Communication

www.molinf.com

molecular informatics

DOI: 10.1002/minf.202000109

Virtual Screening and Design with Machine Intelligence Applied to Pim-1 Kinase Inhibitors

Petra Schneider,[a, b] Martin Welin,[c] Bo Svensson,[c] Björn Walse,[c] and Gisbert Schneider*[a]

Abstract: Ligand-based virtual screening of large compound collections, combined with fast bioactivity determination, facilitate the discovery of bioactive molecules with desired properties. Here, chemical similarity based machine learning and label-free differential scanning fluorimetry were used to rapidly identify new ligands of the anticancer target Pim-1 kinase. The three-dimensional crystal structure complex of human Pim-1 with ligand bound revealed an ATP-competitive binding mode. Generative de novo design with a recurrent neural network additionally suggested innovative molecular scaffolds. Results corroborate the validity of the chemical similarity principle for rapid ligand prototyping, suggesting the complementarity of similarity-based and generative computational approaches.

Keywords: artificial intelligence · crystal structure · de novo design · drug discovery · neural network

Molecular 'scaffold hopping' aims to identify sets of molecules that have markedly different chemical structures but share a certain function of interest, *e.g.*, inhibition of the same enzyme target.[1] Finding such isofunctional compounds expedites fast-follower studies, promotes drug-like compounds with potentially fewer side-effects, and helps re-vitalize stalled discovery projects. There are several computational approaches to facilitate this process by means of virtually screening of large collections of readily available compounds, and by de novo molecular design.[2] Recent advances of computer-assisted molecular design with 'artificial intelligence' (AI) raise the question to which degree the scaffold hopping capabilities of straightforward chemical similarity-based virtual screening, specifically similarity searching, and AI-based generative molecular design overlap.[3] We approached this challenge from a practical perspective, aiming to rapidly identify new inhibitors of the anticancer target Pim-1 (human proviral integration site for MuLV) kinase,[4] a serine/threonine kinase broadly expressed in bone marrow and other tissues.[5] The primary goal of this study was to explore a pragmatic solution to speeding-up hit and lead compound identification, by building on chemical similarity methods and fast experimental validation.

There are many known inhibitors of human Pim-1, rendering both chemical similarity searching and data-hungry machine learning feasible. Therefore, an ensemble similarity approach was pursued (Figure 1a). As a starting point, we collected a reference set of 683 Pim-1 inhibitors (K_i or $IC_{50} < 0.5$ μM) and 2377 experimentally confirmed inactives (K_i or $IC_{50} \geq 0.5$ μM) from the ChEMBL database,[6] and 15,041 bioactive druglike compounds from the manually curated *Collection of Bioactive Reference Analogs* (COBRA).[7] All molecules were represented in terms of cross-correlated, topological two-point pharmacophores (*Chemically Advanced Template Search*, CATS method)[8] and grouped into 200 clusters on a self-organizing map (SOM) by unsupervised learning.[9] Visualization of the resulting compound distributions pointed to cluster (20/9) as a Pim-1 'hot spot' in chemical space (Figure 1b). This cluster contained 168 molecules, 81 of which were known Pim-1 actives and 32 known Pim-1 inactives, corresponding to an active:inactive ratio of approximately 2.5. The centroid vector of a SOM cluster may be interpreted as a consensus pharmacophore of this particular molecular ensemble. This vector has the same dimensionality and chemical meaning as the molecular descriptors of the training compounds. It may be considered a 'virtual ligand' representative of the local compound ensemble. This virtual ligand (centroid) formed by cluster (20/9) served as query for subsequent similarity searching in a collection of 5.35 million commercially available compounds (Table S1). Further analysis and compound selection was restricted to the 100 top-ranking virtual hits as an arbitrary cut-off, emulating a challenging low-budget scenario (Figure 1c, Figure S1, Table S2).

[a] *Dr. P. Schneider, Prof. Dr. G. Schneider*
Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland
E-mail: gisbert@ethz.ch
[b] *Dr. P. Schneider*
inSili.com GmbH, Segantinisteig 3, 8049 Zurich, Switzerland
[c] *Dr. M. Welin, Dr. B. Svensson, Dr. B. Walse*
SARomics Biostructures AB, Medicon Village, SE-223 81 Lund, Sweden

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.202000109

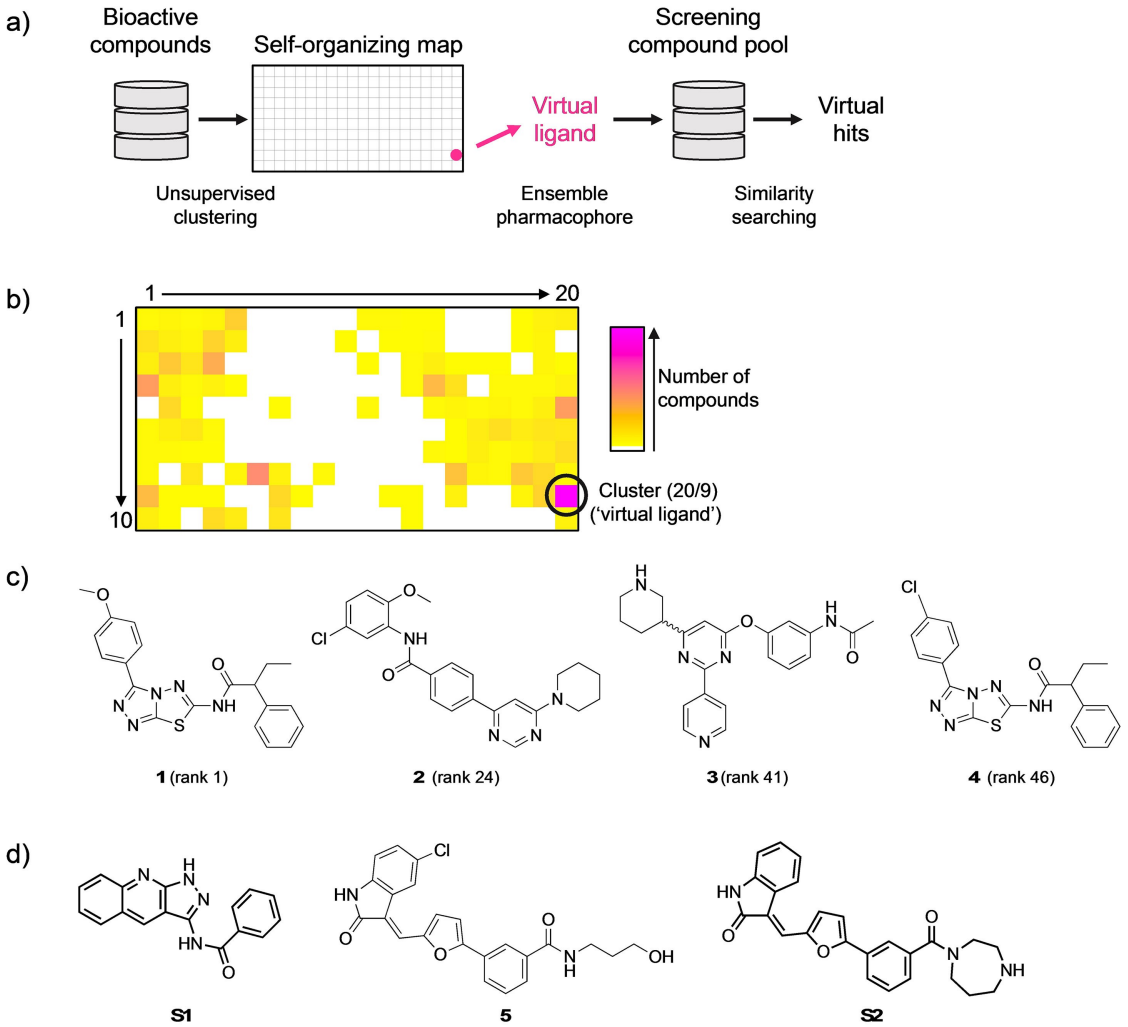

Figure 1. Schematic of the virtual screening approach (a). The self-organizing map (SOM) groups molecules according to their pharmacophore features, implicitly forming pharmacophore models ('virtual ligands') of the compound ensembles for similarity searching. The distribution of known Pim-1 inhibitors on a SOM of druglike chemical space reveals an activity 'hot spot' in cluster (20/9) (b). This map consists of a toroidal array of 20×10 clusters (grid fields). Colouring indicates cluster occupancy (number of compounds per cluster). The ensemble pharmacophore ('virtual ligand') of Cluster (20/9) was selected for similarity searching in the compound pool. Four top-ranking screening compounds ('virtual hits') were tested for Pim-1 binding (c). Chart (d) presents the most frequent scaffold (**S1**) among the virtual hits, the most similar known Pim-1 kinase inhibitor (**5**) to compound **3**, and the only common scaffold (**S2**) among the set of known Pim-1 kinase inhibitors and the *de novo* generated molecules.

Molecular scaffold analysis revealed rich chemical diversity among the set of virtual hits, with scaffold **S1**, a known kinase binding substructure motif,[10] present in 15% of the top-ranking screening compounds (Figure 1d). Importantly, none of these molecular frameworks ('Murcko' scaffolds)[11] were contained in the set of known Pim-1 inhibitors used for SOM training.

Compounds **1**–**4** were manually selected and ordered from the respective suppliers (Figure 1c). Compound **1** was the most similar to the virtual ligand model in terms of its pharmacophore features (CATS method,[8] rank 1 on the list of virtual hits, $CATS_{Distance} = 0.81$). Compounds **2** (rank 24, $CATS_{Distance} = 0.87$) and **3** (rank 41, $CATS_{Distance} = 0.88$) contain chemical scaffolds we considered surprising, especially compound **3** because of its branched scaffold and positively ionizable piperidine moiety. Compound **4** (rank 46, $CATS_{Distance} = 0.88$) was selected to probe the effect of the methoxy group present in compound **1**.

To rapidly obtain experimental information about these compounds as to their potential interaction with Pim-1, label-free differential scanning fluorimetry (DSF) was performed, immediately followed by X-ray crystallography. In this way, sufficient experimental feedback could be obtained without the necessity for extensive biochemical testing during this initial phase of hit identification.

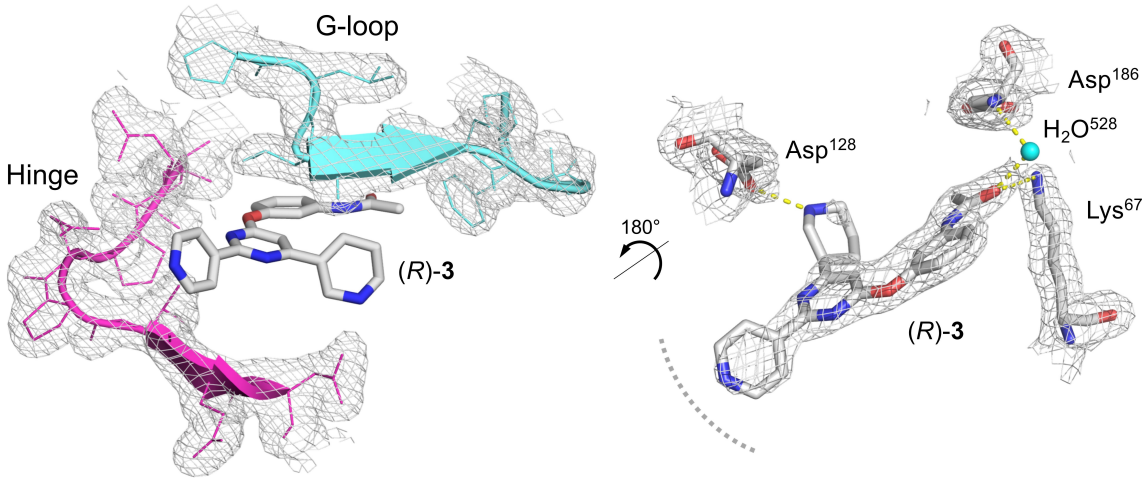

Figure 2. Binding mode of compound (*R*)**3** in the nucleotide pocket of human Pim-1 kinase (PDB ID: 6YKD). Cartoon representation of the structural complex (left) and direct ligand-protein interactions (right). The crystallographically determined electron density (2FoFc map), contoured at 1σ within 2 Å of highlighted residues, is represented as mesh. Directed polar interactions are shown as yellow dotted lines. Solvent-exposure is indicated by the dotted arc.

Compounds **1** and **4** showed interfering fluorescence that prohibited meaningful dose-response determination, and compound **1** was poorly soluble. These were therefore not considered further. DSF revealed stabilization of human Pim-1 by compounds **2** and **3** ($\Delta T_m = 2–3\,^{\circ}C$), suggesting direct ligand interaction. Compound (*RS*)-**3** concentration-dependently stabilized Pim-1 kinase (Figure S2). Compound **5** ($IC_{50} = 25$ nM, ChEMBL-ID: CHEMBL2037202, $CATS_{Distance} = 0.96$) is the most similar known Pim-1 inhibitor to compound **3**.[12] The molecular scaffolds of these two Pim-1 ligands are markedly different, constituting a successful scaffold hop.

For experimental validation of the preliminary DSF result compound (*RS*)-**3** was soaked into apo crystals of human Pim-1 protein. X-ray structure determination confirmed the ligand bound to the canonical ATP binding site of the kinase with a resolution of 1.86 Å (Figure 2). The core structure of the (*R*)-**3** enantiomer could be fitted to the observed electron density in the binding site. In the crystal structure the temperature factors for compound **3** are higher around the pyridine and piperidine moiety, indicating greater flexibility of these rings. The piperidine moiety is positioned with the nitrogen pointing towards Asp^{128}. In this binding mode the pyridine nitrogen of (*R*)-**3** is solvent-exposed. A charge-charge interaction is formed between the positively ionized piperidine nitrogen and Asp^{128}, and the ligand's carbonyl oxygen forms a potential ($d = 3.5$ Å) hydrogen bridge to Lys^{67}. In the ATP-bound state of Pim-1, the function-relevant amino acid residue Lys^{67} participates in canonical nucleotide binding.[13] A second, water-mediated hydrogen bridge seems feasible between the ligand carbonyl and the Asp^{186} main-chain amide.

Having confirmed the ATP-competitive binding mode of the new Pim-1 inhibitor found by similarity-based virtual screening of readily available screening compounds, we performed generative de novo molecular design of potential Pim-1 inhibitors. This machine learning approach provides access to a virtually infinite chemical space and offers the possibility to computationally obtain novel molecules.[2] In contrast to rule-based molecule construction, the generative model samples new molecules from a learned statistical distribution of the training data.[14] Here, we employed a recurrent neural network with long short-term memory (RNN-LSTM) for molecule generation.[15] The underlying chemical language model captures the syntax of bioactive molecules represented as SMILES strings, and emits new SMILES strings that satisfy the constraints of the training set. This approach previously led to the identification of novel compounds with desired bioactivity.[16]

For the present study, an RNN-LSTM model trained on approx. 400k SMILES strings of bioactive compounds[17] was fine-tuned on the set of 683 Pim-1 kinase inhibitors from ChEMBL, and a total of unique SMILES strings were sampled from this optimized generative model. The computer-generated molecules were ranked according to their CATS pharmacophore similarity to the virtual ligand model representing SOM cluster (20/9) (Figure 1b), and the 100 top-ranking molecules were kept for further analysis.

A single common scaffold (**S2**, Figure 1d) was found among the top-ranked de novo generated molecules and the known Pim-1 inhibitors. The nearest neighbour **5** of the confirmed virtual screening hit **3** also features this scaffold **S2**. One might speculate that this particular molecular framework epitomizes essential features of Pim-1 inhibitors. In fact, scaffold **S2** had been optimized as the core structure of potent antiproliferative pan-Pim kinase inhibitors, forming polar interactions with Asp^{128} in the nucleotide binding site.[12] It is noteworthy that compound **3** also contains this

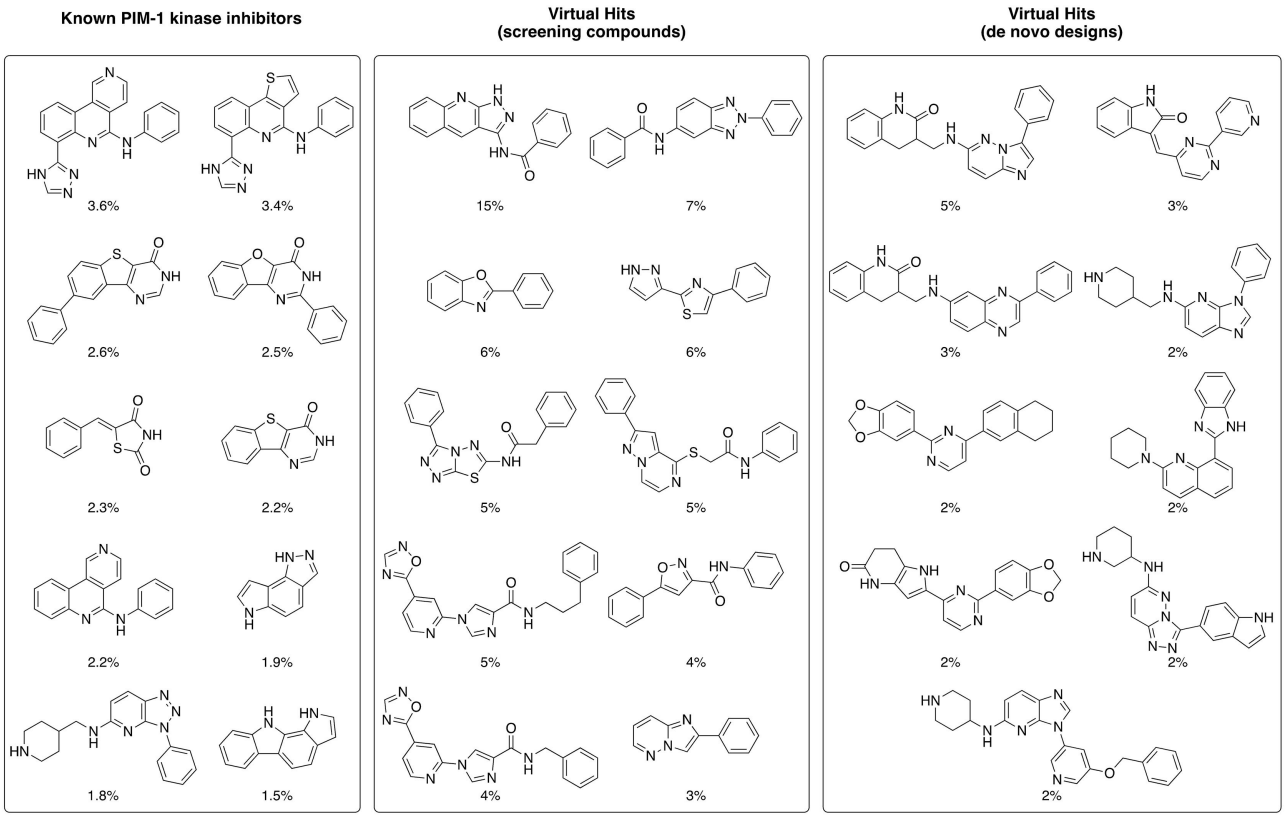

Figure 3. The most frequent atom scaffolds of known Pim-1 kinase inhibitors (*left panel*), the scaffolds of the 100 top-ranking screening compounds (*middle panel*), and the scaffolds of the 100 top-ranking de novo generated molecules (*right panel*). The relative frequency of each scaffold is given in percent.

particular structural feature. The scaffolds of the top-ranking screening compounds and the de novo designs contain several known kinase hinge-binding moieties (Figure 3), embedded in new structural context. The respective compounds may be worthwhile exploring in future kinase drug discovery projects.

Analysis of CATS pharmacophore distances revealed that the known actives from ChEMBL ($CATS_{Distance}$ mean $\pm$ SD $= 1.15 \pm 0.06$) are most dissimilar to the virtual ligand model, followed by the de novo generated molecules (1.02 ± 0.03), and the screening compounds (0.88 ± 0.2). The respective top hits have markedly different chemical structures (Figure 4), which is in agreement with the statistically significant dissimilarity of the distance distributions (Mann-Whitney test, P value $< 10^{-4}$). This result corroborates computational similarity searching for first-pass hit identification in drug discovery.

In conclusion, machine learning methods enabled straightforward virtual screening of existing and computationally-generated molecules. The newly identified Pim-1 inhibitor **3** highlights the efficiency of this ensemble similarity approach. In combination with a fast biophysical binding assay (DSF), the 'virtual ligand' method enabled rapid crystallographic determination of the kinase-ligand complex. With the X-ray structural data and the de novo designs at hand, hit-to-lead expansion and optimization of the initial hit compound by structure-based modeling and design seems straightforward. Importantly, these molecules provide motivated working hypotheses for experiment planning. Given a set of reference compounds, the approach may serve as a template for rapid hit and lead finding in medicinal and biological chemistry, complementing the various methods suitable in 'low-data' scenarios.[17,18]

Experimental

Software. Software for molecular descriptor calculation (CATS),[8] self-organizing map training and analysis (POOMA),[19] target prediction (TIGER),[20] and similarity searching (inSili.com LLC, Zurich, Switzerland) was run on an iMac workstation. *Parameter settings*: CATS: feature correlation over 0–9 bonds, with type scaling; POOMA: 20× 10 toroidal map, Gaussian neighborhood, random initialization, linear decay of the learning rate ($\tau_{initial} = 1$), 10^9 learning cycles. *De novo design*: The RNN-LSTM model for SMILES generation was obtained from GitHub[21] and optimized for Pim-1 kinase, as described.[17] Fine tuning was performed

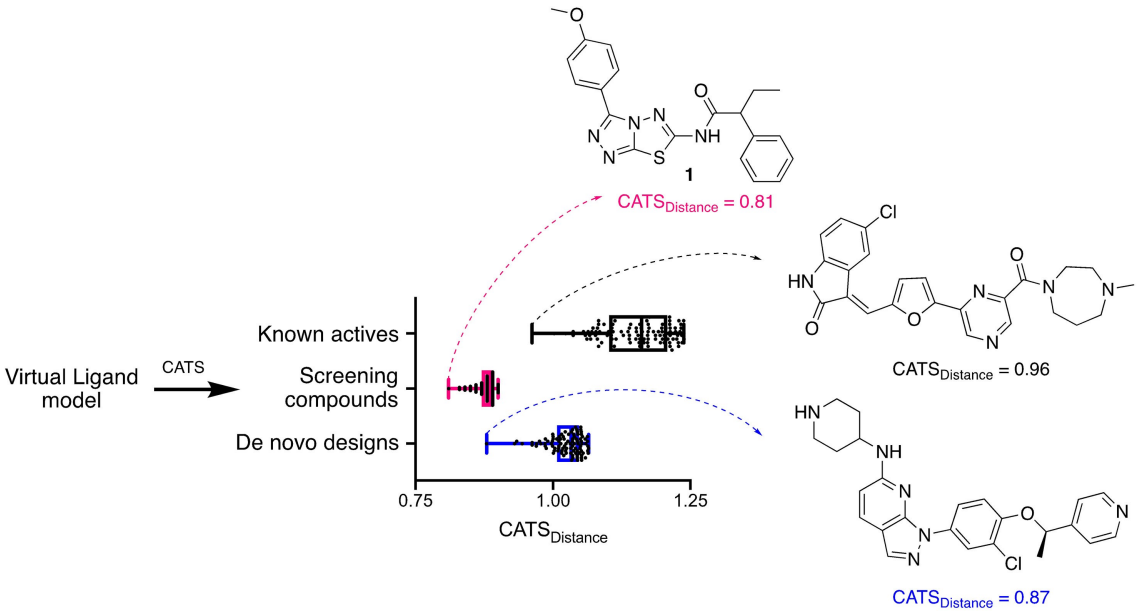

Figure 4. Pharmacophore similarity (CATS topological feature pairs) distributions of the 100 nearest neighbors to the virtual ligand model. The top-ranking chemical structures from the sets of known actives, screening compounds, and the de novo designs are shown. Box plots with median, 50% quartile range, min./max. values (known actives, *black*; screening compounds, *red*; de novo designs, *blue*).

over 40 epochs with 10-fold data augmentation; up to 5,000 new SMILES strings were sampled from each model state after 10, 20, 30, and 40 epochs, at a sampling temperature of 0.7. Generated SMILES strings were standardized and duplicates removed. Data analysis was performed with Prism 8.4.0 software (GraphPad Software LLC, San Diego, CA, USA).

Screening compounds. Compounds **1**, **2**, and **4** (cat. no. E157-4091, L010-0223, G952-3305) were ordered from ChemDiv Inc. (San Diego, CA, USA, www.chemdiv.com), compound **3** (cat. no. LEG 22312959) was ordered from Asinex Corp. (Delft, The Netherlands, www.asinex.com). Solid compound samples were dissolved in 100% dimethyl sulfoxide and used without further purification (>90+/−5% purity as specified by the supplier).

Differential scanning fluorimetry (DSF). Compounds were tested for stabilizing 0.1 mg ml^{-1} of recombinant human Pim-1 kinase (SARomics Biostructures AB, Lund, Sweden), using a Stratagene MX3000P qPCR instrument (ThermoFisher, Waltham, MA, USA) with SYPRO Orange (ThermoFisher, cat. no. S6650) as reporter dye.

Crystallography. Crystallization of Pim-1 kinase was performed at a concentration of 20 mg ml^{-1} in a buffer consisting of 50 mM Tris pH 7.5, 250 mM NaCl and 5 mM DTT. The crystal was grown within a couple of days at 4 °C using an MRC 3-well plate and in a 150+100 nl drop under the following condition: 1 M sodium acetate, 0.1 M imidazole pH 6.3. Apo crystals were soaked with 5 mM of compound **3** for 1 h before they were transferred to a cryosolution containing additional 20% glycerol, and then transferred to liquid N_2 for cryo-cooling. A data set of a soaked Pim-1 crystal was collected at station ID29 in the European Synchrotron Radiation Facility (ESRF, Grenoble, France). The diffraction data were processed using AutoPROC software to 1.86 Å.[22] The published Pim-1 kinase structure (PDB ID: 3R04)[23] served as the search model for structure determination with compound **3**, using Phaser crystallographic software.[24] The structure was refined to convergence using REFMAC5 software,[25] and model building was carried out in Coot.[26] The restraints file for the ligand was generated using Grade [1.2.19] software (Global Phasing Ltd., Cambridge, UK). Data collection and refinement statistics are provided as Supplementary Information.

Supporting Information

Supporting Information to this article is available online from the publisher's site. The X-ray structure of human Pim-1 in complex with compound **3** is available from the Protein Data Bank[27] (www.rcsb.org, PDB ID: 6YKD).

Conflict of Interest

P.S. and G.S. declare a potential financial conflict of interest as the founders of inSili.com GmbH, Zurich, and in their role as consultants to the pharmaceutical industry.

References

[1] a) G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem. Int. Ed.* **1999**, *38*, 2894–2896; *Angew. Chem.* **1999**, *111*, 3068–3070; b) Y. Hu, D. Stumpfe, J. Bajorath, *J. Med. Chem.* **2016**, *59*, 4062–4076.

[2] a) P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow Jr., J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, M. Zentgraf, J. E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann, G. Schneider, *Nat. Rev. Drug Discovery* **2020**, *19*, 353–364; b) G. Schneider, D. Clark, *Angew. Chem. Int. Ed.* **2019**, *58*, 10792–10803.

[3] a) G. Schneider, *Nat. Mach. Intell.* **2019**, *1*, 128–130; b) X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang, *Chem. Rev.* **2019**, *119*, 10520–10594.

[4] a) M. Narlik-Grassow, C. Blanco-Aparicio, A. Carnero, *Med. Res. Rev.* 2014, 34, 136–159; b) X. Zhang, M. Song, J. K. Kundu, M. H. Lee, Z. Z. Liu, *J. Cancer Prev.* **2018**, *23*, 109–116.

[5] A. Eichmann, I. Yuan, C. Breant, K. Alitalo, P. J. Koskinen, *Oncogene* **2000**, *19*, 1215–1224.

[6] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, A. R. Leach, *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

[7] P. Schneider, G. Schneider, *Mol. Inf.* **2003**, *22*, 713–718.

[8] M. Reutlinger, C. P. Koch, D. Reker, N. Todoroff, P. Schneider, T. Rodrigues, G. Schneider, *Mol. Inf.* **2013**, *32*, 133–138.

[9] a) T. Kohonen, *Biol. Cybern.* **1982**, *43*, 59–69; b) P. Schneider, Y. Tanrikulu, G. Schneider, *Curr. Med. Chem.* **2009**, *16*, 258–266.

[10] J. T. Metz, E. F. Johnson, N. B. Soni, P. J. Merta, L. Kifle, P. J. Hajduk, *Nat. Chem. Biol.* **2011**, *7*, 200–202.

[11] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 287–2893.

[12] M. Haddach, J. Michaux, M. K. Schwaebe, F. Pierre, S. E. O'Brien, C. Borsan, J. Tran, N. Raffaele, S. Ravula, D. Drygin, A. Siddiqui-Jain, L. Darjania, R. Stansfield, C. Proffitt, D. Macalino, N. Streiner, J. Bliesath, M. Omori, J. P. Whitten, K. Anderes, W. G. Rice, D. M. Ryckman, *ACS Med. Chem. Lett.* **2011**, *3*, 135–139.

[13] a) R. Padma, L. Nagarajan, *Cancer Res.* **1991**, *51*, 2486–2489; b) M. Friedmann, M. S. Nissen, D. S. Hoover, R. Reeves, N. S. Magnuson, *Arch. Biochem. Biophys.* **1992**, *298*, 594–601.

[14] a) T. Jebara, *Machine Learning: Discriminative and Generative*, Kluwer Academic, Dordrecht, **2004**; b) Y. Xu, K. Lin, S. Wang, L. Wang, C. Cai, C. Song, L. Lai, J. Pei, *Future Med. Chem.* **2019**, *11*, 567–597.

[15] a) S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735–1780; b) F. Grisoni, G. Schneider, *Chimia* **2019**, *73*, 1006–1011.

[16] a) W. Yuan, D. Jiang, D. K. Nambiar, L. P. Liew, M. P. Hay, J. Bloomstein, P. Lu, B. Turner, Q. T. Le, R. Tibshirani, P. Khatri, M. G. Moloney, A. C. Koong, *J. Chem. Inf. Model.* **2017**, *57*, 875–882; b) D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700153; c) D. Merk, F. Grisoni, L. Friedrich, G. Schneider, *Commun. Chem.* **2018**, *1*, 68; d) D. Bruns, D. Merk, K. S. Kumar, M. Baumgartner, G. Schneider, *ChemistryOpen* **2019**, *8*, 1303–1308.

[17] M. Moret, L. Friedrich, F. Grisoni, D. Merk, G. Schneider, *Nat. Mach. Intell.* **2020**, *2*, 171–180.

[18] a) S. K. Saikin, C. Kreisbeck, D. Sheberla, J. S. Becker, A. Aspuru-Guzik, *Expert Opin. Drug Discovery* **2019**, *14*, 1–4; b) I. I. Baskin, *Expert Opin. Drug Discovery* **2019**, *14*, 601–603; c) A. Lin, D. Horvath, G. Marcou, B. Beck, A. Varnek, *J. Comput.-Aided Mol. Des.* **2019**, *33*, 331–343.

[19] G. Schneider, P. Schneider, *Expert Opin. Drug Discovery* **2017**, *12*, 271–277.

[20] P. Schneider, G. Schneider, *Angew. Chem. Int. Ed.* **2017**, *56*, 11520–11524; *Angew. Chem.* **2017**, *129*, 11678–11682.

[21] M. Moret, L. Friedrich, F. Grisoni, D. Merk, G. Schneider, GitHub, ETH Zurich, **2020**, https://github.com/ETHmodlab/virtual_libraries.

[22] V. Vonrhein, C. Flensburg, P. Keller, A. Sharff, O. Smart, W. Paciorek, T. Womack, G. Bricogne, *Acta Crystallogr.* **2019**, *D67*, 293–302.

[23] Y. Xiang, B. Hirth, G. Asmussen, H. P. Biemann, K. A. Bishop, A. Good, M. Fitzgerald, T. Gladysheva, A. Jain, K. Jancsics, J. Liu, M. Metz, A. Papoulis, R. Skerlj, J. D. Stepp, R. R. Wie, *Bioorg. Med. Chem. Lett.* **2011**, *21*, 3050–3056.

[24] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, *J. Appl. Crystallogr.* **2007**, *40*, 658–674.

[25] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2011**, *67*, 355–367.

[26] P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, *Acta Crystllogr. Sect. D Biol. Crystallogr.* **2010**, *66*, 486–501.

[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.

Received: May 8, 2020
Accepted: June 17, 2020
Published online on July 9, 2020

