Expanding the Computational Drug Repositioning Toolbox

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Brown, Adam Samuel. 2017. Expanding the Computational Drug Repositioning Toolbox. Doctoral dissertation, Harvard University, Graduate School of Arts &amp; Sciences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:42061514">http://nrs.harvard.edu/urn-3:HUL.InstRepos:42061514</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>

Expanding the Computational Drug Repositioning Toolbox

A dissertation presented

by

Adam Samuel Brown

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University
Cambridge, Massachusetts

May 2017
Expanding the Computational Drug Repositioning Toolbox

ABSTRACT

Repositioning of previously approved drugs is a promising methodology because it lowers the cost and duration of the drug development pipeline and reduces the likelihood of unforeseen adverse events. Computational repositioning is especially appealing due to the ability to rapidly and scalably screen candidates in silico. However, despite increasing interest in repositioning from both academia and industry, existing computational methods are limited in their generalizability and reproducibility. We hypothesize that by expanding the scope of data accessible for repositioning, while at the same time improving the quality of analytic validation for new methods, repositioning studies can make higher quality predictions of promising repositioning candidates. Here we describe three methods that expand the types of data that can be leveraged for computational repositioning: (1) ksRepo, which expands the scope of gene-based repositioning and allows for the use of any ‘omics modality for disease profiling, (2) MeSHDD, for which we developed a novel drug-drug similarity metric based on overlap between drugs and medical subject heading (MeSH) terms in the biomedical literature, and (3) a novel method for quantitative trait-based repositioning using deep cross-sectional phenotyping studies that explicitly addresses sources of confounding. We also describe our work to survey and improve reproducibility in the repositioning field, and introduce a new standard database, repoDB, that contains over 10,000 drug-disease pairs, including examples of both approved and failed pairs. Together, the methods and resources described here represent an expansion of the computational drug repositioning toolbox, as well as a first step towards enabling reproducible validation of drug repositioning methods.
LICENSE NOTICE

This work is licensed under Creative Commons Attribution 4.0 International (CC-BY 4.0, https://creativecommons.org/licenses/by/4.0/).

You are free to:
Share — copy and redistribute the material in any medium or format
Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you give appropriate credit and provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.
# TABLE OF CONTENTS

Title Page i
Abstract ii
License Notice iii
Acknowledgements v

I. Previous Work in Computational Drug Repositioning 1
II. Generalizing High-Throughput Molecular Phenotype-Based Repositioning 11
III. Tapping Into Prior Biomedical Knowledge for Repositioning 32
IV. Repositioning with Complementary Quantitative Traits in Cross-Sectional and Longitudinal Claims Data 43
V. Enabling Reproducibility in Computational Repositioning 53
VI. Conclusions 68

Literature Cited 76

Appendix A: arrayLasso, a Network-Based Approach to Microarray Interconversion 94
Appendix B: Licences for Reproduction of Published Works 100
Appendix C: Supplemental Methods 102
ACKNOWLEDGEMENTS

Thanks to my advisor, Chirag, for mentoring and pushing me to succeed. Thanks to my group mates, without all of whom the work described herein would not have been possible: Chirag L, Danielle, Jake, Alan, Ben, Andrew, Roger, Braden, Grace, Yeran, Kajal, and Sivateja. Thanks also to my dissertation advisory committee, Soumya, Zak, and Shamil for their expert guidance, and my defense committee, Soumya, Nils, Paul, and Simon. Thanks to Kate, Danny, Maria, and Anne in the BBS office for making it all work. Thanks to my family and friends for their support and encouragement.
1. PREVIOUS WORK IN COMPUTATIONAL DRUG REPOSITIONING

Drug repositioning is the process of identifying new indications for previously approved drugs. Repositioning is a promising methodology because it reduces the cost and duration of the drug development pipeline and reduces the likelihood of adverse events [1–3]. A major goal in repositioning is the development of in silico tools that reduce the number of potential candidate molecules to be tested in in vitro and in vivo validation experiments while also suggesting unlikely and novel possibilities. The following chapter covers three paradigms of computational drug repositioning: (1) high-throughput molecular phenotype-based repositioning, (2) scientific literature-based repositioning, and (3) quantitative trait-based repositioning (Figure 1).

Figure 1. Drug repositioning at a glance. The focus of this thesis is on computational methods that match approved drugs to diseases, rather than drugs to targets, which requires detailed knowledge of the molecular etiology of disease for end-to-end disease identification. The focus of each chapter is shown schematically. Drug repositioning also encompasses high-throughput screening methods in vitro and in vivo.

Key innovations in each of these areas form the core of this thesis, with methodological details of ksRepo (high-throughput molecular phenotype-based), MeSHDD (literature similarity-based)
and complementary phenotype-driven repositioning (phenotype-based) described in Chapters 2, 3, and 4 respectively. We also provide commentary on the state of reproducibility in the computational drug repositioning field, and a potential solution, repoDB, which addresses the current shortcomings in analytic validation of new methods (Chapter 5). Finally, we summarize the contributions of our work to the field and provide insight into open questions and possible future directions for our and other work (Chapter 6).

1.1. Exploiting drug-gene interactions for repositioning

In recent years, understanding of the molecular etiology of disease has enabled the development of specific protein-targeted therapies in a wide range of diseases [4]. The guiding hypothesis for targeted molecular therapy is that by altering the levels or activity of a disease-driving gene product (typically mRNA or protein) in a disease relevant tissue, such therapies can modulate or “reverse” disease [5]. This hypothesis has been leveraged in drug repositioning, as a drug that targets a key protein or pathway can be administered for any disease in which that protein or pathway is dysregulated through protein level changes or mutations (the “one-target-one-drug approach”). For example, drugs inhibiting activating mutations in proto-oncogene B-Raf (v-Raf murine sarcoma viral oncogene homolog B), including vemurafenib and dabrafenib among others, were originally developed for B-Raf mutant melanoma [6] but have since been investigated (and used off-label) in other B-Raf-driven cancers [7,8]. In recent years, repositioning drugs targeting a mutation or protein-level change in a single protein (“target-based repositioning”) accounted for approximately 30% of all new FDA approvals and primarily in oncology, where driver mutations and genes can be identified [9]. A recent example of this type of repositioning is shown by the repositioning of anti-programmed death ligand-1 (PD-L1) antibodies. Anti-PD-L1 antibodies bind to and inactivate PD-L1,
potentially leading to T-cell-mediated cell death of tumor cells [10]. While anti-PD-L1 antibodies were initially only approved for metastatic melanoma and squamous non-small cell lung carcinoma, PD-L1’s expression on the cell surface of a larger number of tumor types has subsequently led to approval in metastatic head and neck squamous cell carcinoma, as well as broad approval for any any unresectable or metastatic solid tumor with DNA mismatch repair deficiencies or a microsatellite instability-high state [10,11].

Owing in part to the success of industry-driven repositioning efforts, a significant focus of the computational drug repositioning literature has been discovery of drug targets (or “drug target interactions” [DTIs]) [12]. Core to such approaches is the notion of drug-drug similarity, whereby “similar” drugs are hypothesized to have similar sets of targets. This approach has been historically applied to chemical structures, in which the distance between two molecules is simplified to the distance between hashed representations of their 2D structural motifs (simplified molecular-input line-entry system, or SMILES)[13,14]. For example, vemurafenib and dabrafenib described above are similar with respect to their SMILES representation, and therefore may interact with similar targets (and in fact, do, binding specifically to mutant BRAF [6]). One common approach is to simply consider the known DTIs catalogued in databases such as DrugBank and the Therapeutics Targets Database (TTD) [15,16]. Both DrugBank and the TTD contain a combination of literature annotated and FDA product label derived drug-protein interactions. Such interactions usually represent the either the intended targets of drugs (“on-target” effects) or known interactions with metabolic enzymes (e.g. the cytochrome P450 family of digestive enzymes [17]), and miss the low effect perturbations in gene expression and protein activity caused by a drug (“off-target” effects) [4]. Perhaps the most well-known method that uses known drug targets is PREDICT [18], which combines structural similarity with shared
DTIs to identify closely related drugs that could be repositioned for complementary diseases. In the original PREDICT paper, the authors highlight cycloleucine, a putative anti-Alzheimer’s disease drugs with weak similarity to currently used off-label Alzheimer’s medications, but known inhibitory interactions with two Alzheimer’s-associated proteins, methionine adenosyltransferase (MAT) and N-methyl-D-aspartate receptor (NMDAR) [18]. A key drawback of known DTI approaches, however, is that their predictions are based heavily on what a given database considers a “target” (e.g. on-target, metabolic, off-target, or some combination); in actuality, there are many significant differences between DrugBank and TTD in terms of both drugs included (2,254 versus 2,071 of approved drugs respectively) and target coverage (4,164 versus 2,589 unique gene targets respectively).

To supplement the inconsistent and often sparse information contained in DTI databases, many methods leverage structural similarity to find novel DTIs. In that vein, a number of structure-based similarity tools have been developed to predict DTIs, with varying degrees of accuracy [14]. Building on these structure-based tools, other methods have incorporated a variety of data modalities to further improve DTI prediction. These include adverse events [19–21], transcriptomic and/or genomic information [22], and a variety of network-based metrics using well-characterized drug-target interactions (see [23–25], reviewed in [26]). A key claim of many DTI identification methodologies is that, beyond understanding the mechanism of action of drugs, DTIs are inherently useful for identifying repositioning opportunities. Such a claim, however, breaks down in complex genetic disease, where a single causal variant or gene may not exist, or for which there may be many, low-penetrance variants to choose from [3,27].
An alternative strategy to the traditional DTI approach is to consider the full space of molecular changes in disease. An extension to the targeted therapy hypothesis would be, in this case, that by altering levels of all or a key subset of gene products that are perturbed in a disease, one can modulate disease through many and small effect size changes to genes that, taken together, cause disease [1]. Information on which genes are perturbed in disease can studied using a wide range of high-throughput ‘omic phenotyping technologies; for example, gene associations in well-studied diseases such as Type II Diabetes Mellitus have been ascertained using GWAS [28], RNA-seq [29], ChIP-seq [30], DNA methylation arrays [31], proteomics [32], and others. In some cases, they have already been deposited into large public databases, such as the National Center for Biotechnology Information’s Gene Expression Omnibus (GEO) [33], among others. In contrast, information on which genes are perturbed through drug exposure is considerably more challenging to obtain.

One approach for surveying drug-based perturbations relies on comparing individual disease RNA-level expression profiles to large databases of pre-generated multi-drug exposure profiles or known gene-drug interactions (hereafter “profile-based methods”) [18,34–36]. Among profile-based methods, the most commonly used is the Broad Connectivity Map (CMAP), initially described in [37], which is an online tool for matching differentially expressed gene sets from a disease of interest (or disease “signature”) to a large library of compound exposure profiles (drug “instances”) generated by the CMAP team in three human cell lines. Their computational method relies on a modified Kolmogorov-Smirnov enrichment test, which they call a “connectivity score,” that expresses the similarity in differential expression between a drug instance and a disease signature. The natural appeal of such a score led to a large number of CMAP studies in targeted disease areas and prompted the subsequent creation of the LINCS
L1000 project (described in [38,39], and reviewed in [40]). The L1000 project expands on the CMAP by adding six additional cell lines and introducing a new, lower cost gene expression platform that relies on high-fidelity analysis of 1000 “landmark genes.” Recent work has further lent credence to profile-based methods, with drugs that reverse disease-induced gene expression profiles having higher efficacy in certain disease models [41].

1.2. Mining the literature for repositioning candidates

While the majority of computational repositioning methodologies rely on genetic data as a hypothesis generation, an emerging source of evidence for repositioning hypothesis generation includes the biomedical literature. Not only does the biomedical literature provide information about currently available drugs and their approved indications, but also contains detailed information about investigations into repositioning. For example, the antidiabetic medication metformin has been the focus of several repositioning studies, disparate indications including hyperlipidemia [42], cystic fibrosis-driven lung [43] and renal pathologies [44], and even cancer [45,46].

In addition to straightforward information about current and candidate indications, the biomedical literature also contains substantial detail on key drug properties, especially for approved drugs. Echoing the DTI prediction studies described above, a number of studies have attempted to catalogue drug-target information from the literature, with two of the most prolific being DrugBank [15] and the Comparative Toxicogenomics Database [47], both of which contain manually curated, literature-derived drug-gene interaction information. Other studies have attempted to supplement manual curation of literature-based DTI information using natural language processing (NLP) methods with the goal of using the information gleaned for drug
repositioning hypothesis generation. For example, the authors of [48] use a rule-based method to detect novel DTIs, where a free text mention in the form of “DRUG pattern GENE” would signify a DTI (patterns could include phrases like “interacts with”, “inhibits”, “activates”, or others). Other methods include co-occurrence based methods (DRUG and GENE are in the same abstract), and machine learning-based methods, which rely on learning complex syntactic relationships between drugs and genes from known DTI mentions (these methods are reviewed in [49]). NLP methods have also been applied to the literature to find drug-drug interactions, in which co-prescription of multiple drugs leads to unforeseen side-effects (reviewed in [50]), and in the “Side Effect Resource” (SIDER), which applies machine learning-based NLP to drug product labels to detect adverse events [51].

Taken together, DTI, drug-drug interaction, and adverse event, and other annotation efforts focus on extracting a single aspect of literature-based drug information. Given their success, a straightforward extension would be to consider latent literature-based knowledge comprehensively. The current gold standard for generalized literature topic annotation is semi-automated topic annotation by the National Library of Medicine’s (NLM) Medical Text Indexer (MTI) [52,53]. The NLM MTI supports the annotation of Medical Subject Heading Terms (MeSH Terms) to all PubMed-tracked biomedical journal articles. MeSH terms are derived from the Unified Medical Language System (UMLS, see [54]), and describe the high-level concepts contained within a given article. MeSH term annotations are available centrally through the NLM managed MEDLINE®, which contains MeSH annotations for over 20 million biomedical articles. Surprisingly, despite the availability of a bulk MeSH term download (see https://www.nlm.nih.gov/databases/download/pubmed_medline.html), there are few MeSH-driven methods for computational repositioning, all of which have utilized heterogeneous
bipartite graph-based techniques to connect drugs and diseases (reviewed in [26]). Specifically, such methods have focused on building bipartite graphs connecting drugs to genes [55] and drugs to diseases [56–58]; to our knowledge, all currently available methods that leverage MeSH terms require the incorporation of other non-MeSH-based prior biomedical knowledge, rather than relying strictly on MeSH terms for repositioning hypothesis generation.

1.3. Observational studies as a source of serendipitous repositioning opportunities

The majority of repositioning methods rely on indirect evidence, such as surrogate molecular biomarkers or side-effect profiles; however, a key challenge in computational repositioning is to provide direct evidence of candidate efficacy in humans. A growing trend in drug repositioning studies is to leverage observational data as a means to find direct associations between drug use and disease outcomes. Observational data can take many forms, with the most popular being (1) electronic medical record or insurance claims data, which often contain prescription histories as well as diagnostic codes, and (2) large epidemiological cohorts with both disease outcomes and prescription drug tracking. Electronic medical records and health claims - both used for administrative purposes, such as billing or keeping patient records [59] - databases typically contain a combination of diagnostic codes (e.g. International Statistical Classification of Diseases and Related Health Problems codes, or ICD codes), prescription drug history information, and, less commonly, quantitative trait data (e.g. physiometric measurements, laboratory values). Claims data have the benefit of immense cohort sizes; for example, the commercially available Truven MarketScan CCAE administrative claims database has claims data for over 170 million patients derived from insurer and governmental insurance (Medicare and Medicaid). An additional benefit of claims data is longitudinal nature of the data, allowing individual patients to be tracked over years or decades.
Large epidemiological cohorts provide an alternative source of observational data to claims-based data. One of the most broadly utilized epidemiological cohorts is the US Centers for Disease Control and Prevention “National Health and Nutrition Examination Surveys” (NHANES), which are bi-annual, cross-sectional, and observational studies. NHANES recruits a large number of participants, selected so as to be representative of the non-institutionalized US population, to answer a number of survey questions pertaining to their medical, psychosocial, and sociodemographic histories. A subset of respondents also receive extensive anthropometric and laboratory testing, including a variety of routine clinical measures. NHANES, and by extension other epidemiological cohorts offer much more modest cohort sizes, but typically offer much higher resolution phenotyping of individuals, and of healthy individuals who otherwise would not have received testing. While NHANES is cross-sectional, ruling out longitudinal analyses, other epidemiological cohorts may offer longitudinal data at the cost of fewer phenotypes or outcomes being ascertained (e.g. the Women’s Health Initiative, see [60]).

Of the two sources of observational data, claims data has been more broadly used for repositioning, stemming from the development of Medication-wide Association Studies in [61]. Ryan et al. studied four disease-related outcomes in two large medical claims datasets (Truven MarketScan CCAE administrative claims database and the GE Centricity EHR database) and predicted drugs that were associated with those outcomes using a self-controlled case series (SCCS) analysis [62], in an attempt to avoid time-invariant confounding by using each patient as their own control. In their SCCS analysis, drugs could either be associated with a decreased risk for a given disease outcome (potential repositioning candidates) or an increased risk (potential adverse event drivers), or neither. Unfortunately, despite confirming that current FDA approved
therapies were associated with improved outcomes and reaffirming several known drug-adverse event relationships, they noted that there was no sufficiently high threshold to eliminate false positive associations. In addition, further work has suggested that hazard modeling in a pharmacoepidemiological context is highly dependent on modeling choices [63,64].

In an attempt to address the high false positive rates seen with medication-wide association studies, recent work has focused on the association between quantitative clinical traits and drug exposure, both in claims data [65] and in NHANES [66]. Quantitative traits are appealing because they are assessed regularly and frequently, as opposed to medical events, which may happen once in a decade, potentially ameliorating the accumulation of time-dependent confounding in longer outcome-based approaches. Further, quantitative traits are used in routine clinical practice to drive decision making and assess disease risk (for example, in the Framingham Heart Study’s “risk score” [67]). Echoing medication-wide association studies, Paik et al. also used SCCS methodologies to associate drugs and quantitative traits [65]. In contrast, Jang et al. used linear regression models to test for the association between drugs and quantitative traits in NHANES [66]. In both cases, the authors then use the associations they found to develop drug-drug similarity measures by comparing the direction and magnitude of effect of all pairwise drug-trait associations. These similarity measures are then used to nominate repositioning candidates for all diseases concurrently [65,66].
2. GENERALIZING HIGH-THROUGHPUT MOLECULAR PHENOTYPE-BASED REPOSITIONING

Despite the promise of target- and profile-based repositioning tools, significant limitations remain for the application of existing tools. The most challenging to mitigate is the requirement for specific data types and formats. All extant methods, to our knowledge, require either expression levels from a single microarray platform, or pre-determined databases of drug-gene interactions (as in L1000), or both. For example, CMAP, which is still in broad use (see [68,69], among others) requires users to input up- and down-regulated probes from the Affymetrix HGU-133A genechip. Even in the recent L1000 initiative, an extension of CMAP, the requirement for up- and down-regulated genes remains [39]. Other requirements include additional disease specific information, including detailed genomic or phenotypic annotations (e.g. [70] among others). Unfortunately, these restrictions prevent investigators from utilizing newer (or older) profiling technologies, such as RNA-seq. While probe-level data can be obtained by back-converting RNA-seq data (which provides transcript- and gene-level data), such conversion typically results in a loss of resolution in both overall gene numbers and transcript resolution (and furthermore, such conversion is typically poor, see Appendix A).

Another emerging -omic modality, DNA methylation, cannot currently be “converted” to gene expression. DNA methylation plays a key role in determining gene expression in concert with a number of other factors, including chromatin modifying enzymes (and chromatin marks themselves) and transcription factors. Because both DNA methylation and these other epigenomic factors vary substantially between cells (and especially in cancer [71,72]), direct prediction of gene expression from DNA methylation is challenging. Recent methods have
attempted to forgo direct prediction of gene expression from DNA methylation in favor of relating differential DNA methylation to differential gene expression, but have struggled to achieve high levels of accuracy [73]. Without high-fidelity gene expression conversion, DNA methylation is currently incompatible with CMAP. Even the newer L1000 tool, which allows users to input gene-, still requires directional input, preventing investigators from using DNA methylation information due to the lack of clear directionality in epigenetic data (Figure 3).

Adding to the difficulties of using currently available tools is the additional drawback of being tied to a single drug profile database. This is again exemplified by the first version of the CMAP, which contains gene expression profiles from a restricted set of approved drugs in only three human cell lines [37]. As noted, with L1000, the team has continued to improve CMAP, however, investigators are limited to 9 cell lines and a subset of approved drugs for which profiling data is available.

To address the limitations of currently available high-throughput molecular phenotype-based repositioning methods, namely, the lack of flexibility in input, we sought to implement a method that generalizes the process. We posited that a generalized computational repositioning tool should have flexibility in the types of data sets and databases that can be used. Specifically, we envisioned such a tool having 1) the ability to interrogate any case/control disease study-derived expression profile, 2) the ability to use any compound database, including those with limited numbers of gene-drug interactions, and 3) an extensible, open-source distribution. In Chapter 2, we propose a generalized tool for computational repositioning that builds on the successes of previous expression-based repositioning tools while allowing greater flexibility for the investigator called ksRepo.
Unlike many popular repositioning tools (e.g. [18,34–36,74]), we provide source code for ksRepo that enables investigators to extend our methodology as new datatypes become available. We demonstrated the promise and extensibility of ksRepo by predicting currently approved drugs for prostate cancer and acute myeloid leukemia (AML) with gene expression from multiple different microarray modalities, and DNA methylation data respectively. In addition we present a case study of ksRepo applied to vestibular schwannoma, a debilitating orphan disease with no approved therapy, for which we demonstrate *in vitro* efficacy of a ksRepo predicted drug. Notably, each of the examples we describe would not be possible with existing tools.

### 2.1 ksRepo development

To develop a flexible and generalized repositioning tool, we began with the CMAP’s KS testing methodology [37]. We then modified their methods to create an “inverse” version of the CMAP implementation in which we compare a single “instance” (complete disease gene profile) to a number of “signatures” (short compound-gene interaction lists) rather than comparing a single signature to a number of instances (Figure 2).

In addition, we focused on drug-gene and disease-gene interactions without directionality to accommodate gene profiles (instances) and compound exposure databases (signatures) with no regulatory component or conflicting regulatory information (for example, in the case of differential methylation at individual CpGs). Lastly, we consider all genes in the ranked instance gene list regardless of significance to ensure overlap between the instance and signatures.
Figure 2. ksRepo methodology. ksRepo uses the Kolmogorov-Smirnov enrichment test to compare a single list of genes that are modulated in disease relative to controls to a database of compound-gene lists.

KS enrichment scores for our method are calculated as follows: Let $n$ be the number of genes in the instance and $t$ be the number of genes in a given signature. Order all $n$ genes in the instance by their differential expression. Construct a vector $V$ of the position $(\in \{1,\ldots,n\})$ of each signature gene in the instance ordered gene list and sort these components in ascending order such that $V(j)$ is the position of gene $j$, where $j \in \{1,\ldots,t\}$. Calculate the following values:

\[
    a = \max_{j=1}^{t} \left[ \frac{j}{t} - \frac{V(j)}{n} \right] \tag{Eqn 2.1}
\]

\[
    b = \max_{j=1}^{t} \left[ \frac{V(j)}{n} - \frac{(j-1)}{t} \right] \tag{Eqn 2.2}
\]

and set $KS = a$ if $a > b$. Else set $KS = -b$. 

14
Both $a$ and $b$ quantify differences in the expected distribution of gene ranks ($\in \{1, \ldots, n\}$) and the observed sample of ranks in the signature. The value $a \in U(0,1)$ and scales inversely with the mean signature rank (mean $V(j)$), with deviations proportional to the standard deviation of signature rank; the value $b$ is the inverse of $a$. If $a > b$, then the mean signature rank is low, corresponding to enrichment, and we assign $a$ as the KS score. If $b > a$, we assign -$b$ as the KS score. In this way, signatures with highly enriched gene sets are assigned highly positive KS scores, while signatures with unenriched or inversely enriched (e.g. very high mean signature rank) are assigned KS scores near zero or negative KS scores respectively.

Because our KS test statistic has no theoretical distribution, we estimated null distribution of test statistics empirically in the following way. First, construct a vector $L$ of the number of genes in each signature. For each unique $\ell$ in $L$, generate 10,000 independent resamples of the instance gene list of length $\ell$ and calculate KS scores for each resample. For each signature, compare the observed signature KS score to the corresponding resample with the same number of genes. Set the $p$ value of that signature as the proportion of resample KS scores that exceed the signature KS score and FDR adjust to correct for multiple hypothesis testing [75].

2.2. Software availability

To enable rapid uptake of our method, we developed an R Package for ksRepo, available on GitHub (https://github.com/adam-sam-brown/ksRepo). While we have attempted to construct a very general tool, we acknowledge that there may be specific use-cases that require modification of the ksRepo source, and therefore provide a full codebase with a permissive license allowing users to make adjustments to our work. We note here that broad distribution of
source code for repositioning methods is rare, hindering the reproducibility and usability of many repositioning platforms.

To further facilitate ease-of-use, the ksRepo package includes a pre-built compound exposure database derived from the Comparative Toxicogenomics Database (CTD) database. The CTD contains manually curated compound-gene interactions (e.g., gene or protein expression influences) for a variety of models organisms and humans, which are collected from the primary literature by trained experts [47]. ksRepo contains a subset of the CTD with human-derived interactions between 1,268 unique drugs and 18,041 unique human genes. Drugs in the ksRepo subset were chosen based on case-insensitive matches between CTD names and names/synonyms for FDA approved drugs downloaded from DrugBank [15]. While we include one option for a compound exposure database, we envision ksRepo being used with any database similar to the CTD (e.g., documented gene-drug interaction associations); we therefore also include scripts for converting the CTD to a ksRepo usable format, which can be straightforwardly adapted for other databases.

2.3. Using ksRepo for gene expression-based repositioning

To demonstrate ksRepo’s applicability in gene expression (the typical use case of extant high-throughput molecular phenotype-based methods), we applied our method to five independent prostate cancer datasets from three distinct microarray platforms, and attempted to detect signal for FDA-approved prostate cancer therapies present in the CTD. We note here that the use of three distinct microarray platforms in itself precludes the use of some methods (e.g. CMAP), but is possible with ksRepo. The prostate cancer datasets used in this study are GSE3868, GSE12348, GSE45016, GSE55945, and GSE6919 [76–80], which were chosen on
the basis of three criteria: (1) the expression profiles were derived from primary prostate cancer
cells, as opposed to cell lines or short-term cultures, (2) there were healthy prostate tissue
controls included in the study, and (3) tissue samples were from fresh-frozen biopsies, and not
preserved (e.g. by formalin). All GEO datasets were accessed through the NCBI GEO portal
and analyzed using the integrated GEO2R tool [33]. As input for GEO2R, we classified each
sample within a GEO series as either normal tissue or tumor tissue. GEO2R provides a list of all
probes (and corresponding gene aliases) ranked according to their degree of differential
expression. We imported all of the results from GEO2R into R [81] and converted all gene
aliases into EntrezGene Identifiers using the org.Hs.eg.db R package [82].

We first identified all FDA-approved prostate cancer therapies using DrugBank [15] and then
determined that out of 11 small-molecule therapies, seven (Bicalutamide, Nilutamide,
Leuprolide, Zoledronic Acid, Docetaxel, Aminoglutethimide, and Estropipate) were also included
in the CTD. We applied ksRepo to the five GEO prostate cancer datasets independently and
determined the FDR-corrected p-values for each of the seven annotated therapies in the CTD
(see §2.2). ksRepo predicted a median of 60 total significant compounds (FDR-corrected
p-value < 0.05) corresponding to around 5% of the dataset, which is similar to other
repositioning strategies [18,37,39,83].

For each of the five prostate cancer datasets we were able to detect significance for between
one and three FDA-approved therapies at a FDR-corrected p-value less than 0.05. In each
case, this represented a significant enrichment for approved therapies (Hypergeometric Test, p
< 0.027, expected number of drugs μ = 0.029). Among compounds, significant prostate cancer
therapies ranked on average in the 3.5th percentile and of the seven therapies, five were
significant for at least one of the five datasets. We did not detect significance for two therapies, Aminoglutethimide and Estropipate; we hypothesize that due to the nature of the microarray datasets we included (tissue from primary, non-metastatic tumors), it is unlikely that we would detect secondary hormone modulatory treatments, which are typically used in treatment refractory patients with metastases [84].

2.4. Using ksRepo for epigenome-based repositioning

As a core design principle for ksRepo, we sought to develop a tool that could incorporate non-RNA expression based profiling of disease, such as epigenomic data. Epigenomic studies have been proposed as a promising data source for personalized medicine [85,86], but to our knowledge no methods have been published that allow the direct translation of epigenomic findings to drug repositioning. The lack of utilization largely stems from the lack of clear directionality at the gene level associated with differential DNA methylation and histone modifications (Figure 3); for instance, intra-exon DNA methylation can lead to upregulation or downregulation of gene expression depending on a variety of factors, including histone modifications [87,88].
A challenge in using ksRepo for DNA methylation is that our package requires a list of genes ranked according to their differential methylation, rather than individual CpG sites. We therefore condensed CpG-level methylation to gene-level differential methylation using a custom pipeline (Figure 4): (1) identify CpGs that are annotated to a given gene (obtained from the manufacturer’s website), (2) weight and combine p-values from CpGs annotated to that gene, and (3) adjust the resulting, combined p-value to take into account correlation between CpGs within a single gene. As a core component of our strategy, we used the EmpiricalBrownsMethod package in R [81,90]. EmpiricalBrownsMethod first combines CpG-level p-values into an omnibus statistic using Fisher’s Method and then calculates an empirical p-value using the correlation structure of the CpGs [91].
We applied our CpG-consolidation pipeline to two genome-wide methylation studies of AML. AML is an ideal case-study for our methodology, as recent work has shown that the AML epigenome evolves independently of mutational burden, and, furthermore, patients with AML respond well to DNA demethylating agents (also known as “epigenetic drugs”) [92–95]. The two methylation studies we selected were GSE58477 and GSE63409, both of which are Illumina Infinium HumanMethylation450 chip-based (hereafter 450K) studies of AML [96,97]. The 450K chip has broad coverage of the genome, encompassing CpGs in 99% of RefSeq genes and
96% of CpG islands [98]. GSE58477 contained data from a total of 62 patients with cytogenically normal AML and 10 CD34+ normal bone marrow aspirate samples [96]. GSE63409 contained data from a total of 15 AML patients of varying FAB subtype classifications (M0: 1, M1: 2, M2: 2, M5: 5, Not Determined: 5) and five sorted, normal bone marrow aspirate samples [97]. Both GEO datasets were accessed through the NCBI GEO portal and analyzed using the GEO2R tool [33]. GEO2R uses the limma package in R [99], an analytic method originally designed for gene expression, which has been used to detect differential methylation [100,101]. As input for GEO2R, we classified each sample within a GEO series as either normal CD34+ bone marrow aspirate or leukemic blast (as indicated by either lack of engraftment only or by a combination of CD34- and lack of engraftment for GSE58477 and GSE63409 respectively). After classification we were left with 62:10 and 11:5 case:control ratios for GSE58477 and GSE63409 respectively.

For each dataset we condensed 407,090 individual CpGs into a list of 21,231 genes ranked by their degree of differential methylation. Some CpGs lacked differential methylation estimates due to poor probe quality; for these CpGs, significance was set as the patient-level median significance for the other CpGs annotated to the same gene. Before continuing with ksRepo analysis, we first wanted to verify that our consolidation technique preserved the differential methylation structure of AML. We therefore subjected significantly differentially methylated genes (those with consolidated FDR significance below 0.05) to pathway enrichment analysis using PANTHER. In short, PANTHER contains manually curated pathway annotations for human genes, as well as built-in analytical tools for reporting statistical enrichment using fisher’s exact test and gene-set enrichment [102]. Similar to the original publication for GSE58477 and GSE63409, our significantly differentially methylated genes were significantly enriched for Wnt
signaling genes (PANTHER Wnt signaling pathway, Bonferroni-corrected \( p < 2 \times 10^{-11} \) and \( p < 3 \times 10^{-20} \) for GSE58477 and GSE63409 respectively). This suggests that, although we lose some resolution by combining CpGs by gene, we do capture major trends in the AML methylation landscape.

Having verified that gene-level consolidation provides similar results to CpG-level differential methylation, we applied ksRepo independently to each GEO dataset. 1,075 drugs had at least one overlapping gene with those annotated in 450K chips and were tested using ksRepo. We identified nine compounds with predicted FDR significance in only one of the datasets, and four that were significant in both. The four compounds significant in both datasets are presented in Table 1.

<table>
<thead>
<tr>
<th>Compound</th>
<th>FDR GSE58477</th>
<th>FDR GSE63409</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALITRETINOIN</td>
<td>( 2 \times 10^{-4} )</td>
<td>&lt; 0.04</td>
</tr>
<tr>
<td>CYTARABINE</td>
<td>( &lt; 2 \times 10^{-16} )</td>
<td>( &lt; 2 \times 10^{-16} )</td>
</tr>
<tr>
<td>PANOBINOSTAT</td>
<td>( &lt; 2 \times 10^{-16} )</td>
<td>( &lt; 2 \times 10^{-16} )</td>
</tr>
<tr>
<td>PROGESTERONE</td>
<td>( &lt; 2 \times 10^{-16} )</td>
<td>( &lt; 2 \times 10^{-16} )</td>
</tr>
</tbody>
</table>

We posited that the most clear demonstration of utility for our method would be to correctly predict significance for the most commonly used AML therapy, cytarabine, which has remained the frontline therapy for over 30 years, and which is broadly effective against all subtypes of AML [103,104]. Indeed, our method did correctly identify cytarabine as highly significant in both datasets examined; however, our method did not identify either daunorubicin or idarubicin, two anthracycline chemotherapeutics that are commonly co-prescribed with cytarabine (FDR = 0.90
and 0.99 for daunorubicin and FDR = 0.80 and 0.86 for idarubicin, in GSE58477 and GSE63409 respectively) [103,105]. We also did not detect significance for decitabine or azacitidine, two commonly off-label prescribed, “epigenetic drugs” (FDR = 0.45 and 0.57 for decitabine and FDR = 0.08 and 0.51 for azacitidine, in GSE58477 and GSE63409 respectively) [86,106].

In addition to cytarabine, ksRepo nominated three novel compounds for use in AML: alitretinoin, panobinostat, and progesterone. Both alitretinoin, a geometric isomer of tretinoin currently used in the treatment of Acute Promyelocytic Leukemia (a subclass of AML) [107], and panobinostat, a histone deacetylase inhibitor commonly used in multiple myeloma [108], have been investigated for use in AML [109–112]. Lastly, while progesterone itself has not, to our knowledge, been suggested for AML treatment, a synthetic progestin, medroxyprogesterone acetate, has recently been suggested as a possible drug repositioning candidate for AML [113].

While our pipeline for epigenome-driven repositioning provided promising results for AML, it does have key limitations. First, as described above, removing the requirement for directionality (e.g. up- or down-regulation) provides substantial generality and allows for the use of CpG methylation data. However, ksRepo may miss some candidates that otherwise would be nominated by methods that explicitly consider direction of gene expression change. We therefore tested for the possibility that ksRepo may miss some candidates that would be nominated by methods that explicitly consider gene-level directionality (see Supplemental Methods). We found two drugs with nominal p-value level of significance (not accounting for multiple hypotheses) in both methylation datasets (irinotecan and sodium nitroprusside). Notably, the directional method did not find significance for cytarabine or any other frontline AML drugs, in concert with previous studies using directional repositioning techniques on gene
expression datasets that also failed to capture significance for cytarabine (as well as daunorubicin, idarubicin, and both epigenetic drugs), and predicted fewer plausible repositioning candidates than ksRepo [114–116]. A second limitation is the requirement that methylation or other epigenomic marks are annotated to a specific gene or genes; however, in the future, as our understanding of extragenic enhancers and other epigenomic features improves, this limitation should diminish. Further investigation into and annotation of other epigenetic modifications will also help to broaden the available inputs for our method. Lastly, we acknowledge that the CTD may contain a relatively small proportion of the true compound-gene interaction space, and future versions of ksRepo will incorporate other data sources to broaden the available databases included in the ksRepo package.

2.5. Prediction and preclinical validation of of Mifepristone for Vestibular Schwannoma

While we have thus far demonstrated the practical utility of ksRepo in using a variety of -omic data for drug repositioning, we have not yet presented a full end-to-end drug repositioning effort supported by ksRepo. To that end, we present a vignette of ksRepo applied to Vestibular Schwannoma (VS) to predict repositioning opportunities (in collaboration with the Stankovic Lab, Massachusetts Eye and Ear Institute, Boston, MA, see Figure 5A). VS is an orphan disease with no FDA-approved therapies, and is the fourth most common intracranial tumor. VS arises from neoplastic Schwann cells of the vestibular nerve and in 95% of patients causes debilitating sensorineural hearing loss (SNHL), tinnitus, and often leads to dizziness and facial paralysis [117]. Bilateral VSs are the hallmark of neurofibromatosis type 2 (NF2), an autosomal dominant disorder caused by inactivation or loss of both alleles of the NF2 gene [118]. If left untreated, growing VSs can compress the brainstem and lead to death. Mutations in the NF2 gene are identified in 100% of NF2-associated VSs and 66% of sporadically arising VSs
Though mechanisms of VS-induced SNHL are multifactorial, with contributions from tumor size, localized or systemic infection, inflammation, and tumor-secreted factors, NF2-associated SNHL often correlates with VS size. This observation suggests that slowing or inhibiting VS growth may not only prolong a patient’s time to surgical intervention, but also minimize or prevent associated SNHL, substantially improving quality of life.

**Figure 5.** Computational repositioning of FDA-approved drugs using ksRepo. A) schematic depicting ksRepo workflow. B) largest meta-analysis to date of genome-wide expression in VS, comprising 80 tumors and yielding 1335 commonly dysregulated genes, 405 of which were found to be significantly differentially expressed after Bonferroni correction for multiple hypothesis testing (p<0.05); analysis with ksRepo yields 8 drugs with high potential for repositioning in sporadic and NF2-associated VS. C) drug classes of repositionable candidates from ksRepo analysis. D) significant enrichment of anti-inflammatory, anti-neoplastic, and hormone-related drugs from all FDA-approved drugs after ksRepo analysis of genome-wide expression in VS (***, p<0.001).
For this VS case-study, we conducted the largest “meta-analysis” of primary human VS tissue to date, comprising genome-wide transcriptional microarray data from 80 tumors and 16 control nerves (Figure 5B, see Supplementary Methods). We applied ksRepo was to differentially expressed genes output from a meta-analysis of two large published datasets, one publicly available (NCBI GEO, GSE39645) and one donated by a collaborator [120]; using ksRepo, we found 36 drugs with significant KS scores. Eight drugs were also significant in a ksRepo analysis of a gene expression meta-analysis from a subset of the 80 tumors derived from patients with known germline NF2 mutations (13 tumors), and were prioritized for preclinical validation (Figure 1C). Out of all FDA-approved drugs, this group of eight was significantly enriched for anti-inflammatory drugs (27.7-fold), hormone-related compounds (13.9-fold), and anti-neoplastic agents (13.6-fold) (Figure 1D).

From the eight drugs, we prioritized mifepristone, a progesterone and glucocorticoid receptor antagonist currently approved by the FDA for use in medical abortion. Mifepristone has several desirable properties for a putative anti-VS drug:

1. mifepristone is able to cross the blood-brain barrier [126,127],

2. mifepristone has been shown in human clinical trials to provide palliative benefits to patients with other intracranial tumors, such as glioblastoma multiforme [126] and meningioma [128–130], as well as systemic cancers including thymic, renal, colon, leukemic, and pancreatic cancers [126,131]

3. in vitro, mifepristone produces antiproliferative effects on cervical [132,133], breast [134,135], endometrial [136,137], ovarian [138–140], gastric [141], bile duct [142], and prostate cancer cells [143–145], regardless of progesterone receptor expression [146]
long-term administration of oral mifepristone is well tolerated by adults and carries only a mild toxicity profile [128]

To investigate whether mifepristone exhibited similar antiproliferative activity in VS, we first studied its impact on the HEI-193 immortalized human schwannoma cells, a commonly used cell line model of VS [147]. In agreement with ksRepo predictions, administration of mifepristone to HEI-193 cells in culture produced a significant, dose-dependent response in metabolic viability (Figure 6A); a significant reduction in cell confluence (Figure 6B-C); and a dramatic decline in cellular proliferation (Figure 6D-F). The effect of mifepristone on HEI-193 cells can be ameliorated by applying this drug in the presence of equimolar progesterone, which restores the metabolic viability of treated cells to levels indistinguishable from untreated controls (Figure 6G). Under mifepristone treatment, HEI-193 cells assume a long, thin, spindle-like shape (Figure 6C). This observation is reported in previous studies of mifepristone treatment of ovarian, breast, prostate, and nerve cells, where such morphological changes are attributed to dysregulated distribution of f-actin and tubulin proteins in the cytoskeleton [148]. Continuous, dynamic actin remodeling is characteristic of NF2-deficient schwannoma cells [149], as the NF2 protein product, merlin, is known to selectively bind f-actin [150]. Phalloidin staining confirms the shrunken, crumpled appearance of f-actin in mifepristone-treated cells (Figure 6H-I).
Figure 6. Mifepristone adversely affects HEI-193 cells in culture. A) the metabolic viability of HEI-193 cells decreases with increasing concentration of mifepristone. B) Confluence remains constant among cells treated with 35 μM mifepristone, while vehicle-treated cells exhibit normal proliferation patterns. C) Cells treated with 35 μM mifepristone display abnormal morphology. D-E) Cells treated with 35 μM mifepristone show decreased proliferation compared to vehicle treated cells. F) Quantification of D-E. G) the simultaneous application of equimolar progesterone (35 μM) restores metabolic viability to levels indistinguishable from vehicle-treated cells.

With promising activity in immortalized cells, we next evaluated the effect of mifepristone on primary human VS cells. Fresh VS tissue samples from eight human patients undergoing tumor resection surgery were collected and schwannoma cells grown in the laboratory according to published protocols [147]. Single-gene sequencing (Illumina MiSeq) of six treated VSs revealed that four of six exhibit novel mutations in the NF2 gene, a fraction consistent with published literature [119]. When applied to primary human VS cells, mifepristone produced a dose-dependent response in metabolic viability and a dramatic reduction in cellular proliferation (Figure 7A-B). Live-cell fluorescence imaging revealed a marked increase in cytotoxicity (Figure
We did not observe a correlation between drug response and NF2 mutation status. To ensure that mifepristone administration did not lead to adverse effects among healthy human Schwann cells, primary human Schwann cells were cultured from eight great auricular nerves harvested from patients undergoing benign parotidectomy or neck dissection. Treatment of these cells with mifepristone did not cause appreciable changes in metabolic viability (Figure 7G). Preliminary testing of clinically reasonable concentrations of other drugs recommended by ksRepo, including adenosine monophosphate, gold sodium thiomalate, succimer, and methylprednisolone showed no effect on the viability of HEI-193 cells, though prednisolone produced modest effects when applied to primary human VS cells.

Together, these results clearly demonstrate the promise of ksRepo to identify plausible candidates for in vitro testing, with broad activity in immortalized and primary human cells. Further validation efforts should be carried out in animal models to confirm the activity shown in the dish; once demonstrated, however, mifepristone is a good candidate for pilot studies in VS sufferers given its mild toxicity and has promise for palliative benefit.
Figure 7. Mifepristone adversely affects primary human VS cells and human-derived arachnoid cells in culture, but leaves primary human Schwann cells unaffected. A) Metabolic viability (MTT) of primary VS cells declines with increasing concentrations of mifepristone. B) Primary VS cells treated with 35 μM mifepristone show decreased proliferation compared to vehicle treated cells. C-E) Primary VS cells treated with 35 μM mifepristone display significantly higher levels of cytotoxicity.

2.6. Conclusions

The results presented in this chapter suggest that ksRepo is successful as a generalized methodology for computational drug repositioning. Even after intentionally reducing the information content presented to our methodology by using a database with a modest number of gene interactors by compound (CTD), we were still able to recover many of the FDA-approved drugs for both prostate cancer and AML. In parallel, by removing the requirement for specific data types and directionality of effect or association (e.g. up- or down-regulation of a gene), we have enabled the use of any microarray or sequencing platform as input, including the use of
epigenomic data for drug repositioning such as DNA methylation. Finally, we provide preclinical evidence for mifepristone in VS, a ksRepo-derived prediction that would not be possible with other computational repositioning approaches. Although we describe here three specific use-cases of our pipeline, our methodology can be adapted to other transcriptomic and epigenomic modalities, such as RNAseq, genome-wide ChIPseq data.

2.6. Contributions

ksRepo was developed by Adam S Brown and Chirag J Patel, with contributions from Isaac Kohane and Sek Won Kong. Gene expression-based repositioning with ksRepo was conducted by Adam S Brown and Chirag J Patel, with contributions from Isaac Kohane and Sek Won Kong. Epigenome-based repositioning with ksRepo was conducted by Adam S Brown and Chirag J Patel. VS preclinical validation was performed in collaboration with Jessica E. Sagers, Sasa Vasilijic, Rebecca Lewis, Mehmet I. Sahin, Lukas D. Landegger, Roy H. Perlis, Isaac S. Kohane, D. Bradley Welling, Chirag J. Patel, and Konstantina M. Stankovic. The original description of ksRepo is published in *BMC Bioinformatics* (© the Authors and reproduced in part here with permission from the publisher, see Appendix B) [151].
3. TAPPING INTO PRIOR BIOMEDICAL KNOWLEDGE FOR DRUG REPOSITIONING

Currently available tools for leveraging the latent information on drugs contained within the biomedical literature for drug repositioning focus either on a single category of information (e.g. extraction of unannotated DTI \[48,58\] or adverse events \[152\]) or leverage Medical Subject Heading Term (MeSH terms) annotations, which describe the high-level concepts contained within a given biomedical research article (see \[54\]), as a source of drug-related information \[55–58\]. Within MeSH-based methods, currently available methods require external sources of prior biomedical knowledge because they link drugs to either genes or diseases, both of which require matching of free-text mentions to external vocabularies or ontologies.

In contrast, a more straightforward approach would be to directly utilize drug-MeSH associations within the literature (excluding genes, diseases or other biomedical entities that require external annotation). Borrowing from the DTI prediction approaches, we hypothesize that drugs that are that have similar drug-MeSH interactions might be similar in potential to treat disease. While drug-drug similarity most commonly refers to structural similarity in DTI prediction (e.g. distance between hashed, binary representations of structure, as used in \[18\] among others), similarity can also refer to molecular phenotypes elicited by a drug (e.g. as described in \[83\]), similarity in side-effect profiles \[19\], or an aggregate of multiple modalities. Drug-drug similarity has been widely applied to a variety of direct and indirect sources of evidence and high predictive power in discovering validated repositioning opportunities \[20,25,74,153,154\].

In this chapter, we describe a method for combining the information contained in MeSH terms with the strength of drug-drug similarity as a repositioning method. We extend methods for
chemical-wise MeSH term enrichment [155] and cluster drugs based on their pairwise similarities. We further describe a methodology for predicting novel indications within drug clusters based on cluster-wise disease enrichment and provide fully commented source code (at http://github.com/adam-sam-brown/MeSHDD) as well as an interactive online tool to aid investigators in generating repositioning hypotheses (at http://apps.chiragjgroup.org/MeSHDD/).

3.1. MeSHDD development

To identify drug-MeSH term overlap, we downloaded the main headings and corresponding chemical items file (which tracks articles referring to specific chemicals) from the MEDLINE® baseline repository for 2013 (accessed January 18, 2016, https://mbr.nlm.nih.gov/Download/), which contained 234,030,670 total MeSH term-article pairs for 20,275,470 unique indexed PubMed articles. Briefly, MeSH main headings describe the high-level topics explored in a given article in the biomedical literature. These terms are applied to articles as they are submitted for PubMed tracking. From this database, we extracted 81,474,709 Drug-MeSH co-occurrences corresponding to 1,629 unique FDA approved drugs catalogued in DrugBank (accessed January 18, 2015, http://www.drugbank.ca/) [15]. To ensure a high degree of specificity in our Drug-MeSH term overlap, we chose to keep those MEDLINE® chemicals with a case-insensitive full-length match to a DrugBank approved drug name.

Using the Drug-MeSH term overlap database, we calculated the enrichment for co-occurrence between each drug and MeSH term (Figure 8A) [155]. To do so, we calculated a hypergeometric p-value using the phyper function in the R programming language [81], which corresponds to the probability of having as many or more Drug-MeSH co-occurrences conditioned on the full
set of Drug-MeSH pairs. To control for multiple testing, we applied the Bonferroni correction using the \texttt{p.adjust} function in R. All associations with a Bonferroni adjusted p-value less than 0.05 were considered significant. Association testing resulted in 251,594 statistically significant pairs (of 42,984,423 total pairs tested for association).

**Figure 8. MeSHDD leverages literature similarity to pair drugs and diseases.** A) Literature similarity is assessed by calculating the bitwise distance between two drugs using their significantly associated MeSH terms. B) clusters are defined from pairwise distances using bootstrapping with 10,000 resamples. C) Repositioning hypotheses are developed by connecting drugs to new, significantly enriched indications.

To cluster each of the 1,629 drugs, we leveraged a binary distance measure, as implemented in the \texttt{dist} function in R. We first converted significant p-values to a binary bits, where Bonferroni-significant entries were considered “on” (a value of 1) and non-significant terms were considered “off” (a value of 0). The binary distance between any two drugs could then be calculated as the proportion of bits for which \textit{only one} drug was “on” amongst those where \textit{at least one} was “on” (see Figure 8A). Highly similar drugs (and those on the diagonal) have
distances close to 0, while those that are dissimilar have values close to 1. Drugs were then clustered using pairwise distances and bootstrapped k-means clustering as implemented in the `clusterboot` function from the `fpc` package in R (Figure 8B) [156]. We used `clusterboot` because it is optimized for large datasets and produces disjoint clusters containing all of the drugs in our database. We performed 100 bootstraps for each value of $k$, the number of clusters, between 10 and 50. This range of potential clusters corresponded to a large window around the commonly used “rule-of-thumb” value for $k$ [157]. Goodness of clustering was assessed using the Jaccard Index [158,159]; the value of $k$ that maximized the mean Jaccard Index was chosen as the optimal $k$ (see Figure 9). Following $k$ selection, we performed 10,000 bootstraps to define robust clusters. Using this process, we determined that the optimal number of drug clusters was 33, producing a median cluster size of 31 drugs.

![Figure 9. Bootstrapped clustering identifies best number of clusters ($k$) to be 33. The number of clusters ($k$) is shown plotted against the mean Jaccard Index among clusters. Error bars are ± standard error of the mean Jaccard Index across bootstraps. For $k = 33$, the mean Jaccard Index is maximized.](image-url)
To identify what indications were enriched in the putative indications in the drug-drug similarity clusters, we downloaded the Therapeutics Target Database (TTD, accessed January 23, 2016, http://bidd.nus.edu.sg/group/cjttd/) [16]. The TTD contains a variety of manually curated information about approved drugs, including Drug-Disease indication information. As before, we selected only those TTD drugs with a case-insensitive full-length match to a DrugBank approved drug name. We then calculated the statistical overrepresentation of each disease in a given cluster using the `phyper` function in R. The resulting p-value corresponds to the probability that a given cluster is enriched for drugs that treat a given disease, conditioned on the full set of disease-drug pairs. Bonferroni correction was applied within each cluster and across clusters to correct for multiple testing. Following correction, all p-values that remained less than 0.05 were considered significant. We detected enrichment for 482 unique diseases of the 2,305 diseases considered (Figure 10). Individual clusters claimed a modest number of indications each, with a median of two significantly enriched indications (among clusters with significant indications). Compared to currently available computational repositioning methods, which typically recommend hundreds or thousands of repositioning opportunities [37,74], MeSHDD on average provides a much tighter set of testable hypotheses.
Figure 10. MeSHDD workflow for drug repositioning using MeSH terms. 1) MeSH terms are downloaded from MEDLINE® Baseline Repository (2013 summary for this study). 2) Drug mentions are downloaded from MEDLINE® Baseline Repository, using the Chemical Items feature. 3) A list of approved drugs is downloaded from DrugBank. 4) The overlap between approved drugs and all MEDLINE® MeSH terms is computed. 5) Each drug-term pair is tested for significance using the hypergeometric test for enrichment. P-values from the test are corrected using the Bonferroni multiple-hypothesis testing method. 6) Drug-drug similarity is measured by binary distance (see Methods). 7) Drug-drug network neighborhoods are defined using bootstrapped k-means with the optimal number of clusters determined by highest mean Jaccard Index. Enrichment for indications is calculated using the hypergeometric test for enrichment. 8) Screenshot from R Shiny Application, showing cluster containing metformin (used in the case study, see Results). Height of cladogram is normalized distance between cluster members.

3.2. Software Availability

To enable investigators to browse the results described in this study, we developed and deployed an R/Shiny application (Figure 11). In ‘drug-centric’ mode, the MeSHDD application allows users to select a drug from the full list of drugs we examine, and then view other drugs in the selected drug’s cluster (displayed as a dendrogram) as well as cluster-enriched disease indications. In addition to the drug-centric mode, the application also allows investigators to
instead select a disease of interest and identify clusters for which that disease is enriched. The application is available at [http://apps.chiragjpgroup.org/MeSHDD/](http://apps.chiragjpgroup.org/MeSHDD/).

Figure 11. MeSHDD Shiny Application Usage. A) Drug-centric MeSHDD repositioning is performed by 1) selecting a drug from the 1426 drugs considered, and 2) viewing and selecting an enriched indication from the table. B) Disease-centric MeSHDD repositioning is performed by 1) selecting a disease of interest from the 622 diseases with enrichment in 1 or more of the 33 clusters, and 2) viewing clusters of drugs enriched for the treatment of that disease. For both A and B, the metformin repositioning case study is highlighted.

To allow for complete reproducibility of our methods and enable investigators to extend and modify MeSHDD, we provide full source code for producing the various datasets upon which MeSHDD is built, as well as the R/Shiny application itself ([https://github.com/adam-sam-brown/MeSHDD](https://github.com/adam-sam-brown/MeSHDD)).

3.3. MeSH-based similarity measure is correlated with therapeutic similarity
As described above, we calculated the pairwise distance between all drug-drug pairs based on overrepresented co-occurring MeSH terms. To assess whether more similar drugs according to our methodology would share more TTD indications, we performed ordinal logit regression using the *ordinal* package in R. Ordinal logit regression accounts for the fact that the number of shared indications between two drugs is an ordinal, rather than continuous. We regressed the binary distance between each pair of drugs on the number of shared indications extracted from the TTD. As expected, we found that the binary distance is strongly negatively correlated with number of shared indications (Ordinal Logit Regression, $\beta_{distance} \approx -21.5$, 95% confidence interval [-21.7, -21.3], $p < 2.2 \times 10^{-16}$), suggesting that high MeSH similarity is predictive of therapeutic similarity and is a potentially useful metric for repositioning.

### 3.4. Case study: repositioning metformin using MeSHDD

To demonstrate MeSHDD's capability, we present an end-to-end case study of hypothesis generation for metformin. We chose metformin because, in addition to being a first-line type 2 diabetes mellitus medication [160], it is an excellent example of successful drug repositioning. In addition to diabetes, metformin has been investigated for the treatment of a number of alternate indications [42,46,161].

To generate repositioning hypotheses for metformin, we first examined the MeSHDD clustering results using the MeSHDD R Shiny Application. As expected, metformin clusters with other known diabetes medications, including the glitzones, pioglitzone and rosiglitazone. Furthermore, MeSHDD correctly predicts both the primary indication, diabetes mellitus, as well as several investigational indications for metformin, including obesity, hyperlipidemia, and hypercholerolemia. Interestingly, the metformin cluster is also enriched for drugs treating cystic
fibrosis (CF), linking CF to metformin. This is striking, as metformin itself is not significantly associated with CF MeSH terms. Metformin is a potent activator of AMP-activated kinase (AMPK), which has recently been implicated in slowing the lung and renal pathologies of CF [43,44]. Despite initial excitement over the prospect of metformin as a well-tolerated AMPK agonist for the treatment of CF, to our knowledge metformin has not yet been tested for CF in a clinical setting (from clinicaltrials.gov, accessed March 3, 2016). Using MeSHDD, we were therefore able to identify a non-obvious and testable repositioning hypothesis for the use of metformin for CF therapy.

3.5. Limitations

While we provide support for MeSH-based drug-drug similarity as a useful approach in drug repositioning, we acknowledge that there are some limitations to our methodology. The primary limitation of our method is the inability to predict drug-disease pairs for diseases in which there are no FDA approved therapies. This problem is common to all drug-drug similarity based methods, and a number of possible extensions have been proposed, including incorporating investigational drugs currently in clinical trials [18,162], and by incorporating disease-disease similarity in an attempt to connect orphan diseases to predictions for well-treated indications [163]. In both cases, however, an additional inferential step is required to bring orphan diseases into similarity-based methods, potentially reducing accuracy. A second possible limitation of our method is the potential for bias in our results due to gross differences in occurrence between drugs; for example, progesterone is annotated to almost 800,000 articles, whereas insulin degludec (long-acting insulin) is only annotated to 162. This discrepancy could potentially bias results by allowing more well-studied drugs to have a higher chance of overlapping with MeSH terms. While this bias has been shown to be modest (around 11% for drugs with lower
annotation levels), we note that reported significance for newer drugs may be deflated. Another possible limitation is the potential for incorrect annotation of MeSH terms and drug mentions. Again, such errors have the potential to more strongly impact less annotated drugs (as a single misannotation accounts for a larger percentage of total annotations), further urging caution in interpreting MeSHDD results for newer drugs. Lastly, as with all computational repositioning studies, results presented here should be considered preliminary, and thus should be thoroughly tested in vitro and in vivo before use in humans.

3.6. Conclusions

In this chapter, we describe MeSHDD a framework for computational drug repositioning using literature-derived drug-drug similarity. Critically, we claim MeSHDD provides an alternate way of searching the biomedical corpus for novel (and existing) uses of approved drugs. We expanded previous methods using curated MeSH terms from MEDLINE® to find drug-MeSH term pairs that were enriched for co-occurrence in the medical literature, and developed a method for calculating pairwise similarities between drugs. Using this methodology, we robustly clustered 1,426 FDA approved drugs and identified within-cluster repositioning opportunities. We demonstrate the utility of MeSHDD with an end-to-end case study for metformin and identify a non-obvious, but supported opportunity for the treatment of cystic fibrosis. All analysis presented here are fully reproducible using open-source code available from GitHub; in addition, we provide free, interactive online tools for exploring the full results of MeSHDD.
3.6 Contributions

MeSHDD was developed by Adam Brown and Chirag Patel. Work described in this chapter is published in *Journal of the American Medical Informatics Association* (© the Authors and reproduced in part here with permission from *JAMIA*, see Appendix B) [164].
4. REPOSITIONING WITH COMPLEMENTARY QUANTITATIVE TRAITS IN CROSS-SECTIONAL AND LONGITUDINAL CLAIMS DATA

An emerging data modality for computational drug repositioning is observational clinical phenotyping from epidemiological cohorts and administrative data (e.g., insurance claims databases and electronic health records [61,63,64]. By leveraging the association between clinical phenotypes and drug prescription history, clinical phenotype-based methodologies can not only provide evidence of efficacy, but also help guide clinical trial design. While methods that consider the association between disease outcomes and drug exposure are challenging to interpret due to time-dependent confounding and high degrees of inflation [61,64], a promising alternative is to use quantitative traits, including physiometric and laboratory measurements, which may be more sensitive to drug exposure, indicative of future disease risk [67], and can be monitored over a shorter period of time [65,66]. Currently available quantitative trait-based methods use a large number of quantitative trait perturbations in concert to propose repositioning candidates. This strategy, however, limits the potential for clinical potential; not only is it unclear whether clinically-relevant phenotypes are driving drug-disease association, but it is also difficult to derive straightforward effect sizes for determining clinical trial size.

One promising alternative is to consider a single or a few quantitative traits’ association with drug prescription history. In doing so, one can not only be certain that the traits chosen are clinically relevant to a disease of interest, but also readily access effect sizes for power considerations in future clinical studies. While such a strategy is appealing, even a study limited to a single disease may be confounded due to disease co-heritability [45], off-label drug usage [165], and variable effects of drugs due to disease severity. In Chapter 4, we propose a novel
framework in which the association between drugs and quantitative traits is assessed in a population without the disease of interest in a non-institutionalized and otherwise healthy individuals participating in a health survey of the United States. Specifically, individuals selected for the survey are representative of the general population mitigating the risk for confounding by indication biases.

To demonstrate the potential of this strategy, we searched for potential antidiabetic drugs in a non-diabetic population derived from the 2005-2012 National Health and Nutrition Examination Survey (NHANES). Using the 5,371 person survey, we tested the association between prescription drug use and two diabetes diagnostics, fasting blood glucose and blood glucose following a 2-hour glucose tolerance test (or, glucose response). By combining associations found in two glycemic traits, we identified a potentially antidiabetic candidate, the antidepressant bupropion (Figure 12A). Notably, other commonly used antidepressants did not show multimodal antidiabetic potential. Because we were able to estimate the effect size of bupropion’s antidiabetic association, we designed a retrospective self-controlled study in a large medical claims dataset (Figure 12B), and again verified that bupropion, but not other commonly prescribed antidepressants, lowered fasting blood glucose.

4.1 Cross-sectional association of bupropion with complementary diabetes phenotypes

The cross-sectional study cohort was derived from a combination of four independent waves of the continuous NHANES: the 2005-2006, 2007-2008, 2009-2010, and the 2011-2012 surveys [166]. For this study, several variables were obtained for each respondent, including: (1) self-reported history of diabetes (field DIQ010 from the respective DIQ questionnaire datasets), (2) fasting blood glucose and fasting time, as well as glucose taken at 2-hours post oral glucose
tolerance test (field LBXGLU and PHASFSTHR, LBXGLT respectively from the respective GLU laboratory datasets) (3) self-reported prescription drug usage at the time of interview (including generic drug names and drug category, as defined in the Lexicon Plus® database, Cerner Multum, Inc, see Figure 12A). Respondents without completed information for any of these fields were excluded from further analysis. To obtain a non-diabetic final cohort for association testing, respondents were filtered to include those with no reported history of diabetes, no use of antidiabetic medications at the time of interview, and normal glycemic status (fasting blood glucose less than 100 mg/dL according to American Diabetes Association guidelines [167]).

After excluding participants with a reported history of diabetes, abnormal fasting blood glucose (including diabetes and prediabetes, according to American Diabetes Association guidelines, >= 100 mg/dL), and who were currently prescribed an antidiabetic drug, we obtained a final NHANES-derived cross-sectional cohort size of 5,371 (see Table 2 for demographic characteristics of the cohort).
Table 2. Demographic breakdown of NHANES (cross-sectional) and claims (longitudinal) cohorts

<table>
<thead>
<tr>
<th></th>
<th>NHANES</th>
<th>Claims Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bupropion (150mg SR)</td>
</tr>
<tr>
<td>N</td>
<td>5,371</td>
<td>386</td>
</tr>
<tr>
<td>Age (Mean [95% CI])</td>
<td>36.7 [36, 37.3]</td>
<td>56.3 [55.1, 57.5]</td>
</tr>
<tr>
<td>Sex (% Female [95% CI])</td>
<td>0.58 [0.57, 0.6]</td>
<td>0.71 [0.66, 0.75]</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% White [95% CI]</td>
<td>0.65 [0.63, 0.67]</td>
<td></td>
</tr>
<tr>
<td>% Black [95% CI]</td>
<td>0.12 [0.1, 0.14]</td>
<td></td>
</tr>
<tr>
<td>% Latino [95% CI]</td>
<td>0.17 [0.15, 0.18]</td>
<td></td>
</tr>
<tr>
<td>% Other Race [95% CI]</td>
<td>0.06 [0.04, 0.08]</td>
<td></td>
</tr>
</tbody>
</table>

Using this normoglycemic cohort, we performed comprehensive association testing between prescription drug use and either fasting blood glucose or blood glucose following an oral glucose tolerance test, adjusting for age, sex, race, and body mass index. Regression coefficients and significance were estimated using the `survey` package in R to account for the stratified design of NHANES [168]. To avoid erroneous associations, drugs with less than 12 prescribed individuals were removed from further analysis (power = 80% to detect an effect with Cohen’s d >= 0.2, see [169]). For the remaining 134 drugs, regression coefficients and significance were obtained and corrected for multiple testing using the False Discovery Rate (FDR) method [75]. Drugs with significant (FDR < 0.2), negative associations with both fasting blood glucose and blood glucose following a 2-hour post oral glucose tolerance test were considered candidate antidiabetic agents (Figure 12A).
Of the 134 prescription drugs with power (out of 1,133 total drugs tracked) to detect an association, the antidepressant bupropion was significantly and negatively associated with both glycemic phenotypes (survey multivariate linear regression $\beta < 0$, FDR < 0.2, Figure 12A). Bupropion was associated with -2.5 mg/dL (95% CI: [-4.4, -0.6]) lower fasting blood glucose, and -10.4 mg/dL (95% CI: [-16.8, -4]) lower blood glucose following an oral glucose tolerance test. We also tested the association between both phenotypes and two commonly used antidepressants, duloxetine and escitalopram, with sufficient populations of exposed individuals (>= 12 users, see above). Neither duloxetine or escitalopram had significant effects on both glycemic phenotypes, suggesting that the association with bupropion is specific (see Figure 12B for a schematic, and Table 3 for antidepressant results).

**Table 3. Antidepressant association with diabetes health in NHANES (cross-sectional)**

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Participants Prescribed (of 5,371)</th>
<th>Observed effect on fasting glucose (mg/dL) [95% CI]</th>
<th>FDR</th>
<th>Observed effect on glucose tolerance (mg/dL) [95% CI]</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bupropion</td>
<td>38</td>
<td>-2.5 [-4.4, -0.6]</td>
<td>0.12</td>
<td>-10.4 [-16.8, -4]</td>
<td>0.07</td>
</tr>
<tr>
<td>Escitalopram</td>
<td>43</td>
<td>-2 [-3.8, -0.1]</td>
<td>0.26</td>
<td>-13.6 [-22.4, -4.9]</td>
<td>0.08</td>
</tr>
<tr>
<td>Duloxetine</td>
<td>24</td>
<td>-1 [-4, 3]</td>
<td>0.84</td>
<td>-8 [-21, 6]</td>
<td>0.63</td>
</tr>
</tbody>
</table>
Figure 12. Quantitative-phenotype based repositioning overview. A) NHANES (cross-sectional) quantitative-phenotype based repositioning workflow. B) Conceptual diagram of claims data-based replication efforts.

4.2 Replication of cross-sectional signal using retrospective self-controlled studies

To replicate the association of bupropion with improved diabetic health, and to simulate a follow-up prospective clinical investigation, we performed a self-controlled study of fasting blood glucose in a large medical claims dataset. We extracted three non-diabetic populations with exposure either to bupropion or to one of two control antidepressants, duloxetine and escitalopram (see Table 2 for demographic characteristics of each drug-exposed cohort).

Cohorts were derived from a large insurance claims database, spanning 9 years and over 50
millions lives around 50 states and territories. To obtain drug-specific non-diabetic cohorts the following restrictions were made: (1) patients did not have any instance of diabetes diagnosis (ICD-9 codes 250-250.93), (2) patients were prescribed one of the study drugs (bupropion, duloxetine, or escitalopram), (3) patients were enrolled in an insurance plan for at least one year before the first prescription date for the respective study drug, (4) patients had a fasting blood glucose measurement up to one year before starting the respective study drug (“pre” measurement, chosen because glucose measurements are typically performed on an annual basis [167]), and between 8 and 60 days after starting (“post” measurement, with a buffer of 8 days to reach steady-state drug concentration) (Figure 12B).

Following cohort creation, we observed that age, sex, and race composition were not significantly different between drug-exposed cohorts (ANOVA, p > 0.1 for age, sex, and race). We then performed self-controlled analysis for each dose of each study drug using a paired t-test between the pre- and post-glucose measurements as previously described [170]. More recent methods have proposed additional corrections to a paired t-test method; however, these methods typically examine longer time periods of exposure (which allows for additional confounders to accumulate), and a population containing cases in addition to healthy controls [171]. In contrast, the method described here relies on a very short analysis window, and excludes cases. Dosage forms of study drugs with fewer than 198 participants were removed from further analysis due to power considerations (assuming a small effect size, Cohen’s d equal to 0.2, requiring power of 80% or greater, see [172] for details). For bupropion, the only dosage form with sufficient individuals was 150mg sustained release with 383 individuals (of 11 total dosage forms), for duloxetine only 60mg was powered with 717 individuals (of 5 total dosage forms), and for escitalopram only 10mg was powered with 206 individuals (of 3 total
dosage forms). Study drugs with significant (t-test p-value < 0.05) and negative associations (improved glucose response) with the pre- to post-drug regimes were considered validated antidiabetic agents. In confirmation of our NHANES results, we found that only bupropion was associated with significantly decreased fasting blood glucose (mean difference -1.88, 95% CI: [-2.91, -0.85], paired t-test, t = -3.6, DF = 385, p < 0.001, see Table 4).

### Table 4. Claims data (longitudinal) confirmation analysis of selected antidepressants

<table>
<thead>
<tr>
<th>Drug Name (Dosage)</th>
<th>Sample Size</th>
<th>Mean difference in fasting glucose (mg/dL) [95% CI]</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bupropion (150mg SR)</strong></td>
<td>386</td>
<td>-1.88 [-2.91, -0.85]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Duloxetine (60mg)</td>
<td>717</td>
<td>0.64 [-0.15, 1.44]</td>
<td>N. S.</td>
</tr>
<tr>
<td>Escitalopram (10mg)</td>
<td>206</td>
<td>0.36 [-1.06, 1.77]</td>
<td>N.S.</td>
</tr>
</tbody>
</table>

### 4.3. Conclusions

To demonstrate our complementary trait-based approach, we predicted repositioning candidates for diabetes using a 5,137 person non-diabetic cohort in NHANES. By using both fasting blood glucose and glucose following an oral glucose tolerance test, we identified a single, high-confidence candidate, the antidepressant bupropion. Bupropion is a treatment for obesity comorbid with diabetes, both alone [173,174] and in combination with naltrexone [175], as well as a monotherapy for comorbid depression and diabetes [176]. What is unclear from the previous studies, however, is the degree to which the observed effects were caused by improvement in body mass index (BMI) and depression, which subsequently may lead to improvement in glycemic status (confounding by indication). In contrast, we explicitly address confounding by adiposity, depression status, and glycemic status: (1) by explicitly adjusting for
BMI in all associations, (2) by testing other commonly used antidepressants associations with improved glycemic health, as measured by fasting glucose and glucose response, and (3) by testing for associations in a non-diabetic and normoglycemic cohort, decreasing the likelihood of confounding scenarios, such as a single upstream process controlling both depression status and diabetic health (e.g. statins lower low-density lipoprotein levels, which in turn lowers heart attack and stroke risk simultaneously).

Another key benefit of our method is the ability to design clinical studies derived from our potential discoveries; we demonstrated this benefit by performing a retrospective self-controlled study in a large medical insurance claims dataset. By designing an experiment by which patients serve as their own control, we avoid time-invariant confounding [170]. We successfully replicated that bupropion alone among the antidepressants considered is associated with lower fasting blood glucose. Despite escitalopram’s significant improvement of glucose response in the NHANES cohort, we did not observe a significant impact on fasting blood glucose in the self-controlled study. This result underscores the importance of combining multiple quantitative traits to achieve high specificity in repositioning candidates. Furthermore, to enable future clinical trials, quantitative trait-driven studies should carefully select phenotypes with the goal of not only identifying repositioning candidates with high specificity, but also with clinical study design in mind, considering both clinical relevance and statistical power.

While we have discussed the potential of combining complementary quantitative traits for drug repositioning, we do note that it has some limitations. Chief among these is the requirement that non-diseased individuals are assayed for quantitative traits. While common diseases by necessity have associated routine diagnostics, for example fasting blood glucose and glucose
tolerance for diabetes or lipid levels and blood pressure for cardiovascular disease, repositioning for rarer diseases may require non-standard tests. We expect that this limitation will diminish over time with the development of birth cohorts (e.g. ALSPAC [177] among others), and large biobanking initiative (e.g. UK Biobank [178] among others), most of which include clinical phenotyping of all participants to complement a variety of ‘omic measurements. A second potential limitation is the requirement for multiple complementary quantitative traits that are associated with the disease of interest for repositioning. For diseases where such phenotypes are not available, further biomarker identification may be required before using our repositioning strategy. We note that any quantitative trait-based methodology would likely require disease-associated phenotypes before producing meaningful and interpretable results. Lastly, because we assess all associations between drugs and phenotypes in a non-diseased population, it will be important to verify any repositioning candidates that arise from this method in disease sufferers.

4.4. Contributions
The work described in this chapter was conducted by Adam Brown, Danielle Rasooly, and Chirag Patel. Results have not been published as of the submission of this thesis, however a preprint is available from bioRxiv (DOI: 10.1101/130799, © the authors).
5. ENABLING REPRODUCIBILITY IN COMPUTATIONAL REPOSITIONING

Despite the availability of dozens of computational repositioning methods, there is no agreement in the field over the proper way to perform validation of in silico predictions, and in fact no systematic review of repositioning validation methodologies. We therefore conducted a literature search for drug repositioning methods with a claim of analytic validation, and present our findings for 39 peer-reviewed articles. Strikingly, we note that not only is there substantial variability in the types and sources of drug-disease relationships, there most commonly used validation strategy rests on the assumption that all potential repositioning candidates are a priori false positives. We posited that to enable reproducibility in the computational repositioning field, a standard database with both true and false positives was needed.

To address this unmet need, we developed repoDB, which contains approved drug-disease pairs derived from DrugCentral [179], as well as failed drug-disease pairs derived from the Clinical Trials Transformation Initiative’s Aggregate Analysis of ClinicalTrials.gov tool [180]. To our knowledge this represents the first database with both true and false positive drug-disease pairs, and represents a step forward towards consistent and reproducible analytic validation in the field.

5.1. A brief review of analytic validation

All computational repositioning methods promise to prioritize repositioning candidates and studies describing these methods typically claim superiority over competing methodologies. To do so, such studies perform analytic validation, whereby they compare the computational results of their methods (and competing methods) to existing biomedical knowledge. A successful
method is one that consistently identifies known associations between drugs and diseases (and for some, fails to identify "wrong" associations). When examining the repositioning literature, however, it is apparent that there are no consistent best practices for comparing studies and for validation of methods.

To gain a better understanding of validation in the computational repositioning field, we searched PubMed for articles in the computational drug repositioning space that claimed to have performed validation of their methodology or pipeline using a boolean search (“(drug repositioning OR drug repurposing) AND (gold standard OR AUC OR receiver operating characteristic OR validation OR validated OR validate),” performed on June 14, 2016, see Figure 13A).
Figure 13. Computational Repositioning Validation Studies. A) The search term used with PubMed to retrieve articles. B) Sources of drug-indication annotation data used in studies retrieved in the literature search. C) Types of validation in studies retrieved in the literature search.

Using this search, we began our analysis with a pool of 213 articles. To further refine our search, we manually reviewed each of the articles, and excluded non-computational papers (e.g.
high-throughput drug screens in cell lines or clinical trials), those not actually in the small molecule/drug field (e.g. articles referring to surgical or dental repositioning), and non-research articles (reviews and book chapters). From the remaining 141 articles, we focused on those that predicted novel indications for drugs. At this point, we excluded 35 articles that focused on target prediction only; *In silico* target prediction studies aim to predict novel, molecular targets for existing and novel drug candidates. We argue that target prediction is still one step removed from drug repositioning; true or predicted molecular targets can be used as part of repositioning methodologies, but do not themselves provide the full repositioning hypothesis from drug to indication. Furthermore, we note that benchmarking such studies is already possible with the wealth of high-throughput drug-target binding screens (see [181]).

From the studies that predicted new indications for existing drugs, we excluded 67 articles that predicted indications for a single drug or disease, and kept those that made predictions for more than one drug and disease. We excluded these single drug and disease studies because they were not designed to be applied broadly, and often contained domain specific knowledge about a particular drug or disease (e.g. GWAS results for a single disease or structure-association relationship studies for a single drug). This resulted in 39 computational repositioning methods articles with predictions spanning multiple drugs and indications and a clear claim of analytic validation (the full list of captured articles is provided in Table S1).

We began our analysis by first examining the types of databases used for analytic validation: we discovered that, although many of the investigators in the 39 studies we examined claim to use a “gold standard,” there is substantial heterogeneity in the source of these standards as well as the types of data they contain (Figure 13B). For example, DrugBank [15] contains information
about only the FDA approved indications for drugs, while the Comparative Toxicogenomics Database (CTD) [47] contains literature-annotated links between drugs and both approved and investigational indications. While DrugBank contains a set of true drug-indication annotations, it misses off-label uses and late-stage clinical trials; on the other hand, the CTD relies on literature annotations and contains drug-indication pairs that have subsequently failed to receive FDA approval. This inconsistency in specificity among databases used for validation is detrimental to reproducibility and may lead to claims on extremely high accuracy.

We next examined the types of analytic validation methodologies used by investigators in computational repositioning. We grouped the 39 captured studies into three classes: 1) validation with a single example or case-study of a single disease area (CSV), 2) sensitivity-based validation only (SV), and 3) both sensitivity- and specificity-based validation (SSV) (Figure 13C). First, of the three validation types, CSV is the least rigorous; each of the four CSV studies reported one to three clinically justifiable predictions (see Table S1). For example, Sirota and colleagues (in [83]) identified cimetidine as a potential therapy for lung adenocarcinoma, which was picked from 2,664 significant predictions (out of more than 16,000 drug-indication pairs tested) on the basis of tolerability. The investigators provided evidence of its success using both the literature and an in vitro study. The inclusion of in vitro evidence lends additional biological credence to their single case study, but analytic evidence of their method’s overall success is lacking. We note here that we are not arguing that in vitro evidence is inferior to analytic validation; biological validation is a requirement for any individual candidate to be advanced in a drug development pipeline. However, successful biological validation of a single repositioning candidate cannot be extrapolated to all predictions made by a method.
Following CSV, SV provides more analytic rigor by measuring the overlap between currently approved or investigational indications for drugs and the indications predicted by a given repositioning method. In contrast to CSV, SV validation methodologies assess the general ability of repositioning methods to make reasonable claims, rather than selecting a single or several high-ranking predictions to test in depth. For example, Jung and Lee (in [182]) examined the overlap between predictions made by their method and both approved drug indications (from a combination of DrugBank [15], PharmGKB [183], and TTD [184]) and investigational indications from ongoing clinical trials (from ClinicalTrials.gov). SV is appealing because investigators only need to have a set of true positives to which to compare their predictions (e.g. all approved or investigational drug indications). A key drawback of SV is the inability to use traditional two-class machine learning (ML) approaches. An alternative is to train one-class classification algorithms on positive examples only; however, to our knowledge no methods in the drug repositioning space have used one-class ML approaches. We emphasize, as in any ML exercise, that investigators should perform cross-validation in which algorithms are fine-tuned on a portion of the data and tested on another; testing using an as yet unseen portion of the data is more representative of future performance than training and testing on the full dataset [185].

Both CSV and SV validation methods are less popular than SSV. SSV is, in theory, the most rigorous type of validation. For our purposes, SSV-based methods include those that directly report sensitivity and specificity (or reported values for positive or negative predictive value), as well as area under the receiver operating characteristic (AUROC, a commonly used method for determining the predictive value of a method reviewed in [186]). For example, Gottlieb and colleagues (in [18]) used a list of approved drug indications (from DailyMed), and determined
how many of their predicted drug indications overlapped with that set (true positives) or did not overlap (false positives); their results are summarized by calculating the AUROC of their predictions. In contrast to sensitivity-only validation methods, methods that rely on both sensitivity and specificity require information about which predicted drug indications are false (false positives). In all of the SSV studies we reviewed, the investigators chose to mark all unannotated drug-indication pairs as false positives. This is troubling for two major reasons. First, the choice of annotation database can substantially impact the sensitivity and specificity estimates. If investigators consistently used a single database of standardized indication information, this issue could be avoided; however, in practice, annotation is derived from a variety of drug information databases and annotation types, from FDA approval, to ongoing clinical trials (Figure 13C). Second, marking unannotated pairs as false suggests that all novel repositioning hypotheses are false positives. This is obviously counterintuitive, as computational repositioning methods should predict novel indications, for which there is no currently annotated association. In addition, this strategy creates a substantial imbalance in the number of true and false positives; such an imbalance has been shown to reduce the accuracy of AUROC and other SSV estimates [187].

5.2. repoDB: a new paradigm in repositioning validation

To address the concerns raised through our review of validation, we present repoDB, a database of approved and failed drugs and their indications. repoDB approved indications were drawn from DrugCentral, which contains United Medical Language System (UMLS) indications mapped from free-text mentions in drug labels [179]. The UMLS is a large biomedical thesaurus that contains information about a wide variety of medical concepts [54]. We downloaded the full DrugCentral PostgreSQL database, and extracted the tables containing DrugBank identifiers,
synonyms for all drugs, and UMLS-mapped indication terms (http://drugcentral.org/, see [179]). DrugCentral provides comprehensive information about approved and investigational drugs, including UMLS-mapped approved indication(s) and, important for the construction of repoDB, all synonyms for a given drug (see Table 5 for database characteristics). DrugCentral uses the OMOP annotation pipeline to map free text drug labels to UMLS terms, which achieves high annotation accuracy (F1 measures around 0.98) [188]. We retrieved all DrugCentral synonyms for all Food and Drug Administration of the United States (FDA) approved drugs. A list FDA approved drugs was derived from DrugBank, a large drug database that is commonly used by computational drug repositioning methods and is frequently updated with new information (see Figure 14A) [15].

Failed indications for repoDB were drawn from the American Association of Clinical Trials Database (the “AACT Database”, Clinical Trials Transformation Initiative, 2016), which contains structured clinical trial records from the National Library of Medicine’s ClinicalTrials.gov service, and includes information about current trial status and interventions (e.g. drugs, lifestyle changes) studied in each trial. We chose to use AACT/ClinicalTrials.gov as our source for trial information because the sponsors of most failed trials do not publish their results in the scientific literature (around 78% fail to publish) [189]. We loaded and parsed the full database in R statistical programming environment [81], and took only those trials that included: 1) an annotated phase between phase 0 and phase 3, 2) a current, overall status of suspended, terminated, or withdrawn, and 3) a MeSH term-mapped intervention (provided by AACT), and 4) a UMLS term-mapped indication (provided by investigators and/or MetaMap analysis of free-text trial descriptions, see Table 5). While the majority of terms are derived from investigator supplied UMLS terms, ClinicalTrials.gov supplements these using the NLM Medical Text Indexer (MTI) to
map text to high confidence MeSH/UMLS (F1 measure around 0.55) [52]. We mapped all annotated interventions to DrugCentral synonyms and excluded trials that were not mappable to at least one approved drug (Figure 14A). Indication information was mapped to UMLS identifiers using the UMLS REST API [54].

**Table 5. Characteristics of databases* using in the construction of repoDB**

<table>
<thead>
<tr>
<th>Source Name</th>
<th>Type of Record</th>
<th>Number of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrugCentral</td>
<td>Indication (UMLS)</td>
<td>41,388</td>
</tr>
<tr>
<td>AACT</td>
<td>Indication (UMLS)</td>
<td>113,571</td>
</tr>
</tbody>
</table>

* Static versions available through figshare (DrugCentral [Full PostgreSQL Database], AACT [Pipe Delimited], Data Citation 1).
Figure 14. repoDB data sources and database characteristics. A) repoDB data were downloaded from two sources: (1) the AACT indexed version of ClinicalTrials.gov for failed indication information, and (2) DrugCentral for approved indication information. AACT drug-indication pairs were filtered to include only failed pairs, and exclude currently approved pairs. B) The repoDB database contains 6,677 approved drug-indication pairs and 4,123 failed drug-indication pairs. Indications are broken into UMLS semantic types, which describe broad categories of disease.
As the final step in creating the repoDB database, we reconciled the approved and failed indication information. We removed all failed trial information for drug-indication pairs that were currently approved: for example, metformin is an FDA-approved drug for diabetes mellitus; there are, however, trials marked as terminated with metformin as a primary intervention (e.g. metformin combination therapies, see NCT00762957) and these trials were removed. After combining the approved and failed indications, we kept only those drug-indication pairs for which the indication fell within a UMLS semantic type related to disease ('Disease or Syndrome', 'Neoplastic Process', 'Pathologic Function', 'Finding', 'Mental or Behavioral Dysfunction', 'Sign or Symptom', 'Injury or Poisoning', 'Congenital Abnormality', 'Acquired Abnormality', and 'Cell or Molecular Dysfunction'). Semantic types describe broad categories of disease as well as other medicine-related concepts; it is therefore necessary to filter out non-disease terms, including those with semantic types such as, “Health Care Related Organization.” The final repoDB database spans 1,571 drugs and 2,051 UMLS disease concepts, accounting for 6,677 approved and 4,123 failed drug-indication pairs (Table 5, Figure 14B, Figure 15). To further assist investigators, we provide a web application (http://apps.chiragjgroup.org/repoDB/) that enables browsing of the full repoDB database and allows users to download either the full database (or portions relevant to their work). repoDB will enable investigators to not only benchmark their computational repositioning methods, but also gain insight into trends in the drug discovery field and avenues that have not yet been explored.

Table 5. Summary of data available for download through repoDB

<table>
<thead>
<tr>
<th>Indication Status</th>
<th>Unique Drugs</th>
<th>Unique Diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approved</td>
<td>1,519</td>
<td>1,229</td>
</tr>
<tr>
<td>Terminated</td>
<td>386</td>
<td>785</td>
</tr>
<tr>
<td>Withdrawn</td>
<td>199</td>
<td>279</td>
</tr>
<tr>
<td>Suspended</td>
<td>77</td>
<td>143</td>
</tr>
</tbody>
</table>
Figure 15. repoDB at-a-glance. Indications were grouped by high-level UMLS “Semantic Types,” which provide insight into broad categories of disease. A. Overlap of approved drugs between semantic types are shown as number of shared drugs. Black indicates the higher overlap and white indicates the lowest (diagonal entries were removed). B. Drugs that are approved for one semantic type and failed in another are shown as above.
5.3. Analytic validation of previous work using repoDB

To illustrate what we envision to be the key use case of repoDB, we present analytic validation of several of the methods described in this thesis. In general, the goal of analytic validation is to determine how well a computational method does at capturing known true positives (as measured by either sensitivity or recall) and excluding known false positives (as measured by either specificity or precision), and is only useful in scenarios where many predictions are being made. With only a single candidate, measures of sensitivity and specificity are binary - either the candidate is, or is not an approved or failed drug for diabetes. For example, the methodology presented in Chapter 4, when applied to diabetes mellitus alone, resulted in a single prediction for a repositioning candidate; because bupropion is not approved for diabetes, giving it a crude sensitivity of 0%, but has not failed for diabetes, giving it a specificity of 100%. Obviously, these results are specious given the limited nature of our predictions, and single-agent studies should not be evaluated using repoDB.

A more interpretable and meaningful analytic validation with repoDB can be performed on our exploration of ksRepo in PC. The first challenge in any evaluation is the conversion of non-UMLS disease names/categories to UMLS terms as used by repoDB. repoDB enables rapid conversion in PC by simply searching for “Prostate” among the available indications, resulting in a list of varying types and severities of PC, including: "Metastatic Prostate Carcinoma", "Malignant neoplasm of prostate", "Adenocarcinoma of prostate", “Prostate cancer recurrent", "Hormone refractory prostate cancer", "Hormone-Resistant Prostate Cancer", "Prostate carcinoma", and "Castration-resistant prostate cancer". repoDB contains a total of 19 unique drugs annotated to PC-related diseases, with 9 approved, 1 suspended, 6 terminated, and 4 withdrawn drug-disease pairs. To avoid unfairly biasing our estimates of sensitivity and
specificity, we removed drugs that were not present in the built-in CTD dataset used to predict repositioning candidates for ksRepo, resulting in 5 approved, 1 suspended, 6 terminated, and 4 withdrawn drug-disease pairs. Using this CTD-compatible set, ksRepo achieved a mean recall of 60% and mean precision of 47% (mean F1-score of 0.51) across the five PC datasets described in Chapter 2. We do not present evaluation of the AML or VS use-cases here because they produce a small number of predictions or an orphan disease indication with no sensitivity data respectively.

While our ksRepo validation studies are informative, they represent a restricted validation for predictions on a single disease class. When designing repoDB, we envisioned users performing analytic validation on all-vs-all datasets, where many drugs are linked to many diseases. To demonstrate this type of validation, we therefore evaluated the full set of predictions made by MeSHDD. For this example, we examined a subset of MeSHDD with case-insensitive matches to drug-disease pairs from repoDB, resulting in a total of 540 repoDB-compatible predictions made by MeSHDD, corresponding to roughly 5% of all repoDB drug-disease pairs. In general, we posit that this number will fluctuate based on the level of specificity in the indication predictions made by a given repositioning tool. Using this subset, MeSHDD achieves a recall of 46% and a precision of 88% (F1-score of 0.60).

repoDB represents, to our knowledge, the first database containing both true positive and true negative drug-indication pairs to enhance repositioning process. We suggest that validation similar to the above examples be conducted for all future methods. As noted, previous estimates of accuracy may be inflated due to lack of true negatives, making methodological comparisons challenging. Furthermore, methods sometimes rely on a single case-study to provide evidence
of validation for tens of thousands of predictions (as in [83]), leaving investigators to wonder how they might choose another “hit” from a sea of predictions. It is therefore important for the field to rigorously evaluate new methods to provide preclinical and clinical investigators with realistic probabilities of success.

5.4. Conclusions

We present here a brief review of the computational drug repositioning field, with a focus on strategies for analytically validating such methods. We describe the three types of validation currently in use, and highlight the issues with both consistency and key assumptions made by each. To address the lack of consistency in the field, and to rectify a common, but unsatisfying assumption, we developed repoDB, a standard database of approved and failed drug-disease pairs. Together, these two efforts help draw attention to the need for reproducibility and consistency in computational drug repositioning.

5.5. Contributions

Work described in this chapter was performed by Adam Brown and Chirag Patel. The review described in this chapter is published in Briefings in Bioinformatics (© the Authors and reproduced in part here with permission from Briefings in Bioinformatics, see Appendix B) [190]. repoDB is described in Nature Scientific Data (© the Authors and reproduced in part here with permission from Briefings in Bioinformatics, see Appendix B) [191].
6. CONCLUSIONS

Repositioning of previously approved drugs for novel indications is a promising methodology because it reduces the cost and duration of the drug development pipeline and reduces the likelihood of unforeseen adverse events. Computational repositioning is especially appealing due to the ability to rapidly screen candidates in silico and to reduce the number of possible repositioning candidates. This dissertation set out to expand the scope and reproducibility of computational drug repositioning. In this final chapter, we review the contributions of the work described here to the field, and open questions that remain to be addressed in future work.

6.1. Contributions

In this thesis, I describe work to expand the computational repositioning toolbox in three distinct areas: high-throughput molecular phenotype-based repositioning, literature similarity-based repositioning, and quantitative trait-based repositioning. In the first area, high-throughput molecular phenotype-based repositioning, we addressed a major limitation of currently available methods, the requirement for specific data types and formats, which prevents investigators from utilizing newer profiling technologies and combining disparate datasets. To allow for more flexibility in the types of omic modalities accessible for repositioning, we developed a generalized tool, ksRepo that enables researchers to combine any type of case/control gene-level experiments with a drug-gene database of their choosing. In the second area, literature similarity-based repositioning, we noted an unmet need for a method that enables investigators to directly compare individual drugs and to investigate their literature-based "similarities", as is commonly done for gene- and structure-based repositioning approaches. To address this need, we built on the MeSH-driven repositioning literature and have developed a novel approach, MeSHDD, that uses a novel drug-drug similarity measure based on MeSH
terms. Finally, we addressed a common concern with all computational repositioning methods, that such methods fail to provide direct evidence of repositioning candidate efficacy in humans. We developed a pipeline that leverages complementary quantitative traits for repositioning in a healthy, US-representative population derived from the National Health and Nutrition Examination Survey (NHANES). We further validated our results using a retrospective case-crossover study in treatment naive patients from a large medical claims dataset.

In addition to the methodological developments described in this thesis, I also examine the state of analytic validation, and by extension reproducibility in computational drug repositioning. Despite substantial interest from academia and industry, fairly comparing the wide array of computational repositioning methods is a challenge due to inconsistencies in method validation in the field. To assess the degree of inconsistency in the field, we reviewed computational repositioning methods and sought to summarize the ways in which such methods were “validated.” We identified a key assumption required for the most common form of analytic validation and developed a resource, repoDB, that bypasses this assumption by providing a standard set of both approved drugs and failed repositioning efforts.

6.2. Open questions in computational drug repositioning

While we have described here four attempts to expand the scope and reproducibility of computational drug repositioning, there are still fundamental questions the remain unaddressed in the computational drug repositioning field. These can be broadly broken into questions of (1) sources of bias in repositioning, (2) the role of multiple testing correction on Type I Error, or the false positive rate of repositioning candidate nomination, and (3) the role of Type II Error, or the false negative rate of candidate nomination.
6.1.1. Sources of Bias

While we have discussed several sources of data-specific bias throughout this dissertation, we emphasize here several of the most impactful forms of bias. One dominant form of bias is sampling bias: older, and more popular drugs are studied more than newer or less popular drugs. This effect can be seen in all three of the major data modalities we’ve discussed in this dissertation: (1) in molecular phenotyping, certain drugs are more likely to have been profiled by the CMAP or L1000, or have more gene-drug interaction annotations in databases due to higher degrees of characterization, (2) in literature annotations, more studied drugs have more publications and therefore more annotations, and (3) in observational studies, people taking a drug are doing so because of a disease or condition, introducing the possibility of confounding by indication (taking a drug cures a disease which then has the effect indirectly).

Both (1) and (2) have the benefit of being, to some extent, self correcting: as drugs are available and studied for more time, their presence in annotation (both literature and molecular) and profiling efforts should increase. Investigators should nonetheless be careful to acknowledge that with increasing numbers of studies comes an increased risk of potential annotation errors; if such errors are systematic, they could further introduce bias into any repositioning methods relying on annotation. In contrast to (1) and (2), (3) is considerably more challenging to address; even after including negative controls to reduce the chance of confounding by indication, methods relying on drug prescriptions are still exposed to bias due to prescribing patterns, co-prescription, and other “hidden” or residual confounders [63].

Another considerable source of bias is reverse causality. Reverse causality is seen when investigators fails to properly account for causality in association studies; for example, a
diabetes drug might paradoxically be associated with *higher* fasting blood glucose in a quantitative trait-based study [192]. Drawing the conclusion that the drug causes higher blood glucose is erroneous, as having higher blood glucose is *what caused the prescription*, and not the other way around. While investigators can, as we have shown in Chapter 4, attempt to explicitly correct for this phenomenon, paucity of deep phenotyping on non-diseased individuals and clinical data for drug exposure on healthy individuals limits the possibility to fully correct for reverse causality in practice. With increasing focus on population-scale deep-phenotyping [178] and clinical data sharing initiatives [193], investigators may, in the near future, be able to further eliminate the impact of reverse causal bias in computational repositioning.

6.2.2. *Type I Error*

A key assumption is prevalent in most repositioning methods: that by correcting for multiple testing at the final step (either with FDR or Bonferroni-type correction [75]), that is repositioning hypothesis generation, the false positive rate is sufficiently controlled and the resulting candidates are significant. This assumption is unsatisfying for a number of reasons: (1) methods that report false positive rates or precision have previously relied on unproven true negatives leading to overly inflated precision estimates (see Chapter 5 and [190]), (2) few methods present both FDR and Bonferroni-corrected significance for their repositioning hypotheses, thus failing to provide sufficient guidance for future investigators, and (3) while not comprehensive, our own methods achieve highly variable levels of precision (see Chapter 5, Section 3), suggesting that different data modalities may produce more or less precise hypotheses (some modalities like drug-related outcomes cannot be sufficiently controlled for Type I Error at all, see [61]). Future work in computational repositioning should therefore carefully examine the role of different multiple testing correction schemes on the resulting precision of repositioning method.
Besides the commonly used Bonferroni and FDR, investigators should include empirically-derived corrections such as permutation-based adjustment when direct case/control or phenotypic comparisons are made (see [194]). In addition, investigators should carefully consider whether independent uses of a method constitute sufficiently distinct investigations as to not warrant being grouped as a “family” of inferences and corrected in tandem (to control for “Family-Wise Error Rate” or FWER) [195]. For example, if a method such as ksRepo were applied to additional datasets in either PC or VS, it could be argued that the original results should be included in any subsequent FWER. Lastly, investigators should be careful to include true negatives (as included in a database such as repoDB) to avoid inflated estimates of precision.

Beyond multiple testing correction of candidate significance, most repositioning methods require additional upstream hypothesis testing. For example, the CMAP and related methodologies ([37,39,83], among others), require users to provide lists of significantly differentially expressed genes; none of these methods, however, provide guidance on what correction methodologies to use to control for multiple testing, or explicitly explore the impact of different multiple testing correction schemes on the downstream repositioning results. Others, like ksRepo [151], rely on external databases of curated drug-gene interactions, which are often themselves derived from corrected or uncorrected literature reports, with no ability to re-filter for specific FDR or Bonferroni cutoffs without individually validated specific interactions (as in CTD [47]). Finally, methods that use indirect evidence like literature-derived MeSH terms (including our MeSHDD approach and others [163,164,196]), also require correction for multiple testing upstream of repositioning hypothesis generation. Future repositioning methodologies should explicitly
examine the upstream assumptions of the data on which their results rely, and if possible, use multiple types of correction to test the sensitivity of their results to different correction schemes.

6.2.3. Type II Error

Unlike Type I Error, there is very little discussion of Type II Error (or the additive inverse of Type II Error probability, statistical power) in the computational repositioning literature. In fact a PubMed search of “(type 2 error OR type II error) AND (computational drug repositioning OR computational drug repurposing)” reveals no articles that discuss both (Accessed June 5, 2017). Why should this be the case? A possible reason would be that, unlike Type I Error, which can be addressed in a post-hoc fashion, power calculations are calculated \textit{a priori}, either using either a theoretical framework or empirical simulation study. For some studies, this is accomplished with established power calculations, as in the method described in Chapter 4; we performed power analysis for multivariate linear regression and paired t-tests for our cross-sectional and longitudinal analyses respectively. Unfortunately, many methods rely on novel test statistics with no current theoretical power calculation (e.g. CMAP’s connectivity score [37], MeSHDD’s similarity measure) or use statistics in a way that violates the assumptions of the test statistic distribution. For example, ksRepo’s KS score violates the theoretical two-sample KS distribution in that genes can be tied in terms of significance.

In those cases, therefore, it is necessary to derive an underlying empirical distribution to estimate power, and \textit{a priori} simulations that vary parameters such as sample size and significance level. To take ksRepo as a concrete example, a potential approach to estimate power would be to repeat the following steps a large number of times:
1. Randomly generate a full length gene list of length $N$ (potentially corresponding to a given microarray or sequencing platform’s expected coverage)

2. Randomly generate permuted compound gene lists of length $j \in \{1 \ldots N-1\}$ to develop a null distribution of KS-scores for a range of potential compound interaction lists

3. For a given compound gene list length $j$, calculate the significance of all $KS \in \{0 \ldots 1\}$ as described for ksRepo

Repeating this process a large number of times would enable the calculation of $\beta$, or the probability of making a Type II Error, for a given $N$, $j$, and confidence level, $\alpha$. To get to this simplified simulation, we need to assume that both the full length gene list and drug-gene interaction lists are sampled randomly from the genome. This is certainly not the case for drug interactions [4], opening the possibility of needing independent distributions for each drug-gene interaction database used for ksRepo, and potentially each case/control dataset. Furthermore, even in this simplified example, we have 3 parameters for simulation: (1) the number of genes in the full length gene list, which may vary based on profiling modality and conditions, (2) the number of genes in a drug-gene interaction list, and (3) the $\alpha$ value at which to call significance, which may vary depending on multiple testing correction strategy as described above. Even simulating a single value of $N$ and corresponding values of $j$ could take minutes to hours for a typical end-user of ksRepo, making comprehensive assessment of the power of ksRepo prohibitively time-intensive.

Beyond ksRepo, increasingly complex and computational intense repositioning methods may have longer run-times and larger spaces of parameters that impact power; for these methods, even focused assessments of power may therefore prove difficult for even high-performance computing environments. Owing to the simulation burden, we therefore advise informaticians to
work closely with statisticians in developing new computational tools. In doing so, they may be able to develop either theoretical justifications of power, or to carefully make simplifying assumptions to enable simulation studies.
LITERATURE CITED


13. Weininger D. SMILES, a chemical language and information system. 1. Introduction to


17 Nakajima M. From the Viewpoint of Drug Metabolism Research. *Yakugaku Zasshi* 2017;**137**:697–705.


77


50 Vilar S, Friedman C, Hripcsak G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Brief Bioinform Published Online First: 17 February 2017. doi:10.1093/bib/bbx010


63 Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model


69 Blom K, Senkowski W, Jarvius M, et al. The anticancer effect of mebendazole may be due to M1 monocyte/macrophage activation via ERK1/2 and TLR8-dependent inflammasome activation. *Immunopharmacol Immunotoxicol* 2017;1–12.


75 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful


88 Lawrence M, Daujat S, Schneider R. Lateral Thinking: How Histone Modifications Regulate


124 Masuda A, Fisher LM, Oppenheimer ML, *et al.* Hearing changes after diagnosis in


147 Dilwali S. *From bench to bedside: elucidating vestibular schwannoma pathobiology to*


159 Hennig C. Dissolution point and isolation robustness: Robustness criteria for general cluster


doi:10.1093/jamia/ocw142


188 Overhage JM, Ryan PB, Reich CG, *et al.* Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012;**19:**54–60.


194 Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014;**15:**335–46.


196 Cheung WA, Ouellette BFF, Wasserman WW. Compensating for literature annotation bias when predicting novel drug-disease relationships through Medical Subject Heading Over-representation Profile (MeSHOP) similarity. *BMCMed Genomics* 2013;**6 Suppl 2:**S3.

197 Chavan SS, Bauer MA, Peterson EA, *et al.* Towards the integration, annotation and association of historical microarray experiments with RNA-seq. *BMC Bioinformatics*
2013; Suppl 14:S4.


APPENDIX A. ARRAYLASSO: A NETWORK-BASED APPROACH TO MICROARRAY INTERCONVERSION

A.1. Abstract

Summary: Robust conversion between microarray platforms is needed to leverage the wide variety of microarray expression studies that have been conducted to date. Currently available conversion methods rely on manufacturer annotations, which are often incomplete, or on direct alignment of probes from different platforms, which often fail to yield acceptable gene-wise correlation. Here, we describe aRrayLasso, which uses the Lasso-penalized generalized linear model to model the relationships between individual probes in different probe sets. We have implemented aRrayLasso in a set of five open-source R functions that allow the user to acquire data from public sources such as GEO, train a set of Lasso models on that data, and directly map one microarray platform to another. aRrayLasso significantly predicts expression levels with similar fidelity to technical replicates of the same RNA pool, demonstrating its utility in the integration of data sets from different platforms.

Availability and Implementation: All functions are available, along with descriptions, at https://github.com/adam-sam-brown/aRrayLasso.

Contact: adambrown@fas.harvard.edu

A.2. Introduction

A pressing issue in translational biology is the ability to reference and utilize historical microarray data sets for large-scale discovery programs [197,198]. The appeal of using historical data sets includes capturing previous investment to construct larger cohorts. Despite interest in both industry and academia [197–199], few groups have attempted to tackle the
problem of platform integration. Current approaches primarily rely upon passing different microarray platforms through a common identifier system, such as EntrezGene IDs, using specially designed packages [197,200,201] or online tools [202]. While these systems work well in cases where manufacturers have maintained annotations of their microarray databases, ID-based conversion methods fail for deprecated and under-maintained microarray platforms. Another approach to convert between platforms is sequenced-based, wherein each sequence tag is aligned to the genome or transcriptome and annotated, rather than relying on manufacturer annotations [203,204]. Unfortunately, it is often the case that sequence-based annotations do not capture the complexity of the transcriptome; for example, in the BRCA1 gene, tags mapping to exon 16 would account for three alternatively spliced transcripts with differential impact on oncogenesis [205].

To address the shortcomings of both annotation- and sequence-based conversion methods, we have developed aRrayLasso, a Lasso-regression based network model (Figure A1). Rather than coercing the source and target microarray platforms into a common intermediate, our method directly predicts the probe expression levels of the target platform. To demonstrate the accuracy of our method, we show that predictions made using aRrayLasso are of similar accuracy to technical replicates using the same mRNA pool. Our methodology allows users to utilize currently available methodologies for integrating cross-experiment microarray data sets (e.g. [198]) and allow for the construction of large-cohort retrospective studies.
\[ f_n \left( s_1, \ldots, s_P \right) \]

where \( f_n \) is the \( n^{th} \) Lasso model and \( s_p \) is a source probe.

Figure A1. Schematic of the aRrayLasso algorithm. aRrayLasso takes in an \( M \times N \) target matrix containing \( M \) samples and \( N \) probes. A Lasso model, \( f_n \), is then constructed for each target probe using all probes in the \( M \times P \) source matrix (\( M \) samples, \( P \) probes).

A.3 Array Lasso Implementation

To convert from a source to a target microarray platform, we chose to model each individual sequence tag in the target platform as a linear combination of all sequence tags from the source platform (see Figure A1). Because microarrays have greater than 10,000 individual probes, we chose to use the Lasso algorithm for generalized linear regression [206]. The Lasso algorithm allows the resulting linear model to be 'sparse' in that only the most relevant and robust (by cross-validation) predictors are assigned non-zero values. This optimization allows the model to outperform similar models that require all predictors to be assigned non-zero coefficients [207]. Lasso is implemented in the R package 'glmnet,' allowing for ease of use [208].

We first generate a list of lasso models for each sequence tag in the target microarray platform. Our implementation can take as input a variety of data formats, including expression matrices, R
expressionSet objects, and Gene Expression Omnibus (GEO) accession numbers [209]. Once the full list of models has been computed, we provide functions that allow either the straightforward prediction of sequence tag values or the validation of the model list by calculation of Pearson product-moment correlation coefficients.

To demonstrate the utility of our methodology, we utilized three data sets: 1) GSE6313, containing C57/B6 adult mouse retina cDNA profiles (see [203]) and 2) GSE7785, containing PANC-1 derived cDNA profiles (see Tan et al., 2003), and 3) GSE4854, containing mouse cortex expression profiles (used for cross-experiment testing, see [210]). Each data set is composed of multiple technical replicates for several distinct microarray platforms. For both datasets, we used aRayLasso to first train models to intraconvert between each individual platform and then predicted intraconversions between each pair of platforms for all technical replicates. To assess the accuracy of our conversions, we calculated the average Pearson's r (APR) between the predicted values and actual experimental values for each platform and replicate..

A.4. Results

To explore the performance of aRayLasso, we began by comparing our method's ability to predict expression to the biological variation between replicates on the same platform. We assessed the degree to which aRayLasso could accurately predict platform interconversions in two datasets, representative of different experimental systems (organism versus cell line), organisms (mouse versus human), and platforms (five platforms in total). For the five platforms tested, aRayLasso predictions are within the technical variation of each microarray platform as compared to technical replicates from the same cDNA pool, even when subjected to multiple
sequential conversions (e.g. ABI to Affymetrix to Amersham and back to ABI, data not shown). In addition, once built, aRrayLasso models can be used between experimental conditions: using the models built on GSE6313, we predicted expression levels in GSE4854 with no significant loss of signal (Pearson’s product-moment correlation, p < 0.38). While the results presented here do not guarantee similar results for all training and testing datasets, these analyses serve as a promising proof of concept. Furthermore, our success with a relatively small dataset suggests that aRrayLasso may reach even higher levels of performance as the size of the datasets involved increases.

A.5. Discussion

In this investigation, we propose a data-driven method for integrating across high-throughput genomic measurement modalities that avoids the use of annotation- or sequence alignment-based tools. We have implemented a Lasso Regression-based modeling approach to model the expression level of each sequence tag in a target microarray as a linear combination of all sequence tags in a source microarray. Our implementation represents a straightforward, easy-to-use, and open-source methodology for conversion between microarray platforms.

While we have demonstrated the success of aRrayLasso in a variety of formats and experimental conditions, we acknowledge that there are some limitations to our methodology. The primary drawback of our method is the need for matched samples in the source and target platforms, the lack of which would require additional experimental effort to create such a resource. In our experience, however, there are a large number of datasets available (i.e. those used in this study) that have matched samples with replicates for a number of popular microarray platforms. A second limitation to our method is in conversion between platforms with
highly divergent probe sets, which lack overlap in gene coverage. In these cases, as with currently available methodologies, our method will fail to provide meaningful conversions and investigators should be wary of aRrayLasso models that do not provide similar correlations to technical replicates. Lastly, while we have shown in one case that inter-experiment conversions are feasible, we caution that systematic technical error in a single experiment may lead to the creation of a biased model - mixing training sets from several individual experiments should help prevent such occurrences. However, when coupled with one of several cross-experiment dataset integration tools (e.g. [198]), aRrayLasso will enable mining of the remarkable and untapped historical pool of microarray datasets for large-scale meta studies for well-powered discovery.

A.6. Contributions
aRrayLasso was developed by Adam Brown and Chirag Patel. The work described in this appendix is published in Bioinformatics (© the Authors and reproduced in part here with permission from Bioinformatics), and is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited [211].
APPENDIX B: LICENCES FOR REPRODUCTION OF PUBLISHED WORKS


“This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.”


Image used with permission from Nature Publishing Group (Nature Biotechnology), RightsLink license number 4076631060648 (Mar 26, 2017), for use in this dissertation in print and electronic formats.


“This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0 . Metadata associated with this Data Descriptor is available at http://www.nature.com/sdata/ and is released under the CC0 waiver to maximize reuse.”
APPENDIX C: SUPPLEMENTAL METHODS

Directional DNA methylation dataset preparation

To develop a directional DNA methylation dataset for validation purposes, we used GSE58477 and GSE63409 (as with our non-directional analyses). As previously described, assigning a single direction to differential methylation across all CpGs annotated to a given gene is not possible due to the complex structure of CpG methylation in human genes [87,88]. We therefore made the simplifying assumption that all hypermethylation at the promoter of genes is associated with gene silencing [212–214]. For each gene, we averaged the methylation of all CpGs annotated as within 200 base pairs of the transcription start site (TSS, annotations provided by Illumina) within each sample, and calculated the degree of promoter differential methylation between the normal bone marrow samples and AML blast samples using a two-sample t-test. We then ranked the t-statistics from most positive to most negative (i.e. from most hypermethylated to most hypomethylated). Using this methodology, we condensed 64,688 TSS-proximal CpGs to t-statistics for 17,730 unique genes. We note here, that, while it is unlikely that all genes follow the promoter methylation association exactly, it is necessary to reduce the complexity of gene-level methylation to clearly assign directionality.

Directional CTD dataset preparation

As the compound-gene interaction database for the directional ksRepo, we again used the CTD [47]; in this case, however, we leveraged the “interaction action” annotations given to all compound-gene interactions in the CTD. We took interactions with the action “increases^expression” and “decreases^expression” to indicate up- and down-regulation respectively of the gene. Using this strategy, we obtained a subset of the CTD containing
human-derived interactions between 453 unique drugs and 18,041 unique human genes. Drugs in the ksRepo subset were chosen based on case-insensitive matches between CTD names and names/synonyms for FDA approved drugs downloaded from DrugBank [15].

**Directional, promoter methylation-based ksRepo analysis**

To determine whether the modified KS statistic used by ksRepo, which does not consider directionality, has an effect on the repositioning candidates identified, we developed a version of ksRepo that uses a more standard KS statistic. To do so, we relied on the methods described in the Connectivity Map online tool. Briefly, this involves calculating KS statistics separately for up- and down-regulated genes ($k_{up}^i$ and $k_{down}^i$ respectively), and calculating a “connectivity score,” which is zero where $k_{up}^i$ and $k_{down}^i$ have the same sign, and equal to $k_{up}^i - k_{down}^i$ otherwise (see [37,151] for detailed methods). Because the connectivity score does not have an empirical distribution, we determined significance by resampling the directional DNA methylation datasets, and recalculating each compound’s connectivity score for each resample. We then set the p-value for each compound to be the proportion of resamples that the connectivity score was higher than the observed value (if the observed compound connectivity was positive), and the proportion of resampled connectivity scores lower than the observed value (if the observed compound connectivity was negative). Among compounds with a resampled p-value less than 0.05, we report those with a positive connectivity score: a positive score is indicative of reversing the methylation-induced silencing of a given gene (for hypermethylation) or the aberrant derepression of a gene (for hypomethylation). As before, we performed ksRepo analysis on GSE58477 and GSE63409 separately, taking those compounds that were significant and had positive connectivity scores in both to be replicated.
VS gene expression dataset preparation

The GEO dataset used in this study is GSE39645, an Affymetrix Human Gene 1.0 ST chip-based gene expression study of VS which contained data from 28 patients with sporadic VS, 3 patients with NF2-associated VS, and 8 control nerve samples [215]. Data for GSE39645 was accessed through the NCBI GEO portal and analyzed using the integrated GEO2R tool [33]. As input for GEO2R, we classified each sample within a GEO series as either normal tissue or VS tissue. The GEO2R analysis was performed on both the full dataset (sporadic and NF2 combined), and a subset of samples containing only NF2-syndromic schwannomas.

GEO2R provides a list of probes and corresponding gene symbols ranked according to their degree of differential expression (as calculated using the limma package in R (see [99]), and includes p-values and t-statistics for differential expression. Following GEO2R analysis, all results were imported into R and probe-level differential expression was consolidated to gene-level differential expression using a custom pipeline: t-statistic values were converted to Cohen’s d statistic values and standard error values [216]. Resulting values were combined by gene using a fixed effects meta-analysis (as implemented in the meta.summaries function from the rmeta package in R [217]). Probes without gene annotations were removed from gene-level consolidation. Following consolidation, significantly differentially expressed genes were taken to be those with a Bonferroni-corrected significance less than 0.05.

In parallel, Raw Affymetrix Human Genome U219 gene expression data (.CEL files) for 36 patients with sporadic VS, 13 patients with NF2 syndrome-associated VS, and 7 control nerves controls was obtained from [120]. CEL files were loaded into R using the justRMA function from the affy package in R [218]. justRMA is an automated tool that both performs normalization using the Robust Multi-Array Average method [219] and also automatically annotates all probes
in the normalized dataset using the Org.Hs.eg.db annotation database package [82].
Normalization was performed on the full dataset and the NF2-associated schwannomas, as
above. Mirroring the GEO2R analysis, each normalized dataset was analyzed using limma and
consolidated to gene-level differential expression using the custom pipeline described above. As
above, significantly differentially expressed genes were taken to be those with a
Bonferroni-corrected significance less than 0.05.

Meta-analysis of 80 VS samples and ksRepo prediction
To robustly determine differential expression between VS and normal tissues, gene-level data
from GSE39645 and [120] were meta-analyzed by first removing genes that were not measured
in both the Affymetrix Human Gene 1.0 ST chip and the Affymetrix Human Genome U219 chip,
and subsequently combining Cohen’s d and standard error values using a fixed-effects
meta-analysis (again using meta.summaries). Meta-analysis was performed for the full
GSE39645 and Agnihotri datasets, as well as for NF2-associated tumors exclusively. Following
meta-analysis, the remaining genes were ranked according to their meta-analytic p-values to
generate a gene list for further analysis using ksRepo. For this analysis, the ksRepo built-in
Comparative Toxicogenomics Database (CTD) dataset was selected. For the full dataset
ksRepo analysis and the NF2-only ksRepo analysis, significant compounds were those for
which the FDR was less than 0.05.

VS specimen collection and primary cell culture
Surgical VS and GAN specimens were collected and processed according to protocols
approved by the Human Studies Committee of Massachusetts General Hospital and
Massachusetts Eye and Ear. Written informed consent was obtained from all subjects prior to
inclusion in this study and all procedures were conducted in accordance with the Helsinki Declaration of 1975. Detailed methods for human surgical specimen collection, processing, and culture are previously published. VS specimens were harvested from patients undergoing surgical tumor resection, and GAN specimens from healthy patients undergoing benign parotidectomy or neck dissection surgery, during which the GAN is routinely sacrificed. Patients who had received radiation therapy prior to surgery were excluded.

Briefly, after surgical resection, VS or GAN tissue was immediately placed in saline solution and transported to the laboratory on ice. Specimens were rinsed with Hank’s Balanced Salt Solution (HBSS, ThermoFisher Scientific), dissected to remove burned tissue and blood vessels, and separated for RNA preservation (RNALater, ThermoFisher Scientific) or primary cell culture. After enzymatic dissolution (collagenase type I, 160 U/mL; hyaluronidase type I-S, 250 U/mL) and trituration with an 18-gauge needle, primary cell culture suspensions were plated on 12 mm coverslips pre-coated with poly-D-lysine and laminin (Neuvitro) and grown in Dulbecco’s Modified Eagle’s Medium (DMEM) and F12-containing medium (ThermoFisher Scientific) consisting of 44% DMEM, 44% F12 nutrient mixture, 10% fetal bovine serum (ThermoFisher Scientific), and 1% of a mixture of penicillin and streptomycin (ThermoFisher Scientific). VS and GAN cultures were incubated at 37 degrees Celsius with 5% carbon dioxide, and culture medium was changed every three days. All downstream procedures were performed on primary cell cultures or collected culture medium at two weeks of age in culture to ensure maximal Schwann and schwannoma cell purity [147].

HEI-193 and Arachnoid Cell Culture
HEI-193 cells are derived from a patient with sporadic bilateral vestibular schwannomas and a history of meningioma; these cells express a splice variant of the merlin protein (encoded by the NF2 gene), but neither typical isoform [220]. HEI-193 cells were cultured in DMEM/F12-containing medium with 10% fetal bovine serum and 1% penicillin and streptomycin mix as described above. Immortalized NF2-null and NF2-expressing arachnoid AC-CRISPR cell lines derived from primary human autopsy specimens were obtained via generous gift from Dr. Vijaya Ramesh at Massachusetts General Hospital. NF2-null and NF2-expressing arachnoid cells were cultured in DMEM with 15% fetal bovine serum and 1% penicillin and streptomycin mix. All cell lines were maintained in an incubator at 37 degrees Celsius with 5% carbon dioxide and treated with drugs 24-36 hours after seeding at 15,000 cells per well in 24-well plates. Phase contrast photos of healthy and drug-treated cultures were taken at 10X using a microscope.

**Mifepristone drug preparation and treatment**

Primary VS and GAN cultures were treated with mifepristone (Sigma Aldrich, lot #WXBC0031V) and progesterone (Sigma Aldrich, lot #SLBQ9723V). Fifteen, 25, 35, and 70 μM mifepristone, and 35 μM progesterone were prepared by suspending the appropriate amount of drug (in powder form) in dimethyl sulfoxide (DMSO). The resulting drug suspension was diluted in culture medium to the concentration of interest, and drug-containing medium was applied to primary VS, GAN, and HEI-193 cells, such that the amount of DMSO applied to cells in culture did not exceed 0.1% (24-well plate, 1 mL medium per well). Cultures were incubated with drug-containing medium or 0.1% DMSO vehicle for 72 hours and then processed for downstream applications.
**VS proliferation assay**

5′bromo-2′-deoxyuridine (BrdU) was added to label proliferating cells in culture 2 hours before fixation in 4% formalin (paraformaldehyde). Cell membranes were permeabilized with 10 minutes of incubation in 1% Triton X-100 and nuclear membranes with 20 minutes in 2N hydrochloric acid (HCl). Cells were blocked in 5% normal horse serum (NHS) and 1% Triton X-100 and incubated with a primary antibody against BrdU (#OBT0030G, AbD Serotec) overnight, followed by incubation with fluorescent anti-rat immunoglobulin G (AlexaFluor, Life Technologies). Cells were stained with Hoechst 33342 (Invitrogen) and phalloidin/f-actin (ThermoFisher Scientific) and coverslips mounted on slides with VectaShield (Vector Laboratories). The ratio of BrdU-positive to Hoechst-positive nuclei was determined after visualization using a Leica Epifluorescence microscope (Leica). Manual counts were performed by J.E.S., who was blinded to treatment conditions by receiving and quantifying image files labeled only with arbitrary numbers.

**VS viability assay**

Metabolic viability of primary VS and HEI-193 cells was assessed using the colorimetric 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) assay (Life Technologies). After 72-hour incubation of cells in a 24-well plate with medium containing drug or vehicle, culture medium was replaced with 362 μL of colorless DMEM and 38 μL of 12 mM MTT in 3-6 random wells. Cells were incubated for an additional 4 hours. The resulting formazan crystals were dissolved in 380 μL of a solution of 100 mg/mL sodium dodecyl sulfate in 0.01 M HCl and incubated for another 4 hours. Optical density (OD) at 570 nm of each well was detected using a spectrophotometer. The average OD value of cells exposed to vehicle (0.1% DMSO) was set to
100% and used to normalize OD values of cells treated with drugs; metabolic viability was then reported as percent change.