commercial as well as freely available—approaches for creating such alignments and extracting a CP have emerged over the years,[6,7] yet it remains a challenging topic and research is still ongoing. In this work we present an approach based on a growing neural gas (GNG)[8] which we named PENG (pharmacophore elucidation by neural gas). The GNG can be seen as an extension of the general neural gas (NG) algorithm presented by Martinetz and Schulten[9] in the early 1990s, which in turn was inspired by self-organizing maps (SOMs).[10] Self-organizing maps are artificial neural networks (ANNs) which are used in unsupervised machine learning. They consist of a predefined number of neurons which are typically arranged in the form of a

Received: October 14, 2014
Published: January 27, 2015

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284−293

two-dimensional rectangular, hexagonal, or toroidal grid. Each neuron has an associated weight vector with the same dimensions as the input data, as well as its position in the grid. During training the neurons (i.e., the associated weight vectors) adapt to the input data; by using a neighborhood function the topology of the input space is preserved; i.e., for a given "winner" neuron (the neuron with the least distance to the presented data point) all of its topological neighbors are adapted, too. Since each neuron has a fixed topological position in the grid, a trained SOM represents a mapping from a high-dimensional to a lower-dimensional (usual 2D) space. These properties make SOMs particularly useful for visualization of high-dimensional data but also for clustering. In contrast to SOMs, a neural gas does not work with a predefined topological arrangement of neurons (but still with a fixed number). Instead of adapting the topological neighbors of a winner neuron, all neurons are adapted, depending on their actual distance to the presented data point. Between the winner and the second neuron an "aging" connection is formed. If the connection is not renewed after a certain time, for example because the second neuron "moved" away, it is removed again. This way the network "learns" the topology of the input data during the course of the training. However, both methods rely on a predefined number of neurons. Finding such a suitable number depends on the input data and may lead to unsatisfying results if chosen wrong. Therefore, Fritzke[8] proposed the GNG method which does not work with a fixed number of neurons but instead creates or removes neurons dynamically. GNGs are used in different application fields, for example, image processing,[11] cluster analysis,[12] and pattern recognition, but are also employed in chemoinformatics-related questions.[13]

We validated our approach by (a) generating models based on co-crystallized ligands, searching a conformation database of these ligands, and calculating the root-mean-square deviation (RMSD) between the best-matching conformation and the co-crystallized pose, (b) performing retrospective virtual screenings and calculating the rate of correctly classified and misclassified compounds, and (c) performing a prospective virtual screening for novel ligands of the leukotriene A4 hydrolase. The leukotriene A_4 hydrolase (LTA4H; EC 3.3.2.6) is a zinc metalloprotease which can act as an epoxide hydrolase as well as an aminopeptidase. It catalyzes the stereoselective hydrolysis of leukotriene A4 to the proinflammatory lipid mediator leukotriene B4 in a two-step reaction, as a part of the arachidonic acid cascade.[14,15] It has previously been shown that the LTA4H preferably cleaves a tripeptide with an N-terminal arginine residue as well as the chemotactic tripeptide Pro-Gly-Pro which leads to its inactivation.[16] LTA4H is an interesting target for the treatment of chronic inflammatory diseases where it plays contrary roles (production and inactivation of chemotactic agents[16]) and is linked to dermatitis, arthritis, cancer, osteoporosis, and atherosclerosis.[17]

METHODS

Pharmacophore Annotation and Output. PENG relies on pharmacophore annotation points as generated by the MOE (Molecular Operating Environment)[18] software and outputs MOE *.ph4* models. In the present work we used the *Unified* annotation scheme inlcuding 19 different features. Internally these annotation points are stored as objects consisting of a coordinate vector v, a 19-dimensional bit-vector encoding the pharmacophore annotation f, and two integer ids for the corresponding molecule and the conformation, respectively.

Growing Neural Gas. For detection of common pharmacophore features a GNG is used. Our implementation follows the original design described by Fritzke[8] with slight modifications. In contrast to the classic neural gas algorithm[9] which operates on a fixed number of neurons (units), the GNG algorithm creates or removes neurons dynamically by competitive Hebbian learning (CHL).[19] As mentioned, each annotation point (i.e., data point/signal) as well as each neuron has two vectors: the coordinate vector, $v \in \mathbb{R}^3$, and the feature vector, $f \in \{0,1\}^{19}$. However, to allow a continuous adaption of the neurons during the training, the feature vector is treated as a real vector, i.e., $[0,1]^{19}$, instead of a binary one. As distance functions the Euclidean distance, $D(x,y)$, between two coordinate vectors x and y, and the Tanimoto similarity, $T(a,b)$, for two feature vectors a and b are used. We introduced a time-dependent factor $\omega_E = e^{(E/E_{max})}$ for the current epoch E and a total of E_{max} epochs, which gives the Euclidean distance more weight in the early phases of the training process. This helps to avoid the adaption of neurons to a "wrong" feature vector while there is still a lot of movement in the coordinate space during the early training. In later phases, with neurons being nearer to their final position in 3D space, the feature vector is then adapted and fine-tuned. In order to get a combined distance value, T is scaled by the factor $D_{max} = \max(\{D(x,y)|x,y \in V_{GNG}\})$ to lie in the range of the Euclidean distance, where V_{GNG} is the set of all coordinate vectors of all data points in this GNG run. The distance between a data point ξ and a neuron s is then calculated as

$$d_{\xi,s} = (1 - T(f_\xi, f_s)(1 - \omega_E))D_{max} + D(v_\xi, v_s)\omega_E$$

where f and v denote the respective feature and coordinate vectors. Likewise, neurons are adapted, i.e., moved toward ξ by fractions ϵ_b and ϵ_n, respectively, of the total distance:

$$\Delta v_s = \epsilon\omega_E(v_\xi - v_s)$$

$$\Delta f_s = \epsilon(1 - \omega_E)(f_\xi - f_s)$$

with $\epsilon = \epsilon_b$ if s is the nearest neuron and $\epsilon = \epsilon_n$ for all direct topological neighbors of s. The default values used are $\epsilon_b = 0.2$ and $\epsilon_n = 0.006$ as in the original work by Fritzke.[8] After running the GNG algorithm for E_{max} epochs, i.e., having randomly presented all data points E_{max} times, several merge steps are performed to remove unnecessary neurons as follows:

(1) Let S_{sorted} be the list of neurons, sorted by coverage. Since each data point $\xi \in P$ belongs to a molecule $m_\xi \in M$, the coverage c_s is calculated as the number of unique molecules w_s associated with a given neuron $s \in S$, i.e., data points with the least distance to the neuron, divided by the total number of molecules:

$$c_s = \frac{|\{m_\xi | \xi \in w_s\}|}{|M|}$$

with

$$w_s = \{\xi | d_{s,\xi} = \min_{t \in S}(\{d_{t,\xi}\})\}$$

(2) Mark all neurons $t \in S_{sorted} | t \neq s, c_s \geq c_t$ for merge with $s \in S_{sorted}$ if their radii r (default 0.1 Å) overlap or their Euclidean distance and Tanimoto similarity are below respectively above two user-defined thresholds t_d and t_s:

$$M_s = \{s\} \cup \{t | r_s + r_t > D(v_s, v_t) \vee (D(v_s, v_t) \leq t_d \wedge T(f_s, f_t) \geq t_s)\} \quad (1)$$

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284−293

Perform the actual merge by setting the coordinate vector of a temporary neuron s_{tmp} to a user-defined percentile (default median) p of all neurons $m \in M_s$. To incorporate the higher coverage of s, the new neuron is "moved" a fraction toward s, though, and its radius is adapted accordingly:

$$v_p = p(\{v_m | m \in M_s\})$$

$$v_{s_{tmp}} = v_p + (v_s - v_p)c_s$$

$$r_{s_{tmp}} = \max(r_s, p(\{D(v_{s_{tmp}}, \xi) | \xi \in w_{s_{tmp}}\}))$$

The feature vector is updated in a similar manner. However, this time the feature vector of neurons within the p Euclidean distance to s_{tmp} is taken into account completely while neurons farther away contribute only a fraction depending on their distance.

$$d_p = p(\{D(v_{s_{tmp}}, v_m) | m \in M_s\})$$

$$f_{sum} = \sum_{\substack{m \in M \\ d'_m = D(v_{s_{tmp}}, v_m)}} \begin{cases} v_m: & d'_m \le d_p \\ v_m e^{-(d'_m - d_p)}: & d'_m > d_p \end{cases}$$

$$f_{s_{tmp}} = \frac{f_{sum}}{|M_s|}$$

Since we use a growing neural gas, all neurons are linked with their topological neighbors. Therefore, these links are reconnected to the temporary neuron s_{tmp}. Let l_m be the list of topological neighbors of a neuron m. Then

$$l_{s_{tmp}} = \bigcup_{m \in M_s} l_m$$

$$l_n = l_n \cup \{s_{tmp}\} \backslash \{m | m \in M_s\}, \quad n \in S$$

reconnects all links. The final step in the merge process is to remove the merged neurons and make s_{tmp} permanent.

$$S = S \backslash M_s \cup s_{tmp}$$

(3) Remove all neurons (including newly created ones by merge) if their coverage is below a user-defined minimum coverage γ_n.

$$S = S \backslash \{t | c_t < \gamma_n, t \in S\}$$

(4) Repeat all steps described above once more. However, this time eq 1 becomes

$$M_s = \{s\} \cup \{t | r_t + d_{s,t} < r_s \text{ or } r_s + d_{s,t} < r_t\}$$

meaning that only neurons will be merged which are fully included in another neuron.

(5) Finalize all neurons by rounding every position of the feature vector to 0 or 1; i.e., make it a bit vector again. The distribution of the values for each position is shown for an exemplary run of the algorithm in Figure S1 in the Supporting Information.

Figure 1 shows an exemplary run of the GNG algorithm on 50 known active inhibitors of the soluble epoxide hydrolase.

Elucidation Algorithm. The elucidation algorithm generates up to *maxPh4* pharmacophore models (in parallel) using the following procedure. It expects a list $L = (V,F)$ of pharmacophore annotation points (signals) and two functions $c: L \to \mathbb{Z}^+$, $m: \mathbb{Z}^+ \to \mathbb{Z}^+$ as input, where $v_\xi \in V$ and $f_\xi \in F$ denote the coordinate and feature vectors of a signal $\xi \in L$ and *c*and*m* are functions mapping each annotation point to the corresponding conformation and subsequently compound identification (id). Additionally active/inactive information may be provided; if it is omitted all compounds are treated as active.

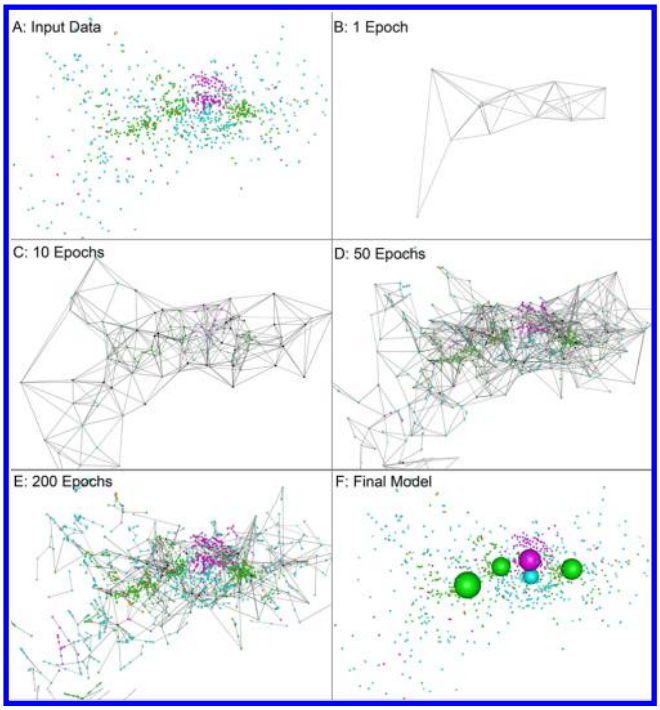


Figure 1. Pharmacophore extraction process with the GNG: (A) Aligned pharmacophore annotation points of the input data; (B–E) GNG after 1, 10, 50, and 200 epochs, respectively (input data not shown); (F) final pharmacophore model after neurons were merged and pruned shown together with the input data.

Repeat until *maxPh4* models are created or no more templates are available:

(1) Choose a random conformation id t with corresponding annotation points (conformation) Ξ_t as a template and mark this conformation id as used.

$$t = \text{rand}(\{c(\xi) | c(\xi) \notin \text{used}, \xi \in L\})$$

$$\Xi_t = \{\xi | c(\xi) = t, \xi \in L\}$$

$$\text{used} = \text{used} \cup t$$

(2) Align all other conformations from different compounds (i.e., $m(c(\xi')) \neq m(t)$) onto this template using the Kabsch algorithm.[20,21] The Kabsch algorithm requires as input the maximum cliques in an association graph and returns the optimal rotation matrices. For clique detection the Bron–Kerbosch algorithm[22,23] is used, which expects an association graph as input. Therefore, we implicitly treat the conformations as undirected, complete graphs, i.e., $V = \Xi_x$ for some conformation x and $E = \{(u,v) | u,v \in V, u \neq v\}$. The nodes are labeled with the feature vector and the edges with the Euclidean distance between two annotation points. Let

$$G = (V, E, \mu, \epsilon) \quad \text{and} \quad G' = (V', E', \mu', \epsilon')$$

with

$$\mu^{(')}(v) = f_v: \quad v \in V^{(')}$$

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284–293

$$\epsilon^{(')}(e) = D(u, v): \quad e = (u, v) \in E^{(')} \wedge u, \quad v \in V^{(')}$$

then the association graph is defined as $G_A = (V_A, E_A)$ with $V_A \subseteq VV'$ and $E_A \subseteq EE'$. A node is inserted in the association graph if both nodes of the parent graphs have the same properties, i.e., in our case some common bits in their feature vector.

$$V_A = \{(v, v') | v \in V, v' \in V', \mu(v) \text{ AND } \mu'(v')\}$$

where AND denotes the bitwise *and* operation. Between two association nodes an edge $e_A \in E_A$ with

$$E_A = \{(u_A, v_A) | u_A = (u, u'), v_A = (v, v')\}$$

is inserted if the following applies:

(a) $u \neq v$, and $u' \neq v'$.
(b) If an edge $e = (u,v) \in E$ exists, then an edge $e' = (u',v') \in E'$ exists and $|\epsilon(e) - \epsilon'(e')| \leq c$ for a user-defined constant c called the *distance cutoff*.
(c) If no edge $e = (u,v) \in E$ exists, then no edge $e' = (u',v') \in E'$ exists.

(3) Let cpds = $\{m(c(\xi)) | \xi \in L\}$ be the ids of all compounds and Ξ_{cpd} the conformation with the lowest RMSD for compound cpd ∈ cpds. Then

$$L_{GNG} = \bigcup_{cpd \in cpds} \Xi_{cpd}$$

is the list of aligned pharmacophore annotation points and $V_{GNG} \subset V$ and $F_{GNG} \subset F$ are the corresponding subsets of coordinate and feature vectors.

Run the GNG algorithm with these subsets. The GNG will return a list of neurons $S = (V'_{GNG}, F'_{GNG})$ (the pharmacophore model) with the same format as the annotation points; i.e., V'_{GNG} is the set of coordinate vectors and F'_{GNG} the set of feature vectors.

(4) Refine the alignments created in step 2 by realigning every Ξ_{cpd} onto the GNG neurons S.

(5) Repeat steps 3 and 4 once.

(6) Refine and score the generated model:

Let

$$P(s, \xi) = (D(v_s, v_\xi) - r_s \leq 0) \wedge (f_s \text{ AND } f_\xi > 0)$$

be a function determining if a neuron matches an annotation point. Then matches(s) is the number of matched annotation points for a neuron $s \in S$.

$$\text{matches}(s) = \sum_{cpd \in cpds} \begin{cases} 1 & \exists \xi \in L_{GNG}: \quad m(c(\xi)) = \text{cpd} \wedge P(s, \xi) \\ 0 & \text{else} \end{cases}$$

(a) Remove neurons not matching at least γ_f percent of compounds.

$$S = S \backslash \left\{ s \middle| \frac{\text{matches}(s)}{|\text{cpds}|} < \gamma_f, s \in S \right\}$$

(b) If neurons were removed, recalculate the alignments as in step 4.

(c) Calculate the mean coverage of the model as

$$\text{mean}(S) = \frac{\sum_{s \in s} \text{matches}(s)}{|\text{cpds}||S|}$$

(d) If active/inactive information was provided, calculate the overall accuracy = m/N, the accuracy on actives $\text{acc}_1 = m_1/N_1$, and the accuracy on inactives $\text{acc}_0 = m_0/N_0$ with $N = N_1 + N_0$ and $m = m_1 + m_0$, where N_1 and N_0 are the number of active/inactive molecules and m_1 and m_0 denote the number of active/inactive molecules matching/not matching the model.

Benchmark Compound Sets and Validation. All compounds have been prepared using the default *wash* routine in the MOE software. As comparison to our method, the *Pharmacophore Elucidation* routine of the MOE software was used with default parameters. Pharmacophore matching was performed with the *Pharmacophore search* of MOE.

Pose Reproduction Set. We used the PharmBench[24] data set to benchmark how well our algorithm performs in the creation of models able to reproduce the crystal structure pose. For each compound up to 50 conformations (RMSD limit, 0.25 Å; energy window, 7.0 kcal/mol) were created with the *Conformation Import* method of the MOE software. For each target the co-crystallized poses of the respective ligands were used to elucidate pharmacophore models. The best-scored model (in terms of accuracy) was selected and used to perform a pharmacophore search on the conformation database for this target. For each ligand the RMSD between the best-matching conformation and the co-crystallized pose was calculated.

Active/Inactive Set. We determined the ability of the generated models to distinguish between active and inactive compounds using targets from the data set described by Sutherland et al.[25] and the directory of useful decoys, enhanced (DUD-E).[26] For each compound up to 25 conformations (RMSD limit, 0.1 Å; energy window, 7.0 kcal/mol) were created with the Conformation Import method of the MOE software. Validation was performed using a 10-fold cross-validation (CV). In each iteration 90% of the active molecules were used to elucidate up to 10 pharmacophore models; the remaining 10% were mixed with the inactives as a test set. With the best-scored model a pharmacophore search on the test set was performed and the number of true positives (matching actives) and false positives (matching inactives) was calculated.

Protein Preparation. The coding sequence of human leukotriene A_4 hydrolase was inserted into the expression vector pET 24(+) (Novagen) with a hexahistidin tag at the C-terminus and a kanamycin resistance. The applied primers were as follows:

LTA4H fwd:

AAGGATCCATGCCCGAGATAGTGGATA

LTA4H rev:

AAACTCGAGGTCCACTTTTAAGTCTTTCCC

The forward primer includes the restriction site for XhoI and the reverse primer for *Bam*HI. After the ligation the vector pET 24(+) with LTA4H coding sequence was transformed into competent *Escherichia coli* (*E. coli*) DH5α cells (Invitrogen). After the evaluation of the clone by sequencing with a T7 promotor and terminator primer (SRD, Bad Homburg), the plasmid was transformed and overexpressed in *E. coli* BL21(DE3)RIPL-Codon Plus cells (Invitrogen). At an OD_{600} of approximately 0.8 the induction was started with 300 μM IPTG (isopropyl-β-D-thiogalactopyranosid, AppliChem). After 4 h the cells were harvested by centrifugation (5500 rpm, 20 min). Approximately 0.5 μg of DNase (AppliChem) and an ethylenediaminetetraacetic acid (EDTA)-free protease inhibitor complete tablet (Roche) were added. For crushing the lysate a cell disrupter (Constant Systems) was used (two times, 1 kbar). The purification of the protein was performed with a 5 mL His-Trap HP column (GE Healthcare) by immobilized metal ion affinity chromatography. A wash and loading buffer containing

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284−293

50 mM Tris, 500 mM NaCl, and 4 mM imidazole with pH 8 was used and the protein was eluted with an amount of 60 mM imidazole. The fractions containing hLTA4H were analyzed by SDS-PAGE (sodium dodecyl sulfate−polyacrylamide gel electrophoresis) and Western blot as well as in-gel digestion followed by mass spectroscopy. For separating contaminants the relevant fractions were pooled and purified with a Superdex200 column (GE Healthcare). A buffer containing 50 mM Tris-HCl and 50 mM NaCl was used for elution of the protein.

LTA4H Assay. The fluorescence-based LTA4H-assay system was performed in black polystyrol 96-well plates. The assays were performed in a buffer containing 50 mM Tris-HCl, 50 mM NaCl, 0.01% Triton-X 100, and a final concentration of 1.4% dimethyl sulfoxide (DMSO). In the first step 10% of the compound solution with 10% DMSO was preincubated with 80 μL of protein for 30 min at room temperature. The protein concentration was adapted to yield an average slope of 0.3 relative fluorescense units (RFU)/s as determined in a separate control. After incubation 10 μL of the nonfluorescent substrate 7-amido-4-methylcoumarine hydrochloride was added with a final concentration of 300 μM. A Tecan fluorescence plate reader (Infinite F200 pro) was used for measuring the fluorescence of the hydrolyzed substrate (excitation at 370 nm and emission at 460 nm) for 30 min (one point every minute) at room temperature. As blank and positive controls, samples without protein or compound, respectively, were used. The evaluation of the reaction was obtained by the slope, and the percentage inhibition of the compounds was calculated based on the following formula:

$$\%\ \text{inhibition} = 100\left(1 - \frac{\text{slope}_{\text{compound}} - \text{slope}_{\text{blank}}}{\text{slope}_{\text{positive}} - \text{slope}_{\text{blank}}}\right)$$

IC_{50} values were determined with the GraphPad Prism Software (5.0) using a sigmoidal dose response curve fit (variable slope with four parameters).

Thermal Shift Assay. The differential scanning fluorimetry (DSF) is a screening method for investigating in vitro interactions between a purified protein and possible ligands. Their binding leads to a change in free enthalpy and thus goes hand in hand with the stability of the protein. The change in free enthalpy depends on the concentration as well as the affinity of the ligand to the target protein.[27] It correlates with the shift in temperature (therefore DSF is also called thermal shift assay) during the temperature-affected unfolding of the protein. The difference of the transition temperature in the presence or absence of an inhibitor can be considered as an indicator of the stabilizing interaction between protein and inhibitor. For the determination of the thermal shift, the fluorescent dye SYPRO Orange (Life Technologies) was used, which has a high, nonspecific binding affinity for hydrophobic regions but is strongly quenched by water. While the protein unfolds, more hydrophobic patches get exposed and can bind the dye, leading to an increase in fluorescence.[27]

The thermal shift assay was carried out in a MicroAmp fast 96-well reaction plate. The assay buffer consists of 50 mM Tris, 50 mM NaCl, pH 8 with a final concentration of 0.01% Triton X-100 and 1% DMSO per well. The total volume per well was 40 μL, consisting of 32 μL of LTA4H with a concentration of 14 μg per well, 4 μL of inhibitor solution, and 4 μL of SYPRO Orange solution (1:200 with MQ-water). As blank and positive controls, samples without protein or compound, respectively, were used. The relative fluorescence intensity (RFU) in dependence of the temperature was recorded with the iCycler IQ single-color real time polymerase chain reaction (PCR). The temperature was raised from 20 to 89.9 °C every 0.24 s by 0.2 °C. With Graph Pad Prism (5.0) the first derivation was formed and smoothed with a second order smoothing function considering six neighbors. Afterward the temperature with maximum fluorescence intensity was determined and the shift calculated as the difference of the positive control and the inhibitor.

Crystallization. For the crystallization of the LTAH4 we adapted the conditions described by Davies et al.[28] Briefly, 1 μL of protein solution containing 10 mM Tris-HCL, 25 mM KCl, and 4−8 mg/mL LTA4H (pH 8) was mixed at different ratios (0.5:1, 1:1, and 1:2) with precipitate solution containing 0.15 M imidazole, 0.1 M sodium acetate pH 6, 14% PEG-8000, and 5 mM $YbCl_3$, and crystallized at 17.4 °C for 2 weeks by hanging drop vapor-diffusion method. The resulting crystals were soaked in the precipitant solution containing 100 μM inhibitor for at least 48 h prior to freezing. For data collection under cryonic conditions the soaked crystals were transferred for 20 s into the cryoprotecting solution (containing precipitant solution complemented with 25% PEG-400) before flash freezing in the liquid nitrogen. X-ray diffraction data were collected on the beamline station X06DA (PXIII) at the Swiss Light Source (SLS, Paul Scherrer Institute, Villingen, Switzerland). All diffraction data were obtained from a single crystal and processed with the XDS software[29,30] package.

RESULTS AND DISCUSSION

Pose Reproduction. For assessing the ability to reproduce the conformation of a co-crystallized ligand we selected the 20 proteins with the highest number of available crystal structures from the PharmBench data set. In the case of our algorithm, up to 20 pharmacophore models were created for each protein; in the case of MOE, no limit was specified. In either case, the model with the highest accuracy was used for pharmacophore search. For each ligand the best-matching pose was retained and the RMSD to the respective crystallized pose calculated. Figure 2 exemplary shows the model with the highest accuracy for target P39900 and the resulting alignment of the compounds. RMSD values were averaged for each protein. Despite the quite high standard deviations, the created models were suitable to find a conformation close to the crystallized pose with average RMSD values below 2 Å in 19 out of 20 cases (Figure 3 and Table 1). In the remaining case no model with

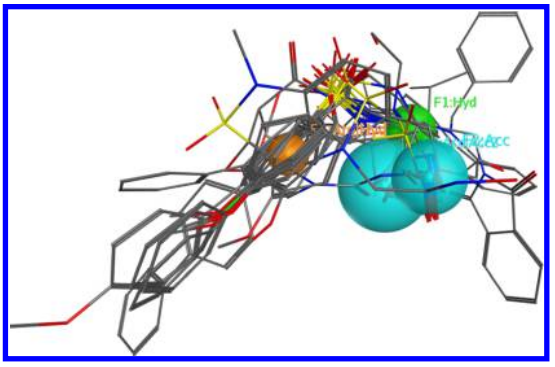


Figure 2. Exemplary pharmacophore model created with our method and resulting alignment of the compounds. Shown is the model with the highest accuracy for target P39900.

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284−293

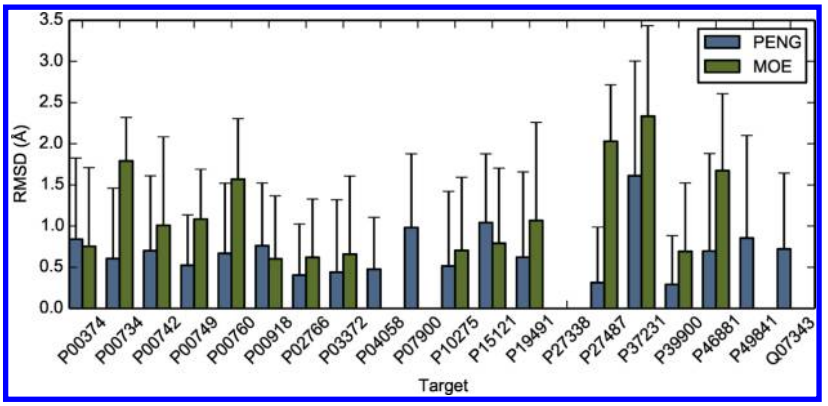


Figure 3. Results of the pose reproduction benchmark set. Shown are the average RMSD values and the positive standard deviation.

Table 1. Results of the Pose Reproduction Benchmark[a]

	PENG			MOE		
protein	hitrate	RMSD	SD	hitrate	RMSD	SD
P00374	1.00	0.84	0.99	1.00	0.75	0.96
P00734	0.85	0.61	0.86	0.88	1.79	0.53
P00742	1.00	0.70	0.91	1.00	1.01	1.08
P00749	0.57	0.52	0.61	0.86	1.08	0.61
P00760	0.86	0.67	0.85	0.98	1.57	0.74
P00918	0.79	0.76	0.76	0.98	0.60	0.76
P02766	0.96	0.40	0.62	0.75	0.62	0.71
P03372	1.00	0.44	0.88	1.00	0.66	0.95
P04058	0.89	0.48	0.63	–	–	–
P07900	0.95	0.98	0.90	–	–	–
P10275	1.00	0.51	0.91	1.00	0.70	0.89
P15121	0.71	1.04	0.84	1.00	0.79	0.91
P19491	1.00	0.62	1.04	1.00	1.07	1.19
P27338	–	–	–	–	–	–
P27487	0.93	0.31	0.68	1.00	2.03	0.69
P37231	0.88	1.61	1.39	0.88	2.34	1.10
P39900	0.93	0.29	0.59	1.00	0.69	0.83
P46881	0.56	0.70	1.18	0.78	1.67	0.93
P49841	0.80	0.85	1.25	–	–	–
Q07343	1.00	0.72	0.92	–	–	–

[a]Shown is the hitrate (i.e., how many compounds were found divided by the total number of compounds), the average RMSD, and the standard deviation. In the cases in which "–" is shown, no model with more than two features could be created.

more than two features could be created, most probably because the prevalent pharmacophore of the actives indeed consists of two aromatic/hydrophobic features. However, MOE was unable to find a model, too.

Active/Inactive. For assessing the ability to separate actives from inactives we used the targets COX2, DHFR, and BZR from the Sutherland data set and FABP4 and INHA from the DUD-E database. For each target the actives were taken from the respective data set and inactives from the DUD-E database. In all cases the percentage of actives recovered is above 50% (Figure 4). However, while the models for targets of the Sutherland data set yield reasonable false positive rates, they are quite high for the DUD-E data set. The most likely reason for this is that the generated models for these targets are too small and therefore too common. Tuning of parameters, especially the distance threshold, t_d, and the percentile p used during the merge step may improve the results.

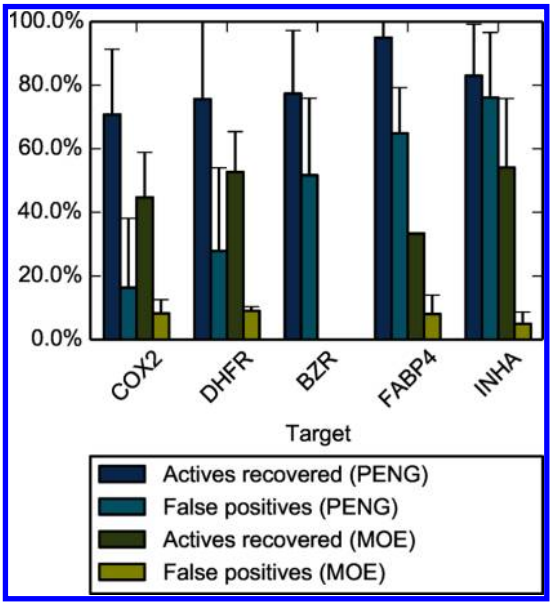


Figure 4. Results of the active/inactive benchmark set. Shown are the percent actives recovered, the percent false positives, and the respective positive standard deviations. For BZR MOE did not find a model.

Prospective Virtual Screening. Additionally we wanted to test PENG in a prospective virtual screening for new ligands of the leukotriene A4 hydrolase. Therefore, we used the co-crystallized ligands from 24 Protein Data Bank (PDB) structures (see Supporting Information) to elucidate pharmacophore models. Additionally an ESshape3D fingerprint model was created with the MOE software. With this model we prefiltered the SPECS database for compounds with an average fingerprint similarity higher than 0.6, leading to 69,106 unique compounds. The three models with the highest accuracy from the pharmacophore elucidation were then used to search the filtered SPECS database. We subsequently focused on the smallest resulting database with 3,136 unique compounds. However, since many hits were very similar, we clustered them using the Butina[31] algorithm as implemented in RDKit[32] using the CATS[33] fingerprint, yielding 113 cluster centroids. From these structures we manually selected 21 compounds (Figure 5) based on diversity and price, ordered, and tested them in vitro. Figure 6B shows the corresponding pharmacophore model with our reference inhibitor **23** (Figure 6A, see Supporting Information for synthesis).

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284–293

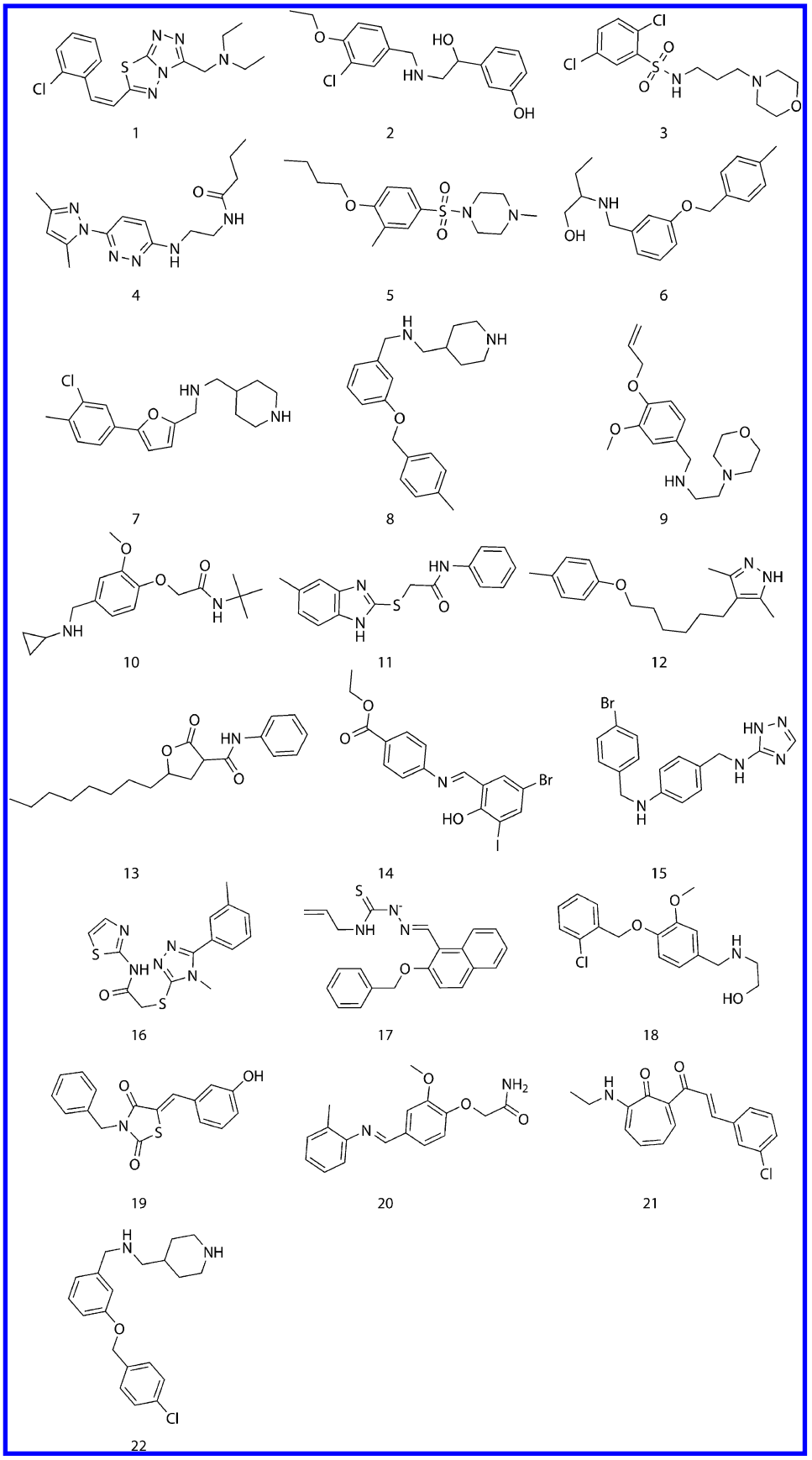


Figure 5. Overview of the ordered compounds (**1**–**21**) as well as one derivative of compound **8** (**22**). The compounds have been depicted to resemble the actual 3D conformation as well as possible using the "flatten" function of MOE.

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284–293

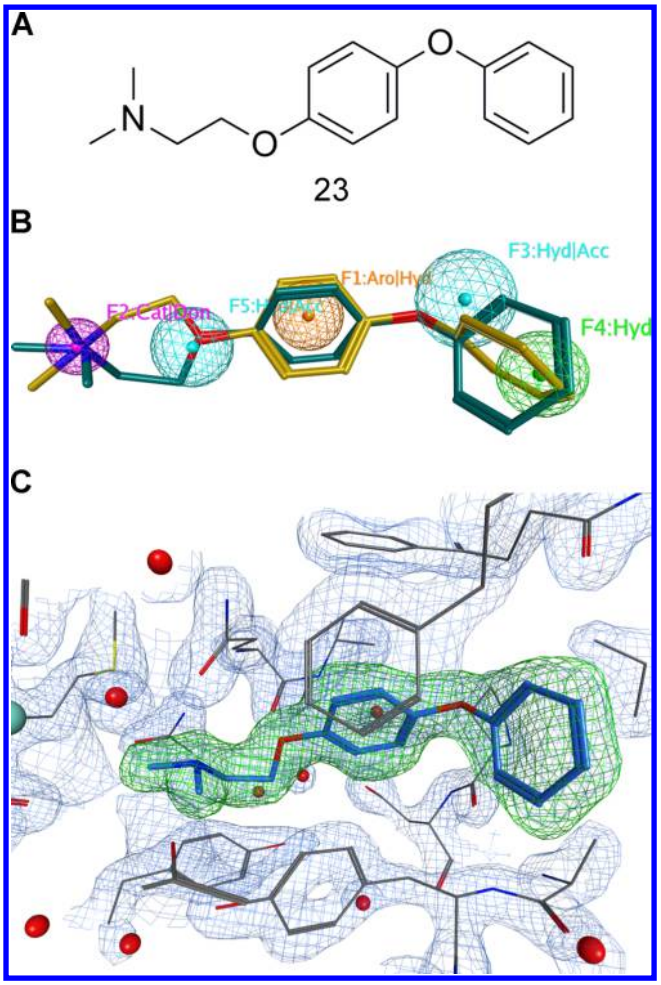


Figure 6. (A) Structure of the reference inhibitor **23**. (B) Pharmacophore model used for selecting compounds with best-matching pose of our reference inhibitor (yellow) and actual crystallized pose (cyan). (C) Crystal structure of the reference ligand in the LTA4H pocket. The F_o-F_c difference map was calculated from the model with the omitted ligand. Shown is the positive F_o-F_c difference map contoured at 3.0σ within 2 Å around the ligand (green) and the $2F_o-F_c$ map of the receptor within 5.5 Å around the ligand, contoured at 1.5σ (blue).

LTA4H Assay. For validation of the assay system we synthesized a reference inhibitor (**23**) adopted from the literature.[34] We measured an IC_{50} value for this compound of 60 nM with a standard deviation of 8, which lies in the range of the ligands described by Penning et al.[34] The selected compounds from the virtual screening were tested in a first step with a final concentration of 10 μM in the LTA4H-assay system. As Figure 7 shows, there is a high deviation between the single measurements for some compounds. A possible explanation for this issue could be the fact that the substances interact with the assay system or the solubility is not adequate. For this reason the five most consistent of these compounds were selected (compounds **8**, **15**, **19**, **20**, and **21**) to determine the IC_{50} value. However, except for compound **8**, no IC_{50} curve could be fitted. Again, the reason could be the solubility of the substances. For that issue other derivatives of compound **8** were ordered. Only one compound (**22**) showed the required potency (inhibition greater 50% at 10 μM) and an IC_{50} curve could be determined (Figure 8). The IC_{50} value of compound **8**

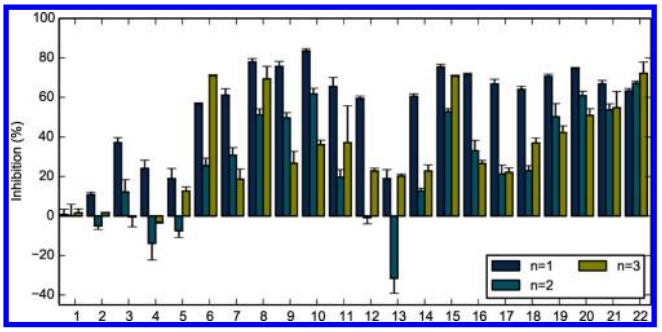


Figure 7. LTA4H assay with compounds at 10 μM. Shown is each repetition of the experiments with the standard deviation of the respective triplicate.

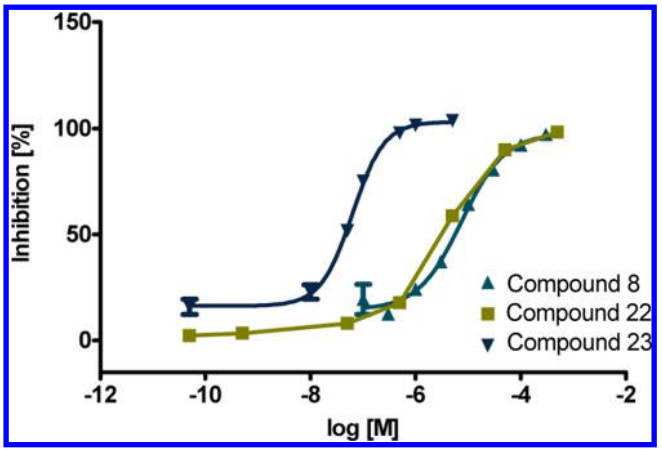


Figure 8. IC_{50} curve of compounds **8**, **22**, and **23**.

Table 2. IC_{50} Values of the Compounds

compound	IC_{50} (μM)	SEM (μM)
8	8.5	2.6
22	3.8	0.7
23	0.060	0.009

was determined as 8.5 μM and of compound **22** as 3.8 μM (Table 2).

Thermal Shift Assay. In a further step the unfolding of the LTA4H was measured by thermal shift assay. Therefore, the inhibitor with a final concentration of 100 μM was used. For the evaluation the respective melting point was determined, and afterward the shift of the positive control and the respective inhibitor was calculated (Table 3). Compound **8** shows a stronger shift than **22**. However, both shifts are below a temperature of 2 °C indicating that there is not a very strong interaction between the protein and the inhibitors. Whereas compound **23** provides the strongest shift with a temperature of 3.6 °C which adverts to a strong interaction between the protein and the inhibitor.

Crystallization. We successfully crystallized compound **23** in the LTA4H binding site. The protein structure was

Table 3. Thermal Shift Values of the Compounds

compound	shift (°C)	SEM (°C)
8	1.9	0.7
22	1.3	0.1
23	3.6	0.3

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284−293

determined at 1.864 Å by molecular replacement with PHASER using corresponding LTA4H protein atomic coordinates as a search model (PDB entry 1GW6). After iterative rounds of model building with COOT[35] into the $2F_o-F_c$ electron density map the model containing polypetide chain and the ligands was refined to a final R and R_{free} factors of 0.1704 and 0.1888, respectively, using the PHENIX software package.[36] Statistics of data collection and structural refinement are summarized in Supporting Information Table S1. The coordinates and structure-factor amplitudes of the structure have been deposited in the Protein Data Bank as entry 5AEN. Figure 6C shows that the pose of the ligand and the positive F_o-F_c electron density difference of the model without ligand are in good agreement.The conformation predicted by the pharmacophore model used for screening matches the crystallized pose with an RMSD of 1.1 Å (Figure 6A). We also tried to crystallize the two hits from the virtual screening (**8** and **22**). However, no significant electron density could be observed, probably due to the lower binding affinity.

CONCLUSION

In the present study we developed a novel method called PENG for automatic elucidation of pharmacophore models based on a growing neural gas algorithm. We have shown that the PENG method is capable of creating meaningful pharmacophore models. The method has been validated retrospectively with co-crystallized ligands from the Pharm-Bench data set and known active ligands from the Sutherland and DUD-E data sets, as well as prospectively by a virtual screening for LTA4H ligands. We established a fluorescence-based assay with a novel substrate for LTA4H. The assay was validated with our reference inhibitor and shows comparable IC_{50} values as similar compounds described formerly by others.[34] Using the PENG method we were able to find novel ligands for LTA4H, including to our knowledge so far unknown chemotypes. The binding mode of the reference ligand determined by X-ray crystallography confirmed the pharmacophore model derived by the PENG method.

ASSOCIATED CONTENT

Supporting Information

Text giving the complete list of PDB codes used in the virtual screening and synthesis of the reference ligand, figure showing the distribution of real values for each bit position, and table listing the data collection and refinement statistics. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: proschak@pharmchem.uni-frankfurt.de.

Author Contributions

#D.M. and S.K.W. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, Sachbeihilfe PR 1405/2-1 and SFB 1039 Teilprojekt A07). We gratefully thank Guido Kirsten (CCG) for the SVL script creating the pharmacophore annotation points. D.M. thanks the German Cancer Research Center (DKFZ)/the German Consortium for Translational Cancer Research (DKTK), R.B. thanks the Else-Kröner-Fresenius Foundation Graduiertenkolleg TRIP for a scholarship. We acknowledge the Paul Scherrer Institut, Villigen, Switzerland for provision of synchrotron radiation beamtime at beamline X06DA (PXIII) of the SLS and also thank Dr. Meitian Wang for assistance.

REFERENCES

(1) Merz, J., Ringe, D., Reynolds, C. H., Eds. *Drug Design*; Cambridge University Press: Cambridge, U.K., 2010.
(2) Ehrlich, P. Über den Jetzigen Stand der Chemotherapie. *Ber. Dtsch. Chem. Ges.* **1909**, *42*, 17−47.
(3) Kier, L. B. Molecular Orbital Calculation of Preferred Conformations of Acetylcholine, Muscarine, and Muscarone. *Mol. Pharmacol.* **1967**, *3*, 487−494.
(4) Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* **1998**, *70*, 1129−1143.
(5) Lemmen, C.; Lengauer, T. Computational Methods for the Structural Alignment of Molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.
(6) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539−558.
(7) Giangreco, I.; Cosgrove, D. A.; Packer, M. J. An Extensive and Diverse Set of Molecular Overlays for the Validation of Pharmacophore Programs. *J. Chem. Inf. Model.* **2013**, *53*, 852−866.
(8) Fritzke, B. A Growing Neural Gas Network Learns Topologies. *Adv. Neural Inf. Process. Syst.* **1995**, *7*, 625−632.
(9) Martinetz, T.; Schulten, K. A "Neural-Gas" Network Learns Topologies. *Artif. Neural Networks* **1991**, *1*, 397−402.
(10) Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **1982**, *43*, 59−69.
(11) Angelopoulou, A.; Psarrou, A.; Rodríguez, J.; Revett, K. In *Computer Vision for Biomedical Image Applications*; Liu, Y., Jiang, T., Zhang, C., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2005; Vol. *3765*, pp 210−219.
(12) Canales, F.; Chacón, M. Progress in Pattern Recognition. In *Image Analysis and Applications*; Rueda, L., Mery, D., Kittler, J., Eds.; Springer: Berlin, Heidelberg, 2007; Vol. *4756*, pp 684−693.
(13) Weisel, M.; Kriegl, J.; Schneider, G. PocketGraph: Graph Representation of Binding Site Volumes. *Chem. Cent. J.* **2009**, *3*, No. P66.
(14) Rudberg, P. C.; Tholander, F.; Thunnissen, M. M. G. M.; Haeggström, J. Z. Leukotriene A4 Hydrolase/Aminopeptidase. Glutamate 271 is a Catalytic Residue with Specific Roles in Two Distinct Enzyme Mechanisms. *J. Biol. Chem.* **2002**, *277*, 1398−1404.
(15) Shim, Y. M. S.; Paige, M. In *Inflammatory Diseases—Immunopathology, Clinical and Pharmacological Bases*; Khatami, M., Ed.; 2012.
(16) Stsiapanava, A.; Olsson, U.; Wan, M.; Kleinschmidt, T.; Rutishauser, D.; Zubarev, R. a.; Samuelsson, B.; Rinaldo-Matthis, A.; Haeggström, J. Z. Binding of Pro-Gly-Pro at the Active Site of Leukotriene A4 Hydrolase/Aminopeptidase and Development of an Epoxide Hydrolase Selective Inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 4227−4232.
(17) Haeggström, J. Z. Leukotriene A4 Hydrolase/Aminopeptidase, the Gatekeeper of Chemotactic Leukotriene B4 Biosynthesis. *J. Biol. Chem.* **2004**, *279*, 50639−50642.
(18) *MOE, Molecular Operating Environment*, 2012.10; Chemical Computing Group, Montreal, Canada. 2012.
(19) Martinetz, T. Competitive Hebbian Learning Rule Forms Perfectly Topology Preserving Maps. *ICANN '93, Proceedings of the International Conference on Artificial Neural Networks*. 1993; pp 427−434.

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284−293

(20) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32*, 922−923.
(21) Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *34*, 827−828.
(22) Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575−577.
(23) Koch, I. Enumerating all Connected Maximal Common Subgraphs in two Graphs. *Theor. Comput. Sci.* **2001**, *250*, 1−30.
(24) Cross, S.; Ortuso, F.; Baroni, M.; Costa, G.; Distinto, S.; Moraca, F.; Alcaro, S.; Cruciani, G. GRID-Based Three-Dimensional Pharmacophores II: PharmBench, a Benchmark Data Set for Evaluating Pharmacophore Elucidation Methods. *J. Chem. Inf. Model.* **2012**, *52*, 2599−2608.
(25) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure-Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.
(26) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.
(27) Niesen, F. H.; Berglund, H.; Vedadi, M. The Use of Differential Scanning Fluorimetry To Detect Ligand Interactions That Promote Protein Stability. *Nat. Protoc.* **2007**, *2*, 2212−2221.
(28) Davies, D. R.; Mamat, B.; Magnusson, O. T.; Christensen, J.; Haraldsson, M. H.; Mishra, R.; Pease, B.; Hansen, E.; Singh, J.; Zembower, D.; Kim, H.; Kiselyov, A. S.; Burgin, A. B.; Gurney, M. E.; Stewart, L. J. Discovery of Leukotriene A4 Hydrolase Inhibitors Using Metabolomics Biased Fragment Crystallography. *J. Chem. Inf. Model.* **2009**, *52*, 4694−4715.
(29) Kabsch, W. Integration, Scaling, Space-Group Assignment and Post-Refinement. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 133−144.
(30) Kabsch, W. XDS. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 125−132.
(31) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Model.* **1995**, *35*, 59−67.
(32) RDKit, Open-Source Cheminformatics. http://www.rdkit.org.
(33) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.
(34) Penning, T. D.; Chandrakumar, N. S.; Chen, B. B.; Chen, H. Y.; Desai, B. N.; Djuric, S. W.; Docter, S. H.; Gasiecki, a. F.; Haack, R. a.; Miyashiro, J. M.; Russell, M. a.; Yu, S. S.; Corley, D. G.; Durley, R. C.; Kilpatrick, B. F.; Parnas, B. L.; Askonas, L. J.; Gierse, J. K.; Harding, E. I.; Highkin, M. K.; Kachur, J. F.; Kim, S. H.; Krivi, G. G.; Villani-Price, D.; Pyla, E. Y.; Smith, W. G. Structure−Activity Relationship Studies on 1-[2-(4-Phenylphenoxy)ethyl]pyrrolidine (SC-22716), a Potent Inhibitor of Leukotriene A(4) (LTA(4)) Hydrolase. *J. Med. Chem.* **2000**, *43*, 721−735.
(35) Emsley, P.; Cowtan, K. Coot: Model-Building Tools for Molecular Graphics. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2126−2132.
(36) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 213−221.

DOI: 10.1021/ci500618u
J. Chem. Inf. Model. 2015, 55, 284−293