Chapter 2

Background

Detailed description of many things. Blurb - brief description of cheminformatics, applications of computational tools to drug discovery. Maybe introduce design-make-teset? This stuff is probably in the introduction.

2.1 Molecular Representation

2.1.1 SMILES

The simplified molecular-input line-entry system (SMILES) [106, 107] 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. 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.

(Table 2.1)

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.

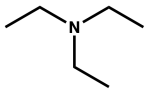
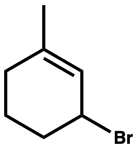
SMILES	Structure
C	CH_4
[Fe2+]	Fe^{2+}
C=O	CH_2O
C#N	HCN (cyan)
CCN(CC)CC	
CC1=CC(CCC1)Br	

Table 2.1 Demonstration of the SMILES language

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 models are used will have molecules represented as SMILES strings. For example, the (blah blah) dataset consists of (blah) SMILES strings alongside the measured (blah) value for each molecule, while USPTO consists of (blah) reaction SMILES strings. 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.

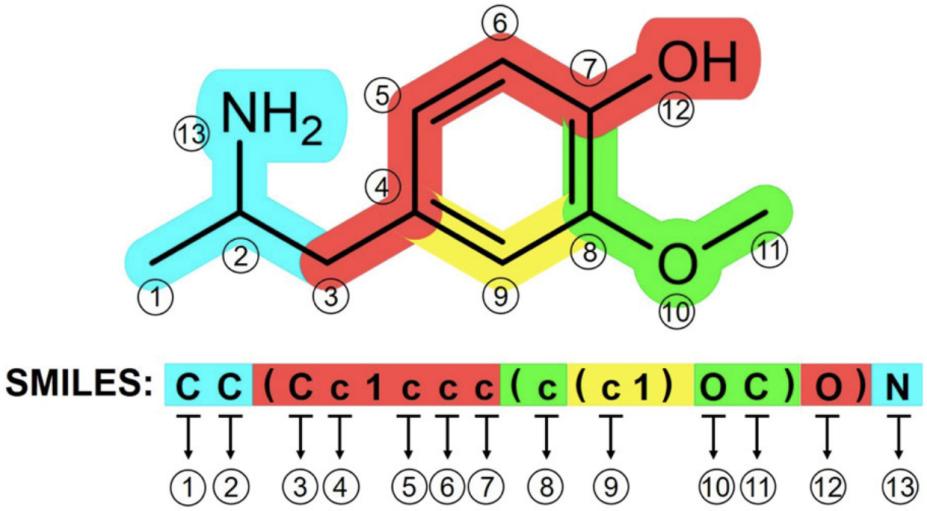


Fig. 2.1 **Illustration of the mapping from chemical structure to SMILES.** Adapted from [49].

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) [35] 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) [53] 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 SMARTS

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. 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=O	wildcard
CCN(CC)CC	reaction SMARTS

Table 2.2 Demonstration of SMARTS patterns

The precise and transparent substructural specification that SMARTS allows has been exploited in a number of applications.

Substructural filters defined in SMARTS have been used [7] to identify undesirable compounds when performing strategic pooling of compounds for high-throughput screening. The REOS (rapid elimination of swill) [8] procedure uses SMARTS to filter out reactive, toxic and otherwise undesirable moieties from databases of chemical structures.

ALADDIN[14] is a pharmacophore matching program that uses SMARTS to define recognition points (e.g. neutral hydrogen bond acceptor) of pharmacophores. A key problem in pharmacophore matching is that functional groups that are likely to be ionised at physiological pH are typically registered in their neutral forms in structural databases. The ROCS shape matching program allows atom types to be defined using SMARTS.[15] (see section 2.1.3)

Scaffold splitting.

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'. For example... Beyond querying for the occurence of substructres, 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 eg EnamineREAL.

Often it is not necessary to fully simulate and understand a chemical reaction and it is sufficient to know the outcome i.e. the major product of it. This is most often the case when experimental organic chemists are willing to validate their synthetic route or when a synthesis planning software uses a reaction prediction model to score its suggestions. This is what is more traditionally referred to as reaction prediction. In these use cases a general purpose model that is able to predict a wide variety of organic reactions with good accuracy is desired. Trained organic chemists usually rationalize reactions based on the reaction mechanisms [16]. These mechanisms can be used to categorize organic reactions and each of these categories can be summarised with the help of so called reaction templates. Figure 2.2 shows a typical general reaction template for the synthesis of an amide using acid chloride and an amine. Here the R_1 and R_2 represent any chemical structures.

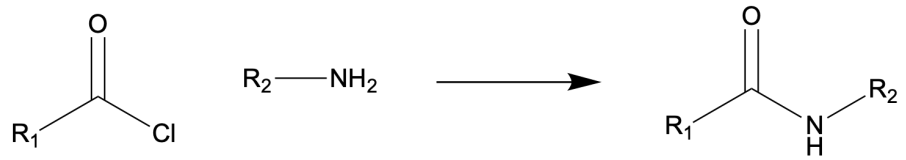


Fig. 2.2 An example of a reaction template for the synthesis of an amide

In addition to virtual library construction, reaction templates can be used for organic reaction prediction by building a catalogue of templates of as many organic reactions as possible. Then given some reactants and reagents as labelled graphs the problem of reaction prediction is transformed into one of subgraph searching to find the best matching general template in the catalogue. When that template is found it can be applied on the input to obtain the predicted outcome of the reaction. This approach was originally proposed and pioneered in the 1980s by E. J. Corey when he used templates for the reverse problem of retrosynthesis [**?**]. The

template-based approach had some success in forward reaction prediction for example as described in Ref [**?**] a template-based model helped design synthetic pathways to a diverse set of 8 drug-like molecules. This method had considerably more success in retrosynthesis though where there does not exist a single correct solution. One of the major limitation of template-based approaches when applied to forward prediction is scalability, meaning that the template library needs to be maintained and every time a new reaction is reported the associated template needs to be added to the template library. 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. IUPAC defines a pharmacophore to be "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response".[1] A pharmacophore model explains how structurally diverse ligands can bind to a common receptor site. Furthermore, pharmacophore models can be used to identify through de novo design or virtual screening novel ligands that will bind to the same receptor.

Typical pharmacophore features include hydrophobic centroids, aromatic rings, hydrogen bond acceptors or donors, cations, and anions. These pharmacophoric points may be located on the ligand itself or may be projected points presumed to be located in the receptor.

In modern computational chemistry, pharmacophores are used to define the essential features of one or more molecules with the same biological activity. A database of diverse chemical compounds can then be searched for more molecules which share the same features arranged in the same relative orientation. Pharmacophores are also used as the starting point for developing 3D-QSAR models. Such tools and a related concept of "privileged structures", which are "defined as molecular frameworks which are able of providing useful ligands for more than one type of receptor or enzyme target by judicious structural modifications",[3] aid in drug discovery.[4]

Use SMARTS to define pharmacophores.

2.1.4 Fingerprints

The similarity-based[4] virtual screening (a kind of ligand-based virtual screening) assumes that all compounds in a database that are similar to a query compound have similar biological activity. Although this hypothesis is not always valid,[5] quite often the set of retrieved compounds is considerably enriched with actives.[6] To achieve high efficacy of similarity-based screening of databases containing millions of compounds, molecular structures are usually represented by molecular screens (structural keys) or by fixed-size or variable-size molecular fingerprints. Molecular screens and fingerprints can contain both 2D- and 3D-information. However, the 2D-fingerprints, which are a kind of binary fragment descriptors, dominate in this area. Fragment-based structural keys, like MDL keys,[7] are sufficiently good for handling small and medium-sized chemical databases, whereas processing of large databases is performed with fingerprints having much higher information density. Fragment-based Daylight,[8] BCI,[9] and UNITY 2D (Tripos[10]) fingerprints are the best known examples. The most popular similarity measure for comparing chemical structures represented by means of fingerprints is the Tanimoto (or Jaccard) coefficient T. Two structures are usually considered similar if T > 0.85 (for Daylight fingerprints). However, it is a common misunderstanding that a similarity of T > 0.85 reflects similar bioactivities in general ("the 0.85 myth").[11]

Tanimoto similarity.

2.2 Computational Approaches

2.2.1 Docking

In the field of molecular modeling, docking is a method which predicts the preferred orientation of one molecule to a second when a ligand and a target are bound to each other to form a stable complex.[1] Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using, for example, scoring functions.

Schematic illustration of docking a small molecule ligand (green) to a protein target (black) producing a stable complex. 0:13 Docking of a small molecule (green) into the crystal structure of the beta-2 adrenergic G-protein coupled receptor (PDB: 3SN6) The associations between biologically relevant molecules such as proteins, peptides, nucleic acids, carbohydrates, and lipids play a central role in signal transduction. Furthermore, the relative orientation of the two interacting partners may affect the type of signal produced (e.g., agonism vs antagonism). Therefore, docking is useful for predicting both the strength and type of signal produced. Molecular docking is one of the most frequently used methods in structure-based drug design, due to its ability to predict the binding-conformation of small molecule ligands to the appropriate

target binding site. Characterisation of the binding behaviour plays an important role in rational design of drugs as well as to elucidate fundamental biochemical processes.[2][3]

One can think of molecular docking as a problem of “lock-and-key”, in which one wants to find the correct relative orientation of the “key” which will open up the “lock” (where on the surface of the lock is the key hole, which direction to turn the key after it is inserted, etc.). Here, the protein can be thought of as the “lock” and the ligand can be thought of as a “key”. Molecular docking may be defined as an optimization problem, which would describe the “best-fit” orientation of a ligand that binds to a particular protein of interest. However, since both the ligand and the protein are flexible, a “hand-in-glove” analogy is more appropriate than “lock-and-key”.[4] During the course of the docking process, the ligand and the protein adjust their conformation to achieve an overall "best-fit" and this kind of conformational adjustment resulting in the overall binding is referred to as "induced-fit".[5] Molecular docking research focuses on computationally simulating the molecular recognition process. It aims to achieve an optimized conformation for both the protein and ligand and relative orientation between protein and ligand such that the free energy of the overall system is minimized.

To perform a docking screen, the first requirement is a structure of the protein of interest. Usually the structure has been determined using a biophysical technique such as X-ray crystallography, NMR spectroscopy or cryo-electron microscopy (cryo-EM), but can also derive from homology modeling construction. This protein structure and a database of potential ligands serve as inputs to a docking program. The success of a docking program depends on two components: the search algorithm and the scoring function. Search algorithm[edit] Main article: Searching the conformational space for docking The search space in theory consists of all possible orientations and conformations of the protein paired with the ligand. However, in practice with current computational resources, it is impossible to exhaustively explore the search space — this would involve enumerating all possible distortions of each molecule (molecules are dynamic and exist in an ensemble of conformational states) and all possible rotational and translational orientations of the ligand relative to the protein at a given level of granularity. Most docking programs in use account for the whole conformational space of the ligand (flexible ligand), and several attempt to model a flexible protein receptor. Each "snapshot" of the pair is referred to as a pose. A variety of conformational search strategies have been applied to the ligand and to the receptor. These include: systematic or stochastic torsional searches about rotatable bonds molecular dynamics simulations genetic algorithms to "evolve" new low energy conformations and where the score of each pose acts as the fitness function used to select individuals for the next iteration.

Ligand flexibility[edit] Conformations of the ligand may be generated in the absence of the receptor and subsequently docked[14] or conformations may be generated on-the-fly in the

presence of the receptor binding cavity,[15] or with full rotational flexibility of every dihedral angle using fragment based docking.[16] Force field energy evaluation are most often used to select energetically reasonable conformations,[17] but knowledge-based methods have also been used.[18] Peptides are both highly flexible and relatively large-sized molecules, which makes modeling their flexibility a challenging task. A number of methods were developed to allow for efficient modeling of flexibility of peptides during protein-peptide docking.[19] Receptor flexibility[edit] Computational capacity has increased dramatically over the last decade making possible the use of more sophisticated and computationally intensive methods in computer-assisted drug design. However, dealing with receptor flexibility in docking methodologies is still a thorny issue.[20] The main reason behind this difficulty is the large number of degrees of freedom that have to be considered in this kind of calculations. Neglecting it, however, in some of the cases may lead to poor docking results in terms of binding pose prediction.[21] Multiple static structures experimentally determined for the same protein in different conformations are often used to emulate receptor flexibility.[22] Alternatively rotamer libraries of amino acid side chains that surround the binding cavity may be searched to generate alternate but energetically reasonable protein conformations.[23][24] Scoring function[edit] Main article: Scoring functions for docking Docking programs generate a large number of potential ligand poses, of which some can be immediately rejected due to clashes with the protein. The remainder are evaluated using some scoring function, which takes a pose as input and returns a number indicating the likelihood that the pose represents a favorable binding interaction and ranks one ligand relative to another. Most scoring functions are physics-based molecular mechanics force fields that estimate the energy of the pose within the binding site. The various contributions to binding can be written as an additive equation:

The components consist of solvent effects, conformational changes in the protein and ligand, free energy due to protein-ligand interactions, internal rotations, association energy of ligand and receptor to form a single complex and free energy due to changes in vibrational modes.[25] A low (negative) energy indicates a stable system and thus a likely binding interaction. Alternative approaches use modified scoring functions to include constraints based on known key protein-ligand interactions,[26] or knowledge-based potentials derived from interactions observed in large databases of protein-ligand structures (e.g. the Protein Data Bank).[27] There are a large number of structures from X-ray crystallography for complexes between proteins and high affinity ligands, but comparatively fewer for low affinity ligands as the latter complexes tend to be less stable and therefore more difficult to crystallize. Scoring functions trained with this data can dock high affinity ligands correctly, but they will also give plausible docked conformations for ligands that do not bind. This gives a large number of false positive hits, i.e., ligands predicted to bind to the protein that actually don't when placed

together in a test tube. One way to reduce the number of false positives is to recalculate the energy of the top scoring poses using (potentially) more accurate but computationally more intensive techniques such as Generalized Born or Poisson-Boltzmann methods.[9]

Simulating the docking process is much more complicated. In this approach, the protein and the ligand are separated by some physical distance, and the ligand finds its position into the protein's active site after a certain number of "moves" in its conformational space. The moves incorporate rigid body transformations such as translations and rotations, as well as internal changes to the ligand's structure including torsion angle rotations. Each of these moves in the conformation space of the ligand induces a total energetic cost of the system. Hence, the system's total energy is calculated after every move. The obvious advantage of docking simulation is that ligand flexibility is easily incorporated, whereas shape complementarity techniques must use ingenious methods to incorporate flexibility in ligands. Also, it more accurately models reality, whereas shape complementary techniques are more of an abstraction. Clearly, simulation is computationally expensive, having to explore a large energy landscape. Grid-based techniques, optimization methods, and increased computer speed have made docking simulation more realistic.

Docking assessment[edit] See also: Critical Assessment of Prediction of Interactions The interdependence between sampling and scoring function affects the docking capability in predicting plausible poses or binding affinities for novel compounds. Thus, an assessment of a docking protocol is generally required (when experimental data is available) to determine its predictive capability. Docking assessment can be performed using different strategies, such as: docking accuracy (DA) calculation; the correlation between a docking score and the experimental response or determination of the enrichment factor (EF);[28] the distance between an ion-binding moiety and the ion in the active site; the presence of induce-fit models. Docking accuracy[edit] Docking accuracy[29][30] represents one measure to quantify the fitness of a docking program by rationalizing the ability to predict the right pose of a ligand with respect to that experimentally observed.[31] Enrichment factor[edit] Docking screens can also be evaluated by the enrichment of annotated ligands of known binders from among a large database of presumed non-binding, "decoy" molecules.[28] In this way, the success of a docking screen is evaluated by its capacity to enrich the small number of known active compounds in the top ranks of a screen from among a much greater number of decoy molecules in the database. The area under the receiver operating characteristic (ROC) curve is widely used to evaluate its performance. Prospective[edit] Resulting hits from docking screens are subjected to pharmacological validation (e.g. IC50, affinity or potency measurements). Only prospective studies constitute conclusive proof of the suitability of a technique for a particular target.[32] In the case of G protein-coupled receptors (GPCRs), which are targets of more

than 30Benchmarking[edit] The potential of docking programs to reproduce binding modes as determined by X-ray crystallography can be assessed by a range of docking benchmark sets. For small molecules, several benchmark data sets for docking and virtual screening exist e.g. Astex Diverse Set consisting of high quality protein-ligand X-ray crystal structures[34] or the Directory of Useful Decoys (DUD) for evaluation of virtual screening performance.[28] An evaluation of docking programs for their potential to reproduce peptide binding modes can be assessed by Lessons for Efficiency Assessment of Docking and Scoring (LEADS-PEP).[35] Applications[edit] A binding interaction between a small molecule ligand and an enzyme protein may result in activation or inhibition of the enzyme. If the protein is a receptor, ligand binding may result in agonism or antagonism. Docking is most commonly used in the field of drug design — most drugs are small organic molecules, and docking may be applied to: hit identification – docking combined with a scoring function can be used to quickly screen large databases of potential drugs in silico to identify molecules that are likely to bind to protein target of interest (see virtual screening). Reverse pharmacology routinely uses docking for target identification. lead optimization – docking can be used to predict in where and in which relative orientation a ligand binds to a protein (also referred to as the binding mode or pose). This information may in turn be used to design more potent and selective analogs. bioremediation – protein ligand docking can also be used to predict pollutants that can be degraded by enzymes.[36][37]

Theory. Virtual screening. Virtual library?

In contrast to high-throughput screening, virtual screening involves computationally screening in silico libraries of compounds, by means of various methods such as docking, to identify members likely to possess desired properties such as biological activity against a given target. In some cases, combinatorial chemistry is used in the development of the library to increase the efficiency in mining the chemical space. More commonly, a diverse library of small molecules or natural products is screened.

2.2.2 FEP?

2.3 Machine Learning

2.3.1 Random Forest

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is

returned.[1][2] Random decision forests correct for decision trees' habit of overfitting to their training set.[3]:587–588 Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.[citation needed] However, data characteristics can affect their performance.[4][5]

Random Forests are a decision tree-based model that use an ensemble of multiple weak regressors to make predictions [47]. Each tree is constructed to find a series of decision boundaries that split the data to minimise the squared deviations between the samples and the sample mean in each branch or leaf of the tree. Predictions are made by averaging the outputs of the different trees when applied to new data. To overcome issues of over-fitting common to decision tree methods, Random Forests use bagging and random subspace projection to reduce the correlation between the trees, improving their generalisation performance.

Examples of RF with morgan fingerprints.

2.3.2 Deep Learning

In contrast to shallow learning, the deep learning revolution of the last decade is built around models that learn their representations from raw data inputs. The workhorse of deep learning is the neural network. At their heart, neural networks are compositions of feature maps that transform the raw input features, x, into a new set of features that are linearly related to their target, y. The prototypical example of a neural network is the multi-layer perceptron (MLP). A l-layer MLP approximates functions f(x) using l successive non-linear feature maps constructed as compositions of affine transformations and non linearities, i.e.

Neural Networks. Loss functions. Optimisation.

$$y = f(x, \theta) \tag{2.1}$$

$$\hat{theta} = \arg\min_{\theta} \mathscr{L}(\theta) \tag{2.2}$$

$$\hat{y} = \sigma(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}) \tag{2.3}$$

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

The remarkable success of deep learning emerges from the fact that neural networks, such as MLPs, can be optimised effectively using first-order gradient-based approaches. Moreover, the necessary gradients can be efficiently calculated using the chain rule by back- propagation

of the training loss. In practice, modern neural networks are implemented inside automatic differentiation frameworks that abstract away the technical burden of implementing back-propagation [71, 72]. In addition, these frameworks are designed to enable the necessary calculations to be carried out on hardware accelerators, such as graphical processing units (GPUs), that dramatically reduce the time for training. The most simple gradient-based optimisation algorithm is gradient-descent. In gradient descent at each step the model's parameters are updated according to

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathscr{L} \tag{2.5}$$

where the learning rate, η, is a hyper-parameter of the optimiser that determines the size of the parameter updates. In practice, the full-batch gradient of the loss, $\nabla_\theta \mathscr{L}(\theta_t)$ is replaced with a stochastic approximation of the gradient calculated on a mini-batch of data randomly sampled from all the available training data. Typically mini-batches are randomly drawn from the training data without replacement until all the training examples have been considered. Each complete cycle through the training set is referred to as an epoch. After each epoch, the training set is shuffled and the process is repeated until the loss has satisfactorily converged. Replacing full-batch gradient descent with mini-batch stochastic gradient descent significantly speeds up optimisation, reducing the amount of computation required to determine the gradient for each step and providing helpful regularisation effects that drive the optimisation towards flatter local basins of attraction in the loss landscape. Further improvements to model optimisation procedures can be achieved by incorporating additional terms such as momentum, learning rate schedules, or adaptive learning rates [73], and additional regularisation procedures such as weight decay, early-stopping, or dropout [74].

In the above description of the MLP, the key requirement is that the model is end- to-end differentiable. This allows the parameters of the model to be optimised by the combination of back-propagation and gradient descent. Accordingly, provided we ensure that all operations in our models have defined derivatives, we can build up novel neural networks architectures with specific inductive biases as compositions of custom differentiable building blocks – notable examples are recurrent neural networks (RNNs) that are designed to handle series data (e.g. Gated-Recurrent-Unit (GRU) [75] and Long-Short-Term-Memory (LSTM) networks [76]) and convolutional neural networks (CNNs) that build in translational invariance for computer vision applications (e.g. LeNet [77], AlexNet [78], and ResNets [79]).

Within materials science, this ability to compose differentiable building blocks into novel architectures has led to the development of a wide variety of message-passing neural networks that operate directly on the atomic coordinates of molecules and materials. Typically these models operate on "radius"-graphs of interconnected local environments determined using a

cutoff radius – the resulting data structure closely mirrors that of Verlet lists used in atomistic simulations [80]. The nodes of the graphs encode atoms with edges encoding interactions or bonds. As with shallow descriptor-based models, it is important to encode the underlying symmetries of the problem into the network architecture. Earlier models ensured SO(3)-invariance by only including position information via the relative distances between connected sites [81–83]. More recently SO(3)-equivariant architectures have been proposed that include angular and higher-order information or relative displacement vectors between atoms to allow construction of message passing operations that maintain equivariance [84].

Examples of NNs with fingerprints and SMILES.

2.3.3 Evaluating Models

ROC-AUC. Accuracy. Enrichment.

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection.[10] The false-positive rate is also known as probability of false alarm[10] and can be calculated as (1 - specificity).

$$\mathrm{EF}(n) = \frac{\mathrm{Hit\,rate(predicted\,top}{-}n)}{\mathrm{Hit\,rate(baseline)}} \tag{2.6}$$

Train-Test splitting

Random split, scaffold split, time split

In machine learning, a common task is the study and construction of algorithms that can learn from and make predictions on data.[1] Such algorithms function by making data-driven predictions or decisions,[2] through building a mathematical model from input data. These input data used to build the model are usually divided in multiple data sets. In particular, three data sets are commonly used in different stages of the creation of the model: training, validation and test sets. The model is initially fit on a training data set,[3] which is a set of examples used to fit the parameters (e.g. weights of connections between neurons in artificial neural networks) of the model.[4] The model (e.g. a naive Bayes classifier) is trained on the training data set using a supervised learning method, for example using optimization methods such as gradient descent or stochastic gradient descent. In practice, the training data set often consists of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), where the answer key is commonly denoted as the target (or label). The current model is run with the training data set and produces a result, which is then compared with the target, for each

input vector in the training data set. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation. Successively, the fitted model is used to predict the responses for the observations in a second data set called the validation data set.[3] The validation data set provides an unbiased evaluation of a model fit on the training data set while tuning the model's hyperparameters[5] (e.g. the number of hidden units—layers and layer widths—in a neural network[4]). Validation datasets can be used for regularization by early stopping (stopping training when the error on the validation data set increases, as this is a sign of over-fitting to the training data set).[6] This simple procedure is complicated in practice by the fact that the validation dataset's error may fluctuate during training, producing multiple local minima. This complication has led to the creation of many ad-hoc rules for deciding when over-fitting has truly begun.[6] Finally, the test data set is a data set used to provide an unbiased evaluation of a final model fit on the training data set.[5] If the data in the test data set has never been used in training (for example in cross-validation), the test data set is also called a holdout data set. The term "validation set" is sometimes used instead of "test set" in some literature (e.g., if the original data set was partitioned into only two subsets, the test set might be referred to as the validation set).[5] Deciding the sizes and strategies for data set division in training, test and validation sets is very dependent on the problem and data available.[7]

2.4 Applications of ML on Drug Discovery

2.4.1 QSAR

2.4.2 Reaction Prediction

Note on bioactivity?

References

[Che] Chemspace: Lead-like compounds.

[2] 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.

[3] 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.

[4] 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

σ

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

[5] 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.

[6] 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.

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

[8] 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.

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

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

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

[12] 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.

[13] 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.

[14] 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*.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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

$$\alpha_{2A}$$

-adrenergic receptor. *Science*, 377(6614):eabn7065.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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*.

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

[47] 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.

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

[49] 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.

[50] 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.

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

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

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

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

[70] 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.

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

[72] 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.

[73] 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.

[74] 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.

[75] 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.

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

[77] 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.

[78] 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*.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

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

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

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

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

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

[92] 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.

[93] 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.

[94] 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.

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

[96] 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.

[97] 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.

[98] 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.

[99] 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*.

[100] 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*.

[101] 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.

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

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.