Social Science, Humanities and Sustainability Research
ISSN 2690-3628 (Print) ISSN 2690-3636 (Online)
Vol. 1, No. 1, 2020
www.scholink.org/ojs/index.php/sshsr

Short Paper

The Trends of De Novo Molecular Designs in the Twenty-First Century: A Mini-Review

Sayan Basak[1*]

[1] Department of Polymer Science and Technology, University of Calcutta, Kolkata, West Bengal, India

[*] Sayan Basak, Department of Polymer Science and Technology, University of Calcutta, Kolkata,700009, West Bengal, India

Received: February 17, 2020 Accepted: March 18, 2020 Online Published: March 24, 2020
doi:10.22158/sshsr.v1n1p27 URL: http://dx.doi.org/10.22158/sshsr.v1n1p27

Abstract

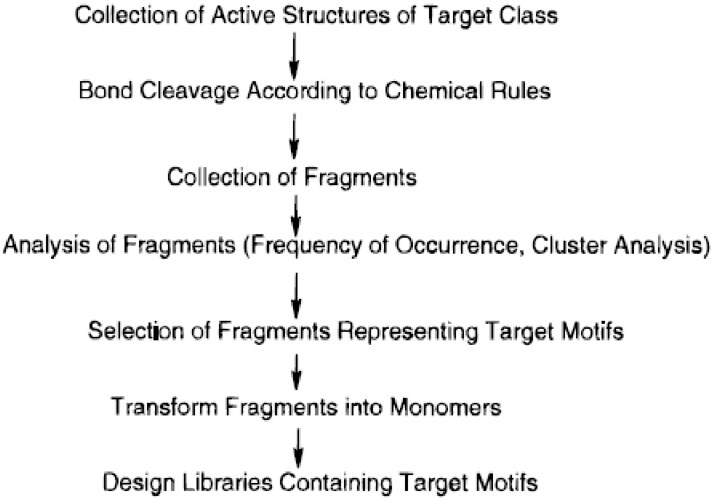
The inception of advanced bioactive agents has driven the growth for sustained drug delivery and the boom of new medicines. The future of the medical and chemical biology relies on the amalgamation of the advanced systematic and analytical techniques, which shall be tethered together with a robust theoretical framework. The de novo drug design is one of such exciting strategies that use computational theories to generate novel molecules with a good affinity to the desired biological target. This mini-review provides a basic overview of the current trends and algorithms, which aids in the advancement of the de novo molecular framework.

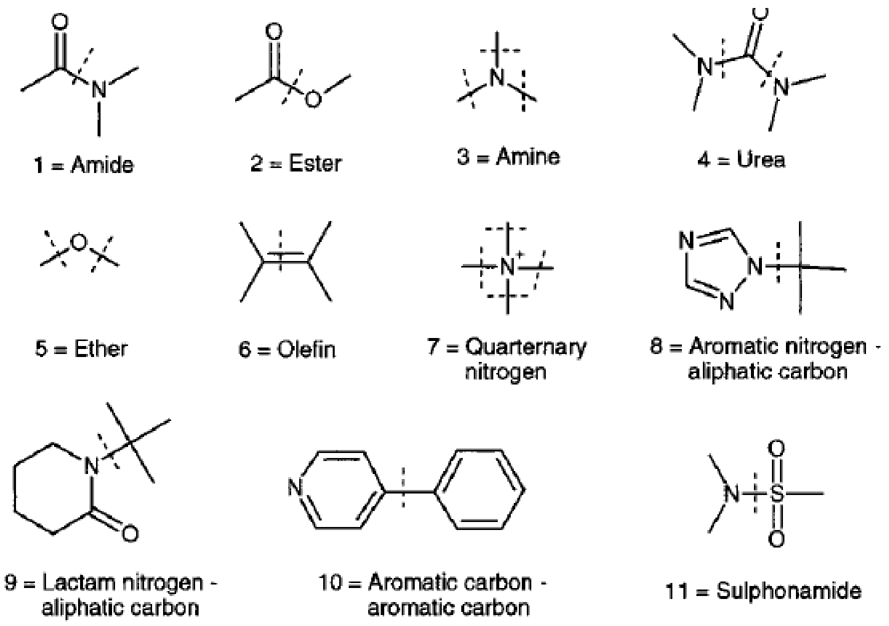
Keywords

De Novo, ChemTS, ChemGE, conditional variational autoencoder, SMILES, Material Science

1. Introduction

The recent advancements of machine learning, coupled with data science, have enabled the scientific route to steer through massive amounts of data and addressing various problems with accuracy and precision (He, Zhang, Ren, & Sun, 2016; Silver et al., 2017). Moreover, using data science has been established to work efficiently by reducing the time taken for a process to complete its operation and thus increasing the overall efficacy of the output. In particular, over the last couple of years, a significant amount of advances have been established in designing superior de novo molecules via machine learning. The ultimate role of devising a De Novo molecular framework is to develop a product via computational sciences, whose physicochemical properties meet arbitrary given requirements (Green et al., 2017; Tabor, Rosch, & Guzik, 2018). However, most of the recent studies solely focus on the computational perspective of designing molecular frameworks, leaving various loopholes in experimental verifications and their physical interpretations. For instance, in the case of polymers informatics, often, the structural complexities and the molecular dynamics posses to be the major disadvantages in synchronizing the theoretical and the experimental results (Ramprasad, Batra, Pilania, Mannodi-Kanakkithodi, & Kim, 2017; Audus & de Pablo, 2017). The conventional method used to prepare molecular designs relies majorly on fragment-based methods (for instance, RECAP) (Figure 1), which aims to synthesize molecules by tethering the known fragments. However, issues like patentability and inferior structural diversity provide a potent setback for these techniques, when implemented. To address these challenges, the idea of de novo molecular generation methods was cultivated, which needed no such fragments to frame the algorithms. These processes use the Simplified Molecular Input Line Entry System strings, along with black-box optimization and deep neural network, to fetch the molecular simulations (Schneider & Fechner, 2005; Lewell, Judd, Watson, & Hann, 1998).


Collection of Active Structures of Target Class
Bond Cleavage According to Chemical Rules
Collection of Fragments
Analysis of Fragments (Frequency of Occurrence, Cluster Analysis)
Selection of Fragments Representing Target Motifs
Transform Fragments into Monomers
Design Libraries Containing Target Motifs



1 = Amide

2 = Ester

3 = Amine

4 = Urea

5 = Ether

6 = Olefin

7 = Quarternary nitrogen

8 = Aromatic nitrogen - aliphatic carbon

9 = Lactam nitrogen - aliphatic carbon

10 = Aromatic carbon - aromatic carbon

11 = Sulphonamide

Figure 1. Top: The Working Mechanism of the RECAP Algorithm. Bottom: The Standard Bond Cleavage Types Employed in the RECAP Process, Reprinted with Permission from (Lewell, Judd, Watson, & Hann, 1998), Copyright J Chem Inf Comput Sci, 1998

This brief review aims to provide a fundamental outline of the recent developments in the De novo molecular designs and their various areas of applications in biomedicine, sustainability developments, and high-performance processing operations.

2. The Current Trends of De Novo Molecular Designs

The goal of developing such advanced materials, especially in the biomedical industry, is to incorporate desired and precise properties into the drugs. However, the integration of desired properties needs optimization of the maximum number of process variables, making this field of science one of the most growing domains (Shoichet, 2004; Scior et al., 2012; Author, 2012). While there are approximately 10^{23}-10^{60} drug like species, the computational sciences have only been able to describe 10^{8} molecules owing to the vast molecular space and accounting for several complex variables. With the evolution in the scientific biome, advanced computer-aided molecular designs are one of the frontiers in developing novel materials because of enhanced calculation methods, cost efficiency, precise, and relatively high accuracy (Cheng, Li, Zhou, Wang, & Bryant, 2012; Reymond, van Deursen, Blum, & Ruddigkeit, 2010). A traditional approach to determine the computationally viable molecules is to access them via the digital library and then experiment with the same, thus reducing the time consumed to trail hundreds of novel units. Molecules in the library may not meet the given criteria. In this case, traditional optimization methods such as a genetic algorithm can be used to enhance further molecular properties beyond the requirements by structural modifications (Cheng, Li, Zhou, Wang, & Bryant, 2012; Reymond, van Deursen, Blum, & Ruddigkeit, 2010; Lim et al., 2018). However, they have a fundamental limitation in terms of efficiency because many trials and errors are inevitable to optimize molecular properties in a substantial molecular space. Kim et al. had recently proposed a novel strategy to design molecular frameworks reinforced on the conditional variational autoencoder (Figure 2). The approach skips the high throughput virtual screening step and uses deep learning-based generative models directly to fabricate molecules having specific target properties. A condition vector is incorporated into the system, which regulates the target properties simultaneously when exposed to a particular environment (Lim et al., 2018). The group demonstrated that it was possible to induce five target properties (MW, LogP, HBD, HBA, and TPSA) having an error range of 10%. Moreover, the property and the behavior of one target property can be instantaneously tuned without disturbing the other working parameters, thus making the algorithm one of the robust techniques to design molecular platforms (Lim et al., 2018).

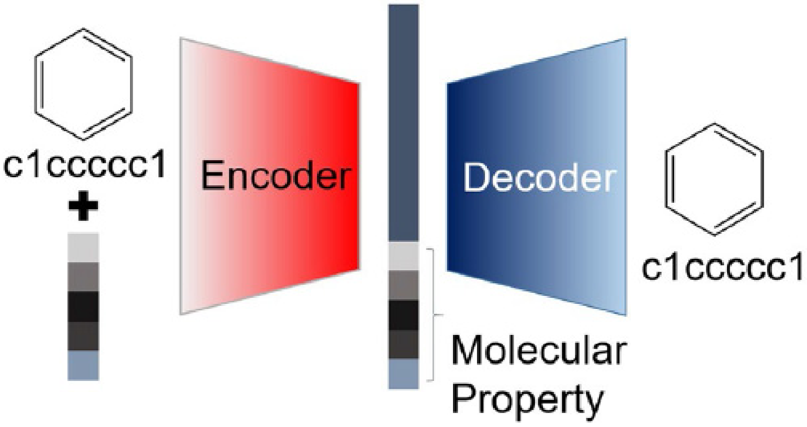


Figure 2. Representation of the Conditional Variational Autoencoder to Develop Efficient Molecular Designs, Reprinted with Permission from (Lim et al., 2018), Copyright BMC, 2018

With the advent of various organo-metallic frameworks, the usages of multiple polymer and polymer hybrids have emerged to be one of the potent candidates for applications in solar cells, organic light-emitting diodes, conductors, sensors and ferroelectrics (Niu, Guo, & Wang, 2015; Kaji et al., 2015; Ueda et al., 2014; Yeung & Yam, 2015). These designs of the organic framework are often accompanied by the optimized designs and predicted drawbacks to generate the most reinforced structure from the "chemical space" for a given environment (Kaji et al., 2015). Various modern theories narrate that often these engineered designs get undermined when compared to neural networks for text and image recognition. The variational autoencoder, which was first developed by Gomez-Bombarelli et al., was refurbished by Kusner et al. to grammar variational autoencoder, which generated the strings multiple times in order the reduce the error and the loss of data due to non-detection of the strings (Kusner, Paige, & Hernández-Lobato, 2017). The improvement made by Segler et al. had demonstrated that long shirt term memory could be used to achieve a recurrent neural network, which accelerated the development of SMILES. The group generated a large number of data points and then used a black-box optimization theory to choose the high performing molecules (Segler et al., 2017). In this context, Tsuda and his colleagues reported communicating a novel Python library (ChemTS) that bridges de novo molecular designs with material science. The set of SMILES strings is represented as a search tree where the *i*th level corresponds to the *i*th symbol (Yang, Zhang, Yoshizoe, Terayama, & Tsuda, 2017). A SMILES route is presumed to be completed when the string travels from the path to the terminal node. The initial root node based search tree was initiated by the Monte Carlo Tree search (a randomized best-fit search method), which effectively creates downstream channels and shallow tree via the rollout optimization (Figure 3). The amalgamation of the Monte Carlo Tree search method resulted in the library to exhibit to perform efficiently than the conventional systems (creating 40 molecules per minute). Moreover, the authors believe that this technique may be excellent in the alloy designing process, which involved the combination of multiple variables aiming for numerous desirable properties (Yang, Zhang, Yoshizoe, Terayama, & Tsuda, 2017).

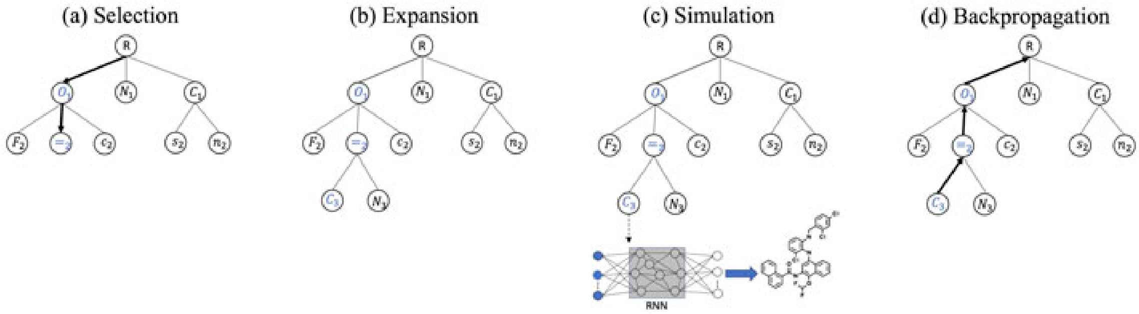


Figure 3. The Monte Carlo Search Technique Consisting of Selection, Expansion, Simulation, and Back Prorogation, Reprinted with Permission from (Yang, Zhang, Yoshizoe, Terayama, & Tsuda, 2017), Copyright Taylor and Francis, 2017

In another instance, Tsuda and his research group developed molecule generators (ChemGE), which have the ability to fabricate multiple molecules with the complement from parallel computation (Yoshikawa, Terayama, Honma, Oono, & Tsuda, 2018). The working principle which ChemGE applies is slightly different from that of the conventional strategies. The novel technique of the population-based grammatical evolution model optimizes the number of unique molecules to be developed (Yoshikawa, Terayama, Honma, Oono, & Tsuda, 2018). The grammatical evolution works upon a given population to optimize a set of strings that operates by a context-free grammar. Such population-based evolutionary methods have been regaining popularity for solving black-box optimization problems, such as hyperparameter optimization and neural network design, because of their inherent concurrency. When contrasted with the programs that operate using SMILES, the mutation operation in ChemGE enables in the probability of developing a higher number of molecules. It is indeed that SMILES inherently poses an organized way to distribute the molecular graph; however, the linear representation of a molecule graph may have various limiting factors (Yoshikawa, Terayama, Honma, Oono, & Tsuda, 2018).

Furthermore, mutation operations in evolution ensure broad diversity throughout the optimization process. The drug-likeness score, which was used to benchmark the fabricated molecule, exhibited that ChemGE can yield a higher number of molecules (because it uses an in-depth learning-based approach) when contrasted with the traditional computational science techniques. Using a parallel probe of 32 cores generated 189 molecules whose docking scores are better than the best molecule in a database (DUD-E18) in 26 hours (Figure 4) (Yoshikawa, Terayama, Honma, Oono, & Tsuda, 2018).

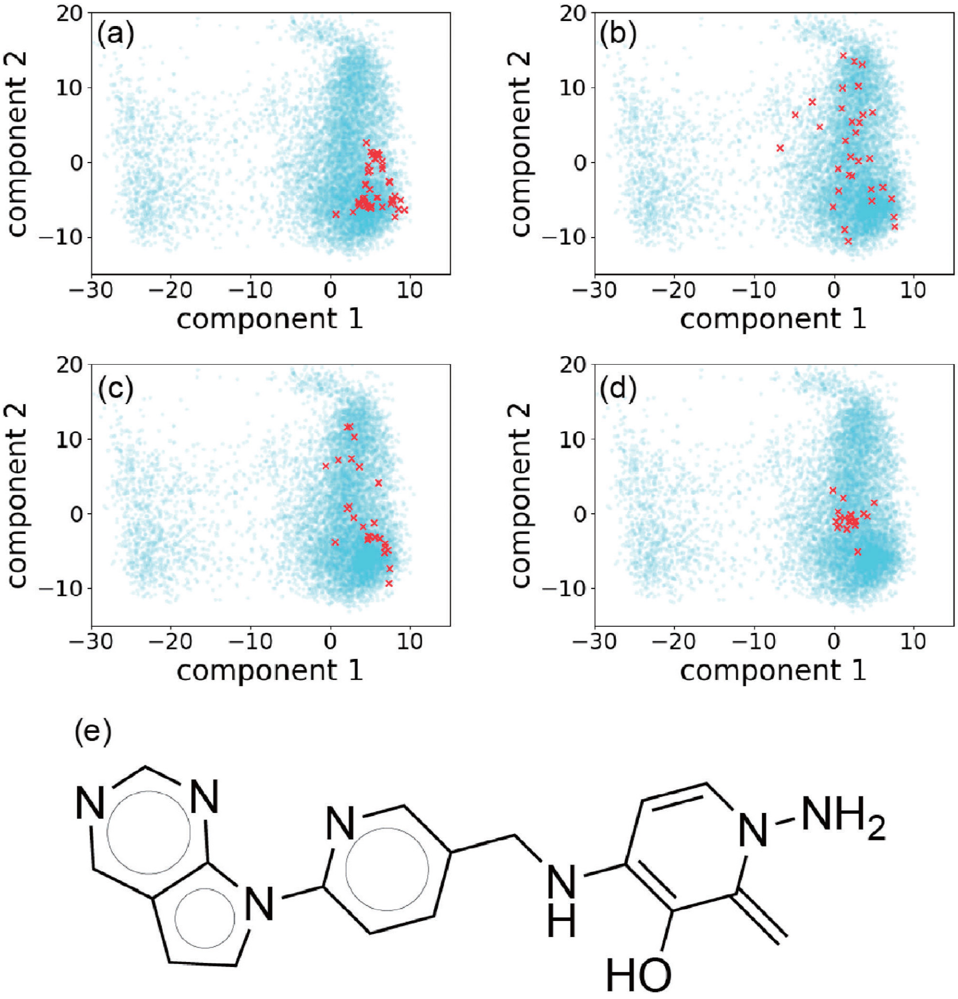


Figure 4. ISOMAP Visualization of the Chemical Space and the Best Molecule. Blue Dots Represent Molecules in the ZINC Database. (e) Shows an Example of Generated Molecules Whose Score Is Better than Known Inhibitors, Reprinted with Permission from (Yoshikawa, Terayama, Honma, Oono, & Tsuda, 2018), arXiv, 2018

Waller et al. long short term memory based recurrent neural networks can be potentially applied to the statistical chemical language model (Segler, Kogej, Tyrchan, & Waller, 2017). With similar physiochemical properties, this algorithm can develop a higher number of new molecules, which can further be extended to create a virtual screening. The model was observed to behave transfer learning when disintegrated into smaller sets of molecules responsive towards a specific target. The program can be a potential for robot conducting synthesis and biological testing as it can autonomously initiate the process owing to the generation of the language model on multiple iterations. One of the prime advantages of using the system is that it provides a bolstered framework to address various molecular generation approaches. Moreover, the model amalgamates the structure generation and optimization, thus behaving as a dual responsive algorithm. Although interpretability is one of the significant drawbacks of the system, the small work step to cast molecule generation as a reinforcement learning problem is something that we all should look forward to (Segler, Kogej, Tyrchan, & Waller, 2017).

4. Conclusions

As chemistry is the language of nature, every day, it is mutating somehow to make our living style better than the previous day. However, in the case of medicinal chemistry, fabricating drug to ensure our well being is challenging (Whitesides, 2015). One of the many challenges in drug design is the vast size of the search space for novel molecules. From a large pool of synthetic chemicals, it is arduous to pick a drug with specific functionality for targeted treatment in a short period. To address the sustainability challenge, modern high-throughput screening techniques allow testing of this molecular space in the laboratory every day. However, the more vast space, the higher the number of trails to be conducted, and hence the more shall become less cost-efficient. Thus, the evolution of computational science developed, which amalgamated theory with experiments to narrow down the search space. De novo drug design is one of the evolving technology that shall disrupt the medical industry once it reaches its peak owing to its potential to fabricate active molecules for tailor-made biological applications.

Conflict of Interest

Sayan Basak declares that he has no conflict of interest.

Human/Animal Rights

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (770-778). IEEE, Las Vegas, NV, USA. https://doi.org/10.1109/CVPR.2016.90

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A. … Hassabis, D. (2017).

Green, M. L. et al. (2017). Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Appl. Phys. Rev.*, *4*, 011105. https://doi.org/10.1063/1.4977487

Tabor, D. P., Rosch, L. M., & Guzik, A. A. (2018). Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater*, *3*, 5-20. https://doi.org/10.1038/s41578-018-0005-z

Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: Recent applications and prospects. *npj Comput. Mater*, *3*, 54. https://doi.org/10.1038/s41524-017-0056-5

Audus, D. J., & de Pablo, J. J. (2017). Polymer informatics: Opportunities and challenges. *ACS Macro Lett.*, *6*, 1078-1082. https://doi.org/10.1021/acsmacrolett.7b00228

Schneider, G., & Fechner, U. (2005). *Nat. Rev. Drug Discovery*, *4*, 649. https://doi.org/10.1038/nrd1799

Lewell, X. Q., Judd, D. B., Watson, S. P., Hann, M. M., & Chem, J. (1998). *Inf. Comput. Sci.*, *38*, 511. https://doi.org/10.1021/ci970429i

Shoichet, B. K. (n.d.). Virtual screening of chemical libraries. *Nature*, *432*(7019), 862. https://doi.org/10.1038/nature03197

Scior, T., Bender, A., Tresadern, G., Medina-Franco, J. L., Martínez-Mayorga, K., … Agrafiotis, D. K. (2012). Recognizing pitfalls in virtual screening: A critical review. *J Chem Inf Model*, *52*(4), 867. https://doi.org/10.1021/ci200528d

Cheng, T., Li, Q., Zhou, Z., Wang, Y., & Bryant, S. H. (2012). Structure-based virtual screening for drug discovery: A problem-centric review. *AAPS J*, *14*(1), 133. https://doi.org/10.1208/s12248-012-9322-0

Reymond, J. L., van Deursen, R., Blum, L. C., & Ruddigkeit, L. (2010). Chemical space as a source for new drugs. *MedChemComm*, *1*(1), 30. https://doi.org/10.1039/c0md00020e

Lim, J. (2018). Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J Cheminform*, *10*, 31. https://doi.org/10.1186/s13321-018-0286-7

Niu, G., Guo, X., & Wang, L. (2015). Review of recent progress in chemical stability of perovskite solar cells. *J MaterChem A.*, *3*(17), 8970-8980. https://doi.org/10.1039/C4TA04994B

Kaji, H. et al. (2015). Purely organic electroluminescent material realizing 100% conversion from electricity to light. *Nat Commun*, *6*, 8476. https://doi.org/10.1038/ncomms9476

Ueda, A., Yamada, S., Isono, T., Kamo, H., Nakao, A., Kumai, R., … Mori, H. (2014). Hydrogen-Bond-Dynamics-Based Switching of Conductivity and Magnetism: A Phase Transition Caused by Deuterium and Electron Transfer in a Hydrogen-Bonded Purely Organic Conductor Crystal. *Journal of the American Chemical Society*. https://doi.org/10.1021/ja507132m

Yeung, M. C. L., & Yam, V. W. W. (2015). Luminescent cation sensors: From host-guest chemistry, supramolecular chemistry to reaction-based mechanisms. *Chem Soc Rev.*, *44*(13), 4192-4202. https://doi.org/10.1039/C4CS00391H

Kusner, M. J., Paige, B., & Hernández-Lobato, J. M. (2017). Grammar variational autoencoder. In *Proceedings of 34th international Conference on Machine Learning (ICML)* (pp. 1945-1954). Sydney.

Segler, M. H. et al. (2017). *Generating focussed molecule libraries for drug discovery with recurrent neural networks*. arXiv preprint arXiv:170101329. https://doi.org/10.1021/acscentsci.7b00512

Yang, X. F., Zhang, J. Z., Yoshizoe, K., Terayama, K., & Tsuda, K. (2017). ChemTS: An Efficient Python Library for de novo Molecular Generation. *Science and Technology of Advanced Materials*. https://doi.org/10.1080/14686996.2017.1401424

Yoshikawa, N., Terayama, K., Honma, T., Oono, K., & Tsuda, K. (2018). P*opulation-based De Novo Molecule Generation, Using Grammatical Evolution*. Chemistry Letters. https://doi.org/10.1246/cl.180665

Segler, M., Kogej, T., Tyrchan, C., & Waller, M. (2017). *Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks*. ACS Central Science. https://doi.org/10.1021/acscentsci.7b00512

Whitesides, G. M. (2015). Reinventing chemistry. *Angew. Chem., Int. Ed.*, *54*, 3196-3209. https://doi.org/10.1002/anie.201410884