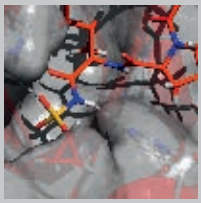
doi:10.2533/chimia.2021.105 Chimia 75 (2021) 105–107



Medicinal Chemistry and Chemical Biology Highlights

Division of Medicinal Chemistry and Chemical Biology

A Division of the Swiss Chemical Society

Augmenting Chemical Space with DNA-encoded Library Technology and Machine Learning

Laura Guasch, Michael Reutlinger, Daniel Stoffler, and Moreno Wichert*

*Correspondence: Dr. M. Wichert, E-mail: moreno.wichert@roche.com, Roche Innovation Center Basel, Grenzacherstrasse 124, CH-4070 Basel

Keywords: Affinity-based selections/compound screening · Combinatorial chemistry · DNA-encoded library technology (DELT) · Machine learning · Neural networks

DNA-encoded library (DEL[1]) technology has emerged as one of the fastest and most cost-effective screening platforms available in industry both for ***hit discovery***[2] as well as more recently for ***druggability*** and ***tractability assessments*** and successive ***prioritization*** of therapeutic targets in the early phase of drug discovery programs.[3]

The key principle of DELs is based on the combinatorial assembly (synthesis) of library members from chemical building blocks (BBs) and the corresponding tagging of each BB with unique DNA sequences (barcodes) in an alternating fashion of chemical reactions and DNA ligations. In analogy to phage display technology,[4] this physical linkage (Fig. 1) of small organic molecules with distinctive ***DNA barcodes*** enables to ***deconvolute*** the ***chemical identity*** (structure) of each and every molecule by next-generation sequencing (NGS) at any time.[5]

Originally proposed by Brenner and Lerner in a theoretical paper in 1992[6] it was not until 2004 as a result of the remarkable advancements in NGS, that several academic groups[7] reduced the technology to practice with multiple implementations of encoding schemes and library designs resulting in distinguished IP space [8] and its commercial exploration by an entire new industry.

Attributable to the specific encoding system, the ***combined set of libraries*** (pool) can be stored in a single test-tube and ***billions of potential ligands can be screened*** as mixtures all at once in a simple, one-day binding experiment (panning) against the target of choice (in general, recombinant protein of high purity and quality is needed). The DNA tags of library members allow for further exponential amplification by polymerase chain reaction (PCR), thus, even minute amounts of binders can be detected and unambiguously identified by deep sequencing after (heat) elution from the target.[9] The obtained sequencing data is evaluated by calculating an ***enrichment ratio (ER)*** or score[10] of preferential binders compared to the background (defined matrix/non-target control) and the results are displayed using dedicated chemical analysis software (*e.g.* TIBCO Spotfire). Identifying patterns or fingerprints (chemical series) within the same library and across different libraries facilitate the discrimination of binding from non-binding library members.

DELT has proven to be robust in delivering novel (and often) radically different ***chemical starting points*** for medicinal chemistry programs ***within Roche*** and also ***externally***. Not surprisingly, DELT takes now a firm place in the screening armamentarium of almost every pharma company (either as an in-house operated platform or accessed through a CRO) with a constantly growing number of success stories[5] (*e.g.* appearance of several ***DELT-derived molecules in the clinic***[11]) and even more players in the market.[12]

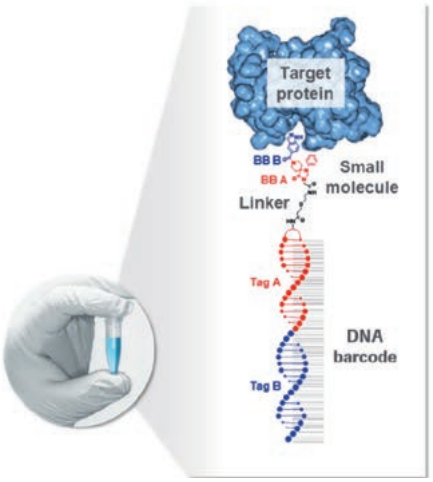


Fig. 1. Schematic representation of a DNA-encoded library member (oligo-compound conjugate) binding to a target protein of interest. The DNA barcode is chemically conjugated to the small molecule (*via* a long linker) and carries the unambiguous structural information of the displayed compound (*e.g.* encoding tags A and B corresponding to the synthesis scheme and identity of building blocks (BBs) A and B of the final structure. Thanks to the DNA, even minute amounts of binders are effectively amplified by PCR and subsequently identified (sequenced and counted) by next-generation sequencing after heat elution from an affinity-based selection experiment with a library mixture input of billions of molecules.

For academic researchers, who are often confronted with budget constraints, DELT offers a convenient and relatively cheap source for accessing tool compounds from large chemical repertoires, *e.g.* to use hits from DELT screens as probes for the elucidation of complex cellular signaling pathways or the analysis of biostructural dynamics and interaction studies.

However, one limiting factor in the whole DELT process is the actual ***hit follow-up***, *i.e.* the binding or activity confirmation with resynthesized hit compounds off-DNA. Hits have to be resynthesized in milligram quantities due to the extremely low abundance of library members in the form of DNA-compound conjugates. No current chemical analytics technique can quantify or characterize these DELT library members in the library pool. The only method which is able to clearly identify molecules in the library is exponential amplification by PCR (and quantify by qPCR) and subsequent decoding of the DNA barcodes by deep sequencing (read and count).

Whereas the DELT screen from target arrival to the processed hit list with ER values may take less than one week, the *de novo* chemical synthesis of these hits without DNA barcodes

Can you show us your Medicinal Chemistry and Chemical Biology Highlight?
Please contact: Dr. Fides Benfatti, E-mail: fides.benfatti@syngenta.com, Syngenta Crop Protection, WST-820-2-15 Schaffhauserstrasse, CH-4332 Stein

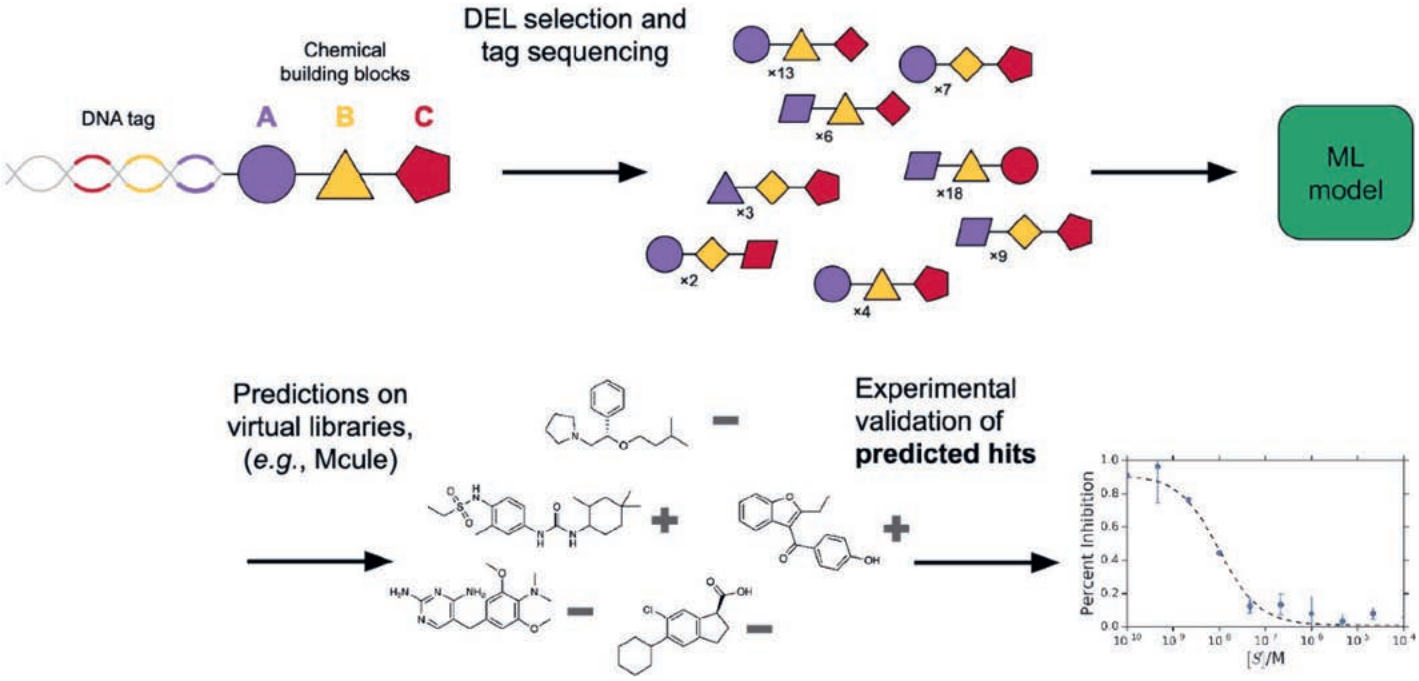


Fig. 2. General concept of machine learning models based on DELT selection data. Starting with a chemically synthesized DNA-encoded library, an affinity-mediated selection is performed against the target of interest, and the DNA tags of binding molecules are PCR-amplified and sequenced following heat elution from the target. The aggregated DELT selection data (disynthon representations) is first used as a training set for machine learning models and subsequently these trained algorithms are run to predict hits from virtual libraries or commercially available catalogs such as provided by Mcule. Predicted hit compounds are ordered or synthesized and tested experimentally to confirm activity in functional assays. Reprinted (adapted) with permission from ref. [15] *J. Med. Chem.* **2020**, *63*, 16, 8857–8866, Publication Date: June 11, 2020, https://doi.org/10.1021/acs.jmedchem.0c00452. Copyright (2020) American Chemical Society.

may take up to 6 months, depending on the number of hits to follow up, the availability of the initial building blocks (starting material), the invested resources and number of chemists allocated to the task. There are two main avenues which are followed to shorten this essential step of biophysical or biochemical ***hit validation***:

1. Using assay methods which are amenable for testing the conjugates as they are present in the library. Such systems have been implemented, for example, by the group of Dario Neri, which uses short LNA duplexes (more stable than DNA) with certain fluorophores incorporated for fluorescence anisotropy experiments or conjugated with biotin for immobilization on a streptavidin-modified gold sensor chips for measuring binding kinetics by surface plasmon resonance (SPR).[13]
2. *In silico* methods are employed for searching close analogs of hits among in-house available compound collections and from vendors catalogues who guarantee fast delivery.[14]

Both methods are currently applied with some success but the throughput for 1) and the hit rate of 2) still limits the efficient use in daily medicinal chemistry lead expansion work.

As mentioned earlier, the sequencing output of DELT screens is typically analyzed by calculating an ER for each individual library member from the sequence count in target selection conditions versus non-target controls. The hits above a certain ER threshold are then visually and interactively inspected library-by-library with the aim to identify structural patterns and chemical motifs of interest.

These extensive analysis efforts limit the throughput of molecules considered, introduce bias, and make it difficult to fully overview and utilize the subtle patterns in the depth of DELT data.

In a recent study[15] through a collaboration between ***ZebiAI***, ***Google Accelerated Science (GAS)*** and ***X-Chem***, researchers presented a method to circumvent two of the current main limitations of DELT: human bias in result analysis and the time-consuming and expensive resynthesis step of compounds off-DNA for hit validation. Accordingly, a combination of ***physical screening data from DELT*** selections was used to build a surprisingly ***effective machine-learning (ML) model*** (Fig. 2). The ML approach allows for the discovery of complex patterns otherwise nearly impossible for a scientist to detect by visual inspection of hundreds of millions of data points derived from DELT selection/sequencing output.

By generating models to targets of interest, the ML algorithm is able to predict activities of collections of compounds that were not in the physical DELs. Hence, the universe of chemical space can be easily explored by sourcing from existing compound collections and vendors at little expense..

ML-supported DELT analysis could also be of value to more accurately predict matrix binders (not binding to the target of interest) and consequently diminish the false-positive rates of DELT screens overall. Once these ML-based similarity searches in existing compound collections have become a robust, well trained and routinely performed method for faster hit finding, one could even envision to leverage the enormous DELT data pool together with sophisticated ML approaches for *de novo* predictions/design of novel compounds to augment the chemical space of the original DELT screening data set to an entire new universe.

In conclusion, DEL technology has become a widely accepted and routinely used method for hit finding across the pharma industry (and academic labs), enabling access to broad chemical diversity through a fast, single-well binding assay thereby complementing other high-throughput screening efforts.

As the number of DELT screens (and the resulting amount of data) is continuously growing, novel ML-based approaches expedite the data analysis to unprecedented levels. We are now

routinely executing DELT screening and analysis and providing novel chemical starting points for our small-molecule research teams to explore.

Received: January 5, 2021

[1] DNA-encoded library technology (DELT), also called encoded library technology (ELT) or DNA-encoded chemical library (DECL).
[2] J. Ottl, L. Leder, J. V. Schaefer, C. E. Dumelin, *Molecules* **2019**, *24*, https://doi.org/10.3390/molecules24081629.
[3] C. A. Machutta, C. S. Kollmann, K. E. Lind, X. P. Bai, P. F. Chan, J. Z. Huang, L. Ballell, S. Belyanskaya, G. S. Besra, D. Barros-Aguirre, R. H. Bates, P. A. Centrella, S. S. Chang, J. Chai, A. E. Choudhry, A. Coffin, C. P. Davie, H. F. Deng, J. H. Deng, Y. Ding, J. W. Dodson, D. T. Fosbenner, E. N. Gao, T. L. Graham, T. L. Graybill, K. Ingraham, W. P. Johnson, B. W. King, C. R. Kwiatkowski, J. Lelievre, Y. Li, X. R. Liu, Q. N. Lu, R. Lehr, A. Mendoza-Losana, J. Martin, L. McCloskey, P. McCormick, H. P. O'Keefe, T. O'Keeffe, C. Pao, C. B. Phelps, H. W. Qi, K. Rafferty, G. S. Scavello, M. S. Steiginga, F. S. Sundersingh, S. M. Sweitzer, L. M. Szewczuk, A. Taylor, M. F. Toh, J. Wang, M. H. Wang, D. J. Wilkins, B. Xia, G. Yao, J. Zhang, J. Y. Zhou, C. P. Donahue, J. A. Messer, D. Holmes, C. C. Arico-Muendel, A. J. Pope, J. W. Gross, G. Evindar, *Nat. Commun*. **2018**, *9*, https://doi.org/ARTN 1622710.1038/ncomms16227.
[4] R. Barderas, E. Benito-Pena, *Anal. Bioanal. Chem.* **2019**, *411*, 2475, https://doi.org/10.1007/s00216-019-01714-4.
[5] a) R. A. Goodnow, C. E. Dumelin, A. D. Keefe, *Nat. Rev. Drug Discov.* **2017**, *16*, 131, https://doi.org/10.1038/nrd.2016.213; b) D. Neri, R. A. Lerner, *Annu. Rev. Biochem.* **2018**, *87*, 479, https://doi.org/10.1146/annurev-biochem-062917-012550; c) R. A. Lerner, D. Neri, Biochem. *Bioph. Res. Co.* **2020**, *527*, 757, https://doi.org/10.1016/j.bbrc.2020.04.080.
[6] S. Brenner, R. A. Lerner, *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 5381, https://doi.org/10.1073/pnas.89.12.5381.
[7] a) Z. J. Gartner, B. N. Tse, R. Grubina, J. B. Doyon, T. M. Snyder, D. R. Liu, *Science* **2004**, *305*, 1601, https://doi.org/10.1126/science.1102629; b) S. Melkko, J. Scheuermann, C. E. Dumelin, D. Neri, *Nat. Biotechnol.* **2004**, *22*, 568, https://doi.org/10.1038/nbt961; c) D. R. Halpin, P. B. Harbury, *PLoS Biol.* **2004**, *2*, E173, https://doi.org/10.1371/journal.pbio.0020173; d) D. R. Halpin, P. B. Harbury, *PLoS Biol*. **2004**, *2*, E174, https://doi.org/10.1371/journal.pbio.0020174; e) D. R. Halpin, J. A. Lee, S. J. Wrenn, P. B. Harbury, *PLoS Biol*. **2004**, *2*, E175, https://doi.org/10.1371/journal.pbio.0020175.
[8] a) M. H. Hansen, P. Blaksjaer, L. K. Petersen, T. H. Hansen, J. W. Hoifeldt, K. V. Gothelf, N. J. V. Hansen, *J. Am. Chem. Soc.* **2009**, *131*, 1322, https://doi.org/10.1021/ja808558a; b) M. A. Clark, R. A. Acharya, C. C. Arico-Muendel, S. L. Belyanskaya, D. R. Benjamin, N. R. Carlson, P. A. Centrella, C. H. Chiu, S. P. Creaser, J. W. Cuozzo, C. P. Davie, Y. Ding, G. J. Franklin, K. D. Franzen, M. L. Gefter, S. P. Hale, N. J. Hansen, D. I. Isreal, J. Jiang, M. J. Kavarana, M. S. Kelley, C. S. Kollmann, F. Li, K. Lind, S. Mataruse, P. F. Medeiros, J. A. Messer, P. Myers, H. O'Keefe, M. C. Oliff, C. E. Rise, A. L. Satz, S. R. Skinner, J. L. Svendsen, L. Tang, K. van Vloten, R. W. Wagner, G. Yao, B. Zhao, B. A. Morgan, *Nat. Chem. Biol.* **2009**, *5*, 647, https://doi.org/10.1038/nchembio.211; c) M. Wichert, N. Krall, W. Decurtins, R. M. Franzini, F. Pretto, P. Schneider, D. Neri, J. Scheuermann, *Nat. Chem.* **2015**, *7*, 241, https://doi.org/10.1038nchem.2158.
[9] W. Decurtins, M. Wichert, R. M. Franzini, F. Buller, M. A. Stravs, Y. X. Zhang, D. Neri, J. Scheuermann, *Nat. Protoc.* **2016**, *11*, 764, https://doi.org/10.1038/nprot.2016.039.
[10] a) J. C. Faver, K. Riehle, D. R. Lancia, J. B. J. Milbank, C. S. Kollmann, N. Simmons, Z. F. Yu, M. M. Matzuk, *ACS Comb. Sci.* **2019**, *21*, 75, https://doi.org/10.1021/acscombsci.8b00116; b) L. Kuai, T. O'Keeffe, C. Arico-Muendel, *Slas. Discov.* **2018**, *23*, 405, https://doi.org/10.1177/2472555218757718.
[11] P. A. Harris, S. B. Berger, J. U. Jeong, R. Nagilla, D. Bandyopadhyay, N. Campobasso, C. A. Capriotti, J. A. Cox, L. Dare, X. Y. Dong, P. M. Eidam, J. N. Finger, S. J. Hoffman, J. Kang, V. Kasparcova, B. W. King, R. Lehr, Y. F. Lan, L. K. Leister, J. D. Lich, T. T. MacDonald, N. A. Miller, M. T. Ouellette, C. S. Pao, A. Rahman, M. A. Reilly, A. R. Rendina, E. J. Rivera, M. C. Schaeffer, C. A. Sehon, R. R. Singhaus, H. H. Sun, B. A. Swift, R. D. Totoritis, A. Vossenkamper, P. Ward, D. D. Wisnoski, D. H. Zhang, R. W. Marquis, P. J. Gough, J. Bertin, *J. Med. Chem.* **2017**, *60*, 1247, https://doi.org/10.1021/acs.jmedchem.6b01751.
[12] Reportlinker.com, 'DNA-Encoded Libraries Platforms and Services Market, 2020-2030', **2020**.
[13] L. Prati, M. Bigatti, E. J. Donckele, D. Neri, F. Samain, *Biochem. Bioph. Res. Co.* 2**020**, *533*, 235, https://doi.org/10.1016/j.bbrc.2020.04.030.
[14] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742, https://doi.org/10.1021/ci100050t.
[15] K. McCloskey, E. A. Sigel, S. Kearnes, L. Xue, X. Tian, D. Moccia, D. Gikunju, S. Bazzaz, B. Chan, M. A. Clark, J. W. Cuozzo, M. A. Guie, J. P. Guilinger, C. Huguet, C. D. Hupp, A. D. Keefe, C. J. Mulhern, Y. Zhang, P. Riley, *J. Med. Chem.* **2020**, *63*, 8857, https://doi.org/10.1021/acs.jmedchem.0c00452.