Chapter 2

Molecular Representations and Computational Methods in Drug Discovery

2.1 Molecular Representation

2.1.1 SMILES

The simplified molecular-input line-entry system (SMILES) [198, 199] is a widely-used text-based description of molecular structure. In SMILES strings, atoms are represented with their chemical symbols and aromatic atoms are denoted in lowercase (see Table 2.1 for examples). Single and aromatic bonds are omitted while for double and triple bonds the specials characters = and # are used. Branches are specified by enclosing them into parentheses. To encode cyclic structures a single bond in the ring is broken and the matching atoms are denoted by numbers. @ characters are used to denote chirality while \ and / characters specify local double bond configurations. Following these rules, a SMILES string is constructed by traversing the nodes of the molecular graph. Depending on the choice of starting node and traversal route there are often multiple valid SMILES representations per molecule, especially for larger molecules. In order to define a single unique SMILES representation for a molecule, known as the 'canonical' SMILES, a deterministic algorithm is used to choose the starting node and traversal route.

Reaction SMILES are a simple extension of SMILES for specifying chemical reactions. Reaction SMILES strings constructed by placing a > character between the SMILES strings of reactants, reagents, and products. If multiple molecules participate in the reactio, their SMILES strings are separated by a period (.) character.

The text-based nature of SMILES strings as well as its expressiveness in encoding the molecular graph alongside stereochemistry results in its widespread use for storing chemical data. In the context of machine learning, the vast majority of molecular datasets where ML

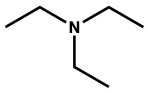
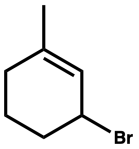
SMILES	Structure
C	CH_4
[Fe2+]	Fe^{2+}
C=O	CH_2O
C#N	HCN (cyan)
CCN(CC)CC	
CC1=CC(CCC1)Br	

Table 2.1 Demonstration of the SMILES language

models are used will have molecules represented as SMILES strings. For example, the ESOL dataset consists of 1128 SMILES strings alongside the measured solubility value for each molecule [39], while USPTO consists of 480k reaction SMILES strings [108]. For text-based ML models such as the Molecular Transformer (see chapter 5), the SMILES strings are directly input to the model, while for other types of models the SMILES strings will be further processed to generate the necessary input features.

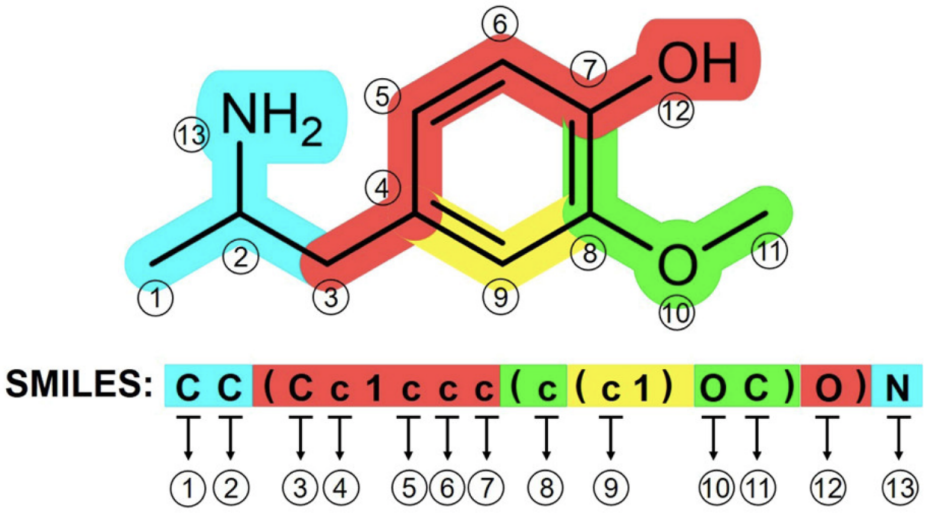


Fig. 2.1 **Illustration of the mapping from chemical structure to SMILES.** Adapted from [89].

While SMILES is by far the most widely used text-based representation of molecules, other representations have been developed and are in use to address some shortcomings in SMILES. For example, the International Chemical Identifier (InChI) [63] string representation, which

has a hierarchical construction for specifying tautomeric/stereochemical/charge states, allows greater precision and flexibility in querying molecules from large chemical databases. Another example is SELF-referencIng Embedded Strings (SELFIES) [96] which is constructed such that every SELFIES string, including random combinations of characters, is a valid molecule. This property is useful for the application of ML models that generate text as output - using SELFIES as the molecular representation, the model always output valid molecules whereas with SMILES that is not guaranteed.

2.1.2 Molecular Substructures

Given a dataset of molecules or chemical reactions encoded with SMILES, we often want to identify molecules or reactions that contain a specific substructure. For example, we may want to identify molecules that contain a specific functional group or reaction that contains a specific reaction center. The standard tool for performing these substructure queries is via SMILES Arbitrary Target Specification (SMARTS) notation [SMA]. The SMARTS line notation is expressive and allows extremely precise and transparent substructural specification and atom typing.

Using many of the same symbols as SMILES, it also allows specification of wildcard atoms and bonds, which allows expressive and precise definitons of substructures and atomic environments for searching chemical databases. One common misconception is that SMARTS-based substructural searching involves matching of SMILES and SMARTS strings. When performing a SMARTS query on a SMILES string, both SMILES and SMARTS strings are first converted to internal graph representations which are then searched for subgraph isomorphism.

SMARTS	Substructure
[C;R]	An aliphatic carbon in a ring
[#6]@[#6]	Two carbons connected by a ring bond
[N;$(NC=[O,S])]	amide or thioamide nitrogen
[N:1][C:2](=[O:3])[N:4]»[N:1][C:2](=[O:3])[C:4]	urea group transforming into an amide

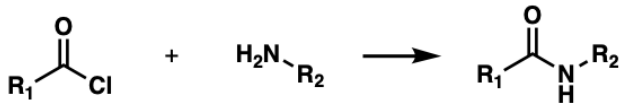
Table 2.2 Examples of SMARTS patterns

The precise substructure specification of SMARTS is useful in many aspects in the drug design process. For example, a common step in assessing the quality of a proposed drug candidate is to perform a SMARTS query to identify if the hit contains any substructures that are likely to produce artifacts in biochemical or cellular assays. These substructures are typically functional groups with a marked propensity to bind to multiple targets, so-called nuisance compounds, which are of little value in drug discovery. Many different sets of these

filters have been compiled in the literature such as REOS (rapid elimination of swill) [194] and PAINS (Pan Assay Interference Compounds) Filters [7]. Similarly, SMARTS queries are used to design 'structural alerts' which flag molecules containing reaction chemical substructures which may lead to undesirable toxicity in the compound itself or its metabolites [102].

Another use of SMARTS is in the labelling of pharmacophores in a molecule. Pharmacophores are an abstract description of the molecular features involved in ligand binding - typical examples of such features are hydrogen bond acceptors/donors and aromatic rings. SMARTS strings are used to map different molecular substructures to particular pharmacophore features and then to query for the presence of these features in a molecule. As with substructure filtering, different companies in the pharmaceutical industry have different/propritary sets of SMARTS strings tailored for their particular use cases. (see section 2.1.3 for details)

Beyond substructures for individual molecules, SMARTS can also be applied to reaction SMILES to capture transformation in substructures. These SMARTS strings for chemical reactions are often referred to in the literature as 'reaction templates' (Fig 2.2). Beyond querrying for matching reactions from a dataset, reaction templates can also be directly applied to a set of molecules to computationally generate a 'reaction product'. This approach is used to generate virtual libraries [195, 153] *in silico* both for more specific usecases such as probing the SAR of a compound as well as the construction of ultra-large commercially available compound libraries eg ZINC [75], EnamineREAL.



`[*:1][C:2](=[O:3])Cl.[N;H2:4][*:5]>>[*:1][C:2](=[O:3])[N;H1:4][*:5]`

Fig. 2.2 An example of a reaction template for the synthesis of an amide from an acid chloride and a primary amine.

In addition to virtual library construction, reaction templates can be used for organic reaction prediction by framing the problem as trying to predict the correct reaction template for a given set of reactants. Starting from a catalogue of possible reaction templates, the best matching general template in the catalogue can be found utilising subgraph searching or machine learning, and applied on the input to obtain the predicted outcome of the reaction. This approach was originally proposed for the reverse problem of retrosynthesis [35] and has had success in forward reaction prediction for the design of synthetic pathways to drug-like molecules [94].

One major limitation of template-based approaches is scalability, as the template library needs to be maintained and updated every time a new reaction is reported. A further problem is that it is often not obvious which parts of the molecule are crucial for a given reaction. This means that given a reaction one can derive a smaller more general template or a larger one that is more specific for the particular reaction. This results in either too many templates matching a particular input resulting in many equally possible reaction outcomes or in the case of larger more specific templates the library will grow very big which results in very slow predictions.

2.1.3 Pharmacophores

A pharmacophore is an abstract description of molecular features that are necessary for molecular recognition of a ligand by a biological macromolecule [87]. A collection of pharmacophores in a geometric configuration is known as a 'pharmacophore model' and it is a representation of the interactions between ligands and the binding site. By their coarse-grained nature, pharmacophore model can explain structurally diverse ligands can bind to a common receptor site, and can be used to identify novel ligands that will bind to the same receptor.

Typical pharmacophore features, characterised by SMARTS strings, include hydrophobic groups, aromatic rings, hydrogen bond acceptors or donors, and charged functional groups.

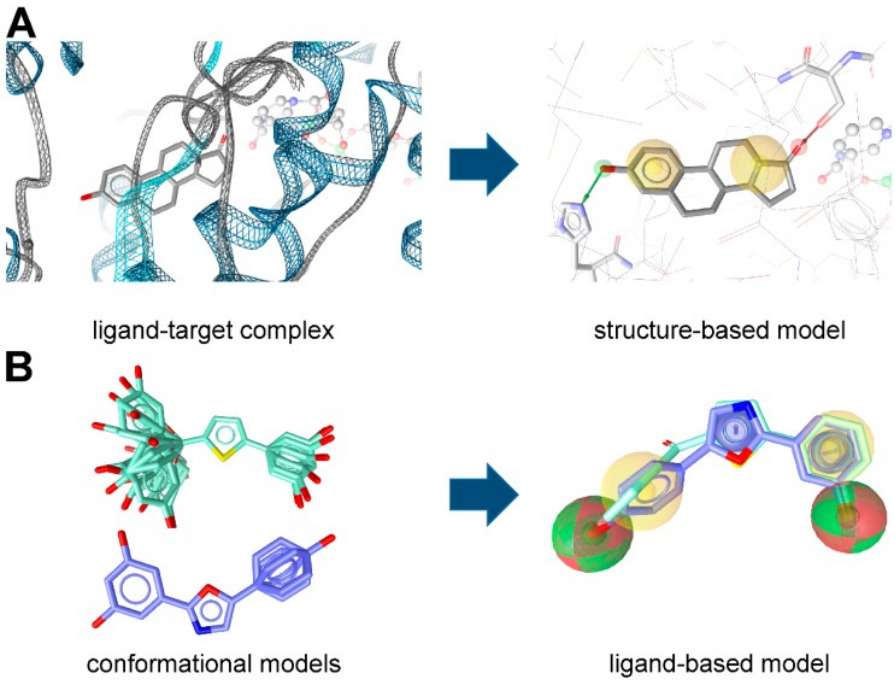


Fig. 2.3 An illustration of structure-based or ligand-based pharmacophore models. Reproduced from [87].

Pharmacophore models can either be constructed using structural data or from a set of active compounds [193]. In the structure-based approach, the pharmacophores are directly inferred from the observed interactions of a molecule and the binding site in experimentally determined ligand-protein complexes. (Fig 2.3A).

In the ligand-based approach, the three-dimensional (3D) structures of known active molecules are aligned and pharmacophores that are found to overlap in space are extracted as the pharmacophore model (Figure 2.3B). Although this approach circumvents the need for structural data which might not be possible to obtain, a downside is that all of the extracted pharmacophore have to be presumed as essential for protein-ligand binding, whereas in the structure-based approach it is possible to identify and discard non-important pharmacophores.

After obtaining a pharmacophore model, it can be used to virtually screen ligands from a database by identifying molecules that share similar pharmacophore features. This approach is known as 'pharmacophore-based virtual screening' and is a useful tool on its own as well as for complementing molecular docking and machine learning approaches [40, 179, 144, 134].

2.1.4 Fingerprints

In order to apply powerful machine learning methods, we require a (typically fixed-length) vector representation of molecules konwn in the literature as 'molecular fingerprints'. The most popular molecular fingerprint is the Morgan fingerprint [124], also known as the Extended-Connectivity FingerPrint (ECFP) [150]. ECFPs are a particular example of 'topological' fingerprints that encode the presence of substructures in a molecule by traversing the molecular graph.

ECFPs are generated via a recursive hashing algorithm that numerically hashes the representation of each atom with those of its neighbours, and again with it's next-nearest neighbours etc until a pre-defined 'radius' is reached. The resulting hash values are then used to generate a fixed-length binary vector of 1/0 bits. The length of the vector is pre-determined and each 1-bit represents a unique substructure that is encountered during the traversal. The radius of the graph traversal is also a pre-defined parameter that controls the size of the substructures that are represented in the fingerprint. The radius is typically set to 2 or 3, and the length of the fingerprint is typically in the range of 1024-4096.

The popularity of the Morgan fingerprint owes to its usefullness in calculating molecular similarity [113]. Intuitively, we would expect two molecules that have 'similar' molecular fingerprints to have similar chemical structures. Numerically, we can quantify the similarity between two molecular fingerprints by the Tanimoto coefficient [200]:

$$\text{Tanimoto}(A,B) = \frac{A \cap B}{A \cup B} = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B} \tag{2.1}$$

where A and B refer to the bit-vector molecular fingerprints of two molecules. The numerator $A \cdot B$ represents the number of bits shared between the two fingerprints, while the denominator represents the total number of unique bits covered by the finerprints. Two structures are usually considered similar if the tanimoto similarity is > 0.4 [9]. Alternate similarity measures exist but a number of comparison studies [185, 8] have shown the Tanimoto similarity to be generally robust and consistently perform well in a variety of applications.

The combination of the Morgan fingerprint with Tanimoto similarity is useful for clustering [24] datasets of similar compounds as well as performing similarity-based virtual screening. Similarity-based virtual screening relies on the similarity property principle (SPP) [83], which states that similar compounds should have similar biological activity. As a guiding strategy this means one could search a database for similar compounds to a known active molecule, and expect those compounds to retain and perhaps have improved biological activity against a target. Although this hypothesis is not always valid in cases known as 'activity cliffs' where small changes in structure cause large changes in biological activity [114], empirically it has been shown that structurally similar compounds are much more likely to be actives compared to dissimilar ones [115]. Performing similarity-based virtual screening in practice involves calculating molecular similarities between known active compounds and unkonwn molecules from a database, then selecting those with the highest similarities; ECFP4 is a consistently well-performing fingerprint for this task [149, 131].

In addition, Morgan fingerprints have also been shown to be versatile as a molecular descriptor for machine learning (ML). Machine learning models learn statistical patterns from data and can be used to make predictions on new data (see section 2.2.2). In the context of drug discovery, ML can be used associate patterns in the molecular fingerprints of the molecules in a dataset with experimentally measured properties. For example, fingerprint-based models have demonstrated success in predicting physical/chemical properties such as solubility [202], biological activities [37] as well as yields and stereoselectivities for chemical reactions [154]. Non-fingerprint based deep learning methods have recently been developed that learn molecular representations directly from molecular graphs, however ECFP-based shallow learning techniques continue to provide a strong, robust baseline to compare against.

2.2 Computational Approaches

2.2.1 Docking

Molecular docking is the process of predicting the binding mode of a small molecule to a protein target, and one of the most frequently used methods in structure-based drug design. [121, 92] The binding mode is the relative orientation of the small molecule in a particular binding site of the protein, which is determined by the shapes of the binding site and molecule, and the physical interactions between the two. The binding mode has a large effect on the strength of the interaction between the small molecule and the protein, known as the binding affinity. The binding affinity, in turn, is a key determinant of the biological activity of the small molecule. The philosophy of structure-based drug design is to experimentally obtain binding modes of molecules and use this information to guide the design of new molecules by docking and choosing molecules that may have more optimal protein-ligand interactions and hence binding affinity.

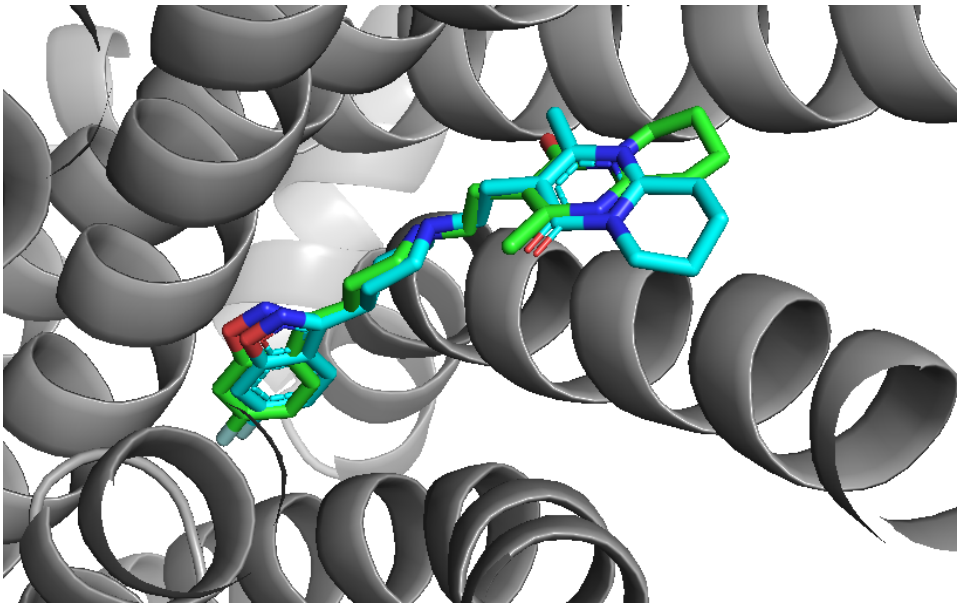


Fig. 2.4 **Example of a docked molecule.** The experimental structure of the ligand risperidone bound to the serotonin 2A receptor is shown in green, with the protein in grey (PDB: 6A93). The structure of the same ligand docked using GOLD is shown in cyan.

The physics of ligand-protein binding are complex, and in reality each ligand will have an ensemble of binding modes. Attempting to accurately simulate the binding process is computationally intractable, and so the goal of molecular docking is to predict the most likely binding mode. In practice this is approached as an opimisation problem, where the coordinates of the ligand and/or protein atoms are adjusted until the 'best-fit' is acheived.

An essential preliminary step to performing molecular docking is obtaining a structure of the protein of interest. Traditionally this means using biophysical techniques such as X-ray crystallography, NMR spectroscopy or cryo-electron microscopy (cryo-EM), but recent development in computational protein structure prediction [85, 201] open the door to performing fully *in-silico* structure-based drug design.

Every docking approach is essentially composed of two parts - conformation scoring and conformation searching. Potential ligand poses are ranked using a scoring function, which are typically physics-based molecular mechanics force fields that estimate the energy of the pose within the binding site. The scoring function can be composed of many components, such as electrostatic interactions, solvent and steric effects, hydrogen bonding, as well as knowledge-based potentials derived from observed interactions from databases of protein-ligand structures [101]. While the accuracy of scoring functions has to be good enough to distinguish good poses from bad ones, major emphasis is put on computational efficiency due to the large number of evaluations required during docking. Thus, scoring functions often involve many assumptions and simplifications to reduce computational costs.

The search space of conformations is imposible to exhaustively explore as in theory it consists of all possible orientations and configurations of the protein paired with the ligand. In practice, usually the whole conformational space of the ligand is searched, while the protein is often treated rigidly. Exploration of the conformational search space is often done using stochastic methods such as Monte Carlo or genetic algorithms which randomly sample the space of conformation parameters (e.g. torsion angles) towards a minimisation of the scoring function.

The wide range of design choices for the scoring function and conformation search results in a large number of different docking algorithms that are in use in the field, such as DOCK [30], Glide [48], AutoDock Vina [44], GOLD [191], and FRED [120]. The relative performance between these docking algorithms are typically retrospectively evaluated by directly comparing predicted binding poses to known crystal structures of ligand-protein complexes. The benchmark datasets used for this purpose are typically high-quality structures of drug-like molecules such as PDBbind [196, 105]. There are also community assessments on the relative prospective performance of different docking approaches [136] and scoring functions [175].

Besides the structural focus of binding pose prediction, increasingly in recent years docking has been used to directly virtually screen large databases of molecules in silico to identify molecules that are likely to bind to protein target of interest [14]. This approach puts the focus on the scoring function, with the rationale that molecules with high docking scores are much more likely to be active than those with low scores. In this scenario, success is defined by

the enrichment of active compounds in the top ranks of a docking screen, measured via the enrichment factor:

$$\mathrm{EF}(n) = \frac{\mathrm{Hit\,rate(predicted\,top}{-}n)}{\mathrm{Hit\,rate(baseline)}} \tag{2.2}$$

where the baseline hit rate is the proportion of actives in the dataset overall, representing the performance of simple random ordering. Different methods are benchmarked by retrospectively evaluating the enrichment factor of known ligands from a large database of presumed non-binding, "decoy" molecules for multiple protein targets - the classic benchmark dataset for this is the Directory of Useful Decoys (DUD) [68, 129].

Prospectively, large-scale virtual screening with molecular docking has had notable successes. A review specifically looking at G protein-coupled receptors (GPCRs) [10], which are the target of more than 30% of all marketed drugs, showed 62 successful virtual screens for 22 unique protein targets belonging to 14 different receptor families in the past decade. Of particular note is that increasing availability of computational resources, together with increases in the sizes of commercially-available make-on-demand compound libraries, has made possible ultra-large virtual screening campaigns against libraries of >100 million compounds [111, 5, 46]. At the same time, limitations in the accuracy of scoring functions and in the modelling of protein flexibility [45, 6] restrict the ability of docking to reliably distinguish active molecules from inactive ones [106, 112], leading to false positives which are exacerbated when screening large libraries [110].

In the absence of exisiting ligand bioactivity measurements for a protein target, virtual screening with molecular docking remains the only computational method of choice and is a default starting point for beginning drug discovery against a brand new target. However, there has been recent results claiming that deep learning models which use neural networks to directly generate ligand binding poses [173, 36] outperform docking algorithms in terms of accuracy. These results are promising and, coupled with continued advancements in protein structure prediction to account for protein flexibility, suggest that deep learning may be a viable alternative to molecular docking for binding pose prediction and virtual screening in the coming years.

2.2.2 Machine Learning

Machine learning (ML) refers to the use of algorithms that 'learns' how to make predictions or decisions based on observed data. By designing models that can learn patterns directly from input data, it is found that ML methods can often surpass manually created algorithms by humans on a wide variety of tasks. In the following paragraphs, we provide an overview

of the ML concepts and ideas needed to understand the research presented in this thesis. We will focus on supervised learning and neglect many important sub-fields such as reinforcement learning and generative models which are more described in greater detail in refs [15, 55].

Very broadly, the aim of machine learning is to learn the parameters θ of a predicative model $y = f(x,\theta)$ that minimise a given cost function, $\mathscr{L}(y,\hat{y})$, where x is a given input, y is the target variable and $\hat{y}$ is the predicted value, i.e. to find the solution

$$\hat{\theta} = \arg\min_{\theta} \mathscr{L}(\theta) \tag{2.3}$$

For regression, which is the modelling of a continuous variable, the most common loss function choice is the squared residuals,

$$\mathscr{L}(y,\hat{y}) = \sum_i (y_i - \hat{y}_i)^2 \tag{2.4}$$

while for binary classification, which is the task of predicting which class an input x belongs to, the most common loss function is the binary cross-entropy loss:

$$\mathscr{L}(y,\hat{y}) = -\sum_i y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \tag{2.5}$$

where the target variable y can be either 0 or 1 while the predicted value $\hat{y}_i$ is the predicted probability of class 1 and $1-\hat{y}_i$ is the predicted probability of class 0.

To find the solution $\hat{\theta}$ for a dataset in practice, we would first divide the input data into training, validation and test sets. The model is initially fit on the training data set, where the data-dependent parameters of the model (e.g. the coefficients of a polynomial regression model) are optimised to minimise the loss function. Afterwards, the fitted model is used to make predictions on the validation data set. The validation data set provides an unbiased estimate of the model's performance on the training data set for the purpose of tuning the non-data-dependent parameters, known as 'hyperparameters', of a model (e.g. the number of degrees to include in a polynomial regression model). This process may be repeated multiple times, with the model's performance on the validation data set being used to select the best hyperparameters. This overall process is known as 'training a model'.

After a model has been trained we use the test dataset, which has never been seen by the model during training, to evaluate the performance of the model. It is important to use the same training and test datasets for a fair comparison of different models, and curated datasets from the literature are commonly used as a benchmark for evaluating the performance of new models.

For regression models, the most common metric used to evaluate the performance of a model is the root mean squared error (RMSE) or the pearson correlation coefficient (PCC). For binary classification models, the most common metric used is the area under the receiver operating characteristic curve (AUC). The receiver operating characteristic (ROC) curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) as the discrimination threshold for classifying one class over the other is varied 2.5. The AUC is the area under the ROC curve, and is a measure of the model's ability to distinguish between the two classes. AUC values range from 0 to 1, with 0.5 indicating a model that is no better than random guessing, and 1 indicating a perfect model.

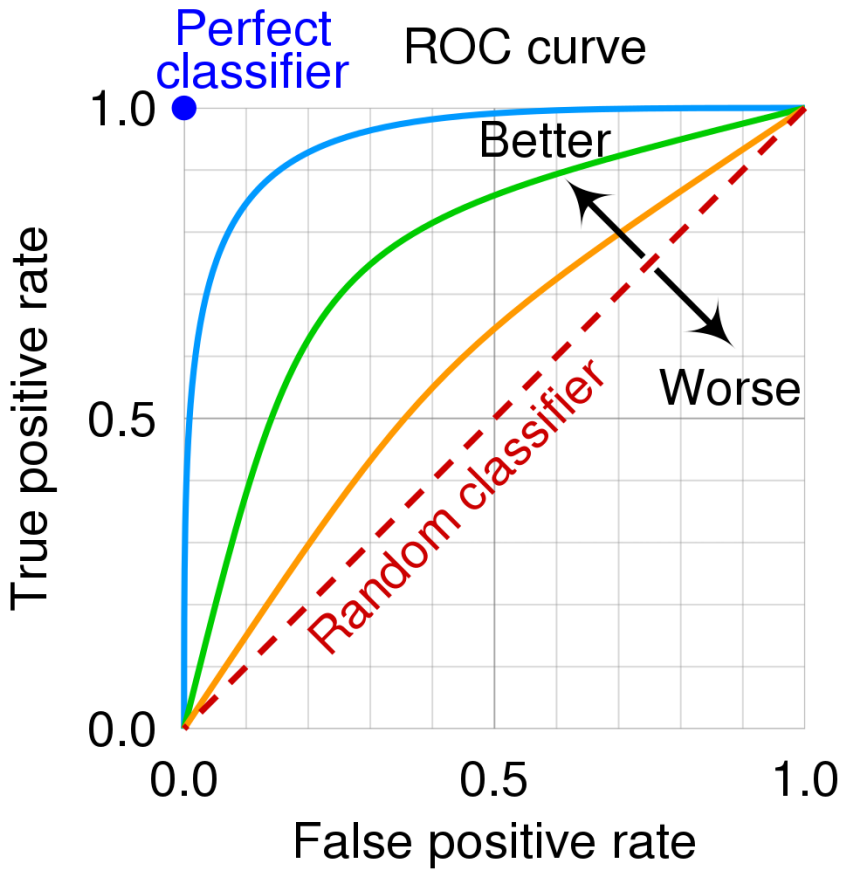


Fig. 2.5 An example Receiver Operating Characteristic (ROC) curve. The diagonal shows the performance of a random classifier. Three example classifiers (blue, orange, green) are shown. Reproduced from [34].

A large number of machine learning techniques have been described and applied for drug discovery, and an overview of them can be found in [12, 13, 192]. In this thesis, two common techniques are used: Random forests and Gaussian processes.

Random Forest

Random Forest models are ensemble learning methods that utilise a large number of decision trees for making predictions [21]. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the output is the mean of the predictions by the individual trees:

$$f(x) = \frac{1}{N}\sum_{i}^{N} f_i(x) \tag{2.6}$$

where f_i is the ith tree in the forest and N is the total number of trees in the forest. Decision trees are constructed to succerssively split the data into branches via 'decision boundaries' (e.g. $x > 1.5$). Decision boundaries are chosen to minimise the square deviations (for regression) or information entropy (for classification) between the samples and the sample mean in each branch or leaf of the tree.

Although extremely computationally efficient and interpretable, individual decision trees are very prone to over-fitting. By using a large number of decision trees each trained on different random subsets of the data (a process known as bootstrap aggregation), random forest models can acheive a lower variance and hence improved performance. To further reduce correlation between the decision trees, random forests use random feature selection, where only a random subset of the features are considered for each decision boundary.

Random forests are relatively easy to use, require little tuning of hyperparameters, and are robust to over-fitting, and thus are a popular general machine learning technique. In particular for drug discovery, random forests have been extensively used with molecular fingerprints features for prediction of properties such as solubility [135], biological activity [177, 122], and toxicity [142].

Gaussian Process

Gaussian Process models are a kernel-based method that utilises what is known as the 'kernel trick' to calculate high-dimensional weighted averages [147]. The kernel trick is the use of a kernel function to calculate the inner product between the features vectors of two datapoints in a high-dimensional feature space without explicitly computing higher-dimensional feature vectors. A typical kernel function is the squared exponential kernel:

$$\mathscr{K}(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2l^2}\right) \tag{2.7}$$

where l is the length scale of the kernel, a hyperparameter that determines how quickly the function can change, and x and x' are the feature vectors of the datapoints in the initial

low-dimensional feature space. Mathematically, the kernel trick enables the construction of models that are, in theory, infinitely complicated with a finite amount of computation [66]. Gaussian Processes utilise the kernel function to calculate the covariance matrix of the Gaussian distribution over the function values. The covariance matrix is then used to calculate the posterior distribution over the function values given the training data, which can then be used to make predictions on new data with associated uncertainty values.

The fact that GPs have few hyperparameters to tune and maintain uncertainty estimates over property values have also led to their use for molecular property prediction [132, 119, 84], in particular incorporating the Tanimoto similarity in the kernel function [178, 57].

Deep Learning

In contrast to methods which use hand-crafted features as input ('shallow learning'), deep learning revolves around learning representations directly from the raw data using neural networks. Neural networks are composed of layers of 'neurons' that successively perform non-linear transformations on their inputs, mimicking the way that biological neurons transfer signals to one another. These transformations are typically of the form:

$$\mathbf{h} = \sigma(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}) \tag{2.8}$$

where $\mathbf{W}$ is a weight matrix, $\mathbf{x}$ is a vector of inputs, $\mathbf{b}$ is a vector of biases and σ is an optional non-linear activation function. The output of the layer is the vector $\mathbf{h}$, which is either input to the next layer or taken as the output of the model. The weights $\mathbf{W}$ and biases $\mathbf{b}$ from all of the layers collectively are the parameters θ of the neural network that are learned by fitting on data.

Deep learning has found remarkable success in a wide range of applications, including computer vision [146], natural language processing [23], speech recognition [160], and bioinformatics [85, 156]. This is because of the ability of neural networks to learn complex, non-linear relationships between inputs and outputs in the presence of large amounts of data. 'Big data' domains are computationally intractable for shallow learning methods, but deep learning can be successfully applied as neural networks can be optimised effectively using gradient-based approaches, such as gradient descent:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathscr{L} \tag{2.9}$$

where at each step t the model's parameters are updated according to the learning rate η, a hyperparameter of the optimiser. Instead of calculating the gradient of the loss $\nabla_\theta \mathscr{L}$ on the full training set, standard practice is to use a stochastic approximation of the gradient

that is calculated from a randomly sampled batch of training data. This significantly speeds up the optimisation process and allows neural networks to be trained on large datasets that would otherwise be intractable. These stochastic gradient steps are iterated repeatedly over the training set until the value of the loss has satisfactorily converged.

The gradient of the loss function with respect to the model parameters can be obtained efficiently by applying the chain rule via a process called 'backpropogation'. Using automatic differentiation frameworks that can be carried out on hardware accelerators, such as graphical processing units (GPUs), the time needed to train neural networks are dramatically reduced [11]. Further improvements to model optimisation can be achieved by incorporating more sophisticated optimisation algorithms, such as Adam [90], as well as the use of regularisation techniques such as dropout [171] and batch normalisation [74], and remains a significant area of active research.

The only constraint on the design of a neural network is that the mathematical operations in the model must have defined derivatives so that the gradient of the loss function can be calculated with backpropagation for efficient training. This results in a zoo of different neural network designs (referred to as 'architectures') that use differentiable building blocks with specific inductive biases tailored to the task at hand. For example, convolutional neural networks (CNNs) utilise built-in translational invariance for computer vision applications (e.g. AlexNet [97], and ResNets [62]) while recurrent neural networks (RNNs) designed for learning temporal dependencies are applied to sequential data such as text [26] and speech [103] (e.g. Gated-Recurrent-Unit (GRU) [28] and Long-Short-Term-Memory (LSTM) networks [65]).

In the context of drug discovery, the challenge of modelling molecular inputs has led to the devlopment of graph-based neural network architectures known as 'graph neural networks' (GNNs). These models are designed to learn representations of molecular graphs that are invariant to the order of the nodes and edges. GNNs have been successfully applied to a range of molecular tasks, including molecular property prediction [202, 52, 116, 203], and predicting reaction templates for input reactants [33]. More standard architectures such as Transformer models developed for natural language processing have also been applied to SMILES, as well as three-dimensional voxel-based CNNs which can be trained on protein-ligand complexes for the prediction of binding affinity [145, 71, 78].

Neural networks can also be used with pre-computed features such as molecular fingerprints. Example applications include the use of bioactivity prediction [37], reaction prediction [197, 169], and the prediction of docking scores [51].

Despite their success in certain molecular tasks, deep learning still has several limitations when it comes to drug discovery. Chief among these is the need for large amounts of training data for strong performance, which can often be costly and time-consuming to obtain. When

only a small amount of data is available, which is typical in the early stages of drug discovery, neural networks may perform worse than simpler shallow learning models [77].

Additionally, neural networks may struggle to generalize to new molecules that are substantially different from the molecules in the training set. This is known as the 'generalization gap' and model performance with typical random-split cross-validation procedures do not accurately reflect the true generalization performance of the model [170]. This has led to a growing movement towards measuring model performance using 'scaffold split' [202, 203] or 'time split' [170] cross-validation, which splits the data into disjoint sets of molecules that are similar in structure or time of data acquisition, respectively. This allows for a more accurate assessment of the generalization gap, but this remains a challenge for applying deep learning and machine learning models in general in drug discovery.

Another challenge is the lack of interpretability of neural network models, which make it difficult to understand the underlying reasons for a model's predictions. Without explainations of model predictions, it becomes difficult to avoid correct predictions for the wrong reasons (the so-called clever Hans effect) [99], avert unfair biases, and gain potentially useful insights from the model. This is a challenge in general for deep learning, but is particularly difficult in drug discovery due to the domain-specific complication of projecting 'explanations' onto molecule representations [79].

Accounting for these limitations are critical for realizing the full potential of applying both shallow and deep learning models to accelerate the design-make-test cycle in drug discovery.

References

[Che] Chemspace: Lead-like compounds.

[SMA] Smarts - a language for describing molecular patterns.

[3] Agarwal, S., Dugar, D., and Sengupta, S. (2010). Ranking chemical structures for drug discovery: a new machine learning approach. *Journal of chemical information and modeling*, 50(5):716–731.

[4] Allen, T. E. H., Wedlake, A. J., Gelžinytė, E., Gong, C., Goodman, J. M., Gutsell, S., and Russell, P. J. (2020). Neural network activation similarity: a new measure to assist decision making in chemical toxicology. *Chem. Sci.*, 11:7335–7348.

[5] Alon, A., Lyu, J., Braz, J. M., Tummino, T. A., Craik, V., O'Meara, M. J., Webb, C. M., Radchenko, D. S., Moroz, Y. S., Huang, X.-P., Liu, Y., Roth, B. L., Irwin, J. J., Basbaum, A. I., Shoichet, B. K., and Kruse, A. C. (2021). Structures of the

$$\sigma$$

2 receptor enable docking for bioactive ligand discovery. *Nature*, 600(7890):759–764.

[6] Antunes, D. A., Devaurs, D., and Kavraki, L. E. (2015). Understanding the challenges of protein flexibility in drug design. *Expert Opinion on Drug Discovery*, 10(12):1301–1313.

[7] Baell, J. B. and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740. PMID: 20131845.

[8] Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20.

[9] Baldi, P. and Nasr, R. (2010). When is chemical similarity significant? the statistical distribution of chemical similarity scores and its extreme values. *Journal of Chemical Information and Modeling*, 50(7):1205–1222.

[10] Ballante, F., Kooistra, A. J., Kampen, S., de Graaf, C., and Carlsson, J. (2021). Structure-based virtual screening for ligands of g protein–coupled receptors: What can molecular docking do for you? *Pharmacological Reviews*, 73(4):1698–1736.

[11] Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43.

[12] Bender, A. and Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today*, 26(2):511–524.

[13] Bender, A. and Cortes-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? part 2: a discussion of chemical and biological data. *Drug Discovery Today*, 26(4):1040–1052.

[14] Bender, B. J., Gahbauer, S., Luttens, A., Lyu, J., Webb, C. M., Stein, R. M., Fink, E. A., Balius, T. E., Carlsson, J., Irwin, J. J., and Shoichet, B. K. (2021). A practical guide to large-scale docking. *Nature Protocols*, 16(10):4799–4832.

[15] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[16] Bjerrum, E. J. (2017). Smiles enumeration as data augmentation for neural network modeling of molecules.

[17] Blakemore, D. C., Castro, L., Churcher, I., Rees, D. C., Thomas, A. W., Wilson, D. M., and Wood, A. (2018). Organic synthesis provides opportunities to transform drug discovery. *Nature chemistry*, 10(4):383.

[18] Boström, J., Brown, D. G., Young, R. J., and Keserü, G. M. (2018). Expanding the medicinal chemistry synthetic toolbox. *Nature Reviews Drug Discovery*.

[19] Botev, Z. I., Grotowski, J. F., and Kroese, D. P. (2010). Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916 – 2957.

[20] Bradshaw, J., Kusner, M. J., Paige, B., Segler, M. H. S., and Hernández-Lobato, J. M. (2019). A generative model for electron paths.

[21] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[22] Brown, N., McKay, B., Gilardoni, F., and Gasteiger, J. (2004). A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *Journal of chemical information and computer sciences*, 44(3):1079–1087.

[23] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

[24] Butina, D. (1999). Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750.

[25] Cannalire, R., Cerchia, C., Beccari, A. R., Di Leva, F. S., and Summa, V. (2020). Targeting sars-cov-2 proteases and polymerase for covid-19 treatment: State of the art and future opportunities. *Journal of medicinal chemistry*.

[26] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

[27] Chodera, J., Lee, A. A., London, N., and von Delft, F. (2020). Crowdsourcing drug discovery for pandemics. *Nature Chemistry*, 12(7):581–581.

[28] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

[29] Clayden, J., Greeves, N., and Warren, S. (2012). *Organic Chemistry*. Oxford University Press, 2nd edition.

[30] Coleman, R. G., Carchia, M., Sterling, T., Irwin, J. J., and Shoichet, B. K. (2013). Ligand pose and orientational sampling in molecular docking. *PLOS ONE*, 8(10):1–19.

[31] Coley, C. W., Eyke, N. S., and Jensen, K. F. (2019a). Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angewandte Chemie - International Edition*, pages 2–38.

[32] Coley, C. W., Green, W. H., and Jensen, K. F. (2018). Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research*, 51(5):1281–1289.

[33] Coley, C. W., Jin, W., Rogers, L., Jamison, T. F., Jaakkola, T. S., Green, W. H., Barzilay, R., and Jensen, K. F. (2019b). A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.*, 10:370–377.

[34] Commons, W. (2018). Receiver operating characteristic (roc) curve. https://commons.wikimedia.org/wiki/File:Roc_curve.svg.

[35] Corey, E. J., Long, A. K., and Rubenstein, S. D. (1985). Computer-assisted analysis in organic synthesis. *Science*, 228(4698):408.

[36] Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. S. (2023). Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*.

[37] Cortés-Ciriano, I. and Bender, A. (2019). Reliable prediction errors for deep neural networks using test-time dropout. *Journal of Chemical Information and Modeling*, 59(7):3330–3339. PMID: 31241929.

[38] Davis, B. J. and Roughley, S. D. (2017). Chapter eleven - fragment-based lead discovery. In Goodnow, R. A., editor, *Platform Technologies in Drug Discovery and Validation*, volume 50 of *Annual Reports in Medicinal Chemistry*, pages 371–439. Academic Press.

[39] Delaney, J. S. (2004). Esol: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005.

[40] Dixon, S. L., Smondyrev, A. M., and Rao, S. N. (2006). Phase: A novel approach to pharmacophore modeling and 3d database searching. *Chemical Biology & Drug Design*, 67(5):370–372.

[41] Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., Brandão-Neto, J., Carbery, A., Davison, G., Dias, A., Downes, T. D., Dunnett, L., Fairhead, M., Firth, J. D., Jones, S. P., Keeley, A., Keserü, G. M., Klein, H. F., Martin, M. P., Noble, M. E. M., O'Brien, P., Powell, A., Reddi, R. N., Skyner, R., Snee, M., Waring, M. J., Wild, C., London, N., von Delft, F., and Walsh, M. A. (2020a). Crystallographic and electrophilic fragment screening of the sars-cov-2 main protease. *Nature Communications*, 11(1):5047.

[42] Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., et al. (2020b). Crystallographic and electrophilic fragment screening of the sars-cov-2 main protease. *Nature communications*, 11(1):1–11.

[43] Duffy, N. P. (2010). Molecular property modeling using ranking. US Patent 7,702,467.

[44] Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. (2021). Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898.

[45] Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A., and Vieth, M. (2004). Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry*, 47(1):45–55.

[46] Fink, E. A., Xu, J., Hübner, H., Braz, J. M., Seemann, P., Avet, C., Craik, V., Weikert, D., Schmidt, M. F., Webb, C. M., Tolmachova, N. A., Moroz, Y. S., Huang, X.-P., Kalyanaraman, C., Gahbauer, S., Chen, G., Liu, Z., Jacobson, M. P., Irwin, J. J., Bouvier, M., Du, Y., Shoichet, B. K., Basbaum, A. I., and Gmeiner, P. (2022). Structure-based discovery of nonopioid analgesics acting through the α_{2A}-adrenergic receptor. *Science*, 377(6614):eabn7065.

[47] Friedel, C. and Crafts, J. (1877). Sur une nouvelle méthode générale de synthèse d'hydrocarbures, d'acétones, etc.

[48] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749.

[49] Gao, W. and Coley, C. W. (2020). The synthesizability of molecules proposed by generative models. *Journal of Chemical Information and Modeling*, 60(12):5714–5723.

[50] Gehrtz, P., Marom, S., Bührmann, M., Hardick, J., Kleinbölting, S., Shraga, A., Dubiella, C., Gabizon, R., Wiese, J. N., Müller, M. P., Cohen, G., Babaev, I., Shurrush, K., Avram, L., Resnick, E., Barr, H., Rauh, D., and London, N. (2022). Optimization of covalent mkk7 inhibitors via crude nanomole-scale libraries. *Journal of Medicinal Chemistry*, 65(15):10341–10356.

[51] Gentile, F., Agrawal, V., Hsing, M., Ton, A.-T., Ban, F., Norinder, U., Gleave, M. E., and Cherkasov, A. (2020). Deep docking: A deep learning platform for augmentation of structure based drug discovery. *ACS Central Science*, 6(6):939–949.

[52] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1263–1272. JMLR.org.

[53] Gironda-Martínez, A., Donckele, E. J., Samain, F., and Neri, D. (2021). Dna-encoded chemical libraries: A comprehensive review with succesful stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279.

[54] 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., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276.

[55] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[56] 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., and Arthanari, H. (2020). An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668.

[57] Griffiths, R.-R., Greenfield, J. L., Thawani, A. R., Jamasb, A. R., Moss, H. B., Bourached, A., Jones, P., McCorkindale, W., Aldrick, A. A., Fuchter, M. J., and Lee, A. A. (2022). Data-driven discovery of molecular photoswitches with multioutput gaussian processes. *Chem. Sci.*, 13:13541–13551.

[58] Guan, Y., Coley, C. W., Wu, H., Ranasinghe, D., Heid, E., Struble, T. J., Pattanaik, L., Green, W. H., and Jensen, K. F. (2021). Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.*, 12:2198–2208.

[59] Hall, R. J., Murray, C. W., and Verdonk, M. L. (2017). The fragment network: A chemistry recommendation engine built using a graph database. *Journal of Medicinal Chemistry*, 60(14):6440–6450.

[60] Hann, M. M., Leach, A. R., and Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of chemical information and computer sciences*, 41(3):856–864.

[61] Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., and Schneider, G. (2012). Dogs: reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol*, 8(2):e1002380.

[62] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

[63] Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. (2013). Inchi - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1):7.

[64] Hermann, J. C., Chen, Y., Wartchow, C., Menke, J., Gao, L., Gleason, S. K., Haynes, N.-E., Scott, N., Petersen, A., Gabriel, S., Vu, B., George, K. M., Narayanan, A., Li, S. H., Qian, H., Beatini, N., Niu, L., and Gan, Q.-F. (2013). Metal impurities cause false positives in high-throughput screening campaigns. *ACS Medicinal Chemistry Letters*, 4(2):197–200.

[65] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[66] Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220.

[67] Howard, J. et al. (2018). fastai. https://github.com/fastai/fastai.

[68] Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801.

[69] Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249.

[70] Ichihara, O., Barker, J., Law, R. J., and Whittaker, M. (2011). Compound design by fragment-linking. *Molecular Informatics*, 30(4):298–306.

[71] Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2018). Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. *Journal of Chemical Information and Modeling*, 58(11):2319–2330.

[72] Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2020). Deep generative models for 3d linker design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995.

[73] Imrie, F., Hadfield, T. E., Bradley, A. R., and Deane, C. M. (2021). Deep generative design with 3d pharmacophoric constraints. *Chem. Sci.*, 12:14577–14589.

[74] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift.

[75] Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., Moroz, Y. S., Mayfield, J., and Sayle, R. A. (2020). Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073.

[76] Jia, X., Lynch, A., Huang, Y., Danielson, M., Lang'at, I., Milder, A., Ruby, A. E., Wang, H., Friedler, S. A., Norquist, A. J., and Schrier, J. (2019). Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature*, 573:251–255.

[77] Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., and Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1):12.

[78] Jiménez, J., Škalič, M., Martínez-Rosell, G., and De Fabritiis, G. (2018). Kdeep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2):287–296.

[79] Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584.

[80] Jin, W., Coley, C. W., Barzilay, R., and Jaakkola, T. (2017). Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):2608–2617.

[81] Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., et al. (2020). Structure of mpro from sars-cov-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293.

[82] Johansson, S., Thakkar, A., Kogej, T., Bjerrum, E., Genheden, S., Bastys, T., Kannas, C., Schliep, A., Chen, H., and Engkvist, O. (2020). Ai-assisted synthesis prediction. *Drug Discovery Today: Technologies*.

[83] Johnson, M. A. and Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. Wiley.

[84] Jorner, K., Brinck, T., Norrby, P.-O., and Buttar, D. (2021). Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.*, 12:1163–1175.

[85] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

[86] Karpov, P., Godin, G., and Tetko, I. V. (2020). Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *Journal of Cheminformatics*, 12(1):17.

[87] Kaserer, T., Beck, K. R., Akram, M., Odermatt, A., and Schuster, D. (2015). Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases. *Molecules*, 20(12):22799–22832.

[88] Kearnes, S. (2021). Pursuing a prospective perspective. *Trends in Chemistry*, 3(2):77–79.

[89] Kim, H., Na, J., and Lee, W. B. (2021). Generative chemical transformer: Neural machine learning of molecular geometric structures from chemical language via attention. *Journal of Chemical Information and Modeling*, 61(12):5804–5814.

[90] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

[91] Kishimoto, A., Buesser, B., Chen, B., and Botea, A. (2019). Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning. In *Advances in Neural Information Processing Systems*, pages 7224–7234.

[92] Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949.

[93] Klein, G., Kim, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT.

[94] Klucznik, T., Mikulak-Klucznik, B., McCormack, M. P., Lima, H., Szymkuć, S., Bhowmick, M., Molga, K., Zhou, Y., Rickershauser, L., Gajewska, E. P., Toutchkine, A., Dittwald, P., Startek, M. P., Kirkovits, G. J., Roszak, R., Adamski, A., Sieredzińska, B., Mrksich, M., Trice, S. L., and Grzybowski, B. A. (2018). Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem*, 4(3):522–532.

[Kovacs et al.] Kovacs, D. P., McCorkindale, W., and Lee, A. A. Molecular Transformer Explainer. https://github.com/davkovacs/MTExplainer.git.

[96] Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (selfies): A 100string representation. *Machine Learning: Science and Technology*, 1(4):045024.

[97] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[Landrum] Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org.

[99] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096.

[100] Lee, A. A., Yang, Q., Sresht, V., Bolgar, P., Hou, X., Klug-McLeod, J. L., Butler, C. R., et al. (2019). Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chemical Communications*, 55(81):12152–12155.

[101] Li, J., Fu, A., and Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdisciplinary Sciences: Computational Life Sciences*, 11(2):320–328.

[102] Limban, C., Nuţă, D. C., Chiriţă, C., Negreș, S., Arsene, A. L., Goumenou, M., Karakitsios, S. P., Tsatsakis, A. M., and Sarigiannis, D. A. (2018). The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicology Reports*, 5:943–953.

[103] Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning.

[104] Liu, Y., Liang, C., Xin, L., Ren, X., Tian, L., Ju, X., Li, H., Wang, Y., Zhao, Q., Liu, H., et al. (2020). The development of coronavirus 3c-like protease (3clpro) inhibitors from 2010 to 2020. *European journal of medicinal chemistry*, page 112711.

[105] Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. (2014). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3):405–412.

[106] Llanos, M. A., Gantner, M. E., Rodriguez, S., Alberca, L. N., Bellera, C. L., Talevi, A., and Gavernet, L. (2021). Strengths and weaknesses of docking simulations in the sars-cov-2 era: the main protease (mpro) case study. *Journal of Chemical Information and Modeling*, 61(8):3758–3770. PMID: 34313128.

[107] Lluch, A. M., Sánchez-Baeza, F., Messeguer, A., Fusco, C., and Curci, R. (1993). Regio- and chemoselective epoxidation of fluorinated monoterpenes and sesquiterpenes by dioxiranes. *Tetrahedron*, 49(28):6299–6308.

[108] Lowe, D. M. (2012). *Extraction of chemical structures and reactions from the literature*. Phd, University of Cambridge.

[109] Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-Decem(Section 2):4766–4775.

[110] Lyu, J., Irwin, J. J., and Shoichet, B. K. (2023). Modeling the expansion of virtual screening libraries. *Nature Chemical Biology*.

[111] Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K., Tolmachev, A. A., Shoichet, B. K., Roth, B. L., and Irwin, J. J. (2019). Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229.

[112] Macip, G., Garcia-Segura, P., Mestres-Truyol, J., Saldivar-Espinoza, B., Ojeda-Montes, M. J., Gimeno, A., Cereto-Massagué, A., Garcia-Vallvé, S., and Pujadas, G. (2022). Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for sars-cov-2 main protease (m-pro) inhibition. *Medicinal research reviews*, 42(2):744–769.

[113] Maggiora, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8):3186–3204.

[114] Maggiora, G. M. (2006). On outliers and activity cliffswhy qsar often disappoints. *Journal of Chemical Information and Modeling*, 46(4):1535–1535.

[115] Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358.

[116] Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. (2018). Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chem. Sci.*, 9:5441–5451.

[117] McCloskey, K., Sigel, E. A., Kearnes, S., Xue, L., Tian, X., Moccia, D., Gikunju, D., Bazzaz, S., Chan, B., Clark, M. A., Cuozzo, J. W., Guié, M.-A., Guilinger, J. P., Huguet, C., Hupp, C. D., Keefe, A. D., Mulhern, C. J., Zhang, Y., and Riley, P. (2020). Machine learning on dna-encoded libraries: A new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866.

[118] McCloskey, K., Taly, A., Monti, F., Brenner, M. P., and Colwell, L. J. (2019). Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences of the United States of America*, 116(24):11624–11629.

[119] McCorkindale, W., Poelking, C., and Lee, A. A. (2020). Investigating 3d atomic environments for enhanced qsar.

[120] McGann, M. (2012). Fred and hybrid docking performance on standardized datasets. *Journal of Computer-Aided Molecular Design*, 26(8):897–906.

[121] Meng, X.-Y., Zhang, H.-X., Mezei, M., and Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157.

[122] Merget, B., Turk, S., Eid, S., Rippmann, F., and Fulle, S. (2017). Profiling prediction of kinase inhibitors: Toward the virtual assay. *Journal of Medicinal Chemistry*, 60(1):474–485.

[123] Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73:1–15.

[124] Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113.

[125] Morreale, F. E., Testa, A., Chaugule, V. K., Bortoluzzi, A., Ciulli, A., and Walden, H. (2017). Mind the metal: A fragment library-derived zinc impurity binds the e2 ubiquitin-conjugating enzyme ube2t and induces structural rearrangements. *Journal of Medicinal Chemistry*, 60(19):8183–8191.

[126] Morris, A., McCorkindale, W., Consortium, T. C. M., Drayman, N., Chodera, J. D., Tay, S., London, N., and Lee, A. A. (2021). Discovery of sars-cov-2 main protease inhibitors using a synthesis-directed de novo design model. *Chem. Commun.*, 57:5909–5912.

[127] Mudrakarta, P. K., Taly, A., Sundararajan, M., and Dhamdhere, K. (2018). Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

[128] Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., et al. (2020). Qsar without borders. *Chemical Society Reviews*, 49(11):3525–3564.

[129] Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594.

[130] Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

[131] O'Boyle, N. M. and Sayle, R. A. (2016). Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics*, 8(1):36.

[132] Obrezanova, O., Csányi, G., Gola, J. M. R., and Segall, M. D. (2007). Gaussian processes: A method for automatic qsar modeling of adme properties. *Journal of Chemical Information and Modeling*, 47(5):1847–1857.

[133] Owen, D. R., Allerton, C. M. N., Anderson, A. S., Aschenbrenner, L., Avery, M., Berritt, S., Boras, B., Cardin, R. D., Carlo, A., Coffman, K. J., Dantonio, A., Di, L., Eng, H., Ferre, R., Gajiwala, K. S., Gibson, S. A., Greasley, S. E., Hurst, B. L., Kadar, E. P., Kalgutkar, A. S., Lee, J. C., Lee, J., Liu, W., Mason, S. W., Noell, S., Novak, J. J., Obach, R. S., Ogilvie, K., Patel, N. C., Pettersson, M., Rai, D. K., Reese, M. R., Sammons, M. F., Sathish, J. G., Singh, R. S. P., Steppan, C. M., Stewart, A. E., Tuttle, J. B., Updyke, L., Verhoest, P. R., Wei, L., Yang, Q., and Zhu, Y. (2021). An oral sars-cov-2 m^{pro} inhibitor clinical candidate for the treatment of covid-19. *Science*, 374(6575):1586–1593.

[134] Pal, S., Kumar, V., Kundu, B., Bhattacharya, D., Preethy, N., Reddy, M. P., and Talukdar, A. (2019). Ligand-based pharmacophore modeling, virtual screening and molecular docking studies for discovery of potential topoisomerase i inhibitors. *Computational and Structural Biotechnology Journal*, 17:291–310.

[135] Palmer, D. S., O'Boyle, N. M., Glen, R. C., and Mitchell, J. B. O. (2007). Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling*, 47(1):150–158.

[136] Parks, C. D., Gaieb, Z., Chiu, M., Yang, H., Shao, C., Walters, W. P., Jansen, J. M., McGaughey, G., Lewis, R. A., Bembenek, S. D., Ameriks, M. K., Mirzadegan, T., Burley, S. K., Amaro, R. E., and Gilson, M. K. (2020). D3r grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of Computer-Aided Molecular Design*, 34(2):99–119.

[137] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076.

[138] Patel, H., Bodkin, M. J., Chen, B., and Gillet, V. J. (2009). Knowledge-based approach to de novo design using reaction vectors. *Journal of chemical information and modeling*, 49(5):1163–1184.

[139] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[140] Perera, D., Tucker, J. W., Brahmbhatt, S., Helal, C. J., Chong, A., Farrell, W., Richardson, P., and Sach, N. W. (2018). A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374):429–434.

[141] Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y., and Jung, S.-H. (2016). An overview of severe acute respiratory syndrome–coronavirus (sars-cov) 3cl protease inhibitors: peptidomimetics and small molecule chemotherapy. *Journal of medicinal chemistry*, 59(14):6595–6628.

[142] Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kolumbin, O. G., Muratov, N. N., and Kuz'min, V. E. (2009). Application of random forest approach to qsar prediction of aquatic toxicity. *Journal of Chemical Information and Modeling*, 49(11):2481–2488.

[PostEra Inc.] PostEra Inc. COVID moonshot. https://postera.ai/covid.

[144] Pradeepkiran, J. A., Reddy, A. P., and Reddy, P. H. (2019). Pharmacophore-based models for therapeutic drugs against phosphorylated tau in alzheimer's disease. *Drug Discovery Today*, 24(2):616–623.

[145] Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling*, 57(4):942–957.

[146] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents.

[147] Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[148] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:1135–1144.

[149] Riniker, S. and Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1):26.

[150] Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.

[151] Saar, K. L., Fearon, D., Consortium, T. C. M., von Delft, F., Chodera, J. D., and Lee, A. A. (2021). Turning high-throughput structural biology into predictive inhibitor design. *bioRxiv*.

[152] Sacha, M., Błaż, M., Byrski, P., Włodarczyk-Pruszyński, P., and Jastrzębski, S. (2020). Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits.

[153] Saldívar-González, F. I., Huerta-García, C. S., and Medina-Franco, J. (2020). Chemoinformatics-based enumeration of chemical libraries: a tutorial. *Journal of Cheminformatics*, 12(1):64.

[154] Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C., and Glorius, F. (2020). A structure-based platform for predicting chemical reactivity. *Chem*, 6(6):1379–1390.

[155] Santanilla, A. B., Regalado, E. L., Pereira, T., Shevlin, M., Bateman, K., Campeau, L.-C., Schneeweis, J., Berritt, S., Shi, Z.-C., Nantermet, P., Liu, Y., Helmy, R., Welch, C. J., Vachal, P., Davies, I. W., Cernak, T., and Dreher, S. D. (2015). Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217):49–53.

[156] Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C. J., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. A. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C. R., Yan, Z., Yao, V., and Treangen, T. J. (2022). Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728.

[157] Schiebel, J., Krimmer, S. G., Röwer, K., Knörlein, A., Wang, X., Park, A. Y., Stieler, M., Ehrmann, F. R., Fu, K., Radeva, N., et al. (2016). High-throughput crystallography: reliable and efficient identification of fragment hits. *Structure*, 24(8):1398–1409.

[158] Schneider, N., Lowe, D. M., Sayle, R. A., and Landrum, G. A. (2015). Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of Chemical Information and Modeling*, 55(1):39–53. PMID: 25541888.

[159] Schneider, P. and Schneider, G. (2016). De novo design at the edge of chaos: Miniperspective. *Journal of medicinal chemistry*, 59(9):4077–4086.

[160] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition.

[161] Schreck, J. S., Coley, C. W., and Bishop, K. J. (2019). Learning retrosynthetic planning through simulated experience. *ACS Central Science*, 5(6):970.

[162] Schuller, M., Correy, G. J., Gahbauer, S., Fearon, D., Wu, T., Díaz, R. E., Young, I. D., Martins, L. C., Smith, D. H., Schulze-Gahmen, U., Owens, T. W., Deshpande, I., Merz, G. E., Thwin, A. C., Biel, J. T., Peters, J. K., Moritz, M., Herrera, N., Kratochvil, H. T., null null, Aimon, A., Bennett, J. M., Neto, J. B., Cohen, A. E., Dias, A., Douangamath, A., Dunnett, L., Fedorov, O., Ferla, M. P., Fuchs, M. R., Gorrie-Stone, T. J., Holton, J. M., Johnson, M. G., Krojer, T., Meigs, G., Powell, A. J., Rack, J. G. M., Rangel, V. L., Russi, S., Skyner, R. E., Smith, C. A., Soares, A. S., Wierman, J. L., Zhu, K., O'Brien, P., Jura, N., Ashworth, A., Irwin, J. J., Thompson, M. C., Gestwicki, J. E., von Delft, F., Shoichet, B. K., Fraser, J. S., and Ahel, I. (2021). Fragment binding to the nsp3 macrodomain of sars-cov-2 identified through crystallographic screening and computational docking. *Science Advances*, 7(16):eabf8711.

[163] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Bekas, C., and Lee, A. A. (2019a). Molecular Transformer - A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583.

[164] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Bekas, C., and Lee, A. A. (2019b). Molecular Transformer - A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583.

[165] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. (2019c). Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

[166] Segler, M. H., Kogej, T., Tyrchan, C., and Waller, M. P. (2018a). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131.

[167] Segler, M. H., Preuss, M., and Waller, M. P. (2018b). Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604.

[168] Segler, M. H. S. (2019). World programs for model-based learning and planning in compositional state and action spaces.

[169] Segler, M. H. S. and Waller, M. P. (2017). Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry – A European Journal*, 23(25):5966–5971.

[170] Sheridan, R. P. (2013). Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4):783–790.

[171] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

[172] Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

[173] Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., and Jaakkola, T. (2022). Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pages 20503–20521. PMLR.

[174] 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., and Jensen, K. F. (2020). Current and future roles of artificial intelligence in medicinal chemistry synthesis. *Journal of Medicinal Chemistry*, 63(16):8667–8682. PMID: 32243158.

[175] Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2019). Comparative assessment of scoring functions: The casf-2016 update. *Journal of Chemical Information and Modeling*, 59(2):895–913.

[176] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, 7:5109–5118.

[177] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958.

[178] Swamidass, S. J., Chen, J., Bruand, J., Phung, P., Ralaivola, L., and Baldi, P. (2005). Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21(suppl_1):359–368.

[179] Temml, V., Voss, C. V., Dirsch, V. M., and Schuster, D. (2014). Discovery of new liver x receptor agonists by pharmacophore modeling and shape-based virtual screening. *Journal of Chemical Information and Modeling*, 54(2):367–371.

[180] Tetko, I. V. (2002). Neural network studies. 4. introduction to associative neural networks. *Journal of Chemical Information and Computer Sciences*, 42(3):717–728. PMID: 12086534.

[181] Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. (2020). State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):1–11.

[182] Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O., and Bjerrum, E. J. (2020). Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.*, 11:154–168.

[183] The COVID Moonshot Consortium (2020). Covid moonshot: open science discovery of sars-cov-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *bioRxiv, doi:10.1101/2020.10.29.339317*.

[184] The COVID Moonshot Consortium, Achdout, H., Aimon, A., Bar-David, E., Barr, H., Ben-Shmuel, A., Bennett, J., Bilenko, V. A., Bilenko, V. A., Boby, M. L., Borden, B., Bowman, G. R., Brun, J., BVNBS, S., Calmiano, M., Carbery, A., Carney, D., Cattermole, E., Chang, E., Chernyshenko, E., Chodera, J. D., Clyde, A., Coffland, J. E., Cohen, G., Cole, J., Contini, A., Cox, L., Cvitkovic, M., Dias, A., Donckers, K., Dotson, D. L., Douangamath, A., Duberstein, S., Dudgeon, T., Dunnett, L., Eastman, P. K., Erez, N., Eyermann, C. J., Fairhead, M., Fate, G., Fearon, D., Fedorov, O., Ferla, M., Fernandes, R. S., Ferrins, L., Foster, R., Foster, H., Gabizon, R., Garcia-Sastre, A., Gawriljuk, V. O., Gehrtz, P., Gileadi, C., Giroud, C., Glass, W. G., Glen, R., Glinert, I., Godoy, A. S., Gorichko, M., Gorrie-Stone, T., Griffen, E. J., Hart, S. H., Heer, J., Henry, M., Hill, M., Horrell, S., Huliak, V. D., Hurley, M. F., Israely, T., Jajack, A., Jansen, J., Jnoff, E., Jochmans, D., John, T., Jonghe, S. D., Kantsadi, A. L., Kenny, P. W., Kiappes, J. L., Kinakh, S. O., Koekemoer, L., Kovar, B., Krojer, T., Lee, A., Lefker, B. A., Levy, H., Logvinenko, I. G., London, N., Lukacik, P., Macdonald, H. B., MacLean, B., Malla, T. R., Matviiuk, T., McCorkindale, W., McGovern, B. L., Melamed, S., Melnykov, K. P., Michurin, O., Mikolajek, H., Milne, B. F., Morris, A., Morris, G. M., Morwitzer, M. J., Moustakas, D., Nakamura, A. M., Neto, J. B., Neyts, J., Nguyen, L., Noske, G. D., Oleinikovas, V., Oliva, G., Overheul, G. J., Owen, D., Pai, R., Pan, J., Paran, N., Perry, B., Pingle, M., Pinjari, J., Politi, B., Powell, A., Psenak, V., Puni, R., Rangel, V. L., Reddi, R. N., Reid, S. P., Resnick, E., Ripka, E. G., Robinson, M. C., Robinson, R. P., Rodriguez-Guerra, J., Rosales, R., Rufa, D., Saar, K., Saikatendu, K. S., Schofield, C., Shafeev, M., Shaikh, A., Shi, J., Shurrush, K., Singh, S., Sittner, A., Skyner, R., Smalley, A., Smeets, B., Smilova, M. D., Solmesky, L. J., Spencer, J., Strain-Damerell, C., Swamy, V., Tamir, H., Tennant, R., Thompson, W., Thompson, A., Tomasio, S., Tsurupa, I. S., Tumber, A., Vakonakis, I., van Rij, R. P., Vangeel, L., Varghese, F. S., Vaschetto, M., Vitner, E. B., Voelz, V., Volkamer, A., von Delft, F., von Delft, A., Walsh, M., Ward, W., Weatherall, C., Weiss, S., White, K. M., Wild, C. F., Wittmann, M., Wright, N., Yahalom-Ronen, Y., Zaidmann, D., Zidane, H., and Zitzmann, N. (2022). Open science discovery of oral non-covalent sars-cov-2 main protease inhibitor therapeutics. *bioRxiv*.

[185] Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., and Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 52(11):2884–2901. PMID: 23078167.

[186] Trnka, T. M. and Grubbs, R. H. (2001). The development of l2x2ruchr olefin metathesis catalysts: An organometallic success story. *Accounts of Chemical Research*, 34(1):18–29. PMID: 11170353.

[187] Ullrich, S. and Nitsche, C. (2020). The sars-cov-2 main protease as drug target. *Bioorganic & Medicinal Chemistry Letters*, page 127377.

[188] Unoh, Y., Uehara, S., Nakahara, K., Nobori, H., Yamatsu, Y., Yamamoto, S., Maruyama, Y., Taoda, Y., Kasamatsu, K., Suto, T., et al. (2022). Discovery of s-217622, a noncovalent oral sars-cov-2 3cl protease inhibitor clinical candidate for treating covid-19. *Journal of Medicinal Chemistry*, 65(9):6499–6512.

[189] Vandenberk, J., Kennis, L. E. J., Van Heertum, A. H. M. T., and Van der Aa, M. J. M. C. (1981). 1,3-dihydro-1-[(1-piperidinyl)alkyl]-2h-benzimidazol-2-one derivatives.

[190] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009.

[191] Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein–ligand docking using gold. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623.

[192] Volkamer, A., Riniker, S., Nittinger, E., Lanini, J., Grisoni, F., Evertsson, E., Rodríguez-Pérez, R., and Schneider, N. (2023). Machine learning for small molecule drug discovery in academia and industry. *Artificial Intelligence in the Life Sciences*, 3:100056.

[193] Vuorinen, A. and Schuster, D. (2015). Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. *Methods*, 71:113–134.

[194] Walters, W., Stahl, M. T., and Murcko, M. A. (1998). Virtual screening—an overview. *Drug Discovery Today*, 3(4):160–178.

[195] Walters, W. P. (2019). Virtual chemical libraries. *Journal of Medicinal Chemistry*, 62(3):1116–1124.

[196] Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The pdbbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980.

[197] Wei, J. N., Duvenaud, D., and Aspuru-Guzik, A. (2016). Neural networks for the prediction of organic chemistry reactions. *ACS Central Science*, 2(10):725–732.

[198] Weininger, D. (1988). SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

[199] Weininger, D., Weininger, A., and Weininger, J. L. (1989). SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101.

[200] Willett, P., Barnard, J. M., and Downs, G. M. (1998). Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996.

[201] Wong, F., Krishnan, A., Zheng, E. J., Stärk, H., Manson, A. L., Earl, A. M., Jaakkola, T., and Collins, J. J. (2022). Benchmarking alphafold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology*, 18(9):e11081.

[202] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530.

[203] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388. PMID: 31361484.

[204] Yang, Y., Zheng, S., Su, S., Zhao, C., Xu, J., and Chen, H. (2020). Syntalinker: automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.*, 11:8312–8322.

[205] Yu, H. S., Modugula, K., Ichihara, O., Kramschuster, K., Keng, S., Abel, R., and Wang, L. (2021). General theory of fragment linking in molecular design: Why fragment linking rarely succeeds and how to improve outcomes. *Journal of Chemical Theory and Computation*, 17(1):450–462. PMID: 33372778.