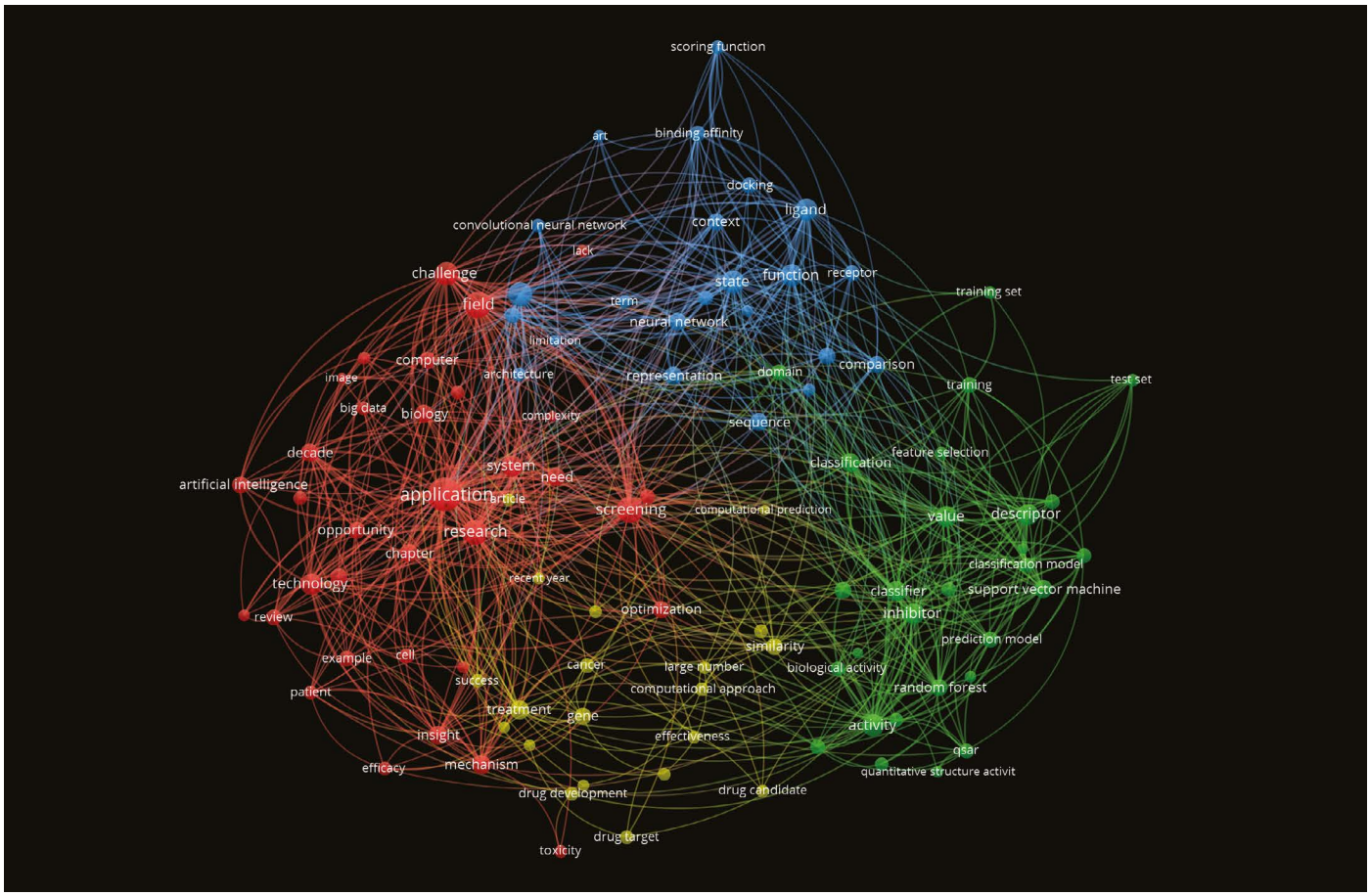


MARKUS KAINDL/VOSVIEWER

**The network connects co-occurring terms in 175 articles published by Springer Nature related to using AI in drug discovery.**

# Getting with the program

## An experiment in auto-generated article summaries.

**By Catherine Armitage and Markus Kaindl**

We present AI-generated summaries of three articles selected from a data set of 175 Springer Nature publications in 2019 and 2020. The data set was derived using the title and abstract search '('artificial intelligence' OR 'AI' OR 'machine learning' OR 'deep learning') AND ('drug discovery' OR 'drug design')' in Publication Year 2020 or 2019.

The results were then filtered according to Springer Nature's article classification model for article type 'original paper' and 'research article' and exported from Digital Science's Dimensions database on 14 October 2020. (Digital Science is a subsidiary of Holtzbrinck Publishing Group, which owns 53% of Springer Nature, publisher of the Nature Index.)

Two of the articles featured ranked highest in the data set on objective metrics for either downloads (Mendez-Lucio *et al.*) or social media attention by Altmetric score (Piazza *et al.*). Madhukar *et al.* was the second-most highly cited article in the data set. We elected to present it instead of the most highly-cited article (D. D. Nguyen *et al. J. Comput. Aided Mol. Des.* **33**, 71–82; 2019) for the sake of consistency and ease of access. Using three articles from *Nature Communications* with a consistent format allows for more direct comparisons between the auto-generated text. They are also open access.

For each we present the authors' original abstract and summaries of it made by two abstractive summarization approaches, Google's T5 and Pegasus. We then show a paper summary by the Allen Institute for Artificial Intelligence's (AI²) SCITLDR, based on the original abstract, introduction and discussion, and summaries of the paper's introduction and discussion by SciBERT using a summary ratio of 0.3. The AI-generated text on pages S139 to S141 has not been modified, so does not conform to *Nature's* usual editorial standards for grammar and style.

More detail about the summary generation process and program sources can be found at go.nature.com/3qd4kxc. We welcome your feedback to natureindex@nature.com with 'Nature Index AI' in the subject line.

## DOWNLOADS

O. Méndez-Lucio *et al*. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature Commun.* **11**, 10 (2020).

### Original abstract

Finding new molecules with a desired biological activity is an extremely difficult task. In this context, artificial intelligence and generative models have been used for molecular de novo design and compound optimization. Herein, we report a generative model that bridges systems biology and molecular design, conditioning a generative adversarial network with transcriptomic data. By doing so, we can automatically design molecules that have a high probability to induce a desired transcriptomic profile. As long as the gene expression signature of the desired state is provided, this model is able to design active-like molecules for desired targets without any previous target annotation of the training compounds. Molecules designed by this model are more similar to active compounds than the ones identified by similarity of gene expression signatures. Overall, this method represents an alternative approach to bridge chemistry and biology in the long and difficult road of drug discovery.

## AI-GENERATED ABSTRACTS

### T5 abstract

a generative model bridges systems biology and molecular design. it can design active-like molecules for desired targets without any previous annotation of the training compounds. this method represents an alternative approach to bridge chemistry and biology in the long and difficult road of drug discovery.

### Pegasus abstract

we report a generative model systems and conditioning method for de novo design of active molecules with desired transcriptomic profile as the desired state. the desired state is provided by the model that is able to design active molecules for desired targets without any previous target annotation of compounds. this method represents an alternative approach to bridge chemistry and to this long road biology in difficult drug discovery.

### AI$^2$ SCITLDR paper

A generative model that bridges systems biology and molecular design, conditioning a generative adversarial network with transcriptomic data.

### SciBERT introduction

The difficulty of the drug discovery process stems from the fact that only a small fraction of the theoretically possible 1060 drug-like molecules are therapeutically relevant. One of the most challenging tasks in this scenario is the hit identification, namely the identification of small molecules with an adequate (but usually weak) activity on a specific target that could then be used as a starting point for the chemical optimization process. Although these methods generate copious bioactivity data, they are not very efficient as the amount of resources required is disproportionally large compared to the small number of hits discovered. Techniques such as virtual screening aim to identify hits from virtual libraries containing large number of molecules, usually by similarity-based searches or by molecular docking. Other generative models such as variational autoencoders (VAE) have been used to automatically optimize molecules to improve their physicochemical and drug-likeness properties. Herein, we report a generative model that bridges systems biology and molecular design. This approach presents several advantages such as the generation of hit-like molecules without the need of previous knowledge of active compounds, biological activity data, or target annotations.

### SciBERT discussion

In conclusion, we reported a method based on conditional generative adversarial networks that proposes new molecules from a particular gene expression signature. Our method offers some advantages over current molecular generative approaches as well as an alternative way to exploit all the information contained in compound-induced gene expression data, in particular the one in L1000 database. Since no previous measurement of biological activity or target annotations are needed for the training data, this approach could potentially be applied to any target, as we have shown herein.

# index

## CITATIONS

N. S. Madhukar *et al.* A Bayesian machine learning approach for drug target identification using diverse data types. *Nature Commun.* **10**, 5221 (2019).

### Original abstract

Drug target identification is a crucial step in development, yet is also among the most complex. To address this, we develop BANDIT, a Bayesian machine-learning approach that integrates multiple data types to predict drug binding targets. Integrating public data, BANDIT benchmarked a ~90% accuracy on 2000+ small molecules. Applied to 14,000+ compounds without known targets, BANDIT generated ~4,000 previously unknown molecule-target predictions. From this set we validate 14 novel microtubule inhibitors, including 3 with activity on resistant cancer cells. We applied BANDIT to ONC201—an anti-cancer compound in clinical development whose target had remained elusive. We identified and validated DRD2 as ONC201's target, and this information is now being used for precise clinical trial design. Finally, BANDIT identifies connections between different drug classes, elucidating previously unexplained clinical observations and suggesting new drug repositioning opportunities. Overall, BANDIT represents an efficient and accurate platform to accelerate drug discovery and direct clinical application.

## AI-GENERATED ABSTRACTS

### T5 abstract

drug target identification is a crucial step in development, yet is also among the most complex. we develop BANDIT, which integrates multiple data types to predict drug binding targets. from this set, we validate 14 novel microtubule inhibitors, including 3 with activity on resistant cancer cells.

### Pegasus abstract

target identification is crucial in drug development. we developed a machine-learning approach to predict drug binding targets. we benchmarked public data to 90% accuracy on 14,000 compounds without known targets. this set was used to validate 14 novel microtubule inhibitors. we applied this to 3201 anticancer compounds in clinical development.

### AI² SCITLDR paper

We develop BANDIT, a Bayesian machine-learning approach that integrates multiple data types to predict drug binding targets.

### SciBERT introduction

It typically takes 15 years and 2.6 billion dollars to go from a small molecule in the lab to an approved drug and for natural products and phenotypic screen derived small molecules, one of the greatest bottlenecks is identifying the targets of any candidate molecules. Traditionally, ligand-based approaches take known binding targets for a given drug and attempt to find other drugs or proteins that are sufficiently similar. However, to achieve high predictive power they require a large input of known binding partners for each tested drug, and therefore can only be used on drugs which have prior comprehensive target information. However these approaches still suffer from a few limitations: 1. To overcome these limitations, we introduce BANDIT, a drug-target prediction platform. BANDIT uses a Bayesian approach to integrate a number of diverse data types in an unbiased manner and provides a platform that allows for simple integration of new datatypes as they become available. Tested on ~2000 different compounds, BANDIT achieves a high accuracy at identifying shared target interactions, uncovers novel targets for the treatment of cancer, and can be used to quickly pinpoint potential therapeutics with novel mechanisms of action to accelerate drug development.

### SciBERT discussion

We have developed BANDIT, an integrative Big-Data approach that combines a set of individually weak features into a single reliable and robust predictor of shared-target drug relationships and individual drug binding targets. Finally, BANDIT can be used on a broader scale to discern mechanisms of approved drugs, characterize the global drug universe landscape, explain existing, yet puzzling, clinical phenotypes, and repurpose drugs for new indications. However, as more information becomes available there are many aspects of the current implementation that can be improved. Furthermore, BANDIT could be used to rapidly screen a large database of compounds and efficiently identify any promising therapeutics that could be further evaluated.

## MOST SOCIAL MEDIA ATTENTION

I. Piazza *et al.* A machine learning-based chemoproteomic approach to identify drug targets and binding sites in complex proteomes. *Nature Commun.* **11**, 4200 (2020).

### Original abstract

Chemoproteomics is a key technology to characterize the mode of action of drugs, as it directly identifies the protein targets of bioactive compounds and aids in the development of optimized small-molecule compounds. Current approaches cannot identify the protein targets of a compound and also detect the interaction surfaces between ligands and protein targets without prior labeling or modification. To address this limitation, we here develop LiP-Quant, a drug target deconvolution pipeline based on limited proteolysis coupled with mass spectrometry that works across species, including in human cells. We use machine learning to discern features indicative of drug binding and integrate them into a single score to identify protein targets of small molecules and approximate their binding sites. We demonstrate drug target identification across compound classes, including drugs targeting kinases, phosphatases and membrane proteins. LiP-Quant estimates the half maximal effective concentration of compound binding sites in whole cell lysates, correctly discriminating drug binding to homologous proteins and identifying the so far unknown targets of a fungicide research compound.

## AI-GENERATED ABSTRACTS

### T5 abstract

chemoproteomics is a key technology to characterize the mode of action of drugs. it directly identifies the protein targets of bioactive compounds and aids in the development of optimized small-molecule compounds.

### Pegasus abstract

mass spectrometry is a key technology to characterize the mode of action of drugs as it identifies protein targets of bioactive compounds. we here develop mass- spectrometry based on limited proteolysis that works in human cells including machine learning to discern drug targets of small molecules and integrate them into a single score to identify their binding sites.

### AI² SCITLDR paper

We develop LiP-Quant, a drug target deconvolution pipeline based on limited proteolysis coupled with mass spectrometry to identify drug

### SciBERT introduction

Unraveling the mechanism of action and molecular target of small molecules remains a major challenge in drug development. However, these target capturing probes may perturb molecular interactions and biological functions. This is achieved through the global detection with liquid chromatography-coupled tandem mass spectrometry (LC-MS) of differential proteolytic patterns produced upon ligand binding. Here we adapt this LiP-based approach to enable the systematic investigation of protein–small-molecule interactions in complex eukaryotic proteomes (e.g. human). We benchmark LiP-Quant by investigating binding of compounds to known human protein targets, addressing the possibility of detecting kinase or phosphatase inhibitors and membrane protein drugs.

### SciBERT discussion

Devising modification-free chemoproteomic strategies that simultaneously probe whole-proteomes will increase efficiency in the drug development pipeline. Although overall the TPP-based approaches examined here are more sensitive, it should be noted that true positive identification rates for LiP-Quant and TPP are virtually identical for the first 25 targets identified. LiP-Quant identifies peptides that undergo structural changes upon compound binding, providing a proxy for drug binding sites and enabling a quantitative assessment of their binding specificity and selectivity through the estimation of EC50s. For instance, we used LiP-Quant with additional validation to discover the potential targets of a research fungicide. We have found that LiP-Quant detects interactions with drugs over a broad range of affinities from nM to μM. EC50s estimated from lysate with LiP-Quant are often higher, by approximately an order of magnitude, than literature-reported values generally measured with recombinant proteins. Further applying the LiP-Quant pipeline to cells that remain intact during drug treatment represents a more biologically relevant experiment, while ensuring proteins maintain their native compartment during compound incubation. As such approaches deviate from the standard LiP-Quant pipeline, they are in their infancy and will require further development to reduce false positives/negatives but they offer a glimpse of potential areas for further development.