Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

Machine learning for target discovery in drug development

Tiago Rodrigues[1] and Gonçalo J. L. Bernardes[1,2]

Abstract

The discovery of macromolecular targets for bioactive agents is currently a bottleneck for the informed design of chemical probes and drug leads. Typically, activity profiling against genetically manipulated cell lines or chemical proteomics is pursued to shed light on their biology and deconvolute drug–target networks. By taking advantage of the ever-growing wealth of publicly available bioactivity data, learning algorithms now provide an attractive means to generate statistically motivated research hypotheses and thereby prioritize biochemical screens. Here, we highlight recent successes in machine intelligence for target identification and discuss challenges and opportunities for drug discovery.

Addresses

[1] Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz, 1649-028 Lisboa, Portugal
[2] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

Corresponding authors: Rodrigues, Tiago (tiago.rodrigues@medicina.ulisboa.pt); Bernardes, Gonçalo J.L (gb453@cam.ac.uk)

Current Opinion in Chemical Biology 2020, **56**:16–22

This review comes from a themed issue on **Next generation therapeutics**

Edited by **Gonçalo Bernardes** and **Raphaël Rodriguez**

For a complete overview see the Issue and the Editorial

https://doi.org/10.1016/j.cbpa.2019.10.003

Keywords

Target identification, Drug discovery, Chemical probes, Machine learning, Chemical proteomics.

Introduction

The future of molecular medicine relies on the identification and validation of small molecule effectors [1–4] despite the advent of biologics [2,5], such as antibody–drug conjugates. Indeed, specific drug–target binding and engagement remains a hallmark for modulation and treatment of diseases. Bioactive matter is rarely selective for a given target but rather engages several other related or unrelated macromolecules [6,7], through what has been coined as polypharmacology or network pharmacology [8,9]. Although in some circumstances, such as cancer, it is desirable to harness the potential of therapeutics that modulate a plethora of targets—ideally belonging to distinct signaling pathways [10]—generally, polypharmacology is unwanted because it is responsible for adverse drug reactions [11,12]. Therefore, factual knowledge of on- and off-targets correlated to efficacy and liabilities of chemical matter is of utmost importance to maximize benefits and mitigate attrition in development pipelines.

The identification and unraveling of pharmacology networks for phenotypic screening hits—either of synthetic or natural origin—arguably remains a bottleneck of modern drug discovery [13–15]. This is a consequence of our current inability to streamline state-of-the-art technologies for target identification, for example chemical proteomics. Nonetheless, advances in machine-intelligence heuristics and hardware and the increasing amount of publicly available chemical-biology data may afford untapped opportunities to speed up discovery programs that to date have stalled at the target identification phase [15]. In this review, we discuss recent progress made in machine learning (for an overview on basic concepts we refer to Refs. [16,17]) as a research hypothesis generator and as a vital method in the chemical biology toolbox for research into drug–target recognition. In addition, we highlight advantages of machine learning–based technologies and outline gaps in our knowledge to hopefully spur on future investigations and democratize its use among chemical biologists.

A winding road to drug target identification

Chemical proteomic approaches remain the gold-standard for the identification of macromolecular counterparts for small bioactive molecules [18–22]. These methods require the chemical modification of the ligand of interest by appending a “tag” moiety that will afterward enable a pull down of the formed ligand–target complexes. Downstream identification of the bound proteins can then be made by mass spectrometry or bespoke analytical methods [20]. By application of this concept, Cravatt et al. were able to identify the mitochondrial carnitine–acylcarnitine translocase SLC25A20 as a functional target of diterpenoid ester ingerol mebutate (**1**), which is a first-in-class treatment for actinic keratosis (Figure 1a) [23]. Other prominent examples include the identification of carbamates as arylacetamide deacetylase-like 1 regulators [24] and a clinical-stage imidazole as a promiscuous lipid hydrolases' inhibitor [25]. Although these studies clearly suggest that chemical proteomics lends itself to scrutinizing drug—target relationships, it becomes apparent that chemical manipulation of the native ligand can disrupt the binding affinity toward relevant on- and off-targets. Moreover, chemical proteomics rarely enables the identification of membrane proteins because of their instability in solution and low copy number [26].

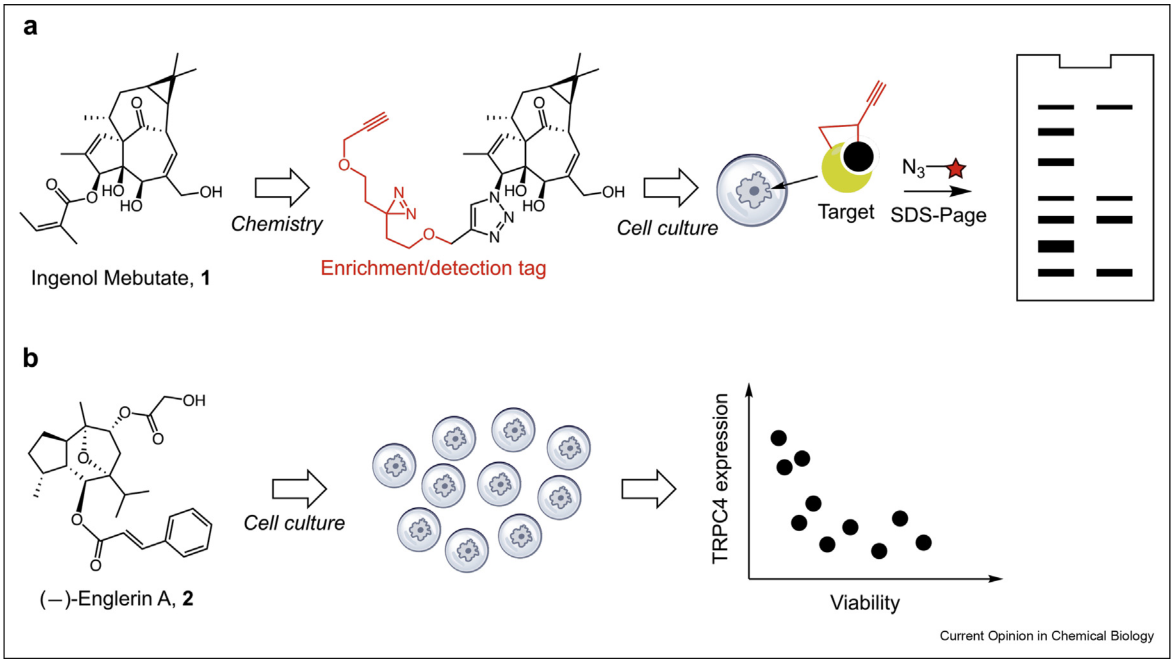
As an alternative, screening of the native ligand against a battery of genetically defined cell lines may offer a solution to identify drug targets on the basis of the observed activity signatures. Indeed, this method has enabled the identification of transient receptor potential canonical channels 4 and 5 as targets for anticancer sesquiterpene (−)-englerin A (**2**) by correlating gene expression with phenotypic changes (Figure 1b) [26,27]. However, like chemical proteomics, this approach is laborious, time consuming, and expensive, which has prompted the search for viable alternatives, such as machine learning.

Machine learning for target identification

With the advent of high-throughput experimentation, a wealth of chemical and biological data has been generated [16,28,29]. Thus, it became impossible for researchers to efficiently analyze all available information and became reasonable to assume that computer algorithms could be employed to sieve through this data to identify latent patterns, which expert human researchers may struggle to identify [30,31]. Indeed, machine learning has recently seen a range of applications in biology, medicine [32—38], chemistry [39—41], and materials science [42—44], which suggests that it can be used to speed up development pipelines and augment human perception. For example, deep-learning algorithms have been devised to predict retrosynthetic pathways to molecules of interest [45] and design new chemical entities that can be scrutinized as hits/drug leads [46—48]. Also, different learning heuristics that leverage chemical structure data have been implemented to identify drug—target interactions that can be exploited in preclinical studies [49,50] and to design experiments [51,52].

The self-organizing maps (SOM)-based prediction of drug equivalence relationships (SPiDER) software uses a neural network-inspired algorithm to discretize the input feature vector onto a so-called feature map in an unsupervised fashion [6]. In practice, this means that drug—target relationships are inferred on the basis of descriptor similarity to reference ligands in the same neuron without explicitly considering the target identity of those reference ligands for the purpose of training or heuristics' validation studies. The method employs a set

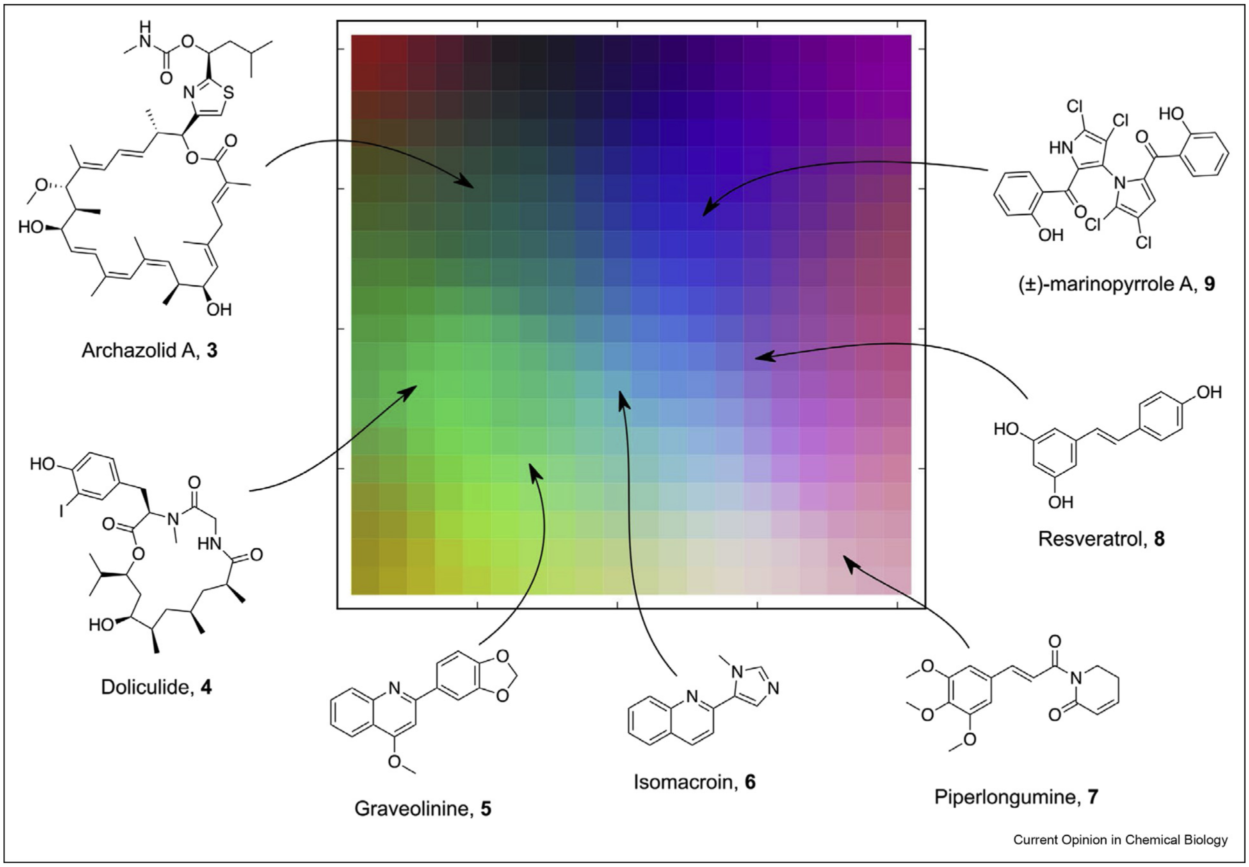
Figure 1



Schematics of two different methods for target identification. (**a**) Identification of targets for ingenol mebutate (**1**) through a chemical proteomics approach. (**b**) Identification of targets for (−)-englerin A (**2**) by means of a phenotypic screen approach.

of topological pharmacophore descriptors (CATS2 [53]) to categorize non-hydrogen atoms. From the autocorrelation of those topological feature pairs in a molecule, a SOM is built. A second, independent SOM is built by using physicochemical descriptors before generating a consensus prediction from both SOMs, which is supported by background statistics. Taken together, the prime goals are to employ descriptors that represent molecules in a sufficiently fuzzy manner and analyze data from disparate vantage points to allow generalization of the method to previously unseen test cases, that is, molecules of potential biological interest. The software tool has been extensively applied to *de novo* designed entities and, especially, natural products of high biological value. Prominent use cases of SPiDER include identification of farnesoid X receptor (EC_{50} = 0.2 μM), peroxisome proliferator-activated receptor gamma (EC_{50} = 8 μM), 5-lipoxygenase (EC_{50} = 11 μM), and microsomal prostaglandin E synthase-1 (EC_{50} = 8 μM) as drug targets for macrolide archazolid A (**3**, Figure 2). Importantly, given the structural difference between **3** and the small molecules used as reference structures by SPiDER, target inference was successfully achieved by exploiting synthetically motivated fragments of **3** as bioactivity blueprints [54]. This appears to be a reasonable approach as an identical strategy was followed to identify the prostanoid 3 receptor (EC_{50} = 6 nM) and the voltage-gated $Ca_v1.2$ channel (IC_{50} = 6 μM) as targets for doliculide, **4**, and **2**, respectively [55,56]. Similarly, fragment-like alkaloids graveolinine (**5**), isomacroin (**6**), and piperlongumine (**7**) could be deorphanized with SPiDER as serotonin 2B receptor (IC_{50} = 12 μM), platelet-derived growth factor receptor alpha (IC_{50} = 25 μM), and transient receptor potential channel vanilloid 2 (EC_{50} = 5 μM) modulators, respectively, which opens new avenues for molecular optimization in hit-to-lead programs [49,57]. More recently, SPiDER has been applied to prioritize *de novo* designed chemical entities fitting a predefined pharmacological profile for synthesis and experimental validation [46,58]. Given that SPiDER is built from two-dimensional descriptors, one may expect a higher rate of false positive predictions for molecules with stereogenic centers relative to achiral molecules, as this information is not taken into account. To better predict the chiral nature of molecular

Figure 2



Current Opinion in Chemical Biology

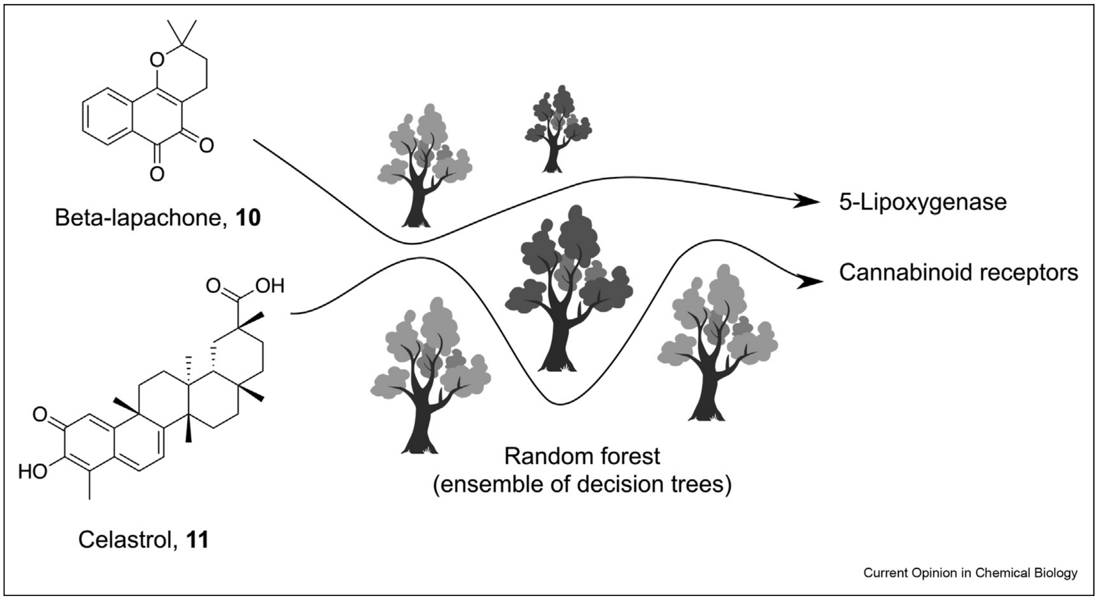
Application of self-organizing maps (SOM) for target deconvolution of phenotypic screen hits. SPiDER and TIGER use the SOM technology to tessellate chemical space and infer targets for bioactive molecules. SPiDER was used to identify targets for archazolid A, doliculide, graveolinine, isomacroin, and piperlongumine, whereas TIGER was used for resveratrol and (±)-marinopyrrole A. TIGER, target inference generator recognition, the extended three-dimensional fingerprint was recently developed and applied to synthetic molecules [59] but remains to be validated with complex natural products (see Figure 3).

The target inference generator method offers target predictions on the basis of SOMs—similar to SPiDER—but differs in the scoring function and the use of a restricted number of topological pharmacophore atom types [60,61]. For flat and neutral natural products, such as resveratrol (**8**) and (±)-marinopyrrole A (**9**), the method was able to confidently predict relationships with targets from different families, pinpointing the polypharmacology that underlies these structures [61,62]. For example, **8** was confirmed as an estrogen receptor ligand (K_i = 0.4–4 μM), whereas **9** was successfully associated to orexin (K_B = 0.3–0.6 μM) and glucocorticoid (K_B = 0.7 μM) receptors, and trypsin (IC_{50} = 3 μM).

The heuristics learn the data structure in an unsupervised way to aggregate molecules with similar feature patterns, yet without knowing if the resulting association is correct in respect to drug target associations. This is frequently assessed through cross-validation studies, hold out test data [63], and adversarial control models [64], as a means of ascertaining the relevance of the obtained models and soundness of the exploited descriptors. Regression-based methods can circumvent some of the limitations of unsupervised and classification methods by affording a predicted affinity value. This however comes at the expense of needing a larger volume of training data, which is often not in hand or is expensive to collect. The molecular ant algorithm [65–67] is a small molecule generator workflow that implements Gaussian process regression to predict affinity values for query ligands and model binding uncertainties toward a range of ChEMBL targets. Thus, the method is well suited to not only prioritize *de novo* designed entities [66] but also repurpose molecules [68] or de-risk preclinical development [69]. The application of other conceptually distinct regression algorithms is viable [50,51], and one can also expect the growing utility of deep-learning architectures in the target prediction space in the near future [70–72].

The DEcRyPT method has been recently reported [73,74] for the prediction of network pharmacology, either as a standalone tool or in combination with SPiDER, by employing random forest technology and CATS2 descriptors. In short, this method harnesses a user-defined number of predictors (decision trees) that independently analyze different portions of the training data, before aggregation of the resulting predicted outputs. When applied to beta-lapachone (**10**), the workflow confidently predicted 5-lipoxygenase as a target. Follow-up studies revealed that **10** was a reversible, allosteric modulator of 5-lipoxygenase (IC_{50} = 240 nM) and also a selective relative to other lipoxygenases and metalloenzymes. Activity of **10** could only be observed in cell-free and cell-based assays in the presence of a reducing agent. The observation suggests that beta-lapachone is reduced *in situ* to its hydroquinone form before modulation of 5-lipoxygenase. Importantly, modulation of 5-lipoxygenase by **10** was crucial for the

Figure 3



Schematics of target identification for beta-lapachone (**10**) and celastrol (**11**) by using the random forest technology implemented in DEcRyPT.

antiproliferative activity of **10** in an acute myeloid leukemia cell line [73]. More recently, a second use case unveiled modulation of cannabinoid receptors by celastrol (**11**), which are also implicated in cancer progression [74].

Outlook

Machine intelligence has recently seen an upsurge of applications in chemical sciences, in particular referring to the design of new chemical entities. The deconvolution of targets for phenotypic screen hits has been a necessary but laborious task [13], which enables the rational design and optimization of chemical matter toward drug leads. Over the years, the large volume of data collected [75] now allows scientists working at the interface of chemistry and biology and equipped with machine learning tools to identify latent patterns worthy of further research. Here, we have shown recent successes in the identification of targets, leveraged by statistical learning algorithms. Naturally, each technique has its own strengths and limitations and domain of applicability, all of which require careful consideration. For example, none of the discussed methods can identify unreported proteins as targets, as these methods are based on the principle that similar ligands (irrespective of the employed descriptor) exert similar biological effects. Similarly, the identification of DNA/RNA binders, modulators of protein—lipid interactions, among others ought to be possible provided that sufficient data are available; this is however not easily accessible, contributing to a preferential exploration of protein targets. One must also bear in mind that no new biology can be uncovered through machine learning, unlike wet laboratory experiments that are able to provide the ground truth. Even for known proteins, there is shortage of ligand data in some cases, which can limit the scope of a new method. Erroneous predictions are also common, especially for poorly studied targets or targets with noisy data available. However, negative, yet validated results offer opportunities to build better-informed machine intelligence and thus should still be reported and not discarded. Alongside healthy skepticism, machine learning for target identification entails an important set of tools to aid decision-making. By filling a gap within the chemical biologists toolbox, we expect machine intelligence to speed up some tasks in drug discovery toward the development of life-changing therapeutics.

Conflict of interest statement

Nothing declared.

Acknowledgements

T.R. is an *Investigador Auxiliar* supported by FCT Portugal (CEECIND/00887/2017). T.R. acknowledges the H2020 (TWINN-2017 ACORN, Grant 807281) and FCT / FEDER (02/SAICT/2017, Grant 28333) for funding. G.J.L.B. is a Royal Society University Research Fellow (URF\R\180019) and a FCT Investigator (IF/00624/2015). The authors thank Dr Vikki Cantrill for her help with the editing of this manuscript.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
* * of outstanding interest

1. van Waterschoot RAB, Parrott NJ, Olivares-Morales A, Lave T, Rowland M, Smith DA: **Impact of target interactions on small-molecule drug disposition: an overlooked area**. *Nat Rev Drug Discov* 2018, **17**:299.

2. Mullard A: **2018 FDA drug approvals**. *Nat Rev Drug Discov* 2019, **18**:85–89.

3. Reker D, Bernardes GJL, Rodrigues T: **Computational advances in combating colloidal aggregation in drug discovery**. *Nat Chem* 2019, **11**:402–418.

4. Ganesh AN, Donders EN, Shoichet BK, Shoichet MS: **Colloidal aggregation: from screening nuisance to formulation nuance**. *Nano Today* 2018, **19**:188–200.

5. Anselmo AC, Gokarn Y, Mitragotri S: **Non-invasive delivery strategies for biologics**. *Nat Rev Drug Discov* 2019, **18**:19–40.

6. * * Reker D, Rodrigues T, Schneider P, Schneider G: **Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus**. *Proc Natl Acad Sci USA* 2014, **111**:4067–4072.
Description of the method most widely used for the identification of drug targets for natural products.

7. * * Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, *et al.*: **Predicting new molecular targets for known drugs**. *Nature* 2009, **462**: 175–181.
Seminal work showing that in silico tools correctly predict pharmacology for phenotypic screen hits and drugs.

8. Hopkins AL: **Network pharmacology: the next paradigm in drug discovery**. *Nat Chem Biol* 2008, **4**:682–690.

9. Hopkins AL: **Network pharmacology**. *Nat Biotechnol* 2007, **25**: 1110–1111.

10. Knight ZA, Lin H, Shokat KM: **Targeting the cancer kinome through polypharmacology**. *Nat Rev Cancer* 2010, **10**: 130–137.

11. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, *et al.*: **Large-scale prediction and testing of drug activity on side-effect targets**. *Nature* 2012, **486**:361–367.

12. Bowes J, Brown AJ, Hamon J, Jarolimek W, Sridhar A, Waldron G, Whitebread S: **Reducing safety-related drug attrition: the use of in vitro pharmacological profiling**. *Nat Rev Drug Discov* 2012, **11**:909–922.

13. Schenone M, Dancik V, Wagner BK, Clemons PA: **Target identification and mechanism of action in chemical biology and drug discovery**. *Nat Chem Biol* 2013, **9**:232–240.

14. Laraia L, Robke L, Waldmann H: **Bioactive compound collections: from design to target identification**. *Chem* 2018, **4**: 705–730.

15. Laraia L, Waldmann H: **Natural product inspired compound collections: evolutionary principle, chemical synthesis, phenotypic screening, and target identification**. *Drug Discov Today Technol* 2017, **23**:75–82.

16. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, *et al.*: **Applications of machine learning in drug discovery and development**. *Nat Rev Drug Discov* 2019, **18**:463–477.

17. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A: **Machine learning for molecular and materials science**. *Nature* 2018, **559**:547–555.

18. Parker CG, Galmozzi A, Wang Y, Correia BE, Sasaki K, Joslyn CM, Kim AS, Cavallaro CL, Lawrence RM, Johnson SR, *et al.*: **Ligand and target discovery by fragment-based screening in human cells**. *Cell* 2017, **168**:527–541 e529.

19. Bar-Peled L, Kemper EK, Suciu RM, Vinogradova EV, Backus KM, Horning BD, Paul TA, Ichu TA, Svensson RU, Olucha J, *et al.*: **Chemical proteomics identifies druggable vulnerabilities in a genetically defined cancer**. *Cell* 2017, **171**: 696–709 e623.

20. Moellering RE, Cravatt BF: **How chemoproteomics can enable drug discovery and development**. *Chem Biol* 2012, **19**:11–22.

21. Matthews ML, He L, Horning BD, Olson EJ, Correia BE, Yates 3rd JR, Dawson PE, Cravatt BF: **Chemoproteomic profiling and discovery of protein electrophiles in human cells**. *Nat Chem* 2017, **9**:234–243.

22. Drewes G, Knapp S: **Chemoproteomics and chemical probes for target discovery**. *Trends Biotechnol* 2018, **36**:1275–1286.

23. * Parker CG, Kuttruff CA, Galmozzi A, Jorgensen L, Yeh CH, Hermanson DJ, Wang Y, Artola M, McKerrall SJ, Josyln CM, *et al.*: **Chemical proteomics identifies SLC25A20 as a functional target of the ingenol class of actinic keratosis drugs**. *ACS Cent Sci* 2017, **3**:1276–1285.

Work describing the use of chemical proteomics to discover a membrane protein as target for a clinical stage drug.

24. Holly SP, Chang JW, Li W, Niessen S, Phillips RM, Piatt R, Black JL, Smith MC, Boulaftali Y, Weyrich AS, *et al.*: **Chemoproteomic discovery of AADACL1 as a regulator of human platelet activation**. *Chem Biol* 2013, **20**:1125–1134.

25. van Esbroeck ACM, Janssen APA, Cognetta 3rd AB, Ogasawara D, Shpak G, van der Kroeg M, Kantae V, Baggelaar MP, de Vrij FMS, Deng H, *et al.*: **Activity-based protein profiling reveals off-target proteins of the FAAH inhibitor BIA 10-2474**. *Science* 2017, **356**:1084–1087.

26. Akbulut Y, Gaunt HJ, Muraki K, Ludlow MJ, Amer MS, Bruns A, Vasudev NS, Radtke L, Willot M, Hahn S, *et al.*: **(-)-Englerin A is a potent and selective activator of TRPC4 and TRPC5 calcium channels**. *Angew Chem Int Ed* 2015, **54**:3787–3791.

27. Carson C, Raman P, Tullai J, Xu L, Henault M, Thomas E, Yeola S, Lao J, McPate M, Verkuyl JM, *et al.*: **Englerin a agonizes the TRPC4/C5 cation channels to inhibit tumor cell line proliferation**. *PLoS One* 2015, **10**, e0127498.

28. Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, Green DV, Hertzberg RP, Janzen WP, Paslay JW, *et al.*: **Impact of high-throughput screening in biomedical research**. *Nat Rev Drug Discov* 2011, **10**:188–195.

29. Häse F, Roch LM, Aspuru-Guzik A: **Next-generation experimentation with self-driving laboratories**. *Trends Chem* 2019, **1**: 282–291.

30. Duros V, Grizou J, Xuan W, Hosni Z, Long DL, Miras HN, Cronin L: **Human versus robots in the discovery and crystallization of gigantic polyoxometalates**. *Angew Chem Int Ed* 2017, **56**:10815–10820.

31. Reker D, Bernardes GJL, Rodrigues T: **Evolving and nano data enabled machine intelligence for chemical reaction optimization**. *ChemRxiv* 2018, https://doi.org/10.26434/chemrxiv.7291205.v7291201.

32. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM: **Machine learning for integrating data in biology and medicine: principles, practice, and opportunities**. *Inf Fusion* 2019, **50**:71–91.

33. Wainberg M, Merico D, Delong A, Frey BJ: **Deep learning in biomedicine**. *Nat Biotechnol* 2018, **36**:829–838.

34. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado G, *et al.*: **End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography**. *Nat Med* 2019, **25**: 954–961.

35. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, Glorot X, O'Donoghue B, Visentin D, *et al.*: **Clinically applicable deep learning for diagnosis and referral in retinal disease**. *Nat Med* 2018, **24**: 1342–1350.

36. Li L, Koh CC, Reker D, Brown JB, Wang H, Lee NK, Liow H-h, Dai H, Fan H-M, Chen L, *et al.*: **Predicting protein-ligand interactions based on bow-pharmacological space and Bayesian additive regression trees**. *Sci Rep* 2019, **9**:7703.

37. Cui J, Hollmen M, Li L, Chen Y, Proulx ST, Reker D, Schneider G, Detmar M: **New use of an old drug: inhibition of breast cancer stem cells by benztropine mesylate**. *Oncotarget* 2017, **8**: 1007–1022.

38. Reker D, Seet M, Pillong M, Koch CP, Schneider P, Witschel MC, Rottmann M, Freymond C, Brun R, Schweizer B, *et al.*: **Deorphaning pyrrolopyrazines as potent multi-target antimalarial agents**. *Angew Chem Int Ed* 2014, **53**:7079–7084.

39. Gromski PS, Henson AB, Granda JM, Cronin L: **How to explore chemical space using algorithms and automation**. *Nat Rev Chem* 2019, **3**:119–128.

40. de Almeida AF, Moreira R, Rodrigues T: **Synthetic organic chemistry driven by artificial intelligence**. *Nat Rev Chem* 2019, **3**, https://doi.org/10.1038/s41570-41019-40124-41570.

41. Coley CW, Thomas 3rd DA, Lummiss JAM, Jaworski JN, Breen CP, Schultz V, Hart T, Fishman JS, Rogers L, Gao H, *et al.*: **A robotic platform for flow synthesis of organic compounds informed by AI planning**. *Science* 2019, **365**.

42. Gomez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel TD, Duvenaud D, Maclaurin D, Blood-Forsythe MA, Chae HS, Einzinger M, Ha DG, Wu T, *et al.*: **Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach**. *Nat Mater* 2016, **15**: 1120–1127.

43. Jensen Z, Kim E, Kwon S, Gani TZH, Roman-Leshkov Y, Moliner M, Corma A, Olivetti E: **A machine learning approach to zeolite synthesis enabled by automatic literature data extraction**. *ACS Cent Sci* 2019, **5**:892–899.

44. Daeyaert F, Ye F, Deem MW: **Machine-learning approach to the design of OSDAs for zeolite beta**. *Proc Natl Acad Sci USA* 2019, **116**:3413–3418.

45. Segler MHS, Preuss M, Waller MP: **Planning chemical syntheses with deep neural networks and symbolic AI**. *Nature* 2018, **555**:604–610.

46. * Button A, Merck D, Hiss JA, Schneider G: **Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis**. *Nat Mach Intell* 2019, **1**: 307–315.

Application of machine learning to prioritize computationally designed small molecules for synthesis and experimental validation.

47. Gomez-Bombarelli R, Wei JN, Duvenaud D, Hernandez-Lobato JM, Sanchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A: **Automatic chemical design using a data-driven continuous representation of molecules**. *ACS Cent Sci* 2018, **4**: 268–276.

48. * Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, *et al.*: **Deep learning enables rapid identification of potent DDR1 kinase inhibitors**. *Nat Biotechnol* 2019, **37**:1038–1040.

Application of machine learning to design small molecules for a specific kinase target.

49. * Baker C, Rodrigues T, Pumroy RA, Conde J, Picard D, Marques MC, de Almeida BP, Samanta A, Sieglitz F, Langini M, *et al.*: **Allosteric antagonist modulation of TRPV2 by piperlongumine impairs glioblastoma progression**. *SSRN Electr J* 2019, https://doi.org/10.2139/ssrn.3402071.

Application of machine learning to deconvolute the polypharmacology of a natural product, with subsequent validation in terms of biochemical assays, structural biology, formulation and in vivo testing.

50. Rodrigues T, Reker D, Welin M, Caldera M, Brunner C, Gabernet G, Schneider P, Walse B, Schneider G: **De novo fragment design for drug discovery and chemical biology**. *Angew Chem Int Ed* 2015, **54**:15079–15083.

51. ** Reker D, Schneider P, Schneider G: **Multi-objective active machine learning rapidly improves structure-activity models and reveals new protein-protein interaction inhibitors**. *Chem Sci* 2016, **7**:3919–3927.

Seminal work on the use of active learning for hit/lead discovery, prediction of affinities for targets of interest and rational design of experiments.

52. Reker D, Schneider G: **Active-learning strategies in computer-assisted drug discovery**. *Drug Discov. Today* 2015, **20**: 458–465.

53. Reutlinger M, Koch CP, Reker D, Todoroff N, Schneider P, Rodrigues T, Schneider G: **Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for 'orphan' molecules**. *Mol Inf* 2013, **32**:133–138.

54. ** Reker D, Perna AM, Rodrigues T, Schneider P, Reutlinger M, Monch B, Koeberle A, Lamers C, Gabler M, Steinmetz H, *et al.*: **Revealing the macromolecular targets of complex natural products**. *Nat Chem* 2014, **6**:1072–1078.
Application of self-organizing maps to discover the targets of a complex macrocyclic natural product and correlate those activities with the previously known anti-proliferative effects.

55. Schneider G, Reker D, Chen T, Hauenstein K, Schneider P, Altmann KH: **Deorphaning the macromolecular targets of the natural anticancer compound doliculide**. *Angew Chem Int Ed* 2016, **55**:12408–12411.

56. Rodrigues T, Sieglitz F, Somovilla VJ, Cal PM, Galione A, Corzana F, Bernardes GJL: **Unveiling (-)-Englerin a as a modulator of L-type calcium channels**. *Angew Chem Int Ed* 2016, **55**:11077–11081.

57. Rodrigues T, Reker D, Kunze J, Schneider P, Schneider G: **Revealing the macromolecular targets of fragment-like natural products**. *Angew Chem Int Ed* 2015, **54**:10516–10520.

58. Bruns D, Gawehn E, Kumar KS, Schneider P, Baumgartner M, Schneider G: **Identification of synthetic activators of cancer cell migration by hybrid deep learning**. *Chembiochem* 2019, https://doi.org/10.1002/cbic.201900346.

59. Axen SD, Huang XP, Caceres EL, Gendelev L, Roth BL, Keiser MJ: **A simple representation of three-dimensional molecular structure**. *J Med Chem* 2017, **60**:7393–7409.

60. Schneider P, Schneider G: **Polypharmacological drug-target inference for chemogenomics**. *Mol Inf* 2018, **37**, e1800050.

61. Schneider P, Schneider G: **De-orphaning the marine natural product (+/-)-marinopyrrole A by computational target prediction and biochemical validation**. *Chem Commun* 2017, **53**: 2272–2274.

62. Schneider P, Schneider G: **A computational method for unveiling the target promiscuity of pharmacologically active compounds**. *Angew Chem Int Ed* 2017, **56**:11520–11524.

63. * Mathai N, Chen Y, Kirchmair J: **Validation strategies for target prediction methods**. *Briefings Bioinf* 2019, https://doi.org/10.1093/bib/bbz026.
Work describing in silico validation strategies applicable to machine learning.

64. * Chuang KV, Keiser MJ: **Adversarial controls for scientific machine learning**. *ACS Chem Biol* 2018, **13**:2819–2821.
Work describing in silico validation strategies applicable to machine learning.

65. Hiss JA, Reutlinger M, Koch CP, Perna AM, Schneider P, Rodrigues T, Haller S, Folkers G, Weber L, Baleeiro RB, *et al.*: **Combinatorial chemistry by ant colony optimization**. *Future Med Chem* 2014, **6**:267–280.

66. Reutlinger M, Rodrigues T, Schneider P, Schneider G: **Multi-objective molecular de novo design by adaptive fragment prioritization**. *Angew Chem Int Ed* 2014, **53**:4244–4248.

67. Reutlinger M, Rodrigues T, Schneider P, Schneider G: **Combining on-chip synthesis of a focused combinatorial library with computational target prediction reveals imidazopyridine GPCR ligands**. *Angew Chem Int Ed* 2014, **53**:582–585.

68. Rodrigues T, Lin YC, Hartenfeller M, Renner S, Lim YF, Schneider G: **Repurposing de novo designed entities reveals phosphodiesterase 3B and cathepsin L modulators**. *Chem Commun* 2015, **51**:7478–7481.

69. Rodrigues T, Hauser N, Reker D, Reutlinger M, Wunderlin T, Hamon J, Koch G, Schneider G: **Multidimensional de novo design reveals 5-HT2B receptor-selective ligands**. *Angew Chem Int Ed* 2015, **54**:1551–1555.

70. Donner Y, Kazmierczak S, Fortney K: **Drug repurposing using deep embeddings of gene expression profiles**. *Mol Pharm* 2018, **15**:4314–4325.

71. Cortes-Ciriano I, Bender A: **KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images**. *J Cheminf* 2019, **11**:41.

72. Cortes-Ciriano I, Bender A: **Reliable prediction errors for deep neural networks using test-time dropout**. *J Chem Inf Model* 2019, **59**:3330–3339.

73. * Rodrigues T, Werner M, Roth J, da Cruz EHG, Marques MC, Akkapeddi P, Lobo SA, Koeberle A, Corzana F, da Silva Junior EN, *et al.*: **Machine intelligence decrypts beta-lapachone as an allosteric 5-lipoxygenase inhibitor**. *Chem Sci* 2018, **9**:6899–6903.
Application of machine learning to deconvolute the phenotypic effects of a natural product and repurposing it as a potential anti-leukemia agent.

74. Rodrigues T, de Almeida BP, Barbosa-Morais NL, Bernardes GJL: **Dissecting celastrol with machine learning to unveil dark pharmacology**. *Chem Commun* 2019, **55**: 6369–6372.

75. ** Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, *et al.*: **ChEMBL: a large-scale bioactivity database for drug discovery**. *Nucleic Acids Res* 2012, **40**:D1100–D1107.
Reference work for a database annotating ligand–receptor relationships, which can be curated and used for building bespoke machine learning methods.