Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus

Daniel Reker[a], Tiago Rodrigues[a], Petra Schneider[a,b], and Gisbert Schneider[a,b,1]

[a]Institute of Pharmaceutical Sciences, Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology, 8093 Zurich, Switzerland; and [b]inSili.com LLC, 8049 Zurich, Switzerland

Edited by Paul Schimmel, The Skaggs Institute for Chemical Biology, La Jolla, CA, and approved February 4, 2014 (received for review October 26, 2013)

De novo molecular design and in silico prediction of polypharmacological profiles are emerging research topics that will profoundly affect the future of drug discovery and chemical biology. The goal is to identify the macromolecular targets of new chemical agents. Although several computational tools for predicting such targets are publicly available, none of these methods was explicitly designed to predict target engagement by de novo-designed molecules. Here we present the development and practical application of a unique technique, self-organizing map–based prediction of drug equivalence relationships (SPiDER), that merges the concepts of self-organizing maps, consensus scoring, and statistical analysis to successfully identify targets for both known drugs and computer-generated molecular scaffolds. We discovered a potential off-target liability of fenofibrate-related compounds, and in a comprehensive prospective application, we identified a multitarget-modulating profile of de novo designed molecules. These results demonstrate that SPiDER may be used to identify innovative compounds in chemical biology and in the early stages of drug discovery, and help investigate the potential side effects of drugs and their repurposing options.

drug design | target prediction | polypharmacology | machine learning | chemical similarity

Computer-assisted de novo molecular design has evolved as a popular source of ideas to combat the perceived lack of new chemical entities (NCEs) in chemical biology and drug discovery (1). We demonstrate that automated de novo design delivers readily synthesizable NCEs with desirable activity profiles. Although receptor-based design operates on a model of a macromolecular binding site, ligand-based methods are either explicitly or implicitly based on the chemical similarity principle (2) without requiring a receptor model (3). Instead, the latter typically uses some measure of chemical or pharmacophore feature similarity to a reference ligand as a fitness function, which aims to generate NCEs as template mimetics via scaffold hopping (4–8). We report the development, implementation, and successful prospective application of an innovative computational technique for the target profiling of de novo-designed molecules. The approach combines the concepts of self-organizing maps (SOMs) (9) for macromolecular target prediction (10), consensus scoring (11), and a statistical evaluation that provides confidence estimates for the predictions.

Predicting polypharmacological activities is a topic relevant to chemical biology and drug discovery, not only to take advantage of inherent drug promiscuity but also to decrease lead compound attrition caused by unfavorable off-target modulation (12). Drug activities on multiple macromolecular targets are responsible for drug side effects (13), but they can also be rationally used to increase drug efficacy (14), repurpose known drugs (15), and design multitarget NCEs (7). For example, Mestres and colleagues recently predicted the inhibition of Pim kinases using a tool compound originating from poly(ADP-ribose) polymerase (PARP) biology (16). Lounkine et al. reported a large-scale off-target prediction for marketed drugs (17). In fact, high target promiscuity appears common among drug-like molecules (18). De novo designed molecules add novelty to the chemical universe and thus risk lying outside the domain of applicability of target prediction tools that exclusively rely on the chemical similarity of molecular substructures. Here, we provide a theoretical framework to address this issue, and we demonstrate the applicability of this framework to prospective de novo design studies. We then describe the development of a ligand-based target prediction algorithm [SOM-based prediction of drug equivalence relationships (SPiDER)] and its experimental validation by finding off-targets of known drugs without strong ligand structural similarity. Finally, we demonstrate the model's applicability for identifying pharmacologically relevant macromolecular targets of de novo-designed NCEs, which we synthesized and tested in vitro.

Results and Discussion

Computer-Based Design of New Chemical Entities. As part of an early discovery program, we generated innovative molecules using our ligand-based de novo design software DOGS (inSili.com LLC) (19). The process by which DOGS generates a compound uses virtual organic synthesis. Compound generation is steered by optimizing the topological similarity between newly generated candidate structures and a template molecule. In this study, amprenavir (Fig. 1), a potent inhibitor of HIV 1 protease (HIVP) (20), served as the template. The virtual synthesis was fueled by a set of 25,144 building blocks and 83 organic reactions (Fig. *2A*). Overall, the automated compound design process generated 2,338 molecules, of which 856 were unique structures with 388 unique Murcko scaffolds (21). From this set of diverse candidates, we selected compound **1** (Fig. 1) based on its chemotype novelty (no entry in the Chemical

Significance

New chemical entities (NCEs) with desired pharmacological and biological activity spectra fuel drug discovery and provide tools for chemical biologists. Computer-assisted molecular design generates novel chemotypes with predictable polypharmacologies. We present the successful application of fully automated de novo drug design coupled with a pioneering approach for target panel prediction to obtain readily synthesizable bioactive compounds. This innovative concept enabled the identification of relevant macromolecular targets of computationally designed NCEs and led to the discovery of previously unknown off-targets of approved drugs.

Author contributions: D.R., T.R., P.S., and G.S. designed research, performed research, analyzed data, and wrote the paper.

Conflict of interest statement: P.S. and G.S. are shareholders of inSili.com LLC, Zurich. G.S. is a shareholder of AlloCyte Pharmaceuticals Ltd., Basel, and a scientific consultant of the pharmaceutical industry.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: gisbert.schneider@pharma.ethz.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1320001111/-/DCSupplemental.

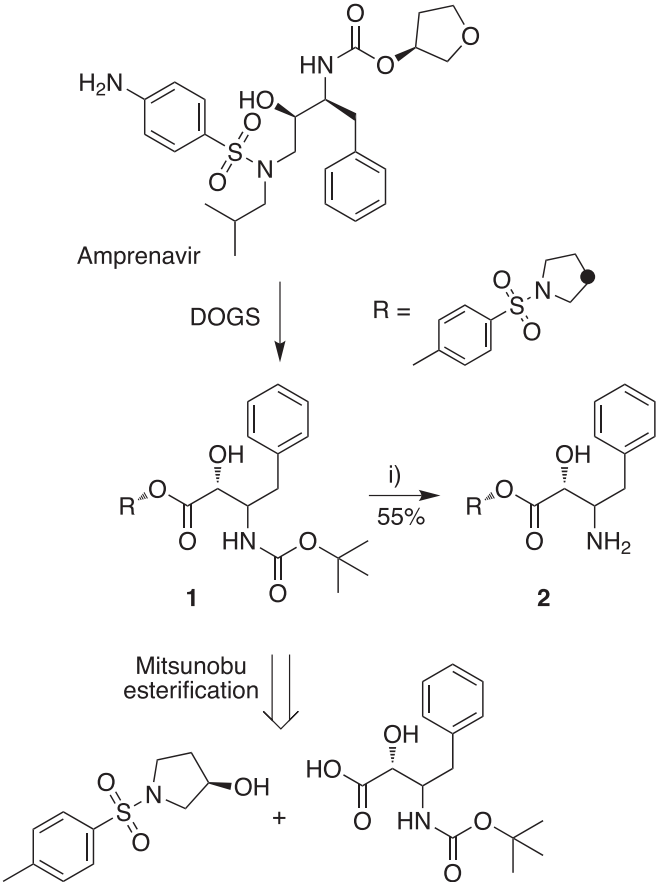

Fig. 1. The de novo design software DOGS generated compound **1** as a potential HIVP inhibitor using amprenavir as the template and suggested a feasible one-step synthetic pathway. Compound **2** was synthesized as a derivative of lead design **1**; (*i*) HCl 4 M/1,4-dioxane 0 °C → rt.

Abstract Service database, www.scifinder.org). Automated ligand docking into the catalytic center of HIVP [Protein Data Bank (PDB) code 1HPV] (22, 23) suggested a reasonable binding pose for compound **1** (GoldScore = 86; Fig. S1). By following the computer-generated synthesis scheme, we obtained **1** through the Mitsunobu esterification of enantiomerically pure building blocks (Fig. 1). We evaluated the inhibitory potency of **1** against HIVP but did not observe any relevant activity (10% inhibition at a compound concentration of 100 μM). We then synthesized and tested derivative **2**, which lacks the *tert*-butoxycarbonyl group of compound **1** and presents a more advantageous docking pose than **1** by potentially interacting with both catalytic Asp25 residues in the dimeric structure (GoldScore = 82; Fig. S1). However, compound **2** was also virtually inactive against HIVP (27% inhibition at 100 μM).

Target Prediction Using Publicly Available Software Tools. Because the applicability of compounds **1** and **2** as anti-HIVP chemotypes appeared limited, we investigated the possibility of exploiting the readily synthesizable NCEs **1** and **2** by leapfrogging to another drug target. Initially, we relied on publicly available target prediction tools. HIVP was the top predicted target for **1** according to the similarity ensemble approach (SEA) (24), fully corroborating the original DOGS design intended to mimic amprenavir (Table S1). The second most confident SEA prediction for **1** was β-secretase-1 (BACE-1), which was also suggested by the semantic link association prediction (SLAP) (25) for amprenavir (Table S2). In addition, the prediction of activity spectra for substances (PASS) (26) predicted that compound **1** would exhibit HIVP and BACE-1 inhibition (Table S3). Finally, the software SuperPred (27), which suggests targets via a pairwise comparison of query molecules to known drugs, identified HIVP inhibitors, including amprenavir, as the drugs most similar to query compound **1** (Table S4). For compound **2**, SuperPred and SEA again advocated HIVP as the drug target (Tables S1 and S4). These results suggested that DOGS retained the essential structural features of amprenavir in the design of compound **1** and in its derivative **2**, which clearly favored HIVP and BACE-1 as the expected targets. In vitro testing revealed that compound **1** was also inactive against BACE-1, thus rendering these target predictions incorrect. We reasoned that structures **1** and **2** may lie outside the domain of applicability of the existing fingerprint- and substructure-based target prediction methods, and therefore, we pursued the development of a novel target prediction method (SPiDER) as a complementary approach with a stronger focus on the prediction of targets for NCEs.

SPiDER Approach. Chemically abstract ("fuzzy") molecular representations, such as pharmacophoric feature descriptors, can be used to find subtle functional relationships between compounds, thereby allowing a molecule to leapfrog onto an unrelated target (28, 29). When used in similarity searches, such fuzzy molecular representations have often demonstrated greater scaffold-hopping potential than atomistic approaches (10, 30). Consequently, we implemented SPiDER as a software tool that builds on fuzzy molecular representations for use with de novo-designed NCEs. We relied on the established concept of SOMs to capture the local domains of model applicability (Fig. 2*A*). SOMs were originally developed as a neural network-inspired heuristic to reduce dimensionality (9), and they have become a workhorse of molecular informatics (31, 32). Using the unsupervised SOM algorithm, we clustered 12,661 manually annotated, pharmaceutically relevant drugs and lead compounds [collection of bioactive reference analogs (COBRA); inSili.com LLC] (33). The resulting 2D map tessellates this reference space into clusters of drug molecules representing local neighborhoods (10, 34–36). A query compound is assigned to exactly one target cluster on this map. For target inference via SPiDER, only the known drugs from this local domain are considered.

One aspect of the SPiDER method is the estimation of the statistical significance of each target prediction. The pairwise Euclidean distances between the query and the local reference drugs are calculated and evaluated according to a background distance distribution. Averaging the false-positive error probabilities for the reference drugs annotated to bind to different targets emulates a false-discovery rate in multiple hypothesis testing and provides a motivated confidence score for the prediction. SPiDER performs the target prediction twice, using the following two different molecular representations for two SOM projections: (*i*) the chemically advanced template search [chemically advanced template search (CATS)] topological pharmacophore descriptor (SOM1) (10, 37) and (*ii*) the Molecular Operating Environment (MOE) physicochemical properties and indices (SOM2) (Chemical Computing Group). Jury consensus predictions are obtained as the average of the two confidence scores. Finally, a background score distribution allows for the *P* value calculation of the jury scores to indicate the significance of an acquired prediction (Fig. 2*A*).

We performed a stratified 10-fold cross-validation (38) to estimate the scope of the SPiDER model. To measure performance, we analyzed the percentage of molecules with all annotated biochemically confirmed targets scored at $P < 5\%$. On average, 10.9 predictions per query compound were statistically significant (Table S5), which is in agreement with other studies that have reported 3–10 targets per drug depending on the target class (39). The

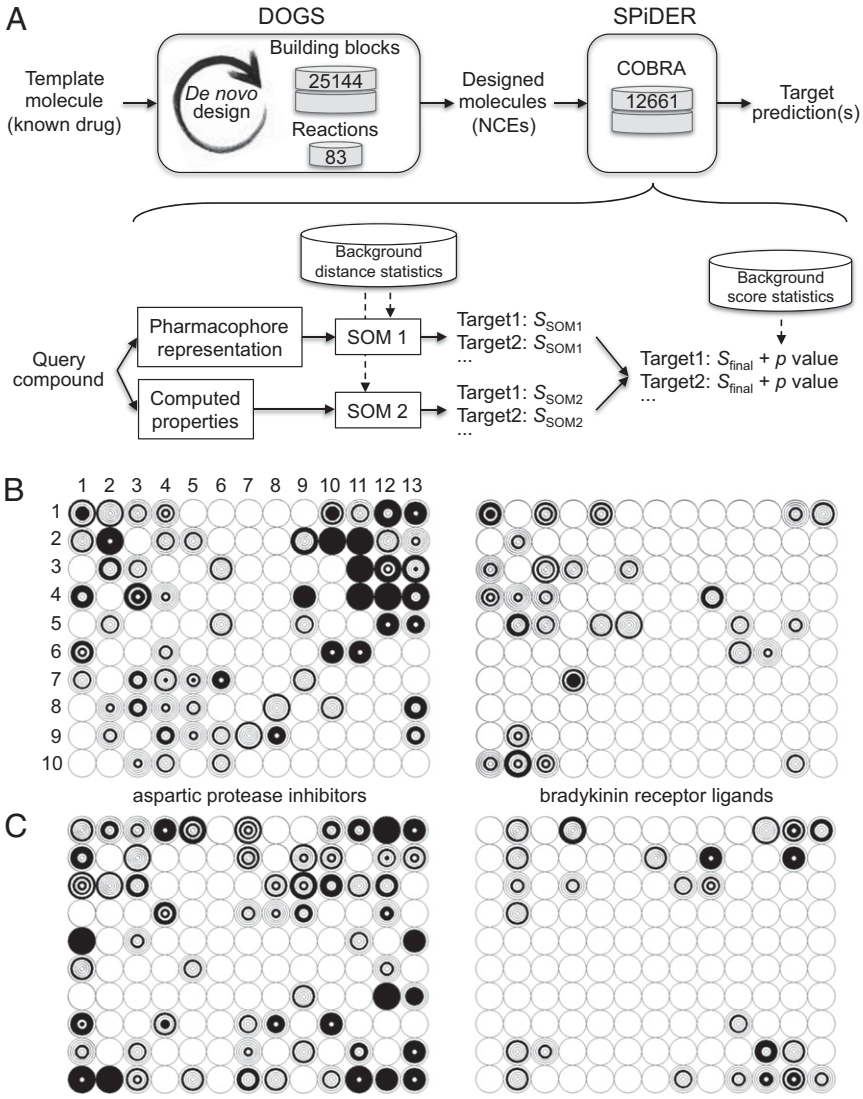

Fig. 2. (*A*) Scheme of the de novo design (DOGS) and SPiDER target prediction. (*B*) SOM depiction of the distribution of aspartic protease inhibitors (*Left*) and bradykinin receptor ligands (*Right*). The compound density in each of the 13 × 10 = 130 clusters (circles) is represented by the black color intensity. The concentric layers in each cluster denote quantile ranges of the compound distribution according to the distance from the centroid. On the CATS-based map (*B*), amprenavir is located in cluster (12/4), and the de novo-designed compounds **1** and **2** are located in cluster (12/5). On the MOE-based map (*C*), amprenavir and **1** are located in cluster (2/10), and compound **2** is in cluster (8/10). The toroidal SOMs were trained using the in-house command line tool *molmap* (56). The SOM size was chosen to have ~100 molecules per cluster, which proved expedient in preliminary experiments. The initial learning step size was 1, and the initial neighborhood update radius was 7 to allow for full signal propagation through the map topology. Training was performed using linear parameter decay and random sampling over 500,000 training cycles (9).

CATS-based (SOM1) prediction alone yielded 41 ± 0.7%, and the MOE-based (SOM2) method yielded 41.3 ± 0.5% complete target profile predictions. To investigate the complementarity of the selected molecular representations, we compared the performance of the individual prediction methods per target. The two molecular representations performed differently for most targets with only weak correlation (Pearson $r^2 = 0.44$; Fig. S2*A*). This finding supports the implementation of a prediction method that combines these multiple models and potentially benefits from their different scopes.

We studied different mathematical functions to merge the prediction scores from SOM1 and SOM2. We found that combination functions sensitive to low scores are the most accurate and yield target profile predictions that are ~65% complete (Table S5). The observed accuracy level is in agreement with the values reported by Hopkins and colleagues, who found that a fingerprint-based model correctly predicted at least one target for 64% of their data (12). We calculated the ROC AUC values to include the global ranking of true positives as a second evaluation criterion. Both the geometric average and the minimum value suffer because predictions made by only one model are neglected, resulting in the loss of low-confidence true-positive predictions (Fig. S2*B*). The arithmetic average performed equivalently in early retrieval and exploited single model predictions, thereby yielding a high ROC AUC value of 0.92 (Table S5).

Off-Target Prediction for Known Drugs via SPiDER. To prospectively probe the applicability of SPiDER, we predicted off-targets for the library of pharmacologically active compounds (LOPAC) collection of compounds (Sigma Aldrich) by mimicking the target prediction for NCEs. For this experiment, we only considered those predictions relying on COBRA reference drugs with a structural Tanimoto similarity <0.2 (Daylight Fingerprints) to the respective LOPAC query. This procedure attempted to find nontrivial, unexpected target relationships between reference drugs that are structurally dissimilar to the respective query molecule (expressed by a low structural Tanimoto similarity) and thus unrecognizable using substructure-based approaches. We investigated LOPAC queries for which SPiDER reported the most confident target predictions with the lowest Tanimoto similarity to the associated reference compounds. Among such high-confidence predictions ($P < 0.001$), we found monoamine oxygenase (MAO) as the top off-target prediction for the serotonin reuptake inhibitor fluoxetine. In full support of the SPiDER prediction, fluoxetine is in fact a MAO inhibitor both in vitro and in vivo (40). Similarly, we experimentally tested the top off-target prediction for fenofibrate ($P < 0.001$, Tanimoto similarity = 0.16 to the nearest

reference compound; Fig. S2*C*) and confirmed the $Na_v1.5$ ion channel (IC_{50} = 69 μM ± 1.2 log units; Fig. 3*B*). Fenofibrate presents natriuretic and cardiac remodeling effects (41, 42) and blocks basolateral KCNQ1 K^+ channels (43). To the best of our knowledge, its binding to Na^+ channels has not been previously reported. $Na_v1.5$ is the principal Na^+ channel isoform in cardiomyocytes, and variants of the $Na_v1.5$-encoding gene have been linked to congenital and acquired long QT syndromes (44). Although fenofibrate demonstrated only a weak affinity to $Na_v1.5$ in the assay, the data suggest that related chemotypes may present similar traits and liabilities.

Target Identification for NCEs via SPiDER. Having validated the SPiDER model for its ability to correctly infer off-targets despite a lack of structural similarity to the reference drugs, we predicted potential targets of de novo-designed compounds **1** and **2**. Although HIVP and BACE-1 were also predicted, SPiDER ranked other targets with higher confidence (Table 1). Similar target profiles were predicted for **1** and **2** that occasionally overlapped the predictions for amprenavir. In contrast with all publicly available prediction models, the top consensus SPiDER prediction for **1** and **2** was the bradykinin B_1 receptor, a G protein-coupled receptor involved in the mechanisms of inflammatory pain (45) and coronary vasomotor function (46). Being confidently predicted and practically exclusive to our approach, we tested compounds **1** and **2** for antagonistic activity toward the B_1 receptor. Although compound **1** presented only modest antagonism (EC_{50} ~ 100 μM; Fig. 4*A*), compound **2** exhibited high affinity and potent concentration-dependent B_1 antagonistic activity (K_B = 3.6 μM; EC_{50} = 17 μM; Fig. 3*A*). The B_1 receptor was correctly predicted by SPiDER based on the structures of reference compounds **3** [B_1 K_i ~ 0.1 nM (47)], **4** [B_1 K_i ~ 0.5 nM (48)], and **5** [B_1 K_i ~ 1.2 μM (49)] (Fig. 4).

SPiDER further suggested that neurokinin 1 (NK1) and vanilloid 1 (TRPV1) were receptors targeted by compounds **1** and **2** with favorable *P* values (Table 1). Again, these represent clinically relevant drug targets exclusively predicted by SPiDER for both NCEs under investigation. In full agreement with the computational analysis, we observed concentration-dependent effects for both compounds. NK1 receptor antagonism (K_B = 15 μM, EC_{50} = 100 μM ± 0.7 log units; Fig. 3*B*) and agonism on the TRPV1 ion channel (EC_{50} = 76 μM ± 0.7 log units; Fig. 3*C*) were measured for compound **2**. To probe whether any of these activities were inherited from the de novo design template, we investigated the effect of amprenavir on the B_1, NK1, and TRPV1 receptors. As correctly recognized by SPiDER, amprenavir did not exhibit activity against the B_1 or TRPV1 receptor up to 100 μM. Its weak NK1 receptor antagonism (EC_{50} > 100 μM) also agrees with the prediction (P = 0.021). Taken together, we successfully and efficiently used SPiDER to determine subtle functional bioactivity traits for both de novo-designed molecular frameworks and their drug template. For further assessment of SPiDER for NCE target prediction, we provide a public version of the model on our webserver (http://modlab-cadd.ethz.ch/software/).

Conclusions. Structural genomics and systems pharmacology greatly benefit from innovations in molecular informatics (50, 51). We implemented and experimentally validated a unique method for ligand-based target prediction that extends the capabilities of related approaches. Our software combines multiple representations of small molecules in a conservative jury approach, thereby reducing false-positive predictions in the high confidence range. The binding association of de novo designs **1** and **2** to the B_1 receptor indicates the explorative nature of the SPiDER approach. Because SPiDER determined all of the correct targets for two-thirds of the test data in retrospective studies, the method is expected to perform at least equivalently to other computational target prediction tools (12, 24, 26, 27). The results of our study demonstrate how the approach can be included in de novo drug design to obtain target-binding confidence in the proposed NCEs or to redirect stalled drug discovery projects.

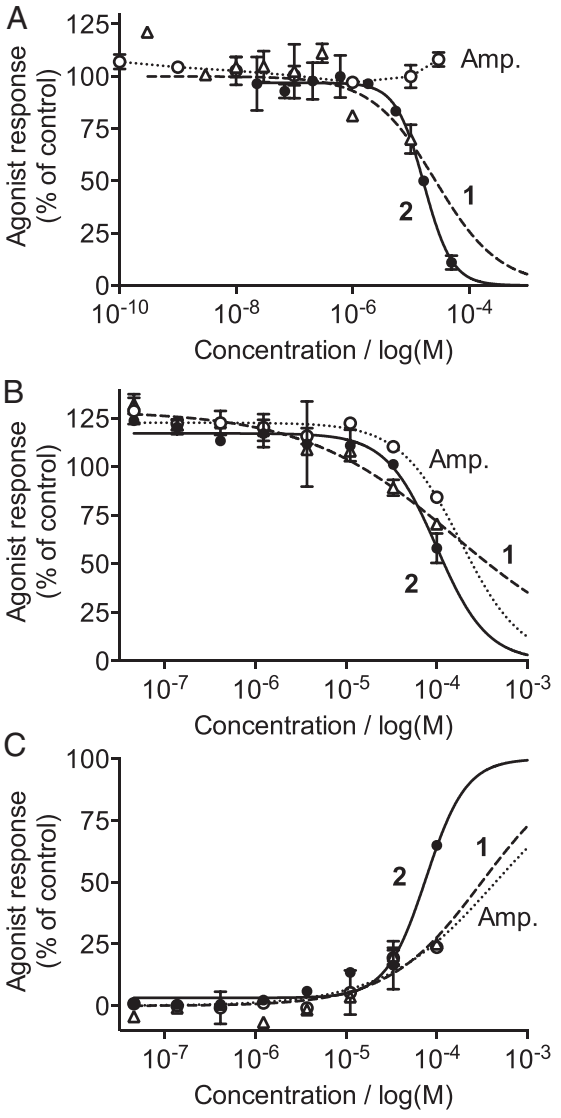

Fig. 3. Effects of compound **1**, compound **2**, and amprenavir (Amp.) on (*A*) bradykinin (B_1) receptor (control agonist: 3 nM LysdesArg^9[Leu^8]-BK; EC_{50} = 0.2 nM), (*B*) NK1 receptor {control agonist: 1 nM [Sar^9,Met(O$_2$)^{11}]-SP, EC_{50} = 0.4 nM}, and (*C*) TRPV1 (control agonist: 1 μM capsaicin, EC_{50} = 7 nM; n = 2, mean and SEM).

Materials and Methods

De Novo Design. DOGS (19) was used for the de novo molecular design with amprenavir as the template, 25,144 molecular building blocks, and 83 coupling reactions (52) (inSili.com LLC). The similarities between the designs and the template were computed using the iterative similarity optimal assignment kernel (ISOAK) method on reduced molecular graph representations (5). Nine independent runs were performed with 200 randomly selected start fragments (ISOAK α values ranging from 0.1 to 0.9). Scaffold occurrences in the set of designed molecules were determined by comparing the canonical simplified molecular input line entry system (SMILES) representations of extracted scaffolds on the KNIME platform v2.6.0 (53) using RDKit nodes (www.rdkit.org) for scaffold extraction and KNIME-native nodes to determine their frequency.

Reference Compounds. A manually curated collection of bioactive reference compounds (COBRA, v12.6, inSili.com LLC) (33) containing 12,661 unique

Table 1. Drug targets predicted by SPiDER for amprenavir, compound 1, and compound 2 with $P < 0.05$

Amprenavir	Compound **1**	Compound **2**
NK_{1-3} (0.021)	Cholecystokinin A/B receptor (0.017)	Opioid δ, κ, μ receptor (0.014)
20S proteasome (0.023)	Bradykinin receptor B_1 (0.018)	Dipeptidyl peptidase IV (0.025)
$Na_v1.7$ channel (0.027)	TRPV1 and 4 (0.020)	CCR1, 3, 5 receptor (0.026)
Aspartic endopeptidase[a,b,c] (0.027)	Vasopressin receptor 1A, 1B, 2 (0.020)	FSH/GnRH receptor (0.027)
Protein tyrosine phosphatase 1B (0.028)	FSH and GnRH receptor (0.028)	TRPV1, 4 and 8 (0.028)
Cholecystokinin A/B receptors (0.029)	Cysteine endopeptidase[d,e] (0.028)	Serine endopeptidase and protease[f,g,h] (0.031)
α4β1 integrin (0.029)	Serine endopeptidase[f,g] (0.033)	Bradykinin receptor B_1 (0.031)
Oxytocin receptor (0.029)	Aspartic endopeptidase[c] (0.033)	Aspartic endopeptidase[c] (0.034)
Serine endopeptidase[f] (0.029)	NK_{1-3} (0.038)	NK_1 and NK_2 (0.035)
Tyrosine kinase[i] (0.033)	Oxytocin receptor (0.042)	Urokinase plasminogen activator surface receptor (0.036)
Cysteine endopeptidase[d,j] (0.034)	$Na^+v1.7$ channel (0.048)	Histone deacetylase (0.040)
Bombesin receptor 1/2 (0.039)		Capsid assembly inhibitor (0.042)
		Serine threonine kinase[k] (0.043)
		α4β1 integrin (0.045)
		RasFTase (0.048)

P values are in parentheses.
[a]Includes cathepsin D, HIV protease, Pol polyprotein, and SIV protease.
[b]Includes endothiapepsin and saccharopepsin (proteinase A).
[c]Includes plasmepsins, renin, and secretase (Abeta secretion).
[d]Includes calpain, cathepsins B, K, L, and S, cruzain, falcipain, picornavirus 3C-like protease, and virus 3C protease.
[e]Includes caspase 1.
[f]Includes elastase, factor Xa, and thrombin.
[g]Includes β-tryptase, chymase, factor IXa, trypsin, tryptase, and urokinase.
[h]Includes factor VIIa.
[i]Includes c-Src, LCK, and VEGFR1.
[j]Includes papain.
[k]Includes CDK, NF-κB, MAPK14, NF-protein kinase B and C, and rho kinase.

pharmacologically active molecules annotated with one or more of 251 biomolecular targets and specific subtype activities was used for training, retrospective, and prospective evaluations of the SPiDER model.

Data Preparation and Molecular Representation. Compounds were provided as SMILES and processed in KNIME v2.6.0 (53) with the MOE "wash" function (2011.10; Chemical Computing Group) using the options "disconnect salts," "remove ion pairs," "deprotonate strong acids," "remove minor component," "protonate strong bases," and "add hydrogen." The 210-dimensional CATS vectors were calculated with an in-house software tool with a maximal correlation distance of 10 bonds and pharmacophoic feature type-sensitive descriptor scaling (10). The following 186 descriptors from MOE were calculated using the MOE "QSAR descriptors" function (MOE 2011.10): simple descriptors, chi indices, BCUT descriptors, PEOE descriptors, Q_charge descriptors, Lipinski rule of fives, Kier indices, GCUT descriptors, Slog*P* descriptors, VSA descriptors, SMR descriptors, and "other 2D descriptors." The raw descriptor values were standardized according to the training data distribution to prevent bias due to different value ranges.

SPiDER Prediction. The respective query molecule was projected onto the two SOMs trained with the corresponding molecular representation of the COBRA compound set according to the winner-takes-all rule (54) using in-house Java code. The projection defined the relevant cluster of reference compounds (the query's local domain). To score a target *C*, only the set C_Θ of coclustered reference compounds labeled to bind this target *C* was considered. Euclidean distances d_{xy} from each reference ligand *x* in C_Θ to the query molecule *y* were calculated in descriptor space and transformed into *P* values $P(D \le d_{xy})$. The empirical probability distribution $P(D)$ of pairwise distances between all reference compounds that are annotated to not bind to the same target was precomputed to yield a distribution of distances between the intertarget compounds as false-positive error estimates for assuming that two molecules at a certain distance have a common biomolecular target. The *P* values were averaged for all ligands in C_Θ, as motivated by the false discovery rate calculations for the multiple hypothesis testing of individual pairwise comparisons (54). The score *S* for target *C* is computed as $S(y,C) = 1 - \sum_{x \in C_\Theta} P(D \le d_{xy}) / |C_\Theta|$, in which a value of *S* close to 1 indicates that the coclustered reference ligands are close to the query compound in descriptor space. Individual scores were obtained for both molecular representations (S_{SOM1}, S_{SOM2}) (Fig. 2*A*) and were combined as an arithmetic average: $S_{final}(y,C) = 0.5\ (S_{SOM1} + S_{SOM2})$. Predicting the targets for all reference compounds generated a distribution of S_{final}, from which *P* values for new predictions were computed based on the likelihood of a certain confidence in the prediction of known drug targets. Note that these *P* values are relatively high due to the likelihood of a more realistic null hypothesis in our background distribution compared with other approaches that report the *P* value as a measure of significance, such as the random shuffling in sequence alignment (55) or the random assembly of ligand sets in the target prediction by SEA (24).

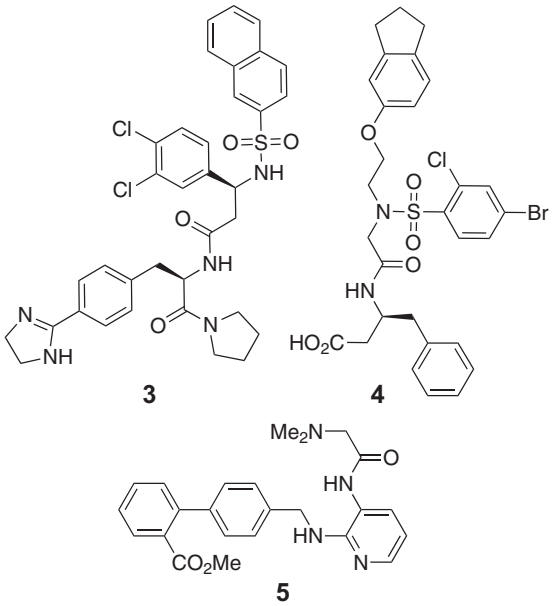

Fig. 4. Reference B_1 receptor antagonists. Compound **3** is the CATS reference for both **1** and **2**. Compound **4** is the MOE reference for **1**, and **5** is the MOE reference for **2**.

Retrospective Evaluation. Stratified 10-fold cross-validation was performed (38). For each fold, the model fitting included standardizing the descriptors according to the training fold, retraining the SOMs, and estimating the background distance distributions. A prediction was considered successful if all annotated targets of a test compound were predicted with a confidence score of $P < 5\%$. Scores for all compounds from the cross-validations were merged to calculate the ROC curves and to integrate them numerically with the trapezoid method for AUC calculation using in-house Java code.

Synthesis of (2*R*,3*R*)-(*R*)-1-Tosylpyrrolidin-3-yl 3-((*tert*-Butoxycarbonyl)Amino)-2-Hydroxy-4-Phenylbutanoate (Compound 1). (2*R*,3*R*)-3-(Boc-amino)-2-hydroxy-4-phenylbutyric acid (1.0 molar equivalent) and (*R*)-1-tosylpyrrolidin-3-ol (2.0 molar equivalent) were dissolved in dry tetrahydrofuran (18 mL/mmol carboxylic acid). Triphenylphosphine (0.8 molar equivalent) and diethyl azodicarboxylate (0.8 molar equivalent) were added to the solution. The mixture was heated under microwaves (150 W) for 30 min. The crude mixture was washed with water and purified via preparative HPLC using a gradient of 50–75% (acetonitrile: H_2O + 0.1% trifluoroacetic acid in each solvent) run over 16 min. White solid, *mp* = 106–107 °C, 42% (Fig. S3*A*).

Synthesis of (2*R*,3*R*)-(*R*)-1-Tosylpyrrolidin-3-yl 3-Amino-2-Hydroxy-4-Phenylbutanoate (Compound 2). Compound **1** (0.04 mmol) was cooled to 0 °C under nitrogen. A 4-N solution of HCl in dioxane (1 mL) was added, and the suspension was stirred at room temperature for 15 min. The solvent was evaporated under a nitrogen stream, and the crude product was purified via preparative HPLC (ACN/H_2O + 0.1% formic acid – 30–95% ACN gradient run over 16 min) to afford **2**. White oil, 55% (Fig. S3*B*).

ACKNOWLEDGMENTS. We thank Dr. Nikolay Todoroff for support in setting up the trial version of the SPiDER tool. This work was supported by the Swiss Federal Institute of Technology Zurich, Deutsche Forschungsgemeinschaft Grant FOR1406TP4, and the OPO Foundation, Zurich. The Chemical Computing Group Inc. (Montreal, Canada) provided a research license for the Molecular Operating Environment. The inSili.com LLC (Zurich, Switzerland) contributed research licenses for the software tools DOGS, *molmap*, and the COBRA database.

1. Bennani YL (2012) Drug discovery in the next decade: Innovation needed ASAP. *Drug Discov Today* 17(Suppl):S31–S44.
2. Johnson MA, Maggiora GM (1990) *Concepts and Applications of Molecular Similarity* (Wiley, New York).
3. Schneider G (2013) *De Novo Molecular Design* (Wiley-VCH, Weinheim, Germany).
4. Hartenfeller M, Schneider G (2011) De novo drug design. *Methods Mol Biol* 672: 299–323.
5. Rodrigues T, et al. (2013) De novo design and optimization of Aurora A kinase inhibitors. *Chem Sci* 4(3):1229–1233.
6. Spänkuch B, et al. (2013) Drugs by numbers: Reaction-driven de novo design of potent and selective anticancer leads. *Angew Chem Int Ed Engl* 52(17):4676–4681.
7. Besnard J, et al. (2012) Automated design of ligands to polypharmacological profiles. *Nature* 492(7428):215–220.
8. Rodrigues T, et al. (2013) Steering target selectivity and potency by fragment-based de novo drug design. *Angew Chem Int Ed Engl* 52(38):10006–10009.
9. Kohonen T (2001) *Self-Organizing Maps* (Springer, New York).
10. Reutlinger M, et al. (2013) Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for 'orphan' molecules. *Molec Inf* 32(2): 133–138.
11. Charifson PS, Corkery JJ, Murcko MA, Walters WP (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* 42(25):5100–5109.
12. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. *Nat Biotechnol* 24(7):805–815.
13. Tatonetti NP, Liu T, Altman RB (2009) Predicting drug side-effects by chemical systems biology. *Genome Biol* 10(9):238.
14. Hopkins AL, Mason JS, Overington JP (2006) Can we rationally design promiscuous drugs? *Curr Opin Struct Biol* 16(1):127–136.
15. Hopkins AL (2007) Network pharmacology. *Nat Biotechnol* 25(10):1110–1111.
16. Antolín AA, Jalencas X, Yélamos J, Mestres J (2012) Identification of pim kinases as novel targets for PJ34 with confounding effects in PARP biology. *ACS Chem Biol* 7(12): 1962–1967.
17. Lounkine E, et al. (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486(7403):361–367.
18. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? *Nat Rev Drug Discov* 5(12):993–996.
19. Hartenfeller M, et al. (2012) DOGS: Reaction-driven de novo design of bioactive compounds. *PLOS Comput Biol* 8(2):e1002380.
20. Porter DJT, Hanlon MH, Carter LH, 3rd, Danger DP, Furfine ES (2001) Effectors of HIV-1 protease peptidolytic activity. *Biochemistry* 40(37):11131–11139.
21. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 39(15):2887–2893.
22. Rose PW, et al. (2013) The RCSB Protein Data Bank: New resources for research and education. *Nucleic Acids Res* 41(Database issue):D475–D482.
23. Kim EE, et al. (1995) Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J Am Chem Soc* 117(3): 1181–1182.
24. Keiser MJ, et al. (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 25(2):197–206.
25. Chen B, Ding Y, Wild DJ (2012) Assessing drug target association using semantic linked data. *PLOS Comput Biol* 8(7):e1002574.
26. Lagunin A, Stepanchikova A, Filimonov D, Poroikov V (2000) PASS: Prediction of activity spectra for biologically active substances. *Bioinformatics* 16(8):747–748.
27. Dunkel M, Günther S, Ahmed J, Wittig B, Preissner R (2008) SuperPred: Drug classification and target prediction. *Nucleic Acids Res* 36(Web Server issue):W55–W59.
28. McGregor MJ, Luo Z, Jiang X (2007) Virtual screening in drug discovery. *Drug Discovery Research: New Frontiers in the Post-Genomic Era*, ed Huang Z (Wiley, New York), pp 63–88.
29. Klenner A, Hartenfeller M, Schneider P, Schneider G (2010) 'Fuzziness' in pharmacophore-based virtual screening and de novo design. *Drug Discov Today Technol* 7(4): e237–e244.
30. Schneider G (2013) De novo design: Hop(p)ing against hope. *Drug Discov Today* 10(4):e453–e460.
31. Schneider P, Tanrikulu Y, Schneider G (2009) Self-organizing maps in drug discovery: Compound library design, scaffold-hopping, repurposing. *Curr Med Chem* 16(3): 258–266.
32. Zupan J, Gasteiger J (1999) *Neural Networks in Chemistry and Drug Design* (Wiley-VCH, Weinheim, Germany).
33. Schneider P, Schneider G (2003) Collection of bioactive reference compounds for focused library design. *QSAR Comb Sci* 22(7):713–718.
34. Schneider P, Schneider G (2004) Navigation in chemical space: Ligand-based design of focused compound libraries. *Chemogenomics in Drug Discovery*, eds Kubinyi H, Müller G (Wiley-VCH, Weinheim, Germany), pp 341–376.
35. Schneider G, Tanrikulu Y, Schneider P (2009) Self-organizing molecular fingerprints: A ligand-based view on drug-like chemical space and off-target prediction. *Future Med Chem* 1(1):213–218.
36. Rupp M, Schneider P, Schneider G (2009) Distance phenomena in high-dimensional chemical descriptor spaces: Consequences for similarity-based approaches. *J Comput Chem* 30(14):2285–2296.
37. Schneider G, Neidhart W, Giller T, Schmid G (1999) 'Scaffold-hopping' by topological pharmacophore search: A contribution to virtua screening. *Angew Chem Int Edu* 38(19):2894–2896.
38. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Molec Inf* 29(6-7):476–488.
39. Antolín AA, Mestres J (2012) Knowledge base for nuclear receptor drug discovery. *Therapeutic Targets: Modulation, Inhibition, and Activation*, eds Botana LM, Loza MI (Wiley, New York), pp 309–326.
40. Holt A, Baker GB (1996) Inhibition of rat brain monoamine oxidase enzymes by fluoxetine and norfluoxetine. *Naunyn Schmiedebergs Arch Pharmacol* 354(1):17–24.
41. Shatara RK, Quest DW, Wilson TW (2000) Fenofibrate lowers blood pressure in two genetic models of hypertension. *Can J Physiol Pharmacol* 78(5):367–371.
42. Lebrasseur NK, et al. (2007) Effects of fenofibrate on cardiac remodeling in aldosterone-induced hypertension. *Hypertension* 50(3):489–496.
43. Bajwa PJ, et al. (2007) Fenofibrate inhibits intestinal Cl- secretion by blocking basolateral KCNQ1 K+ channels. *Am J Physiol Gastrointest Liver Physiol* 293(6):G1288–G1299.
44. Abriel H (2007) Roles and regulation of the cardiac sodium channel Na v 1.5: Recent insights from experimental studies. *Cardiovasc Res* 76(3):381–389.
45. Couture R, Harrisson M, Vianna RM, Cloutier F (2001) Kinin receptors in pain and inflammation. *Eur J Pharmacol* 429(1-3):161–176.
46. Aptecar E, et al. (2006) Coronary vasomotor response to the selective B1-kinin-receptor agonist Des-Arg9-bradykinin in humans. *J Heart Lung Transplant* 25(2):187–194.
47. Ferrari B, et al. (1997) Novel *N*-(arylsuphonyl)amino acid derivatives having bradykinin receptor affinity, Patent WO 9725315.
48. Hart T, Ritchie TJ (2002) Sulfonamide derivatives, Patent WO 02092556.
49. Feng DM, et al. (2005) 2,3-Diaminopyridine as a platform for designing structurally unique nonpeptide bradykinin B1 receptor antagonists. *Bioorg Med Chem Lett* 15(9): 2385–2388.
50. Taboureau O, Baell JB, Fernández-Recio J, Villoutreix BO (2012) Established and emerging trends in computational drug discovery in the structural genomics era. *Chem Biol* 19(1):29–41.
51. Futamura Y, et al. (2012) Morphobase, an encyclopedic cell morphology database, and its use for drug target identification. *Chem Biol* 19(12):1620–1630.
52. Hartenfeller M, et al. (2011) A collection of robust organic synthesis reactions for in silico molecule design. *J Chem Inf Model* 51(12):3093–3098.
53. Berthold MR, et al. (2008) KNIME: The Konstanz information miner. *Data Analysis, Machine Learning and Applications*, eds Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (Springer, Berlin), pp 319–326.
54. Storey JD (2002) A direct approach to false discovery rates. *J R Stat Soc B* 64(3): 479–498.
55. Larkin MA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21): 2947–2948.
56. Schneider G, Wrede P (1998) Artificial neural networks for computer-based molecular design. *Prog Biophys Mol Biol* 70(3):175–222.