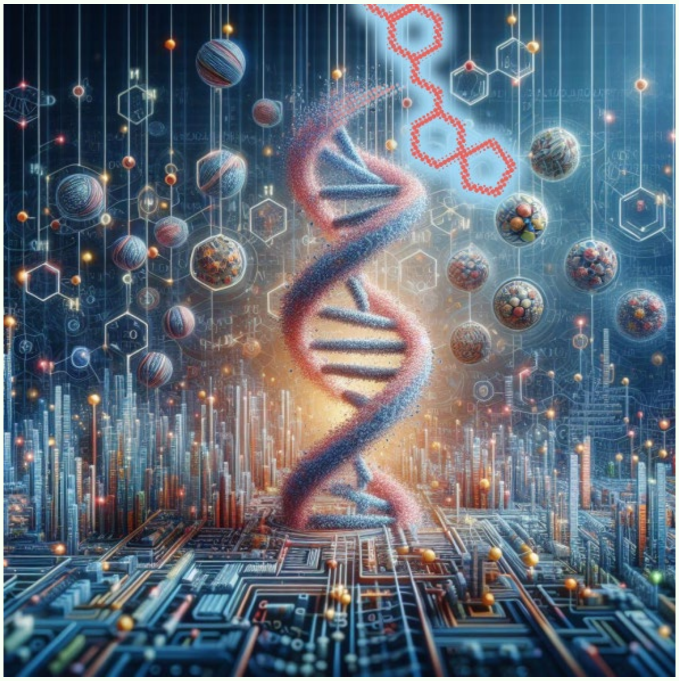
MAY 2024 | VERSION 1

The Beginners Guide to DNA-encoded Libraries



10 Key Concepts to Understand DNA-Encoded Libraries

By Raphael Franzini

If you are reading this, changes are you are enthusiastic about advancing innovations that will change patients' lives through new medicines. You are seeking technologies that can uncover lead molecules beyond what is possible with conventional methods. DNA-encoded libraries (DELs) may be the right approach to propel your discovery efforts forward. However, understanding the potential benefits and pitfalls, and navigating the technical and commercial options associated with DELs can be overwhelming. My mission is to advance DELs and facilitate their widespread adoption by guiding researchers through their initial stages. This guide aims to provide an overview of some key considerations in this exciting field.

Guiding Serendipity

1. Encoding Molecules with DNA

Encoding molecules with DNA is at the heart of DELs, aiming to revolutionize compound discovery by accessing active compounds faster and more affordably than ever before. Like all groundbreaking achievements, DELs build upon previous advancements. They merge the expansive capabilities of combinatorial chemistry with the efficiency of phage display and similar methods to screen vast libraries through simple affinity selection.

Central to this approach is establishing a link between a molecule's physical properties (the phenotype) and a readable DNA barcode (the genotype), a concept borrowed from phage display. DNA serves as an ideal identification barcode because of its unique properties. The ability to amplify minute amounts of DNA through PCR and the precision of high-fidelity DNA sequencing lay the technological foundation of DELs. The advent of high-throughput DNA sequencing has been a game-changer, eliminating the need for iterative selection and amplification cycles used in phage display. This breakthrough enables the screening of libraries containing billions of molecules in just a few hours. What was once an ambitious idea has now become a firmly established approach in drug discovery.

2. Assembling Myriads of Molecules

Creating libraries with millions, billions, or even trillions of molecules, each tagged with a specific DNA code, may sound like an insurmountable task. What makes DEL synthesis possible is an approach called combinatorial synthesis. This involves adding small building blocks in a step-by-step process, combining them in various combinations to exponentially increase the library size. While there are different methods for making DELs, the most common approach nowadays is to add chemical building blocks and DNA codes one after the other in a split-and-pool method.

Thanks to the efforts of numerous research teams, many chemical reactions are available for on-DNA compound synthesis, with DNA-tagging typically done through DNA-ligation. After each round of synthesis and tagging, the DNA-molecule pairs are mixed and then split into separate containers to begin the next cycle of library building. This process allows for extremely large libraries to be created in a manageable number of steps. For instance, consider a DEL made in three cycles, with each cycle adding 1000 different molecules, resulting in a library containing 1 billion molecules.

Today, libraries with tens of billions or even trillions of compounds have been successfully made. One of the great advantages of DELs is that they do not require sophisticated equipment for handling and storage. Regardless of their size, the entire library can be stored frozen in simple tubes. However, it is important not to underestimate the challenge of creating top-quality DELs, as it requires extensive expertise and meticulous attention to detail to ensure the molecules are made and tagged correctly.

3. Blueprint for Effective DELs

At some point, everyone considering using DELs wonders: what makes a library good? This could be when choosing between different commercial options or setting up your own DEL platform. Ultimately, a good DEL is one that gives you the molecule you want. However, this information is not available until you do selections. Nonetheless, there are parameters to consider when comparing library options:

- Library size
- Chemical diversity
- Library geometry
- Physicochemical properties
- Drug-likeness
- Library homogeneity
- Synthetic accessibility

Understanding these aspects will help assess the usefulness of a DEL. When scientists first encounter DELs, they often focused on library size. However, making a library bigger does not necessarily make it better. It is technically easy to make large DELs by using multiple synthetic cycles, but such DELs often contain molecules that are too big, have unfavorable properties, are impure, and are tedious to synthesize. As DEL size increases, it also becomes more challenging to distinguish hits from noise. Alternatively, chemical diversity can be increased by carefully expanding DELs with diverse building blocks and using reactions that produce drug-like molecules. Computational methods can help evaluate chemical diversity, although careful interpretation of chemoinformatic descriptors and linker positioning is necessary to ensure meaningful results.

When starting with DELs, asking the right questions from the outset is crucial. I had the pleasure of co-authoring a review with Dr. Ying Zhang, the Vice-Director of Chemistry at X-Chem, which offers deeper insights into design considerations for DELs: https://link.springer.com/chapter/10.1007/7355_2022_147.

4. Navigating the DEL Selection Process

The most remarkable aspect of DELs is the simplicity of the hit discovery process. DELs are sampled in a straightforward affinity selection protocol. Initially, the target protein is immobilized on a solid support and exposed to the DELs. Binding equilibria are established and a series of wash steps removes conjugates with low target affinity, enriching the molecules that effectively bind to the target. Amplification of the DNA tags and high-throughput DNA sequencing provide quantitative information on this enrichment, facilitating the interpretation and identification of hit compounds. As part of a team of colleagues at ETH Zurich and Philochem, I have published a readily accessible protocol utilizing routine instrumentation for universal application (https://www.nature.com/articles/nprot.2016.039).

There are numerous modified versions of the selection process. While the selection protocols are simple, it is of critical importance to ensure a well-executed selection process for screening success. Any issues with protein immobilization or experimental parameters can result in failure to identify hits or yielding false hits, necessitating extensive follow-up work without producing any active molecules. Meticulous planning and rigorous quality control are thus imperative for a successful DEL experience.

5. Deciphering Sequence Reads into Interpretable Data

The selection process enriches binder-associated DNA codes that are analyzed through high-throughput sequencing, resulting in sequences stored in fasta or fastq formats. These sequences form long lists of individual reads, which need to be converted into an analyzable format and linked to corresponding molecules for further analysis. To achieve this, a simple chemoinformatic platform is required.

This platform transforms the original sequence reads into interpretable data. The output comprises histograms where each encoded library member is linked with the number of sequences found for that molecule. Various normalization and parameterization techniques are applied to facilitate data analysis. Typically, the data is visualized as scatter plots for visual inspection.

For effective visualization and analysis of DEL data and hit structures, DataWarrior is an excellent software tool. Its developers generously provide this software to the scientific community, making it accessible at https://openmolecules.org/datawarrior/.

6. In Search of the Perfect Hit: Strategies for Prioritization

Progress often brings its own challenges. In the case of DELs, their expansion to enormous sizes has boosted hit discovery rates, but this advancement also complicates finding the best molecules. Contemporary DELs often yield thousands of potential hits, far surpassing capabilities to synthesize and test hit compounds.

Identifying the most potent molecules among this abundance of hits requires extensive expertise and a profound understanding of the intricate factors influencing DEL data. While DELs operate under the assumption that DNA-molecule conjugates with high target affinity are enriched the most, there are numerous factors that obscure hit triaging. Uneven distribution among library members, incomplete molecule synthesis, and the presence of side products all distort sequence counts. Moreover, surface immobilization of proteins and DNA attachment can complicate analysis by either enhancing or hindering target engagement. Selection pressures during the process may fail to distinguish between binders of varying affinities, or worse, lead to the attrition of potentially valuable weak binders.

Although some of these uncertainties may be mitigated by strategies such as normalization to control experiments and increasing selection stringency, others remain unavoidable. Effective hit picking demands a holistic approach, considering sequence enrichment, physicochemical properties, and insights from control screens. It also involves identifying structure-activity patterns and leveraging legacy data.

An alternative avenue for advanced hit triaging involves pursuing high-throughput synthesis and testing strategies. Synthesizing compounds on DNAs enables rapid production and exploration of potential side products during library synthesis. Approaches combining affinity binding with mass-spectrometry readouts facilitate swift molecule testing and detection of otherwise obscure actives.

In essence, identifying the optimal hits from DELs is an art that combines experience with a deep understanding of DEL data intricacies. It is not just about finding hits; it's about finding the right ones—an endeavor that requires both skill and insight.

7. From DEL Hit to Clinical Candidates

In drug discovery, hits routinely require optimization to become suitable lead molecules. DELs are no exception, and DEL hits can pose challenges related to their size and other physicochemical properties. Fortunately, often not all building blocks are necessary for effective target engagement. Therefore, one of the first steps in hit optimization is exploring active truncates, which reduces the molecular weight and increases ligand efficiency. Another area for optimization is the chemical structures at the linker attachment site. Modifying these structures can lead to enhancements in potency or improvements in other properties. Structure-enrichment patterns in DEL data can further aid in planning lead development efforts.

Structure-based methods are also valuable for scaffold hopping in DEL hits, a technique our group has explored in a proof-of-concept report (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7530011/). For an overview of lead development examples using DEL-based hits, I recommend an informative review by a team from Janssen Research (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7957921/).

8. The DNA-Attachment Linker: Source of Opportunity and Uncertainty

One recurrent question with DELs is whether the DNA interferes with the selection process. The answer to this question is nuanced. On one hand, the DNA typically points into the solution and is unlikely to undergo sequence-specific interactions with the target.

However, connecting molecules to DNA requires a linker, and this linker can influence how a molecule engages with a protein. Ideally, the linker should point towards the solvent without participating in the binding interaction. Several DEL hits for which X-ray co-crystal structures have been solved fall into this category. However, linkers can also stabilize the interaction of molecules with proteins, which is attractive for lead development. More commonly, the positioning of linkers can restrict accessible binding modes, resulting in many active molecules not being identified as hits. The exact fraction of active molecules that do not register as hits because of the linker is unknown, but our research suggests that for every hit, there are likely many false negatives. In addition to missing potential drug leads, such false negatives can complicate lead development and make machine learning models inaccurate.

DEL linkers can also offer value. There is a rapidly growing number of drug modalities that rely on connecting several molecular modules together. Examples include PROTACs that juxtapose target engagers with ligands of E3 ligases, small-molecule radiochelates, and small-molecule recruiters of the immune system. Converting standard active compounds into such lead molecules is tedious because it requires identifying the correct attachment site and testing panels of active molecules to find the right fit. In contrast, DELs immediately provide the attachment site. Notably, DELs have become the method of choice for companies specializing in PROTACs and have yielded radiopharmaceuticals progressing towards clinical trials.

9. DEL Data Driving AI Innovation

Artificial intelligence (AI) will play a transformative role in drug discovery. Computers can learn structural features and predict lead compounds based on them. However, machine learning (ML) methods work best with extensive datasets, and accessing such data with conventional drug discovery methods is challenging. DELs have the potential to fill this gap and, in combination with AI, transform drug discovery. DELs can provide information on billions of interactions of molecules with target molecules. Learning this data with the right algorithms should enable the prediction of better lead compounds and accelerate drug discovery.

ML is a rapidly advancing field, and its integration with DELs may change. Currently, two main approaches involve supervised learning or ligand alignment.

In supervised learning, each molecule is described in a computer-readable format and assigned a class value (hit or non-hit) or a value such as sequence enrichment. The computer learns which chemical features are associated with activity, building a model for active compounds. Applying the model to libraries of commercially available or synthetically accessible molecules predicts molecules with a high likelihood of being active, which are then acquired and tested. While this approach is conceptually appealing and several exciting success stories have been reported, there are challenges linked to DELs. Standard computational approaches to describing molecules have limitations regarding DEL compounds' large and combinatorial structure and the critical influence of the linker position. The combinatorial structure of DELs is a challenge because the repeat occurrence of certain fragments can mislead the model learning.

Moreover, DEL data is intrinsically unbalanced, skewed towards false negatives, and heteroscedastic, which are all problematic for prediction of active compounds.

Alignment methods do not learn the entire dataset, making them more robust to false negatives and problems associated with data imbalance. Unsupervised machine learning can be leveraged to classify hits according to structural similarity. Hits inside a cluster are aligned to generate consensus models that are used to identify active compounds from virtual libraries. Using structural information and computational binding procedures can further aid in the selection process and provide additional confidence in predicted lead compounds. An example of this approach from our group can be found here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9805525/

Integration with ML provides several advantages over directly pursuing hit compounds:

- Commercially available molecules obviate tedious hit synthesis
- ML expands chemical space beyond that of DELs
- There are no structural limitations on predicted molecules
- The process is very rapid
- Learning from many molecules is more reliable than relying on individual hits

This method has great potential to speed up drug discovery. There are already some success stories, and many companies are working in this field. Overcoming current challenges could lead to even bigger advancements with DELs. It's crucial to understand the details of DEL data for AI methods to succeed, and learning about these aspects will be very helpful for anyone interested in this field.

10. The Best Is Yet to Come

In recent years, DEL research has seen remarkable growth. What was once a niche approach explored by a handful of enthusiasts has now become mainstream in pharmaceutical research. Many powerful DEL hits have been reported, often showcasing new binding modes and some even being first-in-class agents. Numerous other promising leads are being investigated worldwide, though many are not publicly disclosed. Several leads from DELs entering clinical trials have solidified the status of DELs.

Looking ahead, there are significant opportunities on the horizon that have yet to be fully realized. As outlined above, one area of promise is the integration of DELs with computational and ML approaches, while the unique potential for developing PROTACs is apparent. Notably, DELs are increasingly being utilized in selections within living systems, with methods developed for bacteria and mammalian cells. New selection techniques, including those involving photo-cross linkers and covalent DELs, have enabled such advancements. The scope of DEL activities is expanding into new target classes such as GPCRs and RNAs. Furthermore, selections methods sampling for specific activities beyond simple binding are becoming available.

While staying updated of all these developments may be challenging, it is exciting to anticipate the opportunities that will emerge in the coming years. Without a doubt, the best is yet to come...

Interested to learn more?

Franzini.decls@gmail.com

https://delgineering.squarespace.com

Interested to collaborate?

Raphael.franzini@utah.edu

https://franzini-research-group.squarespace.com