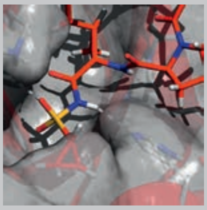
doi:10.2533/chimia.2021.211 Chimia 75 (2021) 211–212 © Swiss Chemical Society



Medicinal Chemistry and Chemical Biology Highlights

Division of Medicinal Chemistry and Chemical Biology
A Division of the Swiss Chemical Society

Chemistry 4.0: How the Digital Revolution is Changing Chemical Research

Marco Stenta*

*Correspondence: Dr. M. Stenta, Syngenta, E-mail: marco.stenta@syngenta.com

Keywords: Direct Design · Generative Models · Inverse Design · Virtual screening

The chemical industry aims at producing new chemical entities (materials, fuels, pharmaceuticals, agrochemicals, *etc.*) with well-controlled properties. This is achieved by iterating through a Design-Make-Test-Analyze (DMTA) cycle[1] where multidisciplinary teams Design, Make, Test new compounds. The Analysis of the results generates insight that influences the Design step of the subsequent cycles. The DMTA cycle is the central unit of an optimization process that continues until one or more stopping criteria are met (*i.e.* performance, safety, cost, *etc.*).

Enhancements in the DMTA optimization process help the industry keep a competitive edge. Accelerating DMTA iterations and reducing their number would decrease the time-to-market for successful projects. Furthermore, a higher information intensity of the DMTA cycles would lower the number of experiments (molecules synthesized and experimental tests) required to validate or disprove a hypothesis and, overall, to achieve the target objectives. Companies achieved significant progress by looking at both the single steps of DMTA and the connection points. Nevertheless, scientists and process engineers still strive to achieve higher effectiveness and efficiency standards to meet our industries' challenges.

According to Chemistry 4.0 manifestos,[2] the digital transformation will profoundly impact the DMTA cycle. Digitalization will contribute innovative solutions to long-standing problems and will provide business opportunities for sustainable growth. The digital transformation is already impacting the Design stage, the medicinal chemists' historical stronghold. Many chemical companies are responding to the increased pressure to innovate with a holistic approach to Design. For example, in Crop Protection, sustainability criteria are considered alongside product performance at very early project stages. However, the usual process to Design struggles to embed the steers from multidisciplinary teams effectively. In particular, there is a 'leak' in the pipeline between Analysis and Design, and the high data volume generated at each DMTA cycle does not fuel Design as it could and should.

If the chemical space[3] were much smaller than it is, Design and Analysis would be unnecessary: a systematic approach could be used to screen each synthesizable compound and assess its performance against a set of properties. However, the chemical space is immense[4] and surprisingly scarce of 'optimal' compounds that possesses the complex holistic profiles we seek.

In the Analysis stage, the scientist builds or refines a series of inference models linking molecular properties and molecular structures. In the Design phase, the scientist uses the inferring models to decide the compound set to inject into the next DMTA cycle. The relationship between Analysis and Design is key to improve the overall DMTA cycle: loss of information leads to poor Design and, thus, waste and missed opportunities.

Borrowing from Kahneman's metaphor,[5] we could conceptualize the art and craft of Design as two separate 'agents' acting in the designer's head. A 'Generative agent' assembles molecular structures based on chemical patterns acquired by experience. A 'Selection agent' filters the ideas based on the qualitative inferring models generated by analyzing data. As an outcome of the interplay between Generative and Selection agents, a set of chemical structures is passed on to Synthesis because they meet qualitative criteria or serve to validate or disprove a hypothesis.

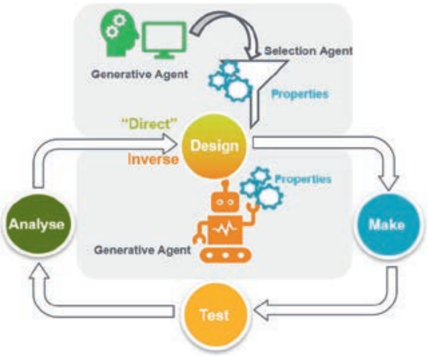


Fig. 1. Direct and Inverse Design approaches in the DMTA cycle.

Early digital approaches to Design mitigate the brain's limitations and the mind's biases to increase the success rate of the DMTA cycle. In the Analysis phase, the data scientist uses software and statistics to build quantitative models linking properties to molecular structure (Quantitative Structure-Property Relationship) and validate specific molecular models (pharmacophore, docking). The medicinal chemist extends his reach by generating a much larger set of compounds (based on hypotheses, molecular models, or cheminformatics software). He selects the most promising ones using quantitative models running on a computer instead of the qualitative ones hosted in his head. The quantitative models are usually arranged in a virtual screening cascade, a multi-stage funnel, analogous to the physical screening cascade. In this **Direct Design** approach, the stages of Generation and Selection are distinct and take place consecutively in a brain or a computer.

The virtual screening enhancement to Direct Design has successfully increased the projects' overall success but it did not fully meet expectations. Improved model quality and higher computational bandwidth have positively impacted virtual screening approaches and constituted vital strategies to support Direct Design better. However, virtual screening suffers from intrinsic limitations that demand a completely different approach to Design.[5] In particular, the tiering of multiple models from coarse/cheap to accurate/expensive introduces an overall error in the virtual screening cascade that might become unbearable. This is especially true when screening extremely large sets of compounds, where rough preliminary filtering mitigates the computational demand at the cost of overall accuracy. In general, virtual screening does not suit well the exploration of the (immense) chemical space. Models derived at the Analysis stage directly impact the Selection portion of Direct Design. At the same time, scientists mediate the influence of Analysis on the Generative agent: the medicinal chemist is still responsible for generating, by combining experience and software, the structural ideas to be screened virtually. With the human acting as the Generative agent's role in Direct Design, there is a substantial risk that information generated in the Test phase is under-used.

The recent developments in Machine Learning propose an alternative to Direct Design to mitigate the information loss between Analysis and Design.[7] The underlying principle is to mimic with an algorithm the subtle interplay between Generative and Selection agents that occurs in the human mind. An algorithm arguably suffers from fewer biases and allows scaling up the whole Design stage. This data-driven approach to Design would augment other design approaches and free up intellectual resources to explore uncharted regions of the 'hypothesis space' scarcely covered by data. The automation of data-driven, model-based Design would indirectly foster hypothesis-driven extrapolative Design to solve complex problems that are still beyond the reach of algorithms.

Early attempts to generate structural recommendations from modelled data were frustrated despite the QSPR community's efforts.[8] The models used in Direct Design contain a functional relationship between molecular structure and property. However, this relationship is uni-directional: the model can predict a molecular property's value given a structure but cannot generate molecular structures from the property's value. The key step in this field was the recent construction of algorithms that generate molecular structures based on the target properties,[9,10] thus effectively reversing the property-structure relationship into what we call **Inverse Design**.

Inverse Design is a form of '*de novo*'[11] design entirely based on algorithms that take as input steers about the desired property and return a set of compounds. The Generative Models embeds (or connect to) external predictive models (QSPR or molecular models) to ensure the generated structures meet the input criteria.[12] This overarching principle is shared by many Inverse Design tools based on different algorithms, molecular representations, learning approach, model embedding, *etc*.[13]

The search strategy limits the computational molecular Design.[14] Intriguingly, some '*de novo*' approaches based on Inverse Design allow, for the first time, an efficient optimization in the chemical-property space by establishing a bi-directional relationship between structure and property. An optimization approach (in the virtual space) overcomes some of the virtual screening pitfalls, since it is compatible with the simultaneous application of expensive and accurate models. Indeed, the optimization approach requires assessing a smaller number of molecular structures to explore a given chemical space, as compared to virtual screening.

In one of the earliest approaches to inverse design[15] the algorithm generates a continuous representation of the discrete molecular space. The resulting Generative model allows perturbing known chemical structures or interpolating between molecules by moving into this open-ended latent space. The continuous representations will enable the use of robust gradient-based optimization to guide the search for optimized functional compounds efficiently.

We observe many academic groups,[16] start-ups,[17] and large companies[13] investing in the development of Generative Models to support Inverse Design. The progress is immense, and we read about successful applications in material and drug design. Interestingly alongside the scientific outcome of the computational research, computer scientists publish the software code used to generate the results. This new, open approach increases reproducibility and allows other scientists to modify and customize the code. Building on top of each other's work is boosting research in the field.

From far above, both Direct Design (based on virtual screening) and Inverse Design (based on Generative Models) look deceptively similar. In both cases, the Design 'box' receives as input a set of models from the Analysis and target values for selected properties. In both cases the Design 'box' returns, as output, a series of molecular structures that stand a higher chance of passing through the biological screening cascade and, ultimately, turn into a viable product. However, there are substantial differences in what happens inside the 'box'. Inverse Design fuels the idea generation directly from the data and increases the value of the Tests. In other terms, Inverse Design, supported by Generative Models, is better linked to Analysis than standard approaches. Data and models impact directly and strongly the 'Generative Agent', not just the 'Selection agent'. As a result, molecular structures are generated that match better the desired property profile and do not require multi-stage virtual screening protocols. The opportunity of optimization in the molecular space allows an unprecedented exploration of the property space that is both more efficient and faster. This constitutes a form of **Artificial Intuition** that improves with data and scales at will. The revolution lies here.

Inverse Design natively embeds multidisciplinary steers into the generation of structural ideas. The advent of Generative models[18] enables us to explore uncharted regions of the chemical/property space. We expect Inverse Design to increasingly improve the effectiveness of DMTA cycles and change the way scientists approach the Design of new chemical entities.

Received: February 5, 2021

[1] A. T. Plowright, C. Johnstone, J. Kihlberg, J. Pettersson, G. Robb, R. A. Thompson, Drug Discov. Today 2012, 17, 56; https://doi.org/10.1016/j.drudis.2011.09.012.

[2] Deloitte, https://ww2.deloitte.com/global/en/pages/consumer-industrial-products/articles/cip-chemistry.html.

[3] J.-L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722; https://doi.org/10.1021/ar500432k.

[4] P. G. Polishchuk, T. I. Madzhidov, A. Varnek, *J. Computer-aided Mol. Des.* **2013**, *27*, 675; https://doi.org/10.1007/s10822-013-9672-4.

[5] D. Kahneman, 'Thinking, fast and slow', Farrar, New York, **2011**.

[6] a) D. Stumpfe, J. Bajorath, *J. Chem. Inform. Model.* **2020**, *60*, 4112; https://doi.org/10.1021/acs.jcim.9b01101; b) T. Scior, A. Bender, G. Tresadern, J. L. Medina-Franco, K. Martínez-Mayorga, T. Langer, K. Cuanalo-Contreras, D. K. Agrafiotis, *J. Chem. Inform. Model.* **2012**, *52*, 867; https://doi.org/10.1021/ci200528d;

[7] A. Tkatchenko, *Nature Commun.* **2020**, *11*, 4125; https://doi.org/10.1038/s41467-020-17844-8.

[8] T. Miyao, H. Kaneko, K. Funatsu, *J. Chem. Inform. Model.* **2016**, *56*, 286; https://doi.org/10.1021/acs.jcim.5b00628.

[9] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, *Mol. Pharmaceut.* **2017**, *14*, 3098; https://doi.org/10.1021/acs.molpharmaceut.7b00346.

[10] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360; https://doi.org/10.1126/science.aat2663.

[11] P. Schneider, G. Schneider, *J. Med. Chem.* **2016**, *59*, 4077; https://doi.org/10.1021/acs.jmedchem.5b01849.

[12] J. Noh, G. H. Gu, S. Kim, Y. Jung, *Chem. Sci.* **2020**, *11*, 4871; https://doi.org/10.1039/D0SC00594K.

[13] T. Blaschke, J. Arús-Pous, H. Chen, C. Margreitter, C. Tyrchan, O. Engkvist, K. Papadopoulos, A. Patronov, *J. Chem. Inform. Model.* **2020**, *60*, 5918; https://doi.org/10.1021/acs.jcim.0c00915.

[14] G. Schneider, *Nature Rev. Drug Discov.* **2010**, *9*, 273; https://doi.org/10.1038/nrd3139.

[15] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268; https://doi.org/10.1021/acscentsci.7b00572.

[16] R. Winter, J. Retel, F. Noé, D.-A. Clevert, A. Steffen, *Bioinform.* **2020**, *36*, 4093; https://doi.org/10.1093/bioinformatics/btaa271.

[17] Y.A. Ivanenkov, A. Zhebrak, D. Bezrukov, B. Zagribelnyy, V. Aladinskiy, D. Polykovskiy, E. Putin, P. Kamya, A. Aliper, A. Zhavoronkov, 'Chemistry42: An AI-based platform for de novo molecular design', 22.01.2021.

[18] Q. Vanhaelen, Y.-C. Lin, A. Zhavoronkov, *ACS Med. Chem. Lett.* **2020**, *11*, 1496; https://doi.org/10.1021/acsmedchemlett.0c00088.