BELKA: The Big Encoded Library for Chemical Assessment

Andrew Blevins[*] [†] Brayden J Halverson [†] Nate Wilkinson[†] Ian K Quigley[†]

belka@leash.bio

Abstract

Small molecule drugs are often discovered using a brute force physical search, wherein scientists test for interactions between candidate drugs and their protein targets in a laboratory setting. As druglike chemical space is large (10^{60}), more efficient methods to search through this space are desirable. To enable the discovery and application of such methods, we generated the Big Encoded Library for Chemical Assessment (BELKA), roughly 3.6B physical binding measurements between 133M small molecules and 3 protein targets using DNA-encoded chemical library technology. We hope this dataset encourages the community to explore methods to represent small molecule chemistry and predict likely binders using chemical and protein target structure.

Keywords

drug discovery, chemistry, biology

1 Competition description

1.1 Background and impact

To develop a small-molecule drug, researchers go through many twists and turns (1). At their core, small molecule drugs are chemicals that interact with cellular protein machinery and affect the functions of this target machinery in some way. Often, drugs are meant to inhibit the activity of single protein targets, and those targets are thought to be involved in a disease process. A classic approach to identify such candidate molecules is to physically make them, one by one, and then expose them to the protein target of interest and test if the two interact. This can be a fairly laborious and time-intensive process.

The US Food and Drug Administration (FDA) has approved roughly 2,000 novel molecular entities (2) in its entire history (3). However, the number of chemicals in druglike chemical space has been estimated to be 10^{60} (4), a space far too big to physically search. There are likely effective treatments for human ailments hiding in that chemical space, and better methods to find such treatments are desirable to us all. Recent advances in ML approaches suggest it might be possible to search chemical space by inference using well-trained computational models rather than running laboratory experiments (5, 6). Similar advances in other fields (7, 8) suggest using ML approaches to search across vast spaces could be a generalizable approach applicable to many domains.

To evaluate potential search methods in small molecule chemistry, competition host Leash Biosciences generated a large corpus of training data. We physically tested some 133M small molecules for their ability to interact with one of three protein targets and include this data as part of the competition.

[*] Lead organizer

[†] Leash Biosciences, Salt Lake City, Utah, USA

Preprint. Under review.

Contestants will predict whether unknown chemical material is likely to bind to those targets. There are a number of methods to make small molecule binding predictions without the use of the training data provided here (e.g. DiffDock, 9), and this contest was designed to allow for such submissions.

Datasets of this size are rare and restricted to large pharmaceutical companies. The current best-curated public dataset of this kind is perhaps bindingdb (10), which, at 2.8M binding measurements, is 1000X smaller than BELKA. We hope that by providing BELKA we will democratize aspects of computational drug discovery, and assist the community in finding new lifesaving medicines.

1.2 Novelty

This competition is new. We are unaware of competitions exploring the same problem by providing large numbers of empirical datapoints. A contest exploring similar themes, but restricted to docking and other simulations, is the Cache Challenge (11). A popular benchmark dataset, PoseBusters (12), is designed to evaluate the location of predicted small molecules in their protein targets, but not to classify or rank candidates on their probability of binding at all.

1.3 Data

All data were generated in-house at Leash Biosciences and the competition will be run by Kaggle. Due to the overlapping nature of DEL chemistry, the test-train splits necessarily shrink the amount of data available during the competition (e.g., for a given building block in the test set, all molecules containing that building block must be removed from the training and validation sets). We are providing roughly 98M training examples per protein, 200K validation examples per protein, and 360K test molecules per protein. These datasets are very imbalanced: roughly 0.5% of examples are classified as hits. Here, examples are small molecules labeled as binders or not; we used 3 rounds of selection in triplicate to identify binders experimentally. Following the competition, Leash will make all the data available for future use (3 targets * 3 rounds of selection * 3 replicates * 133M molecules, or 3.6B measurements).

1.3.1 DELs are libraries of small molecules with unique DNA barcodes covalently attached

Traditional high-throughput screening (13) requires keeping individual small molecules in separate, identifiable tubes and demands a lot of liquid handling to test each one of those against the protein target of interest in a separate reaction. The logistical overhead of these efforts tends to restrict screening collections, called libraries, to 50K-5M small molecules. A scalable solution to this problem, DNA-encoded chemical libraries, was described in 2009 (14). As DNA sequencing got cheaper and cheaper (15), it became clear that DNA itself could be used as a label to identify, and deconvolute, collections of molecules in a complex mixture. DELs leverage this new DNA sequencing technology (16, 17).

These barcoded small molecules are in a pool (a single tube, rather than one tube per small molecule) and are exposed to the protein target of interest in solution. The protein target of interest is then rinsed to remove small molecules in the DEL that don't bind the target, and the remaining binders are collected and their DNA sequenced.

1.3.2 DELs are manufactured by combining different building blocks

An intuitive way to think about DELs is to imagine a Mickey Mouse head as an example of a small molecule in the DEL. We attach the DNA barcode to Mickey's chin. Mickey's left ear is connected by a zipper; Mickey's right ear is connected by velcro. These attachment points of zippers and velcro are analogies to different chemical reactions one might use to construct the DEL.

We could purchase ten different Mickey Mouse faces, ten different zipper ears, and ten different velcro ears, and use them to construct our small molecule library. By creating every combination of these three, we'll have 1,000 small molecules, but we only needed thirty building blocks (faces and ears) to make them. This combinatorial approach is what allows DELs to have so many members: the library in this competition is composed of 133M small molecules. The 133M small molecule library used here, *AMA_014*, was provided by AlphaMa. It has a triazine core and superficially resembles the DELs described in (13).

1.4 Targets

Proteins are encoded in the genome, and those names of the genes encoding those proteins are typically bestowed by their discoverers and regulated by the Hugo Gene Nomenclature Committee. The protein products of these genes can sometimes have different names, often due to the history of their discovery.

We screened three protein targets for this competition.

1.4.1 EPHX2 (sEH)

The first target, epoxide hydrolase 2, is encoded by the EPHX2 genetic locus, and its protein product is commonly abbreviated to "secreted epoxide hydrolase", or sEH. Hydrolases are enzymes that catalyze certain chemical reactions, and EPHX2/sEH also hydrolyzes certain phosphate groups. EPHX2/sEH is a potential drug target for high blood pressure and diabetes progression, and small molecules inhibiting EPHX2/sEH from earlier DEL efforts made it to clinical trials (18).

EPHX2/sEH was also screened with DELs, and hits predicted with ML approaches, in a recent study (5, 6) but the screening data were not published. We included EPHX2/sEH to allow contestants an external gut check for model performance by comparing to these previously-published results.

We screened EPHX2/sEH purchased from Cayman Chemical (19), a life sciences commercial vendor. For those contestants wishing to incorporate protein structural information in their submissions, the amino sequence is positions 2-555 from UniProt entry P34913, the crystal structure can be found in PDB entry 3i28, and predicted structure can be found in AlphaFold2 entry 34913. Additional EPHX2/sEH crystal structures with ligands bound can be found in PDB.

1.4.2 BRD4

The second target, bromodomain 4, is encoded by the BRD4 locus and its protein product is also named BRD4. Bromodomains bind to protein spools in the nucleus that DNA wraps around (called histones) and affect the likelihood that the DNA nearby is going to be transcribed, producing new gene products. Bromodomains play roles in cancer progression and a number of drugs have been discovered to inhibit their activities (20).

BRD4 has been screened with DEL approaches previously but the screening data were not published (21). We included BRD4 to allow contestants to evaluate candidate molecules for oncology indications.

We screened BRD4 purchased from Active Motif (22), a life sciences commercial vendor. For those contestants wishing to incorporate protein structural information in their submissions, the amino acid sequence is positions 44-460 from UniProt entry O60885-1, the crystal structure (for a single domain) can be found in PDB entry 7USK and predicted structure can be found in AlphaFold2 entry O60885. Additional BRD4 crystal structures with ligands bound can be found in PDB.

1.4.3 ALB (HSA)

The third target, serum albumin, is encoded by the ALB locus and its protein product is also named ALB. The protein product is sometimes abbreviated as HSA, for "human serum albumin". ALB, the most common protein in the blood, is used to drive osmotic pressure (to bring fluid back from tissues into blood vessels) and to transport many ligands, hormones, fatty acids, and more (23).

Albumin, being the most abundant protein in the blood, often plays a role in absorbing candidate drugs in the body and sequestering them from their target tissues. Adjusting candidate drugs to bind less to albumin and other blood proteins is a strategy to help these candidate drugs be more effective (24).

ALB has been screened with DEL approaches previously but the screening data were not published (25). We included ALB to allow contestants to build models that might have a larger impact on drug discovery across many disease types. The ability to predict ALB binding well would allow drug developers to improve their candidate small molecule therapies much more quickly than physically manufacturing many variants and testing them against ALB empirically in an iterative process.

We screened ALB purchased from Active Motif (26). For those contestants wishing to incorporate protein structural information in their submissions, the amino acid sequence is positions 25 to 609 from UniProt entry P02768, the crystal structure can be found in PDB entry 1AO6, and predicted structure can be found in AlphaFold2 entry P02768. Additional ALB crystal structures with ligands bound can be found in PDB.

1.5 Tasks and application scenarios

The task is to evaluate the likelihood of a given small molecule binding to the protein targets BRD4, EPHX2/sEH, or ALB/HSA. Accurately predicting binders for BRD4 or EPHX2/sEH could lead to medicines that treat disorders stemming from these protein targets. Accurately predicting ALB/HSA binders could accelerate medicinal chemistry for many candidate drugs. More broadly, this competition will shed light on the feasibility of small molecule-protein target interaction predictions given a large amount of training data.

1.6 Metrics

Here, we treat small molecule binding as a binary classification task. The practical use case of binding prediction would be to chemically synthesize and functionally test top predictions, which can be an expensive and laborious process. To help, we want models that have excellent precision in their top predictions and we care less about the rank ordering of lower predictions. In general, we look at the Top-k precision, usually n=100 or 1000. Top-k precision can be unstable, so we also look at the **Mean Average Precision.**

For this competition we will use **Mean Average Precision averaged over all 3 proteins** computed on a private test set of about 360K molecules. This test set includes molecules made with undisclosed building blocks and molecules synthesized with a completely different set of chemical reactions, and should test the model's ability to generalize to new parts of chemical space.

1.7 Baselines, code, and material provided

A common baseline for predicting the properties of molecules is to train a random forest model on ecfp6 features (27). These features can be derived from RDKIT (28). Tutorial notebooks will be available on Kaggle.

A major opportunity in this contest is for competitors to evaluate different types of molecule representations. Small molecules have been represented with SMILES, graphs, 3D structures, and more (29, 30, 31, 32), including more esoteric methods such as spherical convolutional neural nets (33). We encourage competitors to explore not only different methods of making predictions but also to try different ways of representing the molecules.

1.8 Website, tutorial and documentation

Kaggle will be running this competition. We aim to emulate the Recursion Cellular Image Classification or the Open Problems – Single-Cell Perturbations competitions in terms of background and ease of use. The competition should launch in early April of 2024; Kaggle has already finalized the dataset.

2 Organizational aspects

2.1 Protocol

Kaggle will be running this competition. We are making the test molecules available without binding labels in order for contestants to submit entries with methods that do not use the training data, such as molecular docking. To prevent overfitting, we have curated the test/val/train splits to include random molecules and scaffold splits (we expect some memorization to occur here) but also building block splits and an entirely orthogonal chemical library (predicting on these molecules is much more difficult).

2.2 Rules and Engagement

This competition will follow standard Kaggle rules, with the winners licensing their submission and the source code used to generate the submission under the MIT open source license (see https://opensource.org/licenses/MIT). We also request that contestants not physically screen the test set compounds themselves.

2.3 Schedule and readiness

Our competition is scheduled to start in early April with Kaggle and we intend to run the contest for 90 days. Kaggle is in possession of the final datasets, has generated competition-ready datasets from them, and their team is currently collaborating with us on contest and website copy.

2.4 Competition promotion and incentives

We will be publicizing the contest with press releases and a piece in at least one biotech news publication, and there will be cash awards to the winning teams. Finally, to attract underrepresented groups, we have allocated a side prize for the best-performing student group.

3 Resources

3.1 Organizing team

Andrew Blevins was a data scientist at IMVU and Linkedin before joining Recursion Pharmaceuticals. At Recursion he lead the digital chemistry team and saw the need for better benchmarks for chemical machine learning. He cofounded Leash Biosciences to create these datasets.

Brayden J Halverson is a scientist at Leash Biosciences specializing in DNA-encoded library screens, high-throughput DNA sequencing, and molecular biology. Before Leash, he was a scientist at the University of Utah.

Nathan Wilkinson is a data engineer at Leash Biosciences. Before Leash, he performed a similar role at Stripe and Recursion.

Ian K Quigley cofounded Leash Biosciences to tackle medicinal chemistry with large numbers of small molecule-protein interaction measurements. Before Leash, he worked on molecular methods development and data science at UT Austin, the Salk Institute, and Recursion.

3.2 Resources provided by organizers

Kaggle is providing hosting and computational infrastructure.

3.3 Support requested

None.