Compendium 8

Artificial intelligence in drug discovery

De novo design

This compendium will describe an overview of de novo molecular structure generation or de novo design, moreover strategies for evaluating and analysing de novo libraries.

Two key drug design queries arise - where de novo generation should help answer what to make - and reaction planning is answering how to make it (BioSolveITTutorials, 2021). We focus on de novo design here. We will describe here briefly and consider one method CrEM (Polishchuk, P., 2020) for library compound de novo generation, moreover estimate the drug-likeness using certain similarity metric to a reference (as a prelude to exercise 8a), which include descriptor calculation, similarity maps and similarity coefficients.

As stated in Zhavoronkov (2020): *In the context of early drug discovery, one of the first applications of AI/ML was the quantification of how much chemical has "drug-likeness" attempting to mimic a medicinal chemists intuition of estimating a probability that novel chemical compounds are more likely to become useful as pharmaceuticals. Estimating drug-likeness using AI/ML remains a practical chemical space navigation tool, as the number of possible drug-like chemicals can exceed* 10^{23}. *Regardless of algorithms, such methods do not predict "drug" as an attribute, but rather score a multiparametric similarity, interpreted as "these molecules are similar to other molecules that chemists consider to be of pharmaceutical interest."*

We will also describe another de novo generation method called REINVENT (Blaschke et al., 2022) (however the information is contained in exercise 8b).

Designing optimal molecules de novo covering a region of chemical space

Chemical space, the expanse spanning of all possible molecules, is extremely vast. The chemical or molecular property space can be visualised as a terrain or a landscape (see Figure 1a), like a person looking for the lowest valley, it is similar to finding the most optimal molecules defined by an objective molecular property or chemical space. Maintaining pharmacokinetic properties, whilst design a molecule that has a particular biological response is a common objective; It can be like a needle in a haystack to find a new drug molecule that satisfies the wanted molecular properties and also synthezisability in a lab
De novo design, methods aims to find novel chemical structures to guide the search in the chemical space to a more narrow region of chemical space (see Figure 1b). In contrast, virtual screening, or scanning, of chemical space see Figure 1c) be likened as brute-force search, using compound (or molecule) libraries many of even enormous size. Even if large compound libraries for VS these may still correspond only to a very small part of chemical space.

De novo seeks to generate compound libraries from scratch, that can satisfy the desired molecular profile, and is also called generative chemistry.

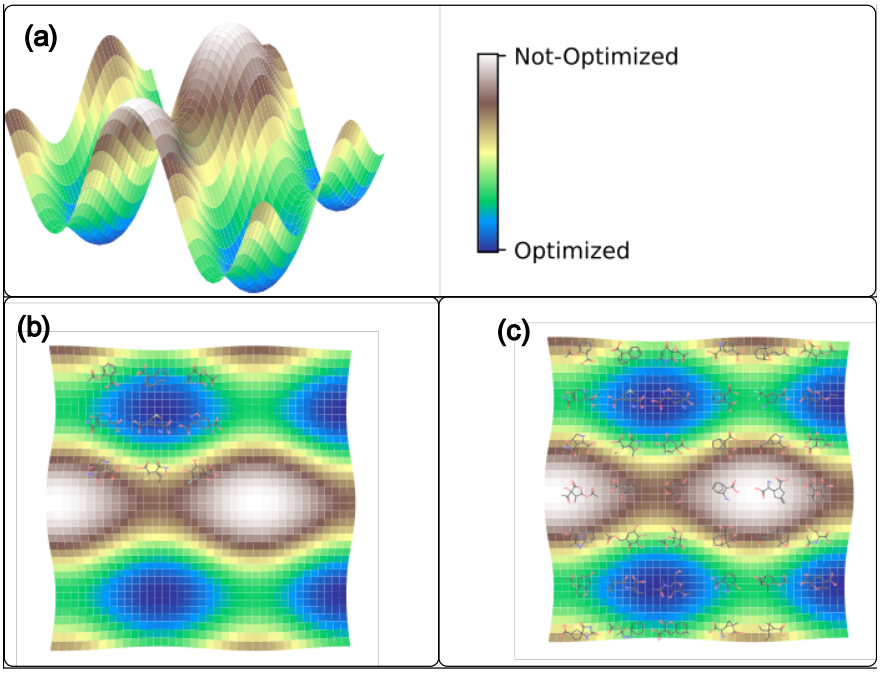

Figure 1: Chemical space landscape schematic chartered with de novo or virtual screening (VS). Here chemical space is arbitrarily likened to a landscape (with an arbitrary mathematical function), one can liken the multiple blue optimized regions meaning similar molecules of e.g. different stereochemistry. Here blue symbolizes the most optimized molecules with regards to a desired molecular profile, whereas towards the upper part of spectrum the one that are not-optimized.
(a) Chemical space landscape, depicting molecular properties, seen from a distance.
(b) De novo generated molecules having the desired profile.
(c) Brute-force search of molecules having a desired molecular profile.

Atom-, fragment-, reaction-based methods of de novo design

(This section is a derivative from Meyers et al. 2021, so cannot be reused, covered briefly in next section rather)

Computational methods for evaluating chemical structures must rely upon a suitable molecular representation, that is, the form in which a molecular structure is seen by a subsequent algorithm. Molecular representation is a broad topic; for example, methods can encode the presence or absence of functional groups, express a molecule as its topological graph, or include 3D information describing bond angles. Among de novo design methods, common molecular representations are text based, such as the simplified molecular input line entry system (SMILES), and graph based where the molecular generator might operate explicitly on the molecular topology. Text-based methods benefit from the huge volume of active research in natural language processing (NLP), whereas graph based approaches embody a more natural representation of molecular structure.

Other influences on choice of representation include whether the molecular representation is discrete (e.g., bitvector), continuous (e.g., vector of floating points), and invertible. Recent reviews of *de novo* design methods have focused on molecular representation through the lens of generative model architecture, whereas it is focused here on the granularity of molecular representation (see Figure 2) because this translates directly to practical aspects of molecular design. In practice, atom-based, fragment-based, and reaction-based approaches have distinct strengths and weaknesses, and many methods blur the boundaries between these classifications.

Content in this section from Meyers et al. (2021) (CC-BY-ND-NC 4.0 license).

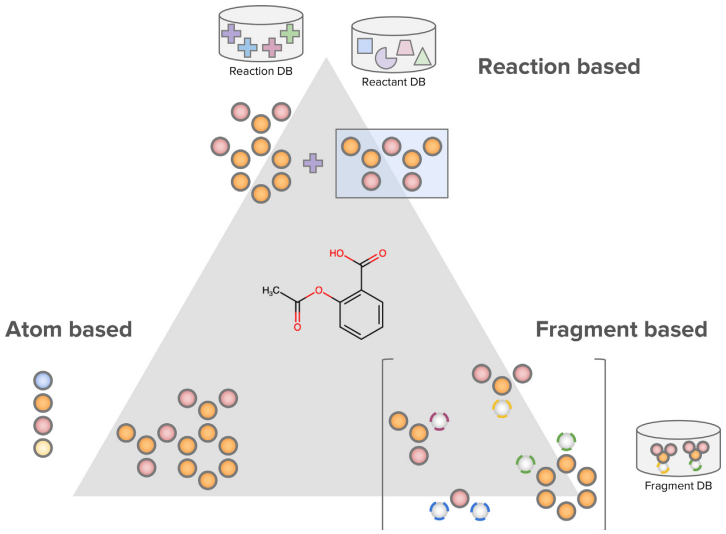

Figure 2: Illustration of the continuum between atom-based, fragment-based, and reaction-based molecular representation paradigms, shown for example molecule aspirin. The atom-based approach is supported by a vocabulary containing a small number of atoms and bonds. The reaction-based approach is supported by dual sets of reactants and reaction rules. Finally, the fragment-based approach is supported by a fragmentation scheme, and a set of interchangeable fragments; gray atoms denote attachment points annotated with a disconnection type (color).

Distinguishing methods and algorithms of de novo design

Machine-learning methods for de novo molecular optimization can be classified into gradient based, and gradient-free machine learning methods, see Table 1. Molecular representation is wide area, one can express a molecule e.g. as a SMILES-string or its 2D or 3D structure graph, as e.g. descriptors of molecule, or e.g. encode and decode structures with or without certain moieties of the molecule. Chemical computational methods such as de novo methods and algorithms rely on suitable molecular representations to process input information to generate an output molecule library.

One may classify methods of de novo methods, depending on what type of molecular representations used to generate compound libraries; often into Atom-, Fragment-, Reaction-based, having particular strength and weaknesses (in addition some classifications like structure and ligand-based), de novo methods may be a mixture or used interchangeably of these classifications. We illustrate the classifications with an example of how chemical structures can be decomposed or acquired information from, e.g. one can make a molecular representation of e.g. aspirin (see only skeletal formula in Figure 4) by (i) encoding each atom and bond between atoms or/and by (ii) making a coarser representation where aspirin is divided in fragments (e.g. its benzene ring with hydrogens and its two other functional groups can be divided into fragments) (iii) or/and one can consider a chemical reaction where aspirin as beginning or end-product (iv) or/and one could gather structural information from binding pockets where aspirin is a ligand to a receptor.

Moreover one can also distinguish these previous molecular representation-based classifications given their optimization algorithm Gradient-base or Gradient-free. Hence given a desired molecular representation, de optimization algorithms guide towards optimal molecule libraries conforming to a calculable objective function (c.f. Figure 1).

For a general overview of Mathematical optimization, gradient-based, gradient-free methods, please see Youtube page AlphaOpt (2017).
For the interested there is a good review on subject in Meyers et al. (2021).

Table 1: Example methods used for de novo molecular design separated by molecular representation that the algorithms are founded upon and whether the de novo methods are based upon Gradient-based or Gradient-free methods.

	Fragment	*Atom*	*Reaction*
Gradient-based	*JT-VAE*	***REINVENT***	*DINGOS*
	…	*GraphINVENT*	…
	…	**…**	…
Gradient free	***CReM***	*MSO*	*AutoGrow4*
	…	…	…

CreM: chemically reasonable mutations framework for structure generation

This is an example of a gradient free and fragment-based de novo design method as described in (Polishchuk, 2020), following is their informative abstract:

> *Structure generators are widely used in de novo design studies and their performance substantially influences an outcome. Approaches based on the deep learning models and conventional atom-based approaches may result in invalid structures and fail to address their synthetic feasibility issues. On the other hand, conventional reaction-based approaches result in synthetically feasible compounds but novelty and diversity of generated compounds may be limited. Fragment-based approaches can provide both better novelty and diversity of generated compounds but the issue of synthetic complexity of generated structure was not explicitly addressed before. Here we developed a new framework of fragment-based structure generation that, by design, results in the chemically valid structures and provides flexible control over diversity, novelty, synthetic complexity and chemotypes of generated compounds. The framework was implemented as an open-source Python module and can be used to create custom workflows for the exploration of chemical space.*

The CrEM library can be accessed at https://github.com/DrrDom/crem. The powerful CreM library has also been implemented and can be used from PlayMolecule (https://playmolecule.com/generative/), some possible structure generation modes are shown in Figure 3. In Exercise 8a, we will generate a compound library using CReM Replace (Mutate) Mode, and also analyze similarity measures to the original reference molecule.

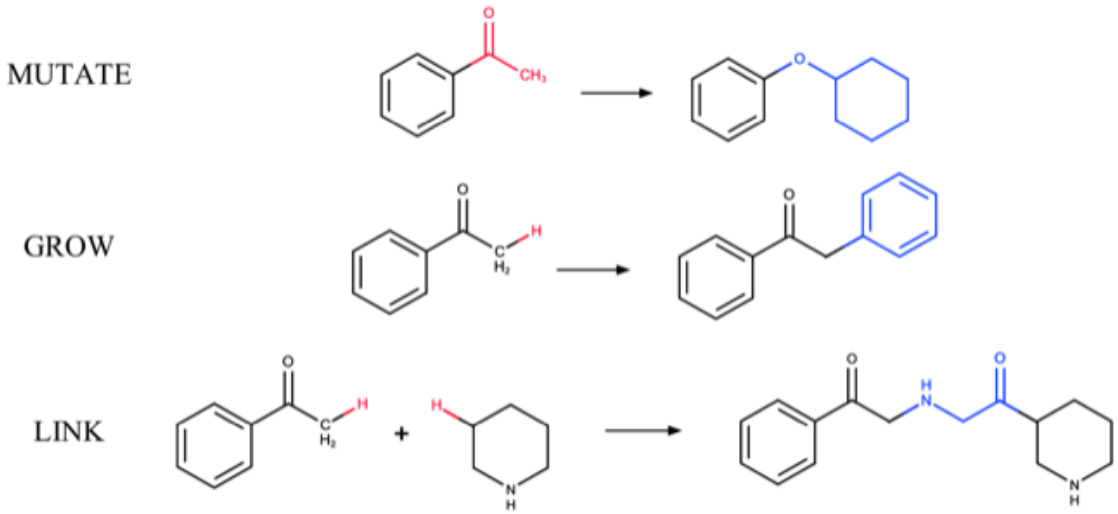

Figure 3: Structure generation modes of CreM: MUTATE (or aka REPLACE), GROW and LINK. Mutate is a replacement of an arbitrarily chosen fragment with another one. GROW is a special case of a MUTATE operation - replacement of a hydrogen with another fragment. LINK is a replacement of hydrogen atoms in two molecules to link them by an appropriate fragment.

REINVENT

This is an example of a gradient-based and atom-based de novo design method as described in (Blaschke, 2020), following is their informative abstract:

In the past few years, we have witnessed a renaissance of the field of molecular de novo drug design. The advancements in deep learning and artificial intelligence (AI) have triggered an avalanche of ideas on how to translate such techniques to a variety of domains including the field of drug design. A range of architectures have been devised to find the optimal way of generating chemical compounds by using either graph- or string (SMILES)-based representations. With this application note, we aim to offer the community a production-ready tool for de novo design, called REINVENT. It can be effectively applied on drug discovery projects that are striving to resolve either exploration or exploitation problems while navigating the chemical space. It can facilitate the idea generation process by bringing to the researcher's attention the most promising compounds. REINVENT's code is publicly available at https://github.com/MolecularAI/Reinvent.

We will in a Jupyter lab notebook do an exercise of this software, with a more thorough description of the steps therein. Please see introduction video about REINVENT, for the extra interested you may read their article Blaschke et al. (2020).

Figure 1: "Reprinted (adapted) with permission from {Blaschke et al. 2020}. Copyright {2020} American Chemical Society."

Molecular representations - SMILES

Structures generated de novo contained in compound libraries can be represented as SMILES or molecular graphs. A thorough review and practical guide for molecular representation used in AI-driven drug discovery has been written (Laurianne et al., 2020). We will review graph -representations in a later compendium.

Since common molecular representations in de novo are text based, such as the simplified molecular input line entry system (SMILES), we introduce it here, see Figure 4. The acronym for simplified molecular-input line-entry system (SMILES), an specification in the form of a line notation for describing the structure of chemical species using short ASCII strings. Most molecule editors can import SMILES strings for conversion back into two-dimensional drawings or three-dimensional models of the molecules.

While other representations has been used in de novo (e.g. InChi format), SMILES remain the de facto representation of choice for chemical language modeling (Grisoni and Schneider, 2022).

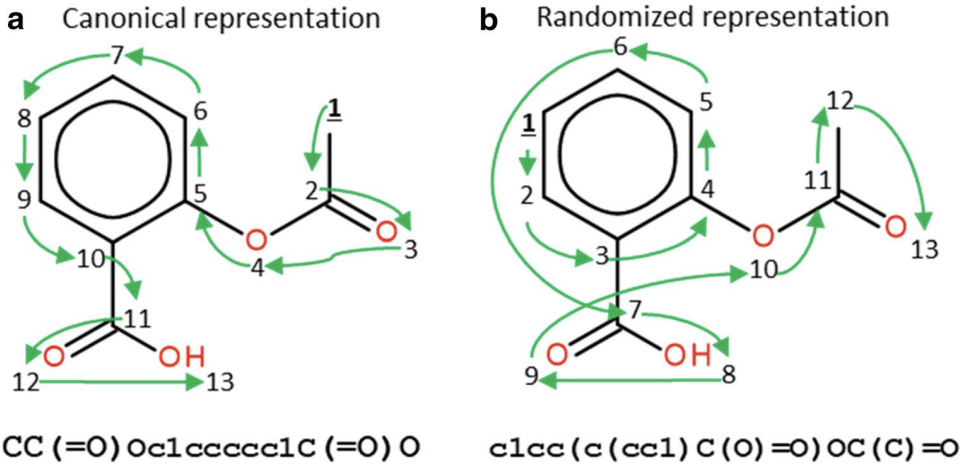

Figure 4: Canonical (a) and randomized (b) SMILES representations of example molecule aspirin. Randomized SMILES correspond to the various representations of a molecule obtained by randomly selecting the starting node in a graph traversal algorithm, thus changing the order of the nodes traversed in the molecular graph. Numbers represent the order of graph traversal, where 1 is the initial node (user defined). Considering **a** as being the canonical representation of aspirin, **b** shows a different ordering of the atoms of the molecule. The final SMILES is one possible SMILES among all the randomized SMILES which can be generated. Green arrows indicate how the molecular graph is traversed. Both SMILES strings shown represent the same molecule but, as the atom numberings are different, the generated SMILES strings are, too. *Figure and caption from (Laurianne et al., 2020).*

Molecular similarity and conversion

There is a general principle that molecules have similar properties if having similar structure. Molecular (or chemical) similarity is a notion to refer to the likeness of chemical compounds with respect to structural or functional qualities, i.e. the effect that the chemical compound has on reaction partners in inorganic or biological settings. Molecular similarity is a very important concept in general in cheminformatics. In drug discovery and de novo compound library analysis, it is an important concept for predicting and designing desirable properties of drug compounds, designing drugs. This concept is used when screening large compound databases such as de novo generated libraries.

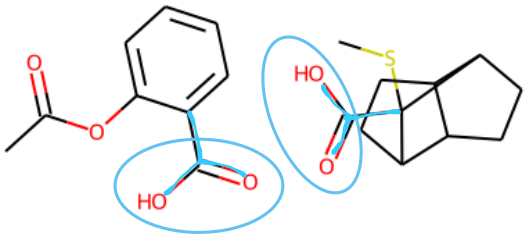

Figure 5: Molecular Similarity of Aspirin, and a de novo generated molecule. Highlighted in blue is the similar moieties of these two molecules.

To make comparisons between molecules one can use fingerprints to encode the structure of a molecule into a fingerprint, an example is the Morgan Fingerprint . Encoding a structure as shown in Figure 6a, which can be done with RDKIT (Landrum, no date), (see adjoining exercise 8a for details). Calculating the fingerprints of a compound library allows to calculate comparison metrics and also descriptors of all compounds in the library.

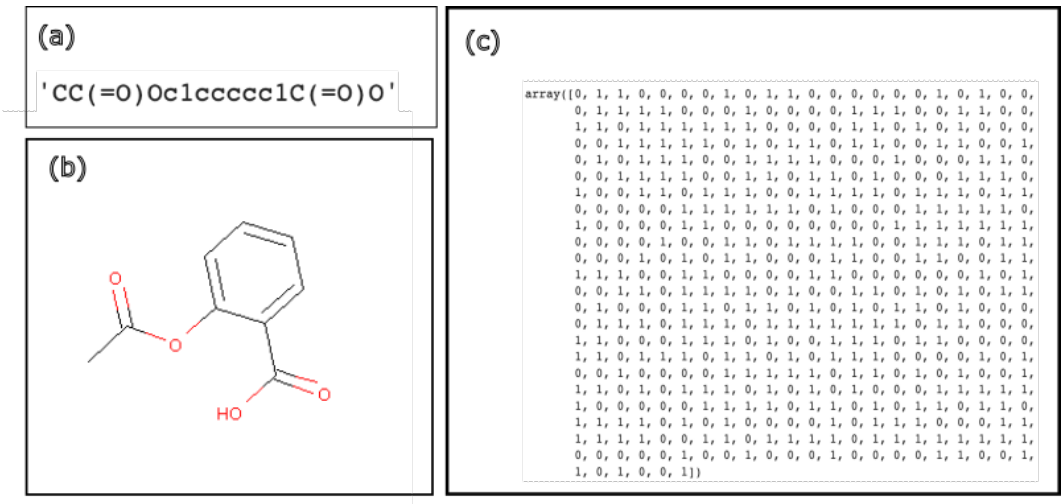

Figure 6: The conversion of SMILES (a) to a molecular object in RDKIT (b) shown as a structure or 2D-graph, that can be converted to a morgan fingerprint with 512 bits (c) represented here as an array or a list of bits i.e. 1's and 0's. In this example the reference molecule is shown.

Quantifying similarity to a reference and a de novo library using their Morgan Fingerprints

Chemical structures are often represented by molecular fingerprints where structural features are converted to either bits in a bit vector or counts in a count vector. This abstract representation allows the computationally efficient handling and comparison of chemical structures. Using such fingerprints, the similarity between two molecules can be calculated in a straightforward manner with simple similarity metrics such as Tanimoto, Dice, and so on. Similarity map is a method that allows comparison of similarity a reference molecule to a de novo generated molecule (Riniker and Landrum, 2013).

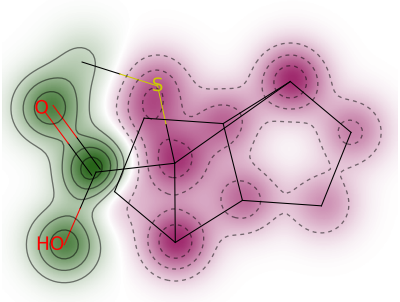

Figure 7: Similarity Map comparison of the same two molecules as in Figure 5, and a de novo generated molecule. Here the green parts indicate a similar chemical moiety between the reference molecule and the de novo molecule; whereas the pink parts indicate a dissimilar moiety.

The Tanimoto Similarity Coefficient (TSC) is a measure of similarity for the two sets of data or in our context two molecules, with a range from 0-1 (0% to 100%) similarity, where the higher percentage means the more similar the two molecules. We can hence use the TSC to compare the reference molecule to all molecules in a de novo library; which can be done in RDkit with only input being the morgan-fingerprint of two molecules compared.

Dice Similarity coefficient is another measure of similarity, which quantifies the similarity between two sets of data, or in this context molecules in a different definition than Tanimoto. For the Dice similarity coefficient, the same measure is used with a range from 0-1 (0% to 100%, where again the higher percentage means the more similar the two molecules.

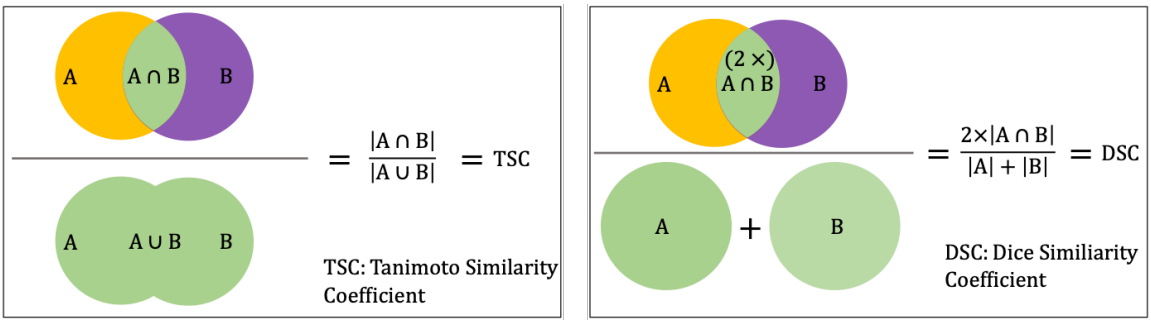

Figure 8: The concept of Tanimoto Similarity Coefficient (or also Tanimoto or Jaccard Coeffiecient or Jaccard Index) (To the left). The concept of Dice Similarity Coefficient) (To the Right). For both similarity coefficients, they range between 1.0 (perfect equality betweenA, B) and 0.0 (null equality between A, B). Here A and B refer respectively to morgan fingerprints of two molecules. Morover for both TSC and DSC as shown here, the green parts are what is included in the division of calculation as is shown in equations.

REFERENCES

AlphaOpt, 2017. Introduction To Optimization.
Available at: https://www.youtube.com/channel/UCkL2HNDjyhrT6hgWjikmQAg
(Accessed: 07 June 2022)

Blaschke, T., Arús-Pous, J., Chen, H., Margreitter, C., Tyrchan, C., Engkvist, O., Papadopoulos, K. and Patronov, A., 2020. REINVENT 2.0: an AI tool for de novo drug design. *Journal of Chemical Information and Modeling*, *60*(12), pp.5918-5922. Available at: https://dx.doi.org/10.1021/acs.jcim.0c00915?ref=pdf (Accessed 07 June 2022)

BioSolveITTutorials, 2021. webinar recording: AI Driven De Novo Design with REINVENT. 04 July. Available at: https://youtu.be/hDuoZDReaQQ (Accessed: 07 June 2022)

Ernst, F.R. and Grizzle, A.J., 2001. Drug-related morbidity and mortality: updating the cost-of-illness model. *Journal of the American Pharmaceutical Association (1996)*, *41*(2), pp.192-199.

David, L., Thakkar, A., Mercado, R. and Engkvist, O., 2020. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, *12*(1), pp.1-22.

Grisoni, F. and Schneider, G., 2022. De Novo Molecular Design with Chemical Language Models. In *Artificial Intelligence in Drug Design* (pp. 207-232). Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1787-8_9#DOI

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

Landrum, G. et al. (no date), Rdkit: Open-source cheminformatics software, URL http://www. rdkit. org/, https://github. com/rdkit/rdkit

Meyers, J., Fabian, B. and Brown, N., 2021. De novo molecular design and generative models. *Drug Discovery Today*, *26*(11), pp.2707-2715. https://doi.org/10.1016/j.drudis.2021.05.019

Polishchuk, P., 2020. CReM: chemically reasonable mutations framework for structure generation. *Journal of Cheminformatics*, *12*(1), pp.1-18. Available at: https://doi.org/10.1186/s13321-020-00431-w (Accessed: 14 June 2022)

Riniker, S. and Landrum, G.A., 2013. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of cheminformatics*, *5*(1), pp.1-7.

Rogers, D.J. and Tanimoto, T.T., 1960. A Computer Program for Classifying Plants: The computer is programmed to simulate the taxonomic process of comparing each case with every other case. *Science*, *132*(3434), pp.1115-1118.

Zhavoronkov, A., Vanhaelen, Q. and Oprea, T.I., 2020. Will artificial intelligence for drug discovery impact clinical pharmacology?. *Clinical Pharmacology & Therapeutics*, *107*(4), pp.780-785.