Trends in Deep Learning for Property-driven Drug Design

Review Article**Author(s):**

Born, Jannis; Manica, Matteo

Publication date:

2021-11

Permanent link:

<https://doi.org/10.3929/ethz-b-000522032>

Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

Current Medicinal Chemistry 28(38), <https://doi.org/10.2174/0929867328666210729115728>

Trends in Deep Learning for Property-driven Drug Design

Jannis Born^{1,2,*} and Matteo Manica¹

¹IBM Research Europe, Zurich, Switzerland; ²Department for Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

Abstract: It is more pressing than ever to reduce the time and costs for the development of lead compounds in the pharmaceutical industry. The co-occurrence of advances in high-throughput screening and the rise of deep learning (DL) have enabled the development of large-scale multimodal predictive models for virtual drug screening. Recently, deep generative models have emerged as a powerful tool to explore the chemical space and raise hopes to expedite the drug discovery process. Following this progress in chemocentric approaches for generative chemistry, the next challenge is to build multimodal conditional generative models that leverage disparate knowledge sources when mapping biochemical properties to target structures.

ARTICLE HISTORY

Received: December 01, 2020

Revised: April 13, 2021

Accepted: May 11, 2021

DOI:

10.2174/0929867328666210729115728

CrossMark

This is an Open Access article published under CC BY 4.0
<https://creativecommons.org/licenses/by/4.0/legalcode>

Here, we call the community to bridge drug discovery more closely with systems biology when designing deep generative models. Complementing the plethora of reviews on the role of DL in chemoinformatics, we specifically focus on the interface of predictive and generative modelling for drug discovery. Through a systematic publication keyword search on PubMed and a selection of preprint servers (*arXiv*, *bioRxiv*, *chemRxiv*, and *medRxiv*), we quantify trends in the field and find that molecular graphs and VAEs have become the most widely adopted molecular representations and architectures in generative models, respectively. We discuss progress on DL for toxicity, drug-target affinity, and drug sensitivity prediction and specifically focus on conditional molecular generative models that encompass multimodal prediction models. Moreover, we outline future prospects in the field and identify challenges such as the integration of deep learning systems into experimental workflows in a closed-loop manner or the adoption of federated machine learning techniques to overcome data sharing barriers. Other challenges include, but are not limited to interpretability in generative models, more sophisticated metrics for the evaluation of molecular generative models, and, following up on that, community-accepted benchmarks for both multimodal drug property prediction and *property-driven* molecular design.

Keywords: Deep learning, drug discovery, generative models, review, QSAR modelling, Machine learning, drug design.

1. INTRODUCTION

The term Eroom's Law was coined to describe the phenomenon that the costs for research and development of new FDA-approved drugs roughly double every nine years since the 1950s [1]. Current estimates are \$2-3 billion dollars [2], and usually, a decade passes by from compound discovery to market release [1].

The chemical space is huge, estimated to contain above 10^{30} potential drug-like molecules [3]. With the large public repository PubChem containing about 60 million unique chemical structures [4], and only ~1500 FDA-approved compounds [5], the total attrition rate is above 99.99%, while only a tiny fraction of the chemical space has been explored.

Deep learning (DL), as a subclass of machine learning (ML) [6], has revolutionized fields like computer vision and natural language processing (NLP) and enjoys increasing adoption in bioinformatics [7] chemoinformatics [8], and medicinal chemistry [9]. DL is da-

*Address correspondence to this author at the IBM Research Europe, Zurich, Switzerland; Tel./Fax: +41 (0)44 724 88 97; E-mail: jab@zurich.ibm.com

ta-hungry and has proven successful in handling data in raw format, thus depicting the ideal tool to deal with the growing wealth of available biochemical activity data that advances in high-throughput screening have bestowed us.

Databases like PubChem [4], BindingDB [10], or GDSC [11] contain invaluable information that, if exploited properly, enables the development of virtual drug screening models that memorize and interpolate pharmacological properties for manifolds of the chemical space and are substantially faster to evaluate than docking methods. While this, in turn, reduces wet-lab experiments and can accelerate drug discovery, it still requires human hypotheses about novel drug candidates. However, generative models like Variational Autoencoders (VAE [12],) have shown great value in navigating the chemical space and proposing *de-novo* molecules with desired physico- or biochemical properties [13, 14]). But less work has been done on explicitly entangling predictive and generative models, specifically on techniques that can conditionally generate molecules for a wide spectrum of biological semantics such as target proteins. This deprives DL of tapping its full potential in drug discovery, and a key challenge is to find techniques that best explore the often multimodal knowledge stored in large-scale predictive models. Usually, naive rejection sampling approaches are adopted to let generative models explore the molecular distributions they learned with unsupervised pre-training from databases like ChEMBL [15]. To fulfill simple chemical properties, these approaches may suffice, but lead discovery is inherently highly multimodal and asks questions like how to *de novo* design a selective ligand for the 3C-like protease of SARS-CoV-2 [16] without any existing binding affinity data? Given the multimodality of the problem, the desired set of molecules are very unlikely to lie in continuous manifolds of the chemical space. Notably, the sheer size of the space and the non-linearity of the desired pharmacological properties render unconditional rejection sampling approaches, the feedback for the generative model is solely based on the outcome of a virtual drug screening, largely impractical. Instead, they call into play the usage of conditional molecular generators, which additionally receive some context about the desired properties or conditions in the form of vectorial representations that directly steer the generative process.

In other words, molecular generative models are large but implicit databases of chemicals. The naturally arising question on how these databases can be queried, especially with multiple biological properties, is largely unsolved.

1.1. Our Contribution

This work aims to review some of the developments interfacing deep predictive and generative models for drug discovery. The current literature body in the field is rich and expands on the role of deep learning for drug discovery [17] and various related subdisciplines thereof (*e.g.*, drug-target interaction [18], drug response prediction [19], pharmacogenomics [20], metabolomics [21], molecular design [22] and molecular simulation [23]). These reviews either exclusively focus on predictive or generative tasks or treat them as separate entities. Instead, a closer entanglement of biochemical property predictors and molecular generative models will be key toward a bright future for deep learning in drug discovery.

Herein, we attempt to fill this gap by highlighting recent progress in DL methodology focusing on two fields:

(1) Multimodal drug property prediction, which encompasses all models that A) map from a joint space of molecules and a secondary modality toward some interaction prediction task, and B) can be evaluated on all entities of that modality. One example is compound-protein interaction prediction.

(2) Biochemical conditional molecular generative models which encompass all models that generate *de novo* molecules based on biochemical information, *e.g.*, about a disease or a protein target.

Overall, we argue that DL might act as a bridge to link the fields of systems biology and drug discovery more closely.

Given the multitude of reviews on related topics, we herein refrain from re-introducing basic concepts and terminology of DL and instead refer the interested reader to the related reviews on the abovementioned subdisciplines.

The relevance of this topic and the rise of ML on predictive and generative modeling in chemoinformatics is illustrated in Fig. (1). The results of a publication keyword search demonstrate an almost 6-fold increase in predictive molecular modeling with ML from 2015 to 2020 and an almost 7- and more than 8-fold increase for generative molecular models and models combining predictive and generative modeling, respectively.

The rest of this manuscript is structured as follows. In section 2 we review recent progress in deep learning on biochemical property prediction and report statistics about the most commonly described tasks as well as molecular representations, based on our publication keyword search. Section 3 discusses common flavors

of deep generative models, reports how much they are adopted for molecular generation tasks, and focuses especially on (biochemical) property-driven molecular generation. We close with section 4, an outlook to the future identifying challenges for the field. The publication keyword search was performed by querying the

APIs of *PubMed*, *arXiv*, *bioRxiv*, *chemRxiv*, and *medRxiv* using synonyms for each keyword. A paper was considered a match when the title or abstract contained at least one of the synonyms for each keyword (see Table S1). The reference date for all calculations was 02.11.2020.

Number of papers in machine learning for molecular modeling

Fig. (1). Venn diagrams of publications in predictive and generative modeling. Keyword searches from PubMed, arXiv, bioRxiv, chemRxiv and medRxiv with the Python package paperscraper. Keywords were “Machine learning” AND “Molecule” AND X where X denotes the term in the plot. Review papers were manually excluded from the intersection (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Fig. (2). Shows a workflow for the development of property-driven generative models and how they can be integrated into the drug discovery and data acquisition process (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

2. DRUG RESPONSE PREDICTION

2.1. Landmarks

DREAM challenges are a notable set of initiatives that enabled remarkable progress in drug response prediction (DRP) tasks, including drug sensitivity [17], drug combination [18], and multi-target and kinase-binding prediction (overview in [19]). But experimental validation of DRP models is rare. One example is the ongoing Open Drug Discovery challenge to find novel antimalarials [206]. By means of predictive ML models, an active ligand was identified that was dismissed by human experts. In another study searching for antibiotics, DL was utilized in a drug repurposing setting to identify Halicin, a compound designed to treat Diabetes, as a broad-spectrum antibiotic [24]. Instead, VirtualFlow is an open-source platform for large-scale virtual screening that was used to evaluate billions of molecules and could identify potent KEAP1 inhibitors [25].

2.1.1. Tasks

We focus on three of the most common Quantitative structure-activity relationship (QSAR) tasks relevant to drug discovery: prediction of 1) toxicity, 2) protein-ligand binding affinity (also called compound-protein interaction, CPI) and 3) drug sensitivity. The statistics on the number of yearly publications for each of those tasks in (Fig. 3 show 1) for all tasks at least a trebling since 2015 and 2) that toxicity is by far the most widely examined task. These tasks share some inherent multimodality; toxicity is defined with respect to an endpoint and a subtype (chronic or acute), binding affinity is a bimodal property of a protein-ligand pair and drug sensitivity describes tumor growth inhibition against a specific cell line.

2.1.2. Problem Formulation and Splitting Strategy

Traditional approaches refrained from explicitly representing both modalities, treating the problem as a multi-label classification/regression for a fixed set of entities from one modality (usually drugs), or thus inherently lacking the ability to predict properties for novel entities of that modality (e.g., [26-28]). Instead, most modern approaches use bimodal deep neural networks and leverage different featurization strategies for molecules, proteins, or tumours. The bimodal nature of the problem formulation gives rise to four complementary paradigms in evaluating model generalization that is essentially determined by the splitting procedure for training and testing data (see Fig. 4 for visualization).

Fig. (3). Number of publications over time on different QSAR tasks. Each task was queried together with the term deep learning and synonyms were used for each keyword. Next to the stable growth of publications in publications on all three compared tasks, a clear trend toward the utilization of preprint servers can be observed (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Data splitting for interaction prediction tasks

Fig. (4). Data splitting strategies for bimodal interaction prediction tasks. Exemplified at the task of compound-protein interaction prediction, the four possible splitting strategies for training and evaluating a bimodal neural network are shown (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

In a naïve splitting strategy, paired samples consisting of two modalities (e.g., a drug and a protein) are build first and thereafter randomly split into train/test

data. This is not only the most lenient, but also the most commonly adopted strategy and it imposes a high risk that the model can perform ostensibly well by merely memorizing training samples (as only the *pair* is unseen, but either the molecule or the protein has been seen before). Instead, in a drug discovery regime, it can be desirable to generalize to new compounds (cf. Fig. 4 top right). Instead, in a drug repurposing (or precision medicine) regime, it is desirable to generalize to new tumor samples (*e.g.*, cell lines). It is, therefore, crucial to carefully consider the planned application when splitting the data. While these regimes are substantially more complex, we recommend that strict splitting (*i.e.*, stratifying by *both* modalities simultaneously) should be adopted for research purposes since 1) it guarantees that generic features are being picked and 2) the true, underlying challenge in drug interaction models is to build multimodal predictive models that can generalize to novel drugs screened against unseen proteins or tumor samples [29]. For further discussion on data splitting techniques, [30].

2.2. Toxicity Prediction

Deep learning has become the *de-facto* standard for toxicity prediction, partly due to the success of the Tox21 Data Challenge, which is the greatest effort to compare computational methods for toxicity data to date, conducted in 2016 [31]. It was based on the Tox21 dataset [32] that currently consists of almost 13k compounds evaluated on a panel of 12 environmental toxicity assays (5 stress response receptors and 7 nuclear receptors). The challenge winner, DeepTox [33], achieved an average ROC-AUC of 0.846, utilized a deep feed-forward neural network comprising several thousand features derived from chemical descriptors, and found that deep learning outperforms traditional methods like SVMs or RF. Their initially used rich set of chemical descriptors can be approximated reasonably well solely based on ECFP4 [34]. Other key insights were the benefit of multitask learning (*i.e.*, sharing weights across different toxicity assays) compared to single-task learning [35, 36], and ensembling (which was later confirmed [37]). The evidence for the value of deep learning in single task settings with small datasets was found less clear [38]. Tox21 is a commonly used dataset to benchmark novel QSAR methods **1** and part of the benchmark MoleculeNet [39]. Going beyond molecular fingerprints, raw inputs such as the molecular inline notation SMILES [40] can improve toxicity prediction [41] (more on SMILES in Section 2.2.), also when coupled with other representations [42].

Traditionally, deep learning models were deemed black boxes, rendering the interpretability of their predictions and the importance of features difficult [43].

One realm with significant progress addresses the general problem of uncertainty modelling in neural networks [44]. To assess confidence of toxicity predictions, snapshot ensembling [45] and conformal prediction [46] were used, but later, Monte Carlo Dropout was shown to achieve the same at a lower computational and implementational cost [47]. Others studied posthoc interpretability methods like Deep Taylor decomposition [48] to identify potential toxicophores [49]. Commonly used datasets include ToxCast [50], ToxBank [51], SuperToxic [52], Toxnet [53] and ClinTox (drug clinical trial failure) and SIDER (adverse drug reactions) from DeepChem [39]. For further reading, [54, 55].

2.3. Protein-ligand Binding Affinity

Understanding protein-ligand binding affinity (or CPI) is key to accelerate drug discovery as over 80% of all FDA-approved drugs exhibit their mechanism of action by acting on specific proteins [56]. CPI can be simulated from the 3D structure in great biochemical detail using molecular docking software such as AutoDock Vina [57]. However, some of the scoring functions available in these programs (*e.g.*, Gibbs free energy) can be outperformed using ML models [58]. Conventionally, DL was used to predict CPI on a per-target basis using single-task models [59]. Data can be obtained from ChEMBL, a source of >15 million bioactivity data points for almost 2 million compound and interaction data of ~8000 protein targets [15].

Integrated (bimodal) deep learning models have become the gold standard in CPI and fall into the area of proteochemometric models [60], *i.e.*, they combine proteomic and ligand information. In 2016, the first attempt used 612k positive interactions (259k compounds and 51k proteins) from STITCH [61], generated negative pairs randomly and utilized binary fingerprints for both proteins and ligands, and used a simple deep feed-forward neural network for a binary classification task [62]. Later research can be separated into two realms; either restricting to sequence-based approaches that usually apply convolutional or recurrent neural networks (CNN/RNN) on more interpretable inputs such as protein primary structure or SMILES sequences [63-67] or structure-based approaches that exploit topological information by either applying 3D CNNs to the binding site [68-71] or by applying CNNs or graph neural networks (GNN) on molecular graphs or protein secondary or tertiary structure [72-76]. Some of these works sought to enhance model interpretability by using neural attention [77], a mechanism to weigh the contributions of individual atoms or residues for the final prediction [64, 72, 73]. For example, the DeepAffinity authors [64] devised a joint attention

mechanism between compound atoms and protein secondary structure elements (SSE). They found that, without explicit supervision on binding site location, the attention scores in binding sites were significantly higher than outside binding sites for a hand-picked set of 3 CPI pairs. The model performed well on a lenient split in predicting pIC_{50} (Pearson's r of 0.86) but failed to generalize to unseen classes of proteins such as tyrosine kinases ($r=0.42$) or ion-channels ($r=0.18$). Instead, MONN [78] predicted not only binding affinity, but also pairwise non-covalent interaction matrix between atoms of compounds and residues of proteins. Their results unravel that the localization of binding sites requires explicit supervision and cannot be learned in a self-supervised manner by analyzing the attention scores on sequential or secondary structures as suggested by case studies in [64, 72, 73]. While it remains unclear how the $\text{MONN}_{\text{single}}$ (affinity-only) model performs in the self-supervised binding site localization setting, it is interesting to note that MONN outperforms structure-based CPI models [68] on the DUD-E benchmark dataset [79]. In the recent Transformer-CPI [80], superior performance to prior work [73, 76] could be achieved using the Transformer architecture [81].

Oftentimes, CPI models remain difficult to compare due to different underlying databases or splitting procedures. MONN created a benchmark dataset of 10k pairs of pairwise non-covalent compound residue interactions. None of the above publications experimentally validated the CPI predictions of their model. This is especially relevant in light of the *hidden ligand bias* [82], *i.e.*, the observation that binding affinity predictions are mostly based on ligand rather than interaction features [80]. The Transformer-CPI authors suggested a label reversal technique to split positive and negative ligand samples between the training and validation set. Promisingly, good generalization to unseen proteins (86% accuracy) was reported [76]. However performance in datasets with only unseen protein and ligands are still to be benchmarked with deep learning models.

The related task of classifying types of ligand-binding sites (*e.g.* heme- vs. nucleotide binding sites) has also been tackled with DL [83]. Commonly used databases include BindingDB [10], PubChem [84], DrugBank [85], STITCH [61], PDBBind [86] or, for docking, DUD [87], DUD-E [79], MUV [88] and DeepDocking [89]. For further reading see [18, 90, 91].

2.4. Drug Response Prediction

Toxicity is not defined on the level of chemical structure but obviously arises from the interaction of the drug in the cellular environment. Due to cancer be-

ing the second leading cause of death worldwide [92] and the high variability in cancer drug response caused by heterogenous genomic and transcriptomic makeup of tumors [93], cancer drug sensitivity prediction is eventually the best-studied subtask of toxicity prediction. Typical deep learning models are bimodal, *i.e.*, they ingest a compound structure alongside a tumor representation (usually bulk single- or multi-omics from cancer cell lines) and predict IC_{50} (half maximal inhibitory concentration) either in regression or a binary classification task.

After many publications on conventional ML methods (for reviews, see [94, 95]), the first DL attempt utilized a fully-connected NN that received a concatenation of physiochemical descriptors and genomic features (sequence and copy number variation data), was trained on the GDSC database [11] and achieved an explained variance of $R^2=72\%$ [96]. Much of the later work conceptually carried over that approach and used Morgan fingerprints and genomic [97] or transcriptomic [98, 99] features. Later, it was demonstrated that better performance ($R^2=86\%$) could be achieved using RMA gene expression, interpretable SMILES sequences, and combinations of 1D convolutions and neural attention [29]. That approach, dubbed PaccMann, also leverages explainability techniques that allow to visualize highly attended genes and compound substructures. The model learns to focus on genes that significantly enrich apoptotic processes and drug substructures that correlate with chemical structure similarity indices in a self-supervised way. Their model is the first to show the benefit of SMILES augmentation [100] for drug-interaction tasks and is also freely available *via* a web service [101]. A recent, historical comparison found PaccMann to perform best and then proposed QSMART, a multi-omics model specifically tailored for protein kinase inhibitors (PKI) that reports higher performance than PaccMann for most cancer sites, but, notably QSMART 1) was only evaluated on PKIs, 2) trained one model per cancer site and is thus not pan-cancer and 3) did not report significances [102]. A quantum leap in biologically inspired drug response prediction has been achieved with DrugCell [103]. While their drug encoding is simplistic (Morgan fingerprint), the key novelty is the incorporation of Gene Ontology [104] to induce structurally fixed connectivity constrained by the hierarchical organization of molecular subsystems (such as signaling pathways, organelles, and cellular processes) based on their previous work bridging systems biology and deep learning [105, 106]. Their ontology-defined network performs similarly to an unconstrained one, suggesting that biologically inspired architectures are particularly useful

in data-limited settings [103]. Unfortunately, their work lacks quantitative comparison to prior research.

A key challenge in drug response prediction is the transfer from preclinical data (cell lines or patient-derived xenografts, PDX) to clinical data (human tumour cells). Specific models exist for PDX data [28]. Moreover, a systematic approach for transfer learning was suggested with AITL [107], a technique for adversarial domain adaptation in both input (-omic) and output space (drug response). In a clinical study, drug-specific models were pre-trained on GDSC and evaluated on small patient cohorts [108]. While DNNs outperformed other ML approaches, performance was significantly above chance in less than 50% of the drugs, indicating that much further research is needed before DL models can be employed in clinical settings.

2.4.1. Feature Selection

Due to the “large p small n ” paradigm and the availability of in omic data analysis, feature selection strategies for drug sensitivity prediction have received growing attention [109, 110]. Most approaches in drug sensitivity prediction are single-omic. However, most comparison studies found transcriptomic data, especially gene expression (bulk RMA or RNASeq), to be superior to genomic features [29, 111, 112]. Generally, multi-omic approaches seem superior to single-omic [28, 102]. While some works consider measurements at whole-genome scale (>20,000 features) [97, 98], typical feature reduction techniques include the COXEN method [113, 114] autoencoders on single [28, 98] or multi-omic data [26] or network propagation [29, 115] that leverages proteomic information from PPI networks (*e.g.* STRING [116]) which was found beneficial for model performance [115, 117]. The proteome of 375 cell lines in the CCLE database [118] was quantified in 2020 [119], which now enables a systematic investigation on the suspected superiority of proteomic data for drug response prediction. Last, incorporating drug perturbation effects can improve drug response prediction [120].

Last, the related task of predicting interactions of two drugs [121] or synergistic effects of anticancer drug combinations has been addressed with deep networks [122-124]. One model is available *via* a web service [122], and others investigated polypharmacy side effects with GCNs [125]. Common drug sensitivity databases include GDSC [11], CCLE [118], NCI-60 [126], PharmacoDB [127] and [128] for PDX data and the NCI Almanac [129], AstraZeneca/DREAM [130] and Merck [131] for drug synergy prediction. For further reading, [19, 132], and for deep learning on

metabolomics and pharmacogenomics, see [21] and [20], respectively.

2.5. Molecular Representations

The ultimate success of QSAR modeling - measured by its relevance for developing novel therapeutics - critically depends on the underlying molecular representation. Throughout the history of computational chemistry, featurization of molecules relied on hand-engineered chemical descriptors [133]. The notion of molecular representations is currently being transformed by the ability of DL methods to obtain them efficiently from data. To give a perspective, (Fig. 5) shows the prevalence of molecular representations as measured by a number of yearly publications in the context of deep learning: molecular fingerprints, SMILES sequences [134], and molecular graphs. Until 2018, the lion's share of publications utilized fingerprints, but the role of SMILES (a text-based inline notation, *e.g.*, c1ccccc1 denotes Benzene) has greatly increased. More importantly, the usage of graphs, processed usually *via* GNNs soared rapidly and surpassed fingerprints in 2020 in the number of publication keyword matches.

Fingerprints are typically 1D binary chemical descriptors where each bit indicates the presence or absence of a feature. For the drug interaction prediction tasks discussed above, ECFPs [34] are most commonly used. ECFPs (or Morgan fingerprints) are topological fingerprints devised for SAR modeling that, albeit humanly uninterpretable, can be processed easily with dense NNs.

Instead, SMILES is a text-based inline notation that allows treating chemistry as a language by adopting techniques from NLP. This was first discovered in 2016 when SMILES outranked fingerprints in molecular property prediction tasks [135]. Rapidly, SMILES were adopted for predicting drug-target interaction [65], chemical reactions [136] and drug sensitivity [29], toxicity [137], and especially also within generative models [138]. SMILES are processed with architectures for sequential inputs such as RNNs, 1D CNNs, or, more recently, transformer-based architectures [81]. SMILES strings are tokenized into atomic units that are passed as one-hot or learned embeddings into the network. Notably, the non-injectivity (or multiplicity) of SMILES enables data augmentation by differently traversing the same molecular graph. This was discovered by [100] and has been confirmed in countless settings (*e.g.* [29, 139]). Quickly, the basic idea of Word2Vec [140], *i.e.*, learning generic embeddings

from large corpora to facilitate downstream prediction tasks, was borrowed by Winter *et al.* [141] and later by SMILES-BERT [142] and SMILES Transformer [143]. Without the notion of pretraining, related ideas on obtaining better SMILES embeddings were explored in SMILES2Vec [144] and SMILES-X [145]. In refinements, it was shown that n -gram encoding of SMILES strings yield significantly compressed sequences while obtaining predictive accuracy surpassing the conventional atom-level tokenization [146]. Incorporating graph structure based on inter-atomic distances improves accuracy [147]. For generative models, the difficulties in generating SMILES lies in 1) their non-locality (*e.g.*, ring-opening/closure symbols) and 2) their potential semantic or syntactic invalidity (due to, *e.g.*, atom valence, which can cause the fraction of valid molecules to be as low as 4% [13]). To circumvent 1) stack-augmented RNNs were proposed and shown to improve validity [14, 148]. To address 2) instead, specific architectures such as the Grammar VAE [149] can operate on a context-free grammar derived from SMILES. Lately, however, SELFIES, a novel, stand-alone string representation for molecular generative models, was proposed that entirely solves the validity problem in SMILES through its internal syntax [150].

Since neither SMILES nor fingerprints can properly capture the topology of a molecule, it seems appropriate to process molecular graphs directly *via* graph-based models after defining atoms as nodes and bonds as edges. The most common example is graph convolutional neural networks (GCNs) that extend the idea of learnable convolution filters to aggregate local information from regular to non-Euclidean irregular grids [151, 152]. Throughout the GCN layers, node representations are updated *via* message passing [153, 154]. Other variants of GNNs include Weave models and Directed acyclic graph (DAG) models, directed graphs with directions of bonds pointing toward a designated central atom [39]. While graph-based representations enjoy more and more success in all forms of QSAR tasks, SMILES are still the gold standard for generative modeling [155], partly because many graph-based generative models do not (yet) scale to sufficiently large molecular graphs. A related, albeit rare approach, is to train QSAR models on images of their Kekule structure [156].

For details on chemical descriptors and graphs, [157] and [39], and for further reading on this chapter, [133].

3. GENERATIVE MODELS FOR *DE-NOVO* DRUG DESIGN

In Section 3.1, we briefly review the most groundbreaking publications in deep generative models for drug discovery. In Section 3.2, we give an in-depth review of conditional deep generative models for molecular design.

3.1. The Generative Models Landscape

In the last years, deep learning has induced a major paradigm shift in computational approaches to drug design, away from discrete, local optimization toward more cohesive and systematic exploration in the chemical space. This has led to a plethora of reviews about the rise of deep learning in, *e.g.*, drug discovery [17], molecular design [22, 155, 158, 159] or, more broadly, molecular biology [160] and generative chemistry [161] to which the reader is referred here.

In 2017, VAEs were proposed to learn continuous representations of molecules in an unsupervised way [13]. A key advantage of VAEs is their smooth latent space, allowing not only sampling of *de-novo* molecules but also their interpolation, thus resembling a manifold of the chemical space. The generative decoder consisted of an RNN (specifically a GRU [162]) that learned the syntax of SMILES and generated SMILES sequences auto-regressively, token by token. Almost concurrently, Segler *et al.* demonstrated that the molecules generated by RNNs could accurately mimic the distributions of physicochemical properties in the underlying training dataset [138] and later be tuned toward specific properties of interest. Another common technique to generate molecules with certain properties is reinforcement learning (RL) which can be viably coupled with RNNs by treating the completion of a SMILES sequence as action and the property of the molecule as a reward and optimize the process *via* model-free [163] and model-based RL [164]. Evidence that RL can tune the generation of molecules toward more synthetically accessible, soluble, or JAK2-inhibiting molecules when compared to a baseline (without RL), while also maintaining high SMILES validity (95%) *via* a designated stack-RNN [165], was given by the policy gradient method proposed in ReLeaSE [14]. Another approach of training RNNs for molecular generation is using adversarial techniques based on GAN [166], which can be coupled with RL to obtain stochastic policies for molecular optimization, as shown in ORGAN [167]. Maybe the most mature study on this matter proposed the GENTRL model to design DDR1 inhibitors [168]. They curated specific datasets, including patent data and (DDR1) kinase inhibitors, generat-

ed 30k molecules, and synthesized and experimentally validated six of them after careful selection. Two of the compounds were active *in-vitro* with $IC_{50} < 20nM$, and one candidate was further found effective *in-vivo* in mice, but later found similar to Ponatinib [169].

While the above approaches are based on SMILES, another promising yet the less mature approach is the direct generation of molecular graphs. Current research can roughly be split into auto-regressive and one-shot generation techniques. In one-shot generation graph generation, the validity of molecular graphs is difficult to guarantee [170, 171]. The more common auto-regressive approaches generate one node [172-174] or sets of nodes [175] or edges [176] at a time, mostly using RNNs. A widely adopted technique is the Junction-Tree VAE [177] which generates tree-structured functional groups which are combined by graph message passing to obtain 100% valid molecular graphs.

To compare the prevalence of different techniques for generative molecular models, (Fig. 6) shows the evolution of the yearly number of publications utilizing VAEs, GANs, and RL. GANs were proposed first in late 2016 [178], RL was the dominating technique in 2017, but at least partially due to work by Gomez-Bombarelli *et al.* [13], VAEs are the most widely adopted technique since 2018. To train molecular generative models, databases like ZINC [179] or ChEMBL [15] are utilized, and new models should be benchmarked against GuacaMol [180] or MOSES [181].

3.2. Conditional Deep Generative Models

With tremendous progress in and much excitement about deep learning in generative chemistry, we believe the next challenge is to build multimodal conditional generative models that leverage knowledge from disparate sources when generating molecules. Driving generative models to optimize physico- or biochemical properties has been shown successful (*e.g.* [13, 14, 182-185]). However, much less work has been done on generating molecules with respect to some semantic, biochemical conditions, especially with adaptable frameworks that are not optimized for a *single* condition (or a *single* multi-objective). Ultimately, we seek generic generative models that can be queried with a semantic context and do not require fine-tuning. Typical forms of these biochemical “starting conditions” in medicinal chemistry include proteins, omics profiles of cells, or existing drugs. Notably, in all these cases, the biochemical properties subject to maximization are multimodal and based on the interaction of the drug candidate: maximizing binding affinity to a protein target, minimizing IC_{50} for a tissue or tumor type, or

Fig. (5). Number of publications that mention different types of molecular representations in the context of deep learning. Molecular descriptors like fingerprints were long the gold standard for QSAR modeling in chemoinformatics, but raw representations like the in-line string notations SMILES and especially molecular graphs have gained popularity in the last years. Notably, 2020 is the first year where graphs were mentioned more frequently than fingerprints. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Fig. (6). Number of publications that mention different types of learning approaches in generative models for drug discovery. The first relevant publications about deep learning for *de-novo* drug discovery emerged in 2017 and mostly utilized reinforcement learning. Since then, variational methods such as Variational Autoencoders developed into the most common learning frameworks followed by adversarial approaches such as GANs. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

maximizing synergistic effects with an existing drug as in polypharmacy. These biochemical properties are notoriously challenging to evaluate *in-silico* and require themselves the presence of multimodal predictive models (like presented in Section 2.). One example of improved entanglement of predictive and generative mod-

els are autofocused oracles that adaptively alter during the generative training [186].

3.2.1. Protein-target-driven Generation

The goal of this task is to generate molecules that bind to a given protein target (site). One early work embedded proteins and ligands with a GCN and a Junction-Tree VAE, respectively, and tried to learn a “direct mapper” from the protein to the ligand while estimating binding affinity and chemical properties [187]. This model operates under the assumption that the direct mapper does not succeed in its objective but rather maps the protein to favorable molecules, thus generating novel ligands on the fly. Two works proposed to generate 3D ligands from protein shape [188, 189]. The first voxelizes the protein pocket and uses adversarial training to generate ligand shapes which are subsequently converted to SMILES and show favorable properties compared with random ZINC molecules in terms of binding affinity and docking [188]. In contrast, Masuda *et al.* embedded protein and ligand into a joint, variational space [189]. In their model, ligand novelty and target affinity are counteracting. All the above papers require CP pairs for training the generative model.

Instead, CogMol [190] trains an independent CPI oracle first and then samples conditionally from a pre-trained SMILES generator by conditioning on attributes such as target affinity and off-target selectivity *via* their previously developed CLaSS technique (Controlled Latent attribute Space Sampling), a conditional rejection sampling approach [191]. Without the availability of binding data, CogMol was applied on 3 novel, SARS-CoV-2 related targets (NSP9, M^{pro}, and RBD), and generated molecules were shown to be low-toxic, drug-like, synthesizable, target-specific and selective, and favorable in docking studies [190]. In a related approach, a pre-trained protein encoder and a molecular decoder were coupled to form a hybrid VAE that was optimized with RL and shown to generate ligands with a high predicted affinity for targets entirely unseen by the generator [192]. Others also used RL [193] or formulated targeted drug generation as translation task and trained on CP pairs from BindingDB using protein primary structure and SMILES and Transformers and RNNs respectively [194, 195].

3.2.2. Transcriptome-driven Generation

Transcriptome data such as gene expression is useful to assess drug efficacy and side effects has been

used for *de-novo* drug identification [196], and advocated a pivotal role for the future of drug discovery [197]. A major advance was achieved by Méndez-Lucio *et al.* [198], who presented a GAN that can be conditioned with gene expression profiles (GEP) to generate molecules that are likely to induce the given profile when administered. The model was trained on the publicly available L1000 CMap dataset [199] of drug-induced GEPs and was shown to generate molecules more similar to active compounds than those identified by naïve GEP similarity comparison. Later, a similar goal was pursued with a novel model, a bidirectional adversarial autoencoder, that learns a joint distribution of molecules and induced changes in GEP and can, besides, also generate GEPs induced by a drug [200].

In contrast, the PaccMann^{RL} model can be conditioned on GEPs with the objective to generate molecules that exhibit low IC₅₀ to target cell lines represented by their GEP [148]. The work uses a hybrid VAE consisting of a GEP encoder and a molecular decoder and is evaluated with a pre-trained drug sensitivity prediction model [29].

3.2.3. Scaffold-driven Generation

More in line with established techniques in medicinal chemistry is the task of generating a molecule given a scaffold. In [201], the deep scaffold decorator was proposed to locally explore the chemical space around a given scaffold by exhaustively slicing molecules into scaffolds and decorations, feeding the scaffold to an RNN, and sampling decoys for each attachment point of the scaffold, either in a single or multi-step decoration. Instead, two related methods did not use SMILES, but graphs, thus overcoming the necessity to define attachment points [202, 203] while one approach can additionally be conditioned on physicochemical properties [203]. Scaffold-driven generative models are also briefly explored in [198, 204, 205]. Others proposed methods to optimize given molecules based on desired properties [206, 207], design molecules from two fragments and their 3D interaction [208], or prototype-based generation [209].

3.2.4. Related Approaches

A flexible method, entangled conditional adversarial autoencoder (ECAAE), was developed by Polykovskiy *et al.* [184]. The idea is based on the observation that standard conditional sampling in VAEs - where a molecule x is converted into its latent code z and one sample from the prior before the decoder samples based on some properties y - assumes independence of y and z . The ECAAE instead samples from a

conditional prior). Among several showcased applications, the most interesting one generated selective JAK3 inhibitors (by specifying γ as high JAK3 and low JAK2 affinity), synthesized the most promising candidate and found good target affinity *in-vitro* [184]. Others explicitly conditioned on physico- or biochemical properties [205, 210-212]. A bioactivity fingerprint for dual inhibition of two proteins was used to generate molecules conditionally [205]. One work generated drug combinations from disease-gene associations [213]. Conditional flows are a novel method for conditional generation from latent variables [214] that can be applied for cross-domain drug discovery tasks such as approximating a drug intervention by converting drugs to cell microscopy images [215].

4. OUTLOOK AND FUTURE CHALLENGES

In this conclusive section, we discuss challenges in deep learning for property-driven drug design, split separately into generative and predictive models.

Closing the loop. The overarching challenge toward more automation in drug discovery is a closer entanglement of predictive and generative ML models with biochemical experiments. Laboratories should integrate generative models into their workflow, and synthesis and experimental results should allow the refinement of these models in a closed-loop, autonomous fashion [216].

As stated by Tabor *et al.* [217], the key domains which should be integrated are (1) ML algorithms (especially generative models), (2) high-throughput virtual screening, (3) automated synthesis planning, and (4) automated laboratories. We have summarized this in our overview figure (*cf.* Fig. 2).

Regarding (1) and (2), we have elaborated in-depth already above.

In automated synthesis planning (3), much progress has been made recently [218-221]. While computer-aided synthesis planning has relied on reaction rules crafted by experts throughout the last decades [222], ML methods based on Monte Carlo tree search [221], or more broadly, reinforcement learning [223], have opened opportunities for closer integration into generative modeling workflows. Recently, coupling reaction prediction models based on the Molecular Transformer [218] with a hypergraph exploration strategy has set a new state-of-the-art for single-step retrosynthesis prediction of all precursors (reactants, catalysts, solvents) [219] and enabled fully autonomous planning of multi-step synthesis routes without any explicit chemical

knowledge. One delicate aspect is that bridging *de novo* generative models with chemical synthesis remains challenging, as synthetic accessibility scores such as SAS [224] or SCScore [225] leave potential for improvement. As molecules produced by generative models are frequently found difficult to synthesize [226], recent efforts have focused on steering generative models to propose synthesizable molecules *a priori* [227]. However, even if synthesizable *de novo* molecules are generated, and feasible synthesis routes are found, the precise action steps for conducting the synthesis have to be extracted tediously and usually with human intervention from plain text such as patents. Vaucher *et al.* [220] recently presented a Transformer-based approach to convert unstructured text-based experimental procedures into sequences of all actions needed to conduct the reactions physically. The contributions from the RXN project, specifically models for reaction prediction [218], multi-step retrosynthesis prediction [219], and conversion of experimental procedures to synthesis actions [220], are publicly available *via* a cloud-based access point (<https://rxn.res.ibm.com>) and a Python package (<https://github.com/rxn4chemistry/rxn4chemistry>). Regarding automated laboratories (4), robotic hardware for conducting reactions has been released lately, for example, a mobile robot searching for photocatalysts using Bayesian search [228] or a platform for synthesizing organic compounds from AI-guided synthesis routes [229].

In sum, however, few works have integrated some of the components (1) – (4), and none have done so entirely. In the landmark work by Zhavoronkov *et al.* [168], a VAE-based deep generative model was combined with extensive *in-silico* screening. In only 46 days, six potential DDR1 inhibitors were manually synthesized according to human-chosen routes, and one was found active in mice. Recently, Das *et al.* [191] evaluated 20 potential antimicrobial peptides proposed by a generative framework called CLaSS and found two potent candidates in only 48 days. In work about antiviral discovery, a suspected ACE2 inhibitor has been proposed by an RL-based deep generative model and synthesized using fully autonomous synthesis planning; however, experimental validation is absent [192].

4.1. Challenges in Predictive Models

4.1.1. Benchmarks

The field would benefit from widely accepted benchmarks for drug interaction prediction tasks. Currently, data preprocessing and splitting are performed

individually by research groups which induce heterogeneity and hamper direct comparisons. Similarly to databases in related fields (*e.g.*, *ImageNet* in computer vision or *IMDB Reviews* in natural language processing), we advise the community to agree upon fixed splits of the databases into train/validation/test. Ideally, these splits should be provided by authorities such as the data holders themselves and facilitate the exploration of different splitting strategies (*cf.* Fig. (3)). Benchmarking all models routinely by their performance in different splitting strategies will accelerate the development of robust models and might aid in finding better inductive biases.

4.1.2. Federated Machine Learning

Although open-sourcing software code and especially data has been proclaimed by ML luminaries since 2007 [230], it is still less common in the field than in related domains of ML, raising concerns about transparency and reproducibility across healthcare applications [120]. We encourage expedite a transformation toward a common practice of publishing both code and, whenever possible, anonymized data (especially in academic, non-commercial settings) in order to validate the proposed methodology and results.

Orchestrating the data acquired by all different industrial players and making them available to the research community in a privacy-conserving way would be a major achievement. Privacy-preserving data science techniques have advanced recently, with Federated Machine Learning (FL) emerging as a particularly promising realm. FL is concerned with decentralized, distributed training of large-scale ML models [231]. Multiple parties can collaboratively train the same model without the data having to leave proprietary servers. One pioneer at the forefront of this development is the EuropeanMELLODDY platform² – a collaborative project to harness the ever-growing arsenal of drug discovery-relevant predictive models and libraries of chemical compounds [232].

4.2. Challenges in Generative Models

4.2.1. Conditional Generation

As we described in the Introduction and Section 3.1, we believe that one key challenge is the development of multimodal conditional generative models that can be controlled with biological contexts (*e.g.*, protein targets, binding sites, or even transcriptomics profiles) and propose potent molecules for a rich set of semantic, biochemical conditions. The first step toward this

goal could be a community-accepted benchmark for conditional molecular design. However, current benchmarks [233] do not yet explicitly address conditional generation.

An unrelated methodological challenge is the development of better conditional generative models. Vanilla methods for structured sequence (SMILES) generation, such as auto-regressive VAEs, impose distributional constraints on the latent variables, most commonly multivariate Gaussians or Gaussian mixtures. This inductive bias creates a model bias that commonly causes “posterior collapse” [234], a phenomenon where the latent code is ignored in favor of the auto-regressive decoder. Practically, this leads to a design choice to either artificially weaken the decoder or changing the training objective [235]. In either way, it is challenging to control the representations learned by the latent variables with auto-regressive sequence generation models. This is especially relevant for a conditional generation where one or multiple modalities are combined into the latent variables. Conditional-flow Variational Autoencoders is a novel model class for structured sequence generation that better captures multimodal conditional distributions [214]. While this model class has not yet entered the domain of molecular design, it is a promising endeavor in the near future.

4.2.2. Interpretability in Generative Models

Interpretability and robustness are notoriously called key requirements for DL models in high-stake domains such as drug discovery [236]. Humanly interpretable access to the model’s decision process is crucial to hone trust in the models. The current efforts by the community have exclusively focused on predictive modeling and largely neglected generative models. However, interpretability in deep molecular generative models is indispensable when we seek to understand why a certain molecule was proposed in a given situation. We should not only rely on posthoc interpretability from downstream predictive models (*e.g.* [237].) but also build transparent generative models. This, however, is a largely unsolved problem in ML that just started to be addressed using (disentangled) representation learning approaches [238, 239]. One interesting approach formulated the task of property-driven molecular generation as the composition of molecular rationales, substructures that are likely to induce a property of interest [240], and others developed a tool for interactive latent space exploration [241].

4.2.3. Expressivity and Reliability

Many generative models rely on SMILES and are trained with small vocabularies, thus limiting the size of learned chemical space. Moreover, the properties of the learned chemical space heavily depend on the type of representation [242]. As we found here, SMILES are the most commonly used representations in molecular generative models. A drawback of SMILES generators is that they are usually auto regressive and based on multinomial sampling inducing “unexpected” stochasticity in that the same input leads to different decoded molecules. In VAEs and related variational methods, this secondary level of stochasticity comes on top of the inherent stochasticity when sampling from the prior to obtain the latent code. It not only significantly hampers reproducibility but also the choice of sampling can have significant impact on the results. While some circumvent this by decoding many times and taking the most frequent valid SMILES [243], it can be argued that stochasticity is a crucial feature to enable a profound exploration of the chemical space. Possible approaches to circumvent this stochasticity are non-autoregressive approaches such as the molecular transformer [143] or, more broadly speaking, masked language models where the generated sequences are obtained *via* beam search.

4.2.4. Evaluation Metrics

The evaluation of generative models remains difficult and a key challenge for the field [22]. The distributions learned by modern generative models cannot be formulated analytically, and validating *de novo* methodology in the wet lab is largely impractical. Common evaluation metrics like novelty, validity, diversity [244], uniqueness, or Frechet Chemnet Distance [245] are seemingly insufficient metrics as a simple algorithm that randomly inserts carbon atoms to molecules from the train set can match state-of-the-art models [246]. Established biochemical metrics are markedly absent, and other metrics like a quantitative estimate of drug-likeness (QED) or the octanol-water partition coefficient (LogP) are equally chemocentric. The current state of the field has been summarized as: “the current evaluations for generative models do not reflect the complexity of real discovery problems” [247].

While molecular docking is computationally expensive, it is arguably the most widely accepted methodology for *in silico* prediction of CPI. Toward alleviating the evaluation concerns in deep molecular generative models, a docking benchmark based on SMINA [248] was recently proposed [233]. The contribution finds that state-of-the-art methods failed to generate

molecules with high docking scores as predicted by SMINA against four targets. The authors call the community to develop prospective *de novo* methods that should at least be able to design molecules that dock well [233].

CONCLUSION

The traditional process of drug and material discovery is currently being transformed by the latest advances in machine and deep learning. This transformation permeates many aspects of the discovery process, such as drug repurposing, *de novo* design, and synthesis planning.

This review focused on one frontier in the field: the closer entanglement of virtual screening and *de novo* molecular design. Most DL approaches for drug design have relied on a combination of 1) a powerful and domain-agnostic molecular generative model that randomly samples from the chemical space and 2) posthoc screening methods to filter out the best candidates according to biochemical properties. This is in stark contrast to the work of medicinal chemists as it does not integrate domain knowledge *a priori* into the design process.

We have reviewed recent progress in deep predictive and generative models, performed publication keyword searches to systematically assess the growth of the field, and identified current challenges, for example: adopting DL to fully automatize laboratory workflows (closing the loop), developing interpretable generative models and adopting federated ML to overcome data privacy concerns.

Conclusively, molecular generative models can be seen as implicit databases of chemicals that can possibly be larger than any enumerated database of the chemical space. To profoundly accelerate the discovery process, new intelligible methods are needed to query the databases with multiple complex properties like target affinity or synthesizability.

CODE AVAILABILITY

The code for the meta-analysis performed in this publication is available publicly at: <https://pypi.org/project/paperscraper/>.

License note (Fig. 1): All printed logos are distributed solely for the purpose of identifying the respective service. All logos are available *via* Creative Commons license, MIT License (TDC), BSD License PyTorch), Apache License (TensorFlow, HuggingFace), were permitted for reprint (RDKit [249]), or are owned by the authors or their employer.

LIST OF ABBREVIATIONS

DL	= Deep Learning
ML	= Machine Learning
FDA	= Food and Drug Administration
QSAR	= Quantitative structure-activity relationship
VAE	= Variational Autoencoder
NLP	= Natural Language Processing
GDSC	= Genomics of Drug Sensitivity in Cancer
DRP	= Drug Response Prediction
CPI	= Compound Protein Interaction
CNN	= Convolutional Neural Network
RNN	= Recurrent Neural Network
GNN	= Graph Neural Network
GCN	= Graph Convolutional Neural Network
NN	= Neural Network
SMILES	= Simplified Molecular-input Line-entry System
RMA	= Robust Multi-Array Average
PDX	= Patient-Derived Xenograft
PPI	= Protein-Protein Interaction
DAG	= Directed Acyclic Graph
GAN	= Generative Adversarial Network
RL	= Reinforcement Learning
CLaSS	= Controlled Latent Attribute Space Sampling
GEP	= Gene Expression Profile
ECAAE	= Entangled Conditional Adversarial Autoencoder
SAS	= Synthetic Accessibility Score
FL	= Federated Machine Learning
QED	= Quantitative Estimate of Drug-likeness

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

- [1] Scannell, J.W.; Blanckley, A.; Boldon, H.; Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.*, **2012**, *11*(3), 191-200. <http://dx.doi.org/10.1038/nrd3681> PMID: 22378269
- [2] DiMasi, J.A.; Grabowski, H.G.; Hansen, R.W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.*, **2016**, *47*, 20-33. <http://dx.doi.org/10.1016/j.jhealeco.2016.01.012> PMID: 26928437
- [3] Polishchuk, P.G.; Madzhidov, T.I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.*, **2013**, *27*(8), 675-679. <http://dx.doi.org/10.1007/s10822-013-9672-4> PMID: 23963658
- [4] Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E.E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **2019**, *47*(D1), D1102-D1109. <http://dx.doi.org/10.1093/nar/gky1033> PMID: 30371825
- [5] Kinch, M.S.; Haynesworth, A.; Kinch, S.L.; Hoyer, D. An overview of FDA-approved new molecular entities: 1827-2013. In: *Drug Discovery Today*; Carney, S., Ed.; Elsevier: Amsterdam, **2014**; 19, pp. 1033-1039. <http://dx.doi.org/10.1016/j.drudis.2014.03.018>
- [6] Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. In: *Nature*; Nature Publishing Group, **2015**; 521, pp. 436-444. <http://dx.doi.org/10.1038/nature14539>
- [7] Li, Y.; Huang, C.; Ding, L.; Li, Z.; Pan, Y.; Gao, X. Deep learning in bioinformatics: introduction, application, and perspective in the big data era. *Methods*, **2019**, *166*, 4-21. <http://dx.doi.org/10.1016/j.ymeth.2019.04.008> PMID: 31022451
- [8] Lo, Y.C.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine learning in chemoinformatics and drug discovery. In: *Drug Discovery Today*; Carney, C.; Overington, J.P.; Bajorath, J., Eds.; Elsevier: Amsterdam, **2018**; 23, pp. 1538-1546. <http://dx.doi.org/10.1016/j.drudis.2018.05.010>
- [9] Struble, T.J.; Alvarez, J.C.; Brown, S.P.; Chytil, M.; Cisar, J.; DesJarlais, R.L.; Engkvist, O.; Frank, S.A.; Greve, D.R.; Griffin, D.J.; Hou, X.; Johannes, J.W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C.A.; Palmer, A.D.; Price, D.J.; Robinson, R.I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, W.H.; Barzilay, R.; Coley, C.W.; Jensen, K.F. Current and future roles of artificial intelligence in medicinal chemistry synthesis. *J. Med. Chem.*, **2020**, *63*(16), 8667-8682. <http://dx.doi.org/10.1021/acs.jmedchem.9b02120> PMID: 32243158

- [10] Gilson, M.K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, **2016**, *44*(D1), D1045-D1053. <http://dx.doi.org/10.1093/nar/gkv1072> PMID: 26481362
- [11] Yang, W.; Soares, J.; Greninger, P.; Edelman, E.J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J.A.; Thompson, I.R.; Ramaswamy, S.; Futreal, P.A.; Haber, D.A.; Stratton, M.R.; Benes, C.; McDermott, U.; Garnett, M.J. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **2013**, *41*(Database issue), D955-D961. <http://dx.doi.org/10.1093/nar/gks1111> PMID: 23180760
- [12] Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv*, **2013**. Preprint paper.
- [13] Gómez-Bombarelli, R.; Wei, J.N.; Duvenaud, D.; Hernández-Lobato, J.M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T.D.; Adams, R.P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.*, **2018**, *4*(2), 268-276. <http://dx.doi.org/10.1021/acscentsci.7b00572> PMID: 29532027
- [14] Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for *de novo* drug design. *Sci. Adv.*, **2018**, *4*(7), eaap7885. <http://dx.doi.org/10.1126/sciadv.aap7885> PMID: 30050984
- [15] Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M.P.; Overington, J.P.; Papadatos, G.; Smit, I.; Leach, A.R. The ChEMBL database in 2017. *Nucleic Acids Res.*, **2017**, *45*(D1), D945-D954. <http://dx.doi.org/10.1093/nar/gkw1074> PMID: 27899562
- [16] Wu, C.; Liu, Y.; Yang, Y.; Zhang, P.; Zhong, W.; Wang, Y.; Wang, Q.; Xu, Y.; Li, M.; Li, X.; Zheng, M.; Chen, L.; Li, H. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sin. B*, **2020**, *10*(5), 766-788. <http://dx.doi.org/10.1016/j.apsb.2020.02.008> PMID: 32292689
- [17] Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today*, **2018**, *23*(6), 1241-1250. <http://dx.doi.org/10.1016/j.drudis.2018.01.039> PMID: 29366762
- [18] Abbasi, K.; Razzaghi, P.; Poso, A.; Ghanbari-Ara, S.; Masoudi-Nejad, A. Deep learning in drug target interaction prediction: current and future perspective. *Curr. Med. Chem.*, **2021**, *28*(11), 2100-2113. <http://dx.doi.org/10.2174/0929867327666200907141016> PMID: 32895036
- [19] Baptista, D.; Ferreira, P.G.; Rocha, M. Deep learning for drug response prediction in cancer. *Brief. Bioinform.*, **2021**, *22*(1), 360-379. <http://dx.doi.org/10.1093/bib/bbz171> PMID: 31950132
- [20] Kalinin, A.A.; Higgins, G.A.; Reamaroon, N.; Sorousmehr, S.; Allyn-Feuer, A.; Dinov, I.D.; Najarian, K.; Athey, B.D. Deep learning in pharmacogenomics: from gene regulation to patient stratification. *Pharmacogenomics*, **2018**, *19*(7), 629-650. <http://dx.doi.org/10.2217/pgs-2018-0008> PMID: 29697304
- [21] Sen, P.; Lamichhane, S.; Mathema, V.B.; McGlinchey, A.; Dickens, A.M.; Khoomrung, S.; Orešič, M. Deep learning meets metabolomics: a methodological perspective. *Brief. Bioinform.*, **2021**, *22*(2), 1531-1542. <http://dx.doi.org/10.1093/bib/bbaa204> PMID: 32940335
- [22] Elton, D.C.; Boukouvalas, Z.; Fuge, M.D.; Chung, P.W. Deep learning for molecular design - a review of the state of the art. In: *Molecular Systems Design and Engineering*; de Pablo, J., Ed.; Royal Society of Chemistry: London, **2019**; pp. 828-849. <http://dx.doi.org/10.1039/C9ME00039A>
- [23] Noé, F.; Tkatchenko, A.; Müller, K.R.; Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.*, **2020**, *71*, 361-390. <http://dx.doi.org/10.1146/annurev-physchem-042018-052331> PMID: 32092281
- [24] Stokes, J.M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N.M.; MacNair, C.R.; French, S.; Carfrae, L.A.; Bloom-Ackermann, Z.; Tran, V.M.; Chiappino-Pepe, A.; Badran, A.H.; Andrews, I.W.; Chory, E.J.; Church, G.M.; Brown, E.D.; Jaakkola, T.S.; Barzilay, R.; Collins, J.J. A deep learning approach to antibiotic discovery. *Cell*, **2020**, *180*(4), 688-702.e13. <http://dx.doi.org/10.1016/j.cell.2020.01.021> PMID: 32084340
- [25] Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.F.; Fischer, P.D.; Coote, P.W.; Padmanabha Das, K.M.; Malets, Y.S.; Radchenko, D.S.; Moroz, Y.S.; Scott, D.A.; Fackeldey, K.; Hoffmann, M.; Iavniuk, I.; Wagner, G.; Arthanari, H. An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, **2020**, *580*(7805), 663-668. <http://dx.doi.org/10.1038/s41586-020-2117-z> PMID: 32152607
- [26] Ding, M.Q.; Chen, L.; Cooper, G.F.; Young, J.D.; Lu, X. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol. Cancer Res.*, **2018**, *16*(2), 269-278. <http://dx.doi.org/10.1158/1541-7786.MCR-17-0378> PMID: 29133589
- [27] Gönen, M.; Margolin, A.A. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multi-task learning. *Bioinformatics*, **2014**, *30*(17), i556-i563. <http://dx.doi.org/10.1093/bioinformatics/btu464> PMID: 25161247
- [28] Sharifi-Noghabi, H.; Zolotareva, O.; Collins, C.C.; Ester, M. MOLLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, **2019**, *35*(14), i501-i509. <http://dx.doi.org/10.1093/bioinformatics/btz318> PMID: 31510700
- [29] Manica, M.; Oskooei, A.; Born, J.; Subramanian, V.; Sáez-Rodríguez, J.; Rodríguez Martínez, M. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Mol. Pharm.*, **2019**, *16*(12), 4797-4806. <http://dx.doi.org/10.1021/acs.molpharmaceut.9b00520> PMID: 31618586
- [30] Schroedl, S. Current methods and challenges for deep learning in drug discovery. *Drug Discov. Today. Technol.*, **2019**, *32-33*(32-33), 9-17. <http://dx.doi.org/10.1016/j.ddtec.2020.07.003> PMID: 33386100
- [31] Huang, R. Xia, Menghang.; Nguyen, D.T.; Zhao, T.; Sakamuru, S.; Zhao, Jinghua.; Shahane, S.A.; Rossoshek, A.; Simeonov, A. Tox21 challenge to build predictive models

- of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Front. Environ. Sci.*, **2016**, 3, 85.
<http://dx.doi.org/10.3389/fenvs.2015.00085>
- [32] Tice, R.R.; Austin, C.P.; Kavlock, R.J.; Bucher, J.R. Improving the human hazard characterization of chemicals: a Tox21 update. In: *Environmental Health Perspectives*; Durham, **2013**; 121, pp. 756-765.
<http://dx.doi.org/10.1289/ehp.1205784>
- [33] Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.*, **2016**, 3, 1-80.
<http://dx.doi.org/10.3389/fenvs.2015.00080>
- [34] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **2010**, 50(5), 742-754.
<http://dx.doi.org/10.1021/ci100050t> PMID: 20426451
- [35] Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task neural networks for QSAR predictions. *arXiv*, **2014**. Preprint paper.
- [36] Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively multitask networks for drug discovery. *arXiv*, **2015**. Preprint paper.
- [37] Zaslavskiy, M.; Jégou, S.; Tramel, E.W.; Wainrib, G. ToxiBlend: virtual screening of toxic compounds with ensemble predictors. *Comput. Toxicol.*, **2019**, 10, 81-88.
<http://dx.doi.org/10.1016/j.comtox.2019.01.001>
- [38] Pu, L.; Naderi, M.; Liu, T.; Wu, H.-C.; Mukhopadhyay, S.; Brylinski, M. eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacol. Toxicol.*, **2019**, 20(1), 2.
<http://dx.doi.org/10.1186/s40360-018-0282-6> PMID: 30621790
- [39] Wu, Z.; Ramsundar, B.; Feinberg, E.N.; Gomes, J.; Geniesse, C.; Pappu, A.S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci. (Camb.)*, **2017**, 9(2), 513-530.
<http://dx.doi.org/10.1039/C7SC02664A> PMID: 29629118
- [40] Weininger, D. SMILES, a chemical language and information system: 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **1988**, 28(1), 31-36.
<http://dx.doi.org/10.1021/ci00057a005>
- [41] Xu, Y.; Pei, J.; Lai, L. Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *J. Chem. Inf. Model.*, **2017**, 57(11), 2672-2685.
<http://dx.doi.org/10.1021/acs.jcim.7b00244> PMID: 29019671
- [42] Paul, A.; Jha, D.; Al-Bahrani, R.; Liao, W.-K.; Choudhary, A.; Agrawal, A. CheMixNet: mixed DNN architectures for predicting chemical properties using multiple molecular representations. *arXiv*, **2020**. Preprint paper.
- [43] Azodi, C.B.; Tang, J.; Shiu, S.H. Opening the black box: interpretable machine learning for geneticists. *Preprints*, **2020**, 2020020239.
<http://dx.doi.org/10.20944/preprints202002.0239.v1>
- [44] Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv*, **2017**, 2017. Preprint paper.
- [45] Cortés-Ciriano, I.; Bender, A. Deep confidence: a computationally efficient framework for calculating reliable prediction errors for deep neural networks. *J. Chem. Inf. Model.*, **2019**, 59(3), 1269-1281.
<http://dx.doi.org/10.1021/acs.jcim.8b00542> PMID: 30336009
- [46] Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: prediction of molecular toxicity with confidence. *Bioinformatics*, **2018**, 34(14), 2508-2509.
<http://dx.doi.org/10.1093/bioinformatics/bty135> PMID: 29522123
- [47] Cortés-Ciriano, I.; Bender, A. Reliable prediction errors for deep neural networks using test-time dropout. *J. Chem. Inf. Model.*, **2019**, 59(7), 3330-3339.
<http://dx.doi.org/10.1021/acs.jcim.9b00297> PMID: 31241929
- [48] Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.*, **2017**, 65, 211-222.
<http://dx.doi.org/10.1016/j.patcog.2016.11.008>
- [49] Webel, H.E.; Kimber, T.B.; Radetzki, S.; Neuenschwander, M.; Nazaré, M.; Volkamer, A. Revealing cytotoxic substructures in molecules using deep learning. *J. Comput. Aided Mol. Des.*, **2020**, 34(7), 731-746.
<http://dx.doi.org/10.1007/s10822-020-00310-4> PMID: 32297073
- [50] Kavlock, R.; Chandler, K.; Houck, K. Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. In: *Chemical Research in Toxicology*; American Chemical Society: Washington D.C., **2012**; 25, pp. 1287-1302.
<http://dx.doi.org/10.1021/tx3000939>
- [51] Kohonen, P.; Benfenati, E.; Bower, D.; Ceder, R.; Crump, M.; Cross, K.; Grafström, R.C.; Healy, L.; Helma, C.; Jeliaskova, N.; Jeliaskov, V.; Maggioni, S.; Miller, S.; Myatt, G.; Rautenberg, M.; Stacey, G.; Willighagen, E.; Wiseman, J.; Hardy, B. The toxbank data warehouse: supporting the replacement of *in vivo* repeated dose systemic toxicity testing. *Mol. Inform.*, **2013**, 32(1), 47-63.
<http://dx.doi.org/10.1002/minf.201200114> PMID: 27481023
- [52] Schmidt, U.; Struck, S.; Gruening, B.; Hossbach, J.; Jaeger, I.S.; Parol, R.; Lindequist, U.; Teuscher, E.; Preissner, R. SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res.*, **2009**, 37(Database issue), D295-D299.
<http://dx.doi.org/10.1093/nar/gkn850> PMID: 19004875
- [53] Fonger, G.C.; Stroup, D.; Thomas, P.L.; Wexler, P. TOXNET: A computerized collection of toxicological and environmental health information. *Toxicol. Ind. Health*, **2000**, 16(1), 4-6.
<http://dx.doi.org/10.1177/074823370001600101> PMID: 10798381
- [54] Wu, Y.; Wang, G. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. *Int. J. Mol. Sci.*, **2018**, 19(8), E2358.
<http://dx.doi.org/10.3390/ijms19082358> PMID: 30103448
- [55] Karim, A.; Singh, J.; Mishra, A.; Dehngangi, A.; Newton, M.A.H.; Sattar, A. Toxicity prediction by multimodal deep learning. In: *Knowledge Management and Acquisition for Intelligent Systems*; Ohara, K.; Bai, Q., Eds.; Springer: Cham, **2019**; 11669, pp. 142-152.
http://dx.doi.org/10.1007/978-3-030-30639-7_12
- [56] Santos, R.; Ursu, O.; Gaulton, A.; Bento, A.P.; Donadi, R.S.; Bologa, C.G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T.I.; Overington, J.P. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.*, **2017**, 16(1), 19-34.
<http://dx.doi.org/10.1038/nrd.2016.230> PMID: 27910877
- [57] Trott, O.; Olson, A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring func-

- tion, efficient optimization, and multithreading. *J. Comput. Chem.*, **2010**, *31*(2), 455-461.
<http://dx.doi.org/10.1002/jcc.21334> PMID: 19499576
- [58] Bitencourt-Ferreira, G.; de Azevedo, W.F. Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes. *Biophys. Chem.*, **2018**, *240*, 63-69.
<http://dx.doi.org/10.1016/j.bpc.2018.05.010> PMID: 29906639
- [59] Bitencourt-Ferreira, G.; de Azevedo, W.F.Jr. Machine learning to predict binding affinity. *Methods Mol. Biol.*, **2019**, *2053*, 251-273.
http://dx.doi.org/10.1007/978-1-4939-9752-7_16 PMID: 31452110
- [60] Van Westen, G.J.P.; Wegner, J.K.; Ijzerman, A.P.; Van Vlijmen, H.W.T.; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm*, **2011**, *2*(1), 16-30.
<http://dx.doi.org/10.1039/C0MD00165A>
- [61] Kuhn, M.; Szklarczyk, D.; Pletscher-Frankild, S.; Blicher, T.H.; von Mering, C.; Jensen, L.J.; Bork, P. STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.*, **2014**, *42*(Database issue), D401-D407.
<http://dx.doi.org/10.1093/nar/gkt1207> PMID: 24293645
- [62] Tian, K.; Shao, M.; Wang, Y.; Guan, J.; Zhou, S. Boosting compound-protein interaction prediction by deep learning. *Methods*, **2016**, *110*, 64-72.
<http://dx.doi.org/10.1016/j.ymeth.2016.06.024> PMID: 27378654
- [63] Xu, Z.; Shi, L.; Wang, Y.; Zhang, J.; Huang, L.; Zhang, C.; Liu, S.; Zhao, P.; Liu, H.; Zhu, L.; Tai, Y.; Bai, C.; Gao, T.; Song, J.; Xia, P.; Dong, J.; Zhao, J.; Wang, F.S. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir. Med.*, **2020**, *8*(4), 420-422.
[http://dx.doi.org/10.1016/S2213-2600\(20\)30076-X](http://dx.doi.org/10.1016/S2213-2600(20)30076-X) PMID: 32085846
- [64] Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, **2019**, *35*(18), 3329-3338.
<http://dx.doi.org/10.1093/bioinformatics/btz111> PMID: 30768156
- [65] Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics*, **2018**, *34*(17), i821-i829.
<http://dx.doi.org/10.1093/bioinformatics/bty593> PMID: 30423097
- [66] Gonczarek, A.; Tomczak, J.M.; Zareba, S.; Kaczmar, J.; Dąbrowski, P.; Walczak, M.J. Interaction prediction in structure-based virtual screening using deep learning. *Comput. Biol. Med.*, **2018**, *100*, 253-258.
<http://dx.doi.org/10.1016/j.compbimed.2017.09.007> PMID: 28941550
- [67] Zhao, L.; Wang, J.; Pang, L.; Liu, Y.; Zhang, J. GANsDTA: predicting drug-target binding affinity using GANs. *Front. Genet.*, **2020**, *10*, 1243.
<http://dx.doi.org/10.3389/fgene.2019.01243> PMID: 31993067
- [68] Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv*, **2015**. Preprint paper.
- [69] Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D.R. Protein-ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.*, **2017**, *57*(4), 942-957.
<http://dx.doi.org/10.1021/acs.jcim.6b00740> PMID: 28368587
- [70] Hassan-Harrirou, H.; Zhang, C.; Lemmin, T. RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *J. Chem. Inf. Model.*, **2020**, *60*(6), 2791-2802.
<http://dx.doi.org/10.1021/acs.jcim.0c00075> PMID: 32392050
- [71] Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. DEEP: Protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J. Chem. Inf. Model.*, **2018**, *58*(2), 287-296.
<http://dx.doi.org/10.1021/acs.jcim.7b00650> PMID: 29309725
- [72] Gao, K.Y.; Fokoue, A.; Luo, H.; Iyengar, A.; Dey, S.; Zhang, P. Proceedings of the 27th International Joint Conference on Artificial Intelligence New York, USA, **2018**, pp. 3371-3377.
- [73] Tsubaki, M.; Tomii, K.; Sese, J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, **2019**, *35*(2), 309-318.
<http://dx.doi.org/10.1093/bioinformatics/bty535> PMID: 29982330
- [74] Lim, J.; Ryu, S.; Park, K.; Choe, Y.J.; Ham, J.; Kim, W.Y. Predicting drug-target interaction using a novel graph neural network with 3D structure-embedded graph representation. *J. Chem. Inf. Model.*, **2019**, *59*(9), 3981-3988.
<http://dx.doi.org/10.1021/acs.jcim.9b00387> PMID: 31443612
- [75] Torng, W.; Altman, R.B. Graph convolutional neural networks for predicting drug-target interactions. *J. Chem. Inf. Model.*, **2019**, *59*(10), 4131-4149.
<http://dx.doi.org/10.1021/acs.jcim.9b00628> PMID: 31580672
- [76] Zheng, S.; Li, Y.; Chen, S.; Xu, J.; Yang, Y. Predicting drug-protein interaction using quasi-visual question answering system. *Nat. Mach. Intell.*, **2020**, *2*, 134-140.
<http://dx.doi.org/10.1038/s42256-020-0152-y>
- [77] Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv*, **2014**. Preprint paper.
- [78] Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; Zeng, J. MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Syst.*, **2020**, *10*(4), 308-322.e11.
<http://dx.doi.org/10.1016/j.cels.2020.03.002>
- [79] Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.*, **2012**, *55*(14), 6582-6594.
<http://dx.doi.org/10.1021/jm300687e> PMID: 22716043
- [80] Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, **2020**, *36*(16), 4406-4414.
<http://dx.doi.org/10.1093/bioinformatics/btaa524> PMID: 32428219
- [81] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J. Attention is all you need. *arXiv*, **2017**. Preprint paper.
- [82] Sieg, J.; Flachsenberg, F.; Rarey, M. In need of bias con-

- trol: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.*, **2019**, 59(3), 947-961.
<http://dx.doi.org/10.1021/acs.jcim.8b00712> PMID: 30835112
- [83] Pu, L.; Govindaraj, R.G.; Lemoine, J.M.; Wu, H.C.; Brylinski, M. DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS Comput. Biol.*, **2019**, 15(2), e1006718.
<http://dx.doi.org/10.1371/journal.pcbi.1006718> PMID: 30716081
- [84] Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S.H. PubChem substance and compound databases. *Nucleic Acids Res.*, **2016**, 44(D1), D1202-D1213.
<http://dx.doi.org/10.1093/nar/gkv951> PMID: 26400175
- [85] Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.*, **2018**, 46(D1), D1074-D1082.
<http://dx.doi.org/10.1093/nar/gkx1037> PMID: 29126136
- [86] Wang, R.; Fang, X.; Lu, Y.; Yang, C.Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.*, **2005**, 48(12), 4111-4119.
<http://dx.doi.org/10.1021/jm048957q> PMID: 15943484
- [87] Huang, N.; Shoichet, B.K.; Irwin, J.J. Benchmarking sets for molecular docking. *J. Med. Chem.*, **2006**, 49(23), 6789-6801.
<http://dx.doi.org/10.1021/jm0608356> PMID: 17154509
- [88] Rohrer, S.G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.*, **2009**, 49(2), 169-184.
<http://dx.doi.org/10.1021/ci8002649> PMID: 19434821
- [89] Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.T.; Ban, F.; Norinder, U.; Gleave, M.E.; Cherkasov, A. Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Cent. Sci.*, **2020**, 6(6), 939-949.
<http://dx.doi.org/10.1021/acscentsci.0c00229> PMID: 32607441
- [90] Huang, R. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Front. Environ. Sci.*, **2016**, 3, 85.
<http://dx.doi.org/10.3389/fenvs.2015.00085>
- [91] Parks, C.; Gaieb, Z.; Amaro, R.E. An analysis of proteochemometric and conformational prediction machine learning protein-ligand binding affinity models. *Front. Mol. Biosci.*, **2020**, 7, 93.
<http://dx.doi.org/10.3389/fmolb.2020.00093> PMID: 32671093
- [92] Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the global burden of disease study 2015. *Lancet*, **2016**, 388(10053), 1459-1544.
[http://dx.doi.org/10.1016/S0140-6736\(16\)31012-1](http://dx.doi.org/10.1016/S0140-6736(16)31012-1) PMID: 27733281
- [93] Garnett, M.J.; Edelman, E.J.; Heidorn, S.J.; Greenman, C.D.; Dastur, A.; Lau, K.W.; Greninger, P.; Thompson, I.R.; Luo, X.; Soares, J.; Liu, Q.; Iorio, F.; Surdez, D.; Chen, L.; Milano, R.J.; Bignell, G.R.; Tam, A.T.; Davies, H.; Stevenson, J.A.; Barthorpe, S.; Lutz, S.R.; Kogera, F.; Lawrence, K.; McLaren-Douglas, A.; Mitropoulos, X.; Mironenko, T.; Thi, H.; Richardson, L.; Zhou, W.; Jewitt, F.; Zhang, T.; O'Brien, P.; Boisvert, J.L.; Price, S.; Hur, W.; Yang, W.; Deng, X.; Butler, A.; Choi, H.G.; Chang, J.W.; Baselga, J.; Stamenkovic, I.; Engelman, J.A.; Sharma, S.V.; Delattre, O.; Saez-Rodriguez, J.; Gray, N.S.; Settleman, J.; Futreal, P.A.; Haber, D.A.; Stratton, M.R.; Ramaswamy, S.; McDermott, U.; Benes, C.H. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **2012**, 483(7391), 570-575.
<http://dx.doi.org/10.1038/nature11005> PMID: 22460902
- [94] De Niz, C.; Rahman, R.; Zhao, X.; Pal, R. Algorithms for drug sensitivity prediction. *Algorithms*, **2016**, 9(4), 77.
<http://dx.doi.org/10.3390/a9040077>
- [95] Cortés-Ciriano, I.; Mervin, L.H.; Bender, A. Current trends in drug sensitivity prediction. *Curr. Pharm. Des.*, **2016**, 22(46), 6918-6927.
<http://dx.doi.org/10.2174/1381612822666161026154430> PMID: 27784247
- [96] Menden, M.P.; Iorio, F.; Garnett, M.; McDermott, U.; Benes, C.H.; Ballester, P.J.; Saez-Rodriguez, J. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One*, **2013**, 8(4), e61318.
<http://dx.doi.org/10.1371/journal.pone.0061318> PMID: 23646105
- [97] Chang, Y.; Park, H.; Yang, H.J.; Lee, S.; Lee, K.Y.; Kim, T.S.; Jung, J.; Shin, J.M. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci. Rep.*, **2018**, 8(1), 8857.
<http://dx.doi.org/10.1038/s41598-018-27214-6> PMID: 29891981
- [98] Li, M.; Wang, Y.; Zheng, R.; Shi, X.; Li, Y.; Wu, F.X.; Wang, J. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **2021**, 18(2), 575-582.
<http://dx.doi.org/10.1109/TCBB.2019.2919581> PMID: 31150344
- [99] Zhu, Y.; Bretin, T.; Evrard, Y.A.; Partin, A.; Xia, F.; Shukla, M.; Yoo, H.; Doroshov, J.H.; Stevens, R.L. Ensemble transfer learning for the prediction of anti-cancer drug response. *Sci. Rep.*, **2020**, 10(1), 18040.
<http://dx.doi.org/10.1038/s41598-020-74921-0> PMID: 33093487
- [100] Bjerrum, E.J. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv*, **2017**. Preprint paper.
- [101] Cadow, J.; Born, J.; Manica, M.; Oskooei, A.; Rodríguez Martínez, M. PaccMann: a web service for interpretable anticancer compound sensitivity prediction. *Nucleic Acids Res.*, **2020**, 48(W1), W502-W508.
<http://dx.doi.org/10.1093/nar/gkaa327> PMID: 32402082
- [102] Huang, L.-C.; Yeung, W.; Wang, Y.; Cheng, H.; Venkat, A.; Li, S.; Ma, P.; Rasheed, K.; Kannan, N. Quantitative structure-mutation-activity relationship tests (QSMART) model for protein kinase inhibitor response prediction. *BMC Bioinformatics*, **2020**, 21(1), 520.
<http://dx.doi.org/10.1186/s12859-020-03842-6> PMID: 33183223
- [103] Kuenzi, B.M.; Park, J.; Fong, S.H.; Sanchez, K.S.; Lee, J.; Kreisberg, J.F.; Ma, J.; Ideker, T. Predicting drug response and synergy using a deep learning model of human cancer

- cells. *Cancer Cell*, **2020**, 38(5), 672-684.e6.
<http://dx.doi.org/10.1016/j.ccell.2020.09.014> PMID: 33096023
- [104] Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; Harris, M.A.; Hill, D.P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.C.; Richardson, J.E.; Ringwald, M.; Rubin, G.M.; Sherlock, G. Gene ontology: tool for the unification of biology. *Nat. Genet.*, **2000**, 25(1), 25-29.
<http://dx.doi.org/10.1038/75556> PMID: 10802651
- [105] Ma, J.; Yu, M.K.; Fong, S.; Ono, K.; Sage, E.; Demchak, B.; Sharan, R.; Ideker, T. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods*, **2018**, 15(4), 290-298.
<http://dx.doi.org/10.1038/nmeth.4627> PMID: 29505029
- [106] Yu, M.K.; Ma, J.; Fisher, J.; Kreisberg, J.F.; Raphael, B.J.; Ideker, T. Visible machine learning for biomedicine. *Cell*, **2018**, 173(7), 1562-1565.
<http://dx.doi.org/10.1016/j.cell.2018.05.056> PMID: 29906441
- [107] Sharifi-Noghabi, H.; Peng, S.; Zolotareva, O.; Collins, C.C.; Ester, M. AITL: adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics. *Bioinformatics*, **2020**, 36(Suppl. 1), i380-i388.
<http://dx.doi.org/10.1093/bioinformatics/btaa442> PMID: 32657371
- [108] Sakellaropoulos, T.; Vougas, K.; Narang, S.; Koinis, F.; Kotsinas, A.; Polyzos, A.; Moss, T.J.; Piha-Paul, S.; Zhou, H.; Kardala, E.; Damianidou, E.; Alexopoulos, L.G.; Aifantis, I.; Townsend, P.A.; Panayiotidis, M.I.; Sfikakis, P.; Bartek, J.; Fitzgerald, R.C.; Thanos, D.; Mills Shaw, K.R.; Petty, R.; Tsirigos, A.; Gorgoulis, V.G. A deep learning framework for predicting response to therapy in cancer. *Cell Rep.*, **2019**, 29(11), 3367-3373.e4.
<http://dx.doi.org/10.1016/j.celrep.2019.11.017> PMID: 31825821
- [109] Koras, K.; Juraeva, D.; Kreis, J.; Mazur, J.; Staub, E.; Szczurek, E. Feature selection strategies for drug sensitivity prediction. *Sci. Rep.*, **2020**, 10(1), 9377-9377.
<http://dx.doi.org/10.1038/s41598-020-65927-9> PMID: 32523056
- [110] Clyde, A. A systematic approach to featurization for cancer drug sensitivity predictions with deep learning. *arXiv*, **2020**. Preprint paper.
- [111] Costello, J.C.; Heiser, L.M.; Georgii, E.; Gönen, M.; Menden, M.P.; Wang, N.J.; Bansal, M.; Ammad-ud-din, M.; Hintsanen, P.; Khan, S.A.; Mpindi, J.P.; Kallioniemi, O.; Honkela, A.; Aittokallio, T.; Wennerberg, K.; Collins, J.J.; Gallahan, D.; Singer, D.; Saez-Rodriguez, J.; Kaski, S.; Gray, J.W.; Stolovitzky, G. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, **2014**, 32(12), 1202-1212.
<http://dx.doi.org/10.1038/nbt.2877> PMID: 24880487
- [112] Iorio, F.; Knijnenburg, T.A.; Vis, D.J.; Bignell, G.R.; Menden, M.P.; Schubert, M.; Aben, N.; Gonçalves, E.; Barthorpe, S.; Lightfoot, H.; Cokelaer, T.; Greninger, P.; van Dyk, E.; Chang, H.; de Silva, H.; Heyn, H.; Deng, X.; Egan, R.K.; Liu, Q.; Mironenko, T.; Mitropoulos, X.; Richardson, L.; Wang, J.; Zhang, T.; Moran, S.; Sayols, S.; Soleimani, M.; Tamborero, D.; Lopez-Bigas, N.; Ross-Macdonald, P.; Esteller, M.; Gray, N.S.; Haber, D.A.; Stratton, M.R.; Benes, C.H.; Wessels, L.F.A.; Saez-Rodriguez, J.; McDermott, U.; Garnett, M.J. A landscape of pharmacogenomic interactions in cancer. *Cell*, **2016**, 166(3), 740-754.
<http://dx.doi.org/10.1016/j.cell.2016.06.017> PMID: 27397505
- [113] Lee, J.K.; Havaleshko, D.M.; Cho, H.; Weinstein, J.N.; Kaldjian, E.P.; Karpovich, J.; Grimshaw, A.; Theodorescu, D. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proc. Natl. Acad. Sci. USA*, **2007**, 104(32), 13086-13091.
<http://dx.doi.org/10.1073/pnas.0610292104> PMID: 17666531
- [114] Zhu, Y.; Brettin, T.; Evrard, Y.A.; Xia, F.; Partin, A.; Shukla, M.; Yoo, H.; Doroshow, J.H.; Stevens, R.L. Enhanced co-expression extrapolation (COXEN) gene selection method for building anti-cancer drug response prediction models. *Genes (Basel)*, **2020**, 11(9), 1070.
<http://dx.doi.org/10.3390/genes11091070> PMID: 32933072
- [115] Oskooei, A.; Manica, M.; Mathis, R.; Martínez, M.R. Network-based biased tree ensembles (NetBiTE) for drug sensitivity prediction and drug sensitivity biomarker identification in cancer. *Sci. Rep.*, **2019**, 9(1), 15918.
<http://dx.doi.org/10.1038/s41598-019-52093-w> PMID: 31685861
- [116] Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K.P.; Kuhn, M.; Bork, P.; Jensen, L.J.; von Mering, C. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **2015**, 43, D447-D452.
<http://dx.doi.org/10.1093/nar/gku1003> PMID: 25352553
- [117] Zhang, F.; Wang, M.; Xi, J.; Yang, J.; Li, A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci. Rep.*, **2018**, 8(1), 3355.
<http://dx.doi.org/10.1038/s41598-018-21622-4> PMID: 29463808
- [118] Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehár, J.; Kryukov, G.V.; Sonkin, D.; Reddy, A.; Liu, M.; Murray, L.; Berger, M.F.; Monahan, J.E.; Morais, P.; Meltzer, J.; Korejwa, A.; Jané-Valbuena, J.; Mapa, F.A.; Thibault, J.; Bric-Furlong, E.; Raman, P.; Shipway, A.; Engels, I.H.; Cheng, J.; Yu, G.K.; Yu, J.; Aspesi, P., Jr; de Silva, M.; Jagtap, K.; Jones, M.D.; Wang, L.; Hatton, C.; Palesscandolo, E.; Gupta, S.; Mahan, S.; Sougnez, C.; Onofrio, R.C.; Liefeld, T.; MacConaill, L.; Winckler, W.; Reich, M.; Li, N.; Mesirov, J.P.; Gabriel, S.B.; Getz, G.; Ardlie, K.; Chan, V.; Myer, V.E.; Weber, B.L.; Porter, J.; Warmuth, M.; Finan, P.; Harris, J.L.; Meyerson, M.; Golub, T.R.; Morrissey, M.P.; Sellers, W.R.; Schlegel, R.; Garraway, L.A. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **2012**, 483(7391), 603-607.
<http://dx.doi.org/10.1038/nature11003> PMID: 22460905
- [119] Nusinow, D.P.; Szpyt, J.; Ghandi, M.; Rose, C.M.; McDonald, E.R., III; Kalocsay, M.; Jané-Valbuena, J.; Gelfand, E.; Schweppe, D.K.; Jedrychowski, M.; Golji, J.; Porter, D.A.; Rejtar, T.; Wang, Y.K.; Kryukov, G.V.; Stegmeier, F.; Erickson, B.K.; Garraway, L.A.; Sellers, W.R.; Gygi, S.P. Quantitative proteomics of the cancer cell line encyclopedia. *Cell*, **2020**, 180(2), 387-402.e16.
<http://dx.doi.org/10.1016/j.cell.2019.12.023> PMID: 31978347
- [120] Rampasek, L.; Hidru, D.; Smirnov, P.; Haibe-Kains, B.; Goldenberg, A. Dr.VAE: Drug response variational autoencoder. *Bioinformatics*, **2019**, 35(19), 3743-3751.
<http://dx.doi.org/10.1093/bioinformatics/btz158> PMID: 31978347

- 30850846
- [121] Huang, K.; Xiao, C.; Hoang, T. N.; Glass, L. M.; Sun, J. CASTER: predicting drug interactions with chemical substructure representation. *AAAI*, **2019**, 34(01), 702-709. <http://dx.doi.org/10.1609/aaai.v34i01.5412>
- [122] Preuer, K.; Lewis, R.P.I.; Hochreiter, S.; Bender, A.; Bulusu, K.C.; Klambauer, G. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, **2018**, 34(9), 1538-1546. <http://dx.doi.org/10.1093/bioinformatics/btx806> PMID: 29253077
- [123] Chen, G.; Tsoi, A.; Xu, H.; Zheng, W.J. Predict effective drug combination by deep belief network and ontology fingerprints. *J. Biomed. Inform.*, **2018**, 85, 149-154. <http://dx.doi.org/10.1016/j.jbi.2018.07.024> PMID: 30081101
- [124] Xia, F.; Shukla, M.; Brettin, T.; Garcia-Cardona, C.; Cohn, J.; Allen, J.E.; Maslov, S.; Holbeck, S.L.; Doroshow, J.H.; Evrard, Y.A.; Stahlberg, E.A.; Stevens, R.L. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics*, **2018**, 19(Suppl. 18), 486. <http://dx.doi.org/10.1186/s12859-018-2509-3> PMID: 30577754
- [125] Zitnik, M.; Agrawal, M.; Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, **2018**, 34(13), i457-i466. <http://dx.doi.org/10.1093/bioinformatics/bty294> PMID: 29949996
- [126] Shoemaker, R.H. The NCI60 human tumour cell line anti-cancer drug screen. *Nat. Rev. Cancer*, **2006**, 6(10), 813-823. <http://dx.doi.org/10.1038/nrc1951> PMID: 16990858
- [127] Smirnov, P.; Kofia, V.; Maru, A.; Freeman, M.; Ho, C.; El-Hachem, N.; Adam, G.A.; Ba-Alawi, W.; Safikhani, Z.; Haibe-Kains, B. PharmacDB: an integrative database for mining *in vitro* anticancer drug screening studies. *Nucleic Acids Res.*, **2018**, 46(D1), D994-D1002. <http://dx.doi.org/10.1093/nar/gkx911> PMID: 30053271
- [128] Gao, H.; Korn, J.M.; Ferretti, S.; Monahan, J.E.; Wang, Y.; Singh, M.; Zhang, C.; Schnell, C.; Yang, G.; Zhang, Y.; Balbin, O.A.; Barbe, S.; Cai, H.; Casey, F.; Chatterjee, S.; Chiang, D.Y.; Chuai, S.; Cogan, S.M.; Collins, S.D.; Dammasa, E.; Ebel, N.; Embry, M.; Green, J.; Kauffmann, A.; Kowal, C.; Leary, R.J.; Lehar, J.; Liang, Y.; Loo, A.; Lorenzana, E.; Robert McDonald, E., III; McLaughlin, M.E.; Merkin, J.; Meyer, R.; Naylor, T.L.; Patawaran, M.; Reddy, A.; Röelli, C.; Ruddy, D.A.; Salangsang, F.; Santacrose, F.; Singh, A.P.; Tang, Y.; Tinetto, W.; Tobler, S.; Velazquez, R.; Venkatesan, K.; Von Arx, F.; Wang, H.Q.; Wang, Z.; Wiesmann, M.; Wyss, D.; Xu, F.; Bitter, H.; Atadja, P.; Lees, E.; Hofmann, F.; Li, E.; Keen, N.; Cozens, R.; Jensen, M.R.; Pryer, N.K.; Williams, J.A.; Sellers, W.R. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.*, **2015**, 21(11), 1318-1325. <http://dx.doi.org/10.1038/nm.3954> PMID: 26479923
- [129] Holbeck, S.L.; Camalier, R.; Crowell, J.A.; Govindharajulu, J.P.; Hollingshead, M.; Anderson, L.W.; Polley, E.; Rubinstein, L.; Srivastava, A.; Wilsker, D.; Collins, J.M.; Doroshow, J.H. The national cancer institute almanac: a comprehensive screening resource for the detection of anti-cancer drug pairs with enhanced therapeutic activity. *Cancer Res.*, **2017**, 77(13), 3564-3576. <http://dx.doi.org/10.1158/0008-5472.CAN-17-0489> PMID: 28446463
- [130] Menden, M.P.; Wang, D.; Mason, M.J.; Szalai, B.; Bulusu, K.C.; Guan, Y.; Yu, T.; Kang, J.; Jeon, M.; Wolfinger, R.; Nguyen, T.; Zaslavskiy, M.; Jang, I.S.; Ghazoui, Z.; Ahn, M.E.; Vogel, R.; Neto, E.C.; Norman, T.; Tang, E.K.Y.; Garnett, M.J.; Veroli, G.Y.D.; Fawell, S.; Stolovitzky, G.; Guinney, J.; Dry, J.R.; Saez-Rodriguez, J. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.*, **2019**, 10(1), 2674. <http://dx.doi.org/10.1038/s41467-019-09799-2> PMID: 31209238
- [131] O'Neil, J.; Benita, Y.; Feldman, I.; Chenard, M.; Roberts, B.; Liu, Y.; Li, J.; Kral, A.; Lejnine, S.; Loboda, A.; Arthur, W.; Cristescu, R.; Haines, B.B.; Winter, C.; Zhang, T.; Bloecher, A.; Shumway, S.D. An unbiased oncology compound screen to identify novel combination strategies. *Mol. Cancer Ther.*, **2016**, 15(6), 1155-1162. <http://dx.doi.org/10.1158/1535-7163.MCT-15-0843> PMID: 26983881
- [132] Adam, G.; Rampášek, L.; Safikhani, Z.; Smirnov, P.; Haibe-Kains, B.; Goldenberg, A. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis. Oncol.*, **2020**, 4, 19. <http://dx.doi.org/10.1038/s41698-020-0122-1> PMID: 32566759
- [133] Chuang, K.V.; Gunsalus, L.M.; Keiser, M.J. Learning molecular representations for medicinal chemistry. *J. Med. Chem.*, **2020**, 63(16), 8705-8722. <http://dx.doi.org/10.1021/acs.jmedchem.0c00385> PMID: 32366098
- [134] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **1988**, 28(1), 31-36. <http://dx.doi.org/10.1021/ci00057a005>
- [135] Jastrzebski, S.; Leśniak, D.; Czarnecki, W. M. Learning to smile (S). *arXiv*, **2016**. Preprint paper.
- [136] Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci. (Camb.)*, **2018**, 9(28), 6091-6098. <http://dx.doi.org/10.1039/C8SC02339E> PMID: 30090297
- [137] Hirohara, M.; Saito, Y.; Koda, Y.; Sato, K.; Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics*, **2018**, 19(Suppl. 19), 526. <http://dx.doi.org/10.1186/s12859-018-2523-5> PMID: 30598075
- [138] Segler, M.H.S.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.*, **2018**, 4(1), 120-131. <http://dx.doi.org/10.1021/acscentsci.7b00512> PMID: 29392184
- [139] Tetko, I.V.; Karpov, P.; Bruno, E.; Kimber, T.B.; Godin, G. In: Augmentation Is What You Need!, proceeding of the 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17-19, 2019; Tetko, I.; Kůrková, V.; Karpov, P.; Theis, F., Eds; ICANN 2019: Workshop and Special Sessions. Lecture Notes in Computer Science, Springer: Cham, **2019**, Vol. 11731, pp. 831-835. http://dx.doi.org/10.1007/978-3-030-30493-5_79
- [140] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Distributed

- representations of words and phrases and their compositionality. *arXiv*, **2013**. Preprint paper.
- [141] Winter, R.; Montanari, F.; Noé, F.; Clevert, D.A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci. (Camb.)*, **2018**, *10*(6), 1692-1701. <http://dx.doi.org/10.1039/C8SC04175J> PMID: 30842833
- [142] Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. In: *SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction*, Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, September 4, **2019**, ACM: New York, United States, pp. 429-436. <http://dx.doi.org/10.1145/3307339.3342186>
- [143] Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C.A.; Bekas, C.; Lee, A.A. Molecular transformer-a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.*, **2019**, *5*(9), 1572-1583. <http://dx.doi.org/10.1021/acscentsci.9b00576> PMID: 31572784
- [144] Goh, G.B.; Hodas, N.O.; Siegel, C.; Vishnu, A. SMILES2Vec: an interpretable general-purpose deep neural network for predicting chemical properties. *arXiv*, **2017**. Preprint paper.
- [145] Lambard, G.; Gracheva, E. SMILES-X: autonomous molecular compounds characterisation for small datasets without descriptors. *Mach. Learn. Sci. Technol.*, **2020**, *1*(2), 025004. <http://dx.doi.org/10.1088/2632-2153/ab57f3>
- [146] Li, X.; Fourches, D. Smiles pair encoding: a data-driven substructure tokenization algorithm for deep learning. *J. Chem. Inf. Model.*, **2021**, *61*(4), 1560-1569. <http://dx.doi.org/10.1021/acs.jcim.0c01127> PMID: 33715361
- [147] Maziarka, L.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S. Molecule attention transformer. *arXiv*, **2020**. Preprint paper.
- [148] Born, J.; Manica, M.; Oskooei, A.; Cadow, J.; Markert, G.; Rodríguez Martínez, M. PaccMann^{RL}: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience*, **2021**, *24*(4), 102269. <http://dx.doi.org/10.1016/j.isci.2021.102269> PMID: 33851095
- [149] Kusner, M.J.; Paige, B.; Hernández-Lobato, J.M. In: *Grammar Variational Autoencoder*, Proceedings of the 34th International Conference on Machine Learning, August 6-11, **2017**, Vol. 70, pp. 1945-1954.
- [150] Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. *Mach. Learn. Sci. Technol.*, **2020**, *1*(4), 045024. <http://dx.doi.org/10.1088/2632-2153/aba947>
- [151] Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R.P. Convolutional networks on graphs for learning molecular fingerprints. *arXiv*, **2015**. Preprint paper.
- [152] Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv*, **2016**. Preprint paper.
- [153] Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural message passing for quantum chemistry. *arXiv*, **2017**. Preprint paper.
- [154] Withnall, M.; Lindelöf, E.; Engkvist, O.; Chen, H. Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction. *J. Cheminform.*, **2020**, *12*(1), 1-18. <http://dx.doi.org/10.1186/s13321-019-0407-y> PMID: 33430988
- [155] Schwalbe-Koda, D.; Gómez-Bombarelli, R. Generative models for automatic chemical design. In: *Machine Learning Meets Quantum Physics. Lecture Notes in Physics*; Schütt, K.T.; Chmiela, S.; von Lilienfeld, O.A.; Tkatchenko, K.; Müller, K.-R., Eds.; Springer: Cham, **2020**; 968, pp. 445-467. http://dx.doi.org/10.1007/978-3-030-40245-7_21
- [156] Cortés-Ciriano, I.; Bender, A. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *J. Cheminform.*, **2019**, *11*(1), 41. <http://dx.doi.org/10.1186/s13321-019-0364-5> PMID: 31218493
- [157] Koutsoukas, A.; Paricharak, S.; Galloway, W.R.; Spring, D.R.; Ijzerman, A.P.; Glen, R.C.; Marcus, D.; Bender, A. How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inf. Model.*, **2014**, *54*(1), 230-242. <http://dx.doi.org/10.1021/ci400469u> PMID: 24289493
- [158] Xu, Y.; Lin, K.; Wang, S.; Wang, L.; Cai, C.; Song, C.; Lai, L.; Pei, J. Deep learning for molecular generation. *Future Med. Chem.*, **2019**, *11*(6), 567-597. <http://dx.doi.org/10.4155/fmc-2018-0358> PMID: 30698019
- [159] Jørgensen, P.B.; Schmidt, M.N.; Winther, O. Deep generative models for molecular science. *Mol. Inform.*, **2018**, *37*(1-2), 1700133. <http://dx.doi.org/10.1002/minf.201700133> PMID: 29405647
- [160] Lopez, R.; Gayoso, A.; Yosef, N. Enhancing scientific discoveries in molecular biology with deep generative models. *Mol. Syst. Biol.*, **2020**, *16*(9), e9198. <http://dx.doi.org/10.15252/msb.20199198> PMID: 32975352
- [161] Vanhaelen, Q.; Lin, Y.C.; Zhavoronkov, A. The advent of generative chemistry. *ACS Med. Chem. Lett.*, **2020**, *11*(8), 1496-1505. <http://dx.doi.org/10.1021/acsmmedchemlett.0c00088> PMID: 32832015
- [162] Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv*, **2014**. Preprint paper.
- [163] Jaques, N.; Gu, S.; Bahdanau, D.; Miguel Hernández-Lobato, J.; Turner, R.E.; Eck, D. Sequence tutor: conservative fine-tuning of sequence generation models with kl-control. *arXiv*, **2017**. Preprint paper.
- [164] Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.*, **2017**, *9*(1), 48. <http://dx.doi.org/10.1186/s13321-017-0235-x> PMID: 29086083
- [165] Joulin, A.; Mikolov, T. Inferring algorithmic patterns with stack-augmented recurrent nets. *arXiv*, **2015**. Preprint paper.
- [166] Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv*, **2014**. Preprint paper.
- [167] Guimaraes, G.L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P.L.C.; Aspuru-Guzik, A.A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. *ArXiv*, **2017**. Preprint paper.
- [168] Zhavoronkov, A.; Ivanenkov, Y.A.; Aliper, A.; Veselov,

- M.S.; Aladinskiy, V.A.; Aladinskaya, A.V.; Terentiev, V.A.; Polykovskiy, D.A.; Kuznetsov, M.D.; Asadulaev, A.; Volkov, Y.; Zhodus, A.; Shayakhmetov, R.R.; Zhebrak, A.; Minaeva, L.I.; Zagribelnyy, B.A.; Lee, L.H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.*, **2019**, 37(9), 1038-1040. <http://dx.doi.org/10.1038/s41587-019-0224-x> PMID: 31477924
- [169] Walters, W.P.; Murcko, M. Assessing the impact of generative AI on medicinal chemistry. *Nat. Biotechnol.*, **2020**, 38(2), 143-145. <http://dx.doi.org/10.1038/s41587-020-0418-2> PMID: 32001834
- [170] Simonovsky, M.; Komodakis, N. In: *GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders*, 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018; Simonovsky, M.; Komodakis, N., Eds., Springer: Cham, Vol. 11139, pp. 412-422. http://dx.doi.org/10.1007/978-3-030-01418-6_41
- [171] Kipf, T.N.; Welling, M. Variational graph auto-encoders. *arXiv*, **2016**. Preprint Paper.
- [172] De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv*, **2018**, 2018, arXiv:1805.11973.
- [173] Li, Y.; Vinyals, O.; Dyer, C.; Pascanu, R.; Battaglia, P. Learning deep generative models of graphs. *arXiv*, **2018**, 2018, arXiv:1803.03324.
- [174] You, J.; Liu, B.; Ying, Z.; Pande, V.; Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. *arXiv*, **2018**, 2018, arXiv:1806.02473.
- [175] Liao, R.; Li, Y.; Song, Y.; Wang, S.; Nash, C.; Hamilton, W.L.; Duvenaud, D.; Urtasun, R.; Zemel, R.S. Efficient graph generation with graph recurrent attention networks. *arXiv*, **2019**, 2019, arXiv:1910.00760.
- [176] Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A.L. Constrained graph variational autoencoders for molecule design. *arXiv*, **2018**, 2018, arXiv:1805.09076.
- [177] Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *arXiv*, **2018**, 2018, arXiv:1802.04364.
- [178] Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, **2017**, 8(7), 10883-10890. <http://dx.doi.org/10.18632/oncotarget.14073> PMID: 28029644
- [179] Sterling, T.; Irwin, J.J. ZINC 15-ligand discovery for everyone. *J. Chem. Inf. Model.*, **2015**, 55(11), 2324-2337. <http://dx.doi.org/10.1021/acs.jcim.5b00559> PMID: 26479676
- [180] Brown, N.; Fiscato, M.; Segler, M.H.S.; Vaucher, A.C. GuacaMol: benchmarking models for *de novo* molecular design. *J. Chem. Inf. Model.*, **2019**, 59(3), 1096-1108. <http://dx.doi.org/10.1021/acs.jcim.8b00839> PMID: 30887799
- [181] Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Ar-tamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Front. Pharmacol.*, **2020**, 11, 565644. <http://dx.doi.org/10.3389/fphar.2020.565644> PMID: 33390943
- [182] Kwon, Y.; Yoo, J.; Choi, Y.-S.; Son, W.-J.; Lee, D.; Kang, S. Efficient learning of non-autoregressive graph variational autoencoders for molecular graph generation. *J. Cheminform.*, **2019**, 11(1), 70. <http://dx.doi.org/10.1186/s13321-019-0396-x> PMID: 33430985
- [183] Domenico, A.; Nicola, G.; Daniela, T.; Fulvio, C.; Nicola, A.; Orazio, N. *De novo* drug design of targeted chemical libraries based on artificial intelligence and pair-based multi-objective optimization. *J. Chem. Inf. Model.*, **2020**, 60(10), 4582-4593. <http://dx.doi.org/10.1021/acs.jcim.0c00517> PMID: 32845150
- [184] Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled conditional adversarial autoencoder for *de novo* drug discovery. *Mol. Pharm.*, **2018**, 15(10), 4398-4405. <http://dx.doi.org/10.1021/acs.molpharmaceut.8b00839> PMID: 30180591
- [185] Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A.V.; Aliper, A.; Zhavoronkov, A. Adversarial threshold neural computer for molecular *de novo* design. *Mol. Pharm.*, **2018**, 15(10), 4386-4397. <http://dx.doi.org/10.1021/acs.molpharmaceut.7b01137> PMID: 29569445
- [186] Fannjiang, C.; Listgarten, J. Autofocused oracles for model-based design. *arXiv*, **2020**, (2020), arXiv:2006.08052.
- [187] Aumentado-armstrong, T. Latent molecular optimization for targeted therapeutic design. *arXiv*, **2018**.
- [188] Skalic, M.; Sabbadin, D.; Sattarov, B.; Sciabola, S.; De Fabritiis, G. From target to drug: generative modeling for the multimodal structure-based ligand design. *Mol. Pharm.*, **2019**, 16(10), 4282-4291. <http://dx.doi.org/10.1021/acs.molpharmaceut.9b00634> PMID: 31437001
- [189] Masuda, T.; Ragoza, M.; Koes, D.R. Generating 3D molecular structures conditional on a receptor binding site with deep generative models. *arXiv*, **2020**, 2020, arXiv:2010.14442.
- [190] Chenthamarakshan, V.; Das, P.; Hoffman, S.C.; Strobelt, H.; Padhi, I.; Lim, K.W.; Hoover, B.; Manica, M.; Born, J.; Laino, T.; Mojsilovic, A. CogMol: Target-specific and selective drug design for covid-19 using deep generative models. *arXiv*, **2020**, 2020, arXiv:2004.01215.
- [191] Das, P.; Sercu, T.; Wadhawan, K.; Padhi, I.; Gehrmann, S.; Cipicigan, F.; Chenthamarakshan, V.; Strobelt, H.; Dos Santos, C.; Chen, P.-Y.; Yang, Y.Y.; Tan, J.P.K.; Hedrick, J.; Crain, J.; Mojsilovic, A. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.*, **2021**, 5(6), 613-623. <http://dx.doi.org/10.1038/s41551-021-00689-x> PMID: 33707779
- [192] Born, J.; Manica, M.; Cadow, J.; Markert, G.; Mill, N.A.; Filipavicius, M.; Janakaraman, N.; Cardinale, A.; Laino, T.; Martínez, M.R. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Mach. Learn. Sci. Technol.*, **2021**, 2(2), 025024. <http://dx.doi.org/10.1088/2632-2153/abe808>
- [193] Krishnan, S.R.; Bung, N.; Bulusu, G.; Roy, A. Accelerating *de novo* drug design against novel proteins using deep learning. *J. Chem. Inf. Model.*, **2021**, 61(2), 621-630. <http://dx.doi.org/10.1021/acs.jcim.0c01060> PMID:

- 33491455
- [194] Grechishnikova, D. Transformer neural network for protein specific *de novo* drug generation as machine translation problem. *bioRxiv*, **2019**. <http://dx.doi.org/10.1101/863415>
- [195] Ghanbargpour, A.; Lill, M.A. Seq2Mol: automatic design of *de novo* molecules conditioned by the target protein sequences through deep neural networks. *arXiv*, **2020**.
- [196] Verbist, B.; Klambauer, G.; Vervoort, L.; Talloen, W.; Shkedy, Z.; Thas, O.; Bender, A.; Göhlmann, H.W.; Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project. *Drug Discov. Today*, **2015**, *20*(5), 505-513. <http://dx.doi.org/10.1016/j.drudis.2014.12.014> PMID: 25582842
- [197] Dopazo, J. Genomics and transcriptomics in drug discovery. *Drug Discov. Today*, **2014**, *19*(2), 126-132. <http://dx.doi.org/10.1016/j.drudis.2013.06.003> PMID: 23773860
- [198] Méndez-Lucio, O.; Baillif, B.; Clevert, D.A.; Rouquié, D.; Wichard, J. *De novo* generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.*, **2020**, *11*(1), 10. <http://dx.doi.org/10.1038/s41467-019-13807-w> PMID: 31900408
- [199] Subramanian, A.; Narayan, R.; Corsello, S.M.; Peck, D.D.; Natoli, T.E.; Lu, X.; Gould, J.; Davis, J.F.; Tubelli, A.A.; Asiedu, J.K.; Lahr, D.L.; Hirschman, J.E.; Liu, Z.; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I.C.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O.M.; Piccioni, F.; Johnson, S.A.; Lyons, N.J.; Berger, A.H.; Shamji, A.F.; Brooks, A.N.; Vrcic, A.; Flynn, C.; Rosains, J.; Takeda, D.Y.; Hu, R.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Greenside, P.; Gray, N.S.; Clemons, P.A.; Silver, S.; Wu, X.; Zhao, W.N.; Read-Button, W.; Wu, X.; Haggarty, S.J.; Ronco, L.V.; Boehm, J.S.; Schreiber, S.L.; Doench, J.G.; Bittker, J.A.; Root, D.E.; Wong, B.; Golub, T.R. A next generation connectivity map: 11000 platform and the first 1,000,000 profiles. *Cell*, **2017**, *171*(6), 1437-1452.e17. <http://dx.doi.org/10.1016/j.cell.2017.10.049> PMID: 29195078
- [200] Shayakhmetov, R.; Kuznetsov, M.; Zhebrak, A.; Kadurin, A.; Nikolenko, S.; Aliper, A.; Polykovskiy, D. Molecular generation for desired transcriptome changes with adversarial autoencoders. *Front. Pharmacol.*, **2020**, *11*, 269. <http://dx.doi.org/10.3389/fphar.2020.00269> PMID: 32362822
- [201] Arús-Pous, J.; Patronov, A.; Bjerrum, E.J.; Tyrchan, C.; Reymond, J.L.; Chen, H.; Engkvist, O. SMILES-based deep generative scaffold decorator for *de-novo* drug design. *J. Cheminform.*, **2020**, *12*(1), 38. <http://dx.doi.org/10.1186/s13321-020-00441-8> PMID: 33431013
- [202] Li, Y.; Hu, J.; Wang, Y.; Zhou, J.; Zhang, L.; Liu, Z. Deep-scaffold: a comprehensive tool for scaffold-based *de novo* drug discovery using deep learning. *J. Chem. Inf. Model.*, **2020**, *60*(1), 77-91. <http://dx.doi.org/10.1021/acs.jcim.9b00727> PMID: 31809029
- [203] Lim, J.; Hwang, S.Y.; Moon, S.; Kim, S.; Kim, W.Y. Scaffold-based molecular design with a graph generative model. *Chem. Sci. (Camb.)*, **2019**, *11*(4), 1153-1164. <http://dx.doi.org/10.1039/C9SC04503A> PMID: 34084372
- [204] Zheng, S.; Yan, X.; Gu, Q.; Yang, Y.; Du, Y.; Lu, Y.; Xu, J. QBMG: quasi-biogenic molecule generator with deep recurrent neural network. *J. Cheminform.*, **2019**, *11*(1), 5. <http://dx.doi.org/10.1186/s13321-019-0328-9> PMID: 30656426
- [205] Li, Y.; Zhang, L.; Liu, Z. Multi-objective *de novo* drug design with conditional graph generative model. *J. Cheminform.*, **2018**, *10*(1), 33. <http://dx.doi.org/10.1186/s13321-018-0287-6> PMID: 30043127
- [206] Maziarka, L.; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Danel, T.; Warchol, M. Mol-CycleGAN: a generative model for molecular optimization. *J. Cheminform.*, **2020**, *12*(1), 2. <http://dx.doi.org/10.1186/s13321-019-0404-1> PMID: 33431006
- [207] Zhou, Z.; Kearnes, S.; Li, L.; Zare, R.N.; Riley, P. Optimization of molecules *via* deep reinforcement learning. *Sci. Rep.*, **2019**, *9*(1), 10752. <http://dx.doi.org/10.1038/s41598-019-47148-x> PMID: 31341196
- [208] Imrie, F.; Bradley, A.R.; van der Schaar, M.; Deane, C.M. Deep generative models for 3D linker design. *J. Chem. Inf. Model.*, **2020**, *60*(4), 1983-1995. <http://dx.doi.org/10.1021/acs.jcim.9b01120> PMID: 32195587
- [209] Harel, S.; Radinsky, K. Prototype-based compound discovery using deep generative models. *Mol. Pharm.*, **2018**, *15*(10), 4406-4416. <http://dx.doi.org/10.1021/acs.molpharmaceut.8b00474> PMID: 30063142
- [210] Lim, J.; Ryu, S.; Kim, J.W.; Kim, W.Y. Molecular generative model based on conditional variational autoencoder for *de novo* molecular design. *J. Cheminform.*, **2018**, *10*(1), 31. <http://dx.doi.org/10.1186/s13321-018-0286-7> PMID: 29995272
- [211] Kang, S.; Cho, K. Conditional molecular design with deep generative models. *J. Chem. Inf. Model.*, **2019**, *59*(1), 43-52. <http://dx.doi.org/10.1021/acs.jcim.8b00263> PMID: 30016587
- [212] Kotsias, P.-C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E.J. Direct steering of *de novo* molecular generation using descriptor conditional recurrent neural networks (cRNNs) *ChemRxiv*, **2019**. <http://dx.doi.org/10.26434/chemrxiv.9860906.v2>
- [213] Karimi, M.; Hasanzadeh, A.; Shen, Y. Network-principled deep generative models for designing drug combinations as graph sets. *Bioinformatics*, **2020**, *36*(Suppl. 1), i445-i454. <http://dx.doi.org/10.1093/bioinformatics/btaa317> PMID: 32657357
- [214] Bhattacharyya, A.; Hanselmann, M.; Fritz, M.; Schiele, B.; Straehle, C.-N. Conditional flow variational autoencoders for structured sequence prediction. *arXiv*, **2019**, 2019, arXiv:1908.09008.
- [215] Yang, K.; Goldman, S.; Jin, W.; Lu, A.; Barzilay, R.; Jaakkola, T.; Uhler, C. Improved conditional flow models for molecule to image synthesis. *arXiv*, **2020**. Preprint paper.
- [216] Gromski, P.S.; Henson, A.B.J.; Granda, J.M.; Cronin, L. How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.*, **2019**, *3*, 119-128. <http://dx.doi.org/10.1038/s41570-018-0066-y>

- [217] Tabor, D.P.; Roch, L.M.; Saikin, S.K.; Kreisbeck, C.; Sheberla, D.; Montoya, J.H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; Amador-Bedolla, C. Brabec, C.J.; Maruyama, B.; Persson, K.A.; Aspuru-Guzik, A. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.*, **2018**, *3*, 5-20. <http://dx.doi.org/10.1038/s41578-018-0005-z>
- [218] Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Bekas, C.; Lee, A.A. Molecular transformer for chemical reaction prediction and uncertainty estimation. *ACS Cent. Sci.*, **2019**, *5*(9), 1572-1583. <http://dx.doi.org/10.1021/acscentsci.9b00576> PMID: 31572784
- [219] Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V.H.; Haeuselmann, R.A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci. (Camb.)*, **2020**, *11*(12), 3316-3325. <http://dx.doi.org/10.1039/C9SC05704H> PMID: 34122839
- [220] Vaucher, A.C.; Zipoli, F.; Geluykens, J.; Nair, V.H.; Schwaller, P.; Laino, T. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.*, **2020**, *11*(1), 3601. <http://dx.doi.org/10.1038/s41467-020-17266-6> PMID: 32681088
- [221] Segler, M.H.S.; Preuss, M.; Waller, M.P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, **2018**, *555*(7698), 604-610. <http://dx.doi.org/10.1038/nature25978> PMID: 29595767
- [222] Coley, C.W.; Green, W.H.; Jensen, K.F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.*, **2018**, *51*(5), 1281-1289. <http://dx.doi.org/10.1021/acs.accounts.8b00087> PMID: 29715002
- [223] Schreck, J.S.; Coley, C.W.; Bishop, K.J.M. Learning retrosynthetic planning through simulated experience. *ACS Cent. Sci.*, **2019**, *5*(6), 970-981. <http://dx.doi.org/10.1021/acscentsci.9b00055> PMID: 31263756
- [224] Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.*, **2009**, *1*(1), 8. <http://dx.doi.org/10.1186/1758-2946-1-8> PMID: 20298526
- [225] Coley, C.W.; Rogers, L.; Green, W.H.; Jensen, K.F. SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.*, **2018**, *58*(2), 252-261. <http://dx.doi.org/10.1021/acs.jcim.7b00622> PMID: 29309147
- [226] Gao, W.; Coley, C.W. The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.*, **2020**, *60*(12), 5714-5723. <http://dx.doi.org/10.1021/acs.jcim.0c00174> PMID: 32250616
- [227] Bradshaw, J.; Paige, B.; Kusner, M.J.; Segler, M.H.S.; Hernández-Lobato, J.M. Barking up the right tree: an approach to search over molecule synthesis DAGs. *arXiv*, **2020**. Preprint paper.
- [228] Burger, B.; Maffettone, P.M.; Gusev, V.V.; Aitchison, C.M.; Bai, Y.; Wang, X.; Li, X.; Alston, B.M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R.S.; Cooper, A.I. A mobile robotic chemist. *Nature*, **2020**, *583*(7815), 237-241. <http://dx.doi.org/10.1038/s41586-020-2442-2> PMID: 32641813
- [229] Coley, C.W.; Thomas, D.A., III; Lummiss, J.A.M.; Jaworski, J.N.; Breen, C.P.; Schultz, V.; Hart, T.; Fishman, J.S.; Rogers, L.; Gao, H.; Hicklin, R.W.; Plehiers, P.P.; Byington, J.; Piotti, J.S.; Green, W.H.; Hart, A.J.; Jamison, T.F.; Jensen, K.F. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science*, **2019**, *365*(6453), aax1566. <http://dx.doi.org/10.1126/science.aax1566> PMID: 31395756
- [230] Sonnenburg, S.; Braun, M.L.; Ong, C.S.; Bengio, S.; Bottou, L.; Holmes, G.; LeCun, Y.; Muller, K.-R.; Pereira, F.; Rasmussen, C.E. The need for open source software in machine learning Klaus-Robert M " uller Gunnar Rätsch. *J. Mach. Learn. Res.*, **2007**, *8*, 2443-2466.
- [231] Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: concept and applications. *ACM Trans. Intell. Syst. Technol.*, **2019**, *10*(2), 1-19. <http://dx.doi.org/10.1145/3298981>
- [232] Burki, T. Pharma blockchains AI for drug development. *Lancet*, **2019**, *393*(10189), 2382. [http://dx.doi.org/10.1016/S0140-6736\(19\)31401-1](http://dx.doi.org/10.1016/S0140-6736(19)31401-1) PMID: 31204669
- [233] Cieplinski, T.; Danel, T.; Podlowska, S.; Jastrzebski, S. We should at least be able to design molecules that dock well. *arXiv*, **2020**. Preprint paper.
- [234] van den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural discrete representation learning. *arXiv*, **2017**. Preprint paper.
- [235] Razavi, A.; van den Oord, A.; Poole, B.; Vinyals, O. Preventing posterior collapse with delta-VAEs. *arXiv*, **2019**. Preprint paper.
- [236] Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.*, **2020**, *2*(10), 573-584. <http://dx.doi.org/10.1038/s42256-020-00236-4>
- [237] Preuer, K.; Klambauer, G.; Rippmann, F.; Hochreiter, S.; Unterthiner, T. Interpretable deep learning in drug discovery. In: *Explainable, A.I. Interpreting, Explaining and Visualizing Deep Learning*; Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L.K.; Müller, L.R., Eds.; Springer: Cham, **2019**; 11700, pp. 331-345. http://dx.doi.org/10.1007/978-3-030-28954-6_18
- [238] Bengio, Y.; Courville, A.; Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2013**, *35*(8), 1798-1828. <http://dx.doi.org/10.1109/TPAMI.2013.50> PMID: 23787338
- [239] Higgins, I.; Matthey, L.; Pal, A.; Burgees, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-VAE: learning basic visual concepts with a constrained variational framework. *Int. Conf. Learn. Represent.*, **2017**.
- [240] Jin, W.; Barzilay, R.; Jaakkola, T. Multi-objective molecule generation using interpretable substructures. *arXiv*, **2020**.
- [241] Sercu, T.; Gehrmann, S.; Strobelt, H.; Das, P.; Padhi, I.; Santos, C.D.; Wadhawan, K.; Chenthamarakshan, V. Interactive visual exploration of latent space (IVELS) for peptide auto-encoder model selection. *ICLR Workshop about Deep Generative Models for Highly Structured Data*, **2019**.
- [242] Bjerrum, E.J.; Sattarov, B. Improving chemical autoencoder latent space and molecular *de novo* generation diversity with heteroencoders. *Biomolecules*, **2018**, *8*(4), 131. <http://dx.doi.org/10.3390/biom8040131> PMID: 30380783
- [243] Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.;

- Chen, H. Application of generative autoencoder in *de novo* molecular design. *Mol. Inform.*, **2018**, 37(1-2), 1700123. <http://dx.doi.org/10.1002/minf.201700123> PMID: 29235269
- [244] Benhenda, M. ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? *arXiv*, **2017**.
- [245] Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *J. Chem. Inf. Model.*, **2018**, 58(9), 1736-1741. <http://dx.doi.org/10.1021/acs.jcim.8b00234> PMID: 30118593
- [246] Renz, P.; Van Rompaey, D.; Wegner, J.K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discov. Today. Technol.*, **2019**, 32-33, 55-63. <http://dx.doi.org/10.1016/j.ddtec.2020.09.003> PMID: 33386095
- [247] Coley, C.W.; Eyke, N.S.; Jensen, K.F. Autonomous discovery in the chemical sciences part ii: outlook. *Angew. Chem. Int. Ed. Engl.*, **2020**, 59(52), 23414-23436. <http://dx.doi.org/10.1002/anie.201909989> PMID: 31553509
- [248] Koes, D.R.; Baumgartner, M.P.; Camacho, C.J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.*, **2013**, 53(8), 1893-1904. <http://dx.doi.org/10.1021/ci300604z> PMID: 23379370
- [249] Landrum, G. The RDKit Documentation - The RDKit 2020.09.1 documentation. **2020**. Available at : <http://www.rdkit.org/docs/> (Accessed date: Mar 16, 2021).