Harnessing The CRISPR Data Revolution to Uncover The Secrets of Double-Strand DNA Repair

Seale, C.F.

DOI
10.4233/uuid:b2c6cef6-3a79-4d83-940b-3705d1782ee9

Publication date
2025

Document Version
Final published version

Citation (APA)
Seale, C. F. (2025). *Harnessing The CRISPR Data Revolution to Uncover The Secrets of Double-Strand DNA Repair*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:b2c6cef6-3a79-4d83-940b-3705d1782ee9

Harnessing the CRISPR Data Revolution to Uncover the Secrets of Double-Strand DNA Repair

Colm Fintan Seale

Harnessing the CRISPR Data Revolution to Uncover the Secrets of Double-Strand DNA Repair

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board of Doctorates,
to be defended publicly on
Thursday 30th October 2025 at 10:00 hours

by

Colm Fintan SEALE

Master of Science in Computer Science,
Delft University of Technology, Delft,
born in Laois, Ireland.

This dissertation has been approved by the promoters.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T Reinders,	Delft University of Technology, *promotor*
Dr. J.S. Pinho de Gonçalves,	Delft University of Technology, *copromotor*
Independent members:	
Prof. dr. L.F.A. Wessels,	Delft University of Technology
Prof. dr. ir. A.H.C. van Kampen,	University of Amsterdam
Prof. dr. ir. D. de Ridder,	Wageningen University and Research
Dr. H.J.G. van de Werken,	Erasmus University Medical Center
Prof. dr. M. Tijsterman,	Leiden University Medical Center
Prof. dr. ir. R.L. Lagendijk,	Delft University of Technology, *reserve member*

Printed by Ridderprint | https://www.ridderprint.nl
Cover designed by Eric Lardenoije | info@lardenoije.art

ISBN: 978-94-6518-105-9

An electronic version of this dissertation is available at
http://repository.tudelft.nl

To my parents, Kathleen and Gerry,
you are my foundation and my inspiration,
and you are sorely missed.

Contents

Summary

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technology has transformed molecular biology by enabling a strategy for precise, efficient, and relatively simple **genome editing**. Guided by a small strand of RNA, CRISPR locates specific DNA sequences within the genome and introduces **double-strand breaks** (DSBs). A typical cell can detect and fix the damage by invoking one of several **DNA repair pathways**. However, repair is not error-free and often introduces **mutations**. The mutagenic nature of repair pathways can be leveraged to disrupt **genes** or regulatory elements with high specificity, providing a powerful tool for gaining insights into gene function. Researchers can also generate **datasets of mutations** left behind after DSB induction and repair within different **genomic contexts** to learn more about the mutagenic effects of DNA repair. In this thesis, we explore challenges and novel approaches for analysing **large-scale** datasets of mutations and gene essentiality generated via CRISPR technology.

We begin by addressing ways to improve the efficacy of **template-free CRISPR editing**, a gene-editing approach that leverages the more prevalent cellular repair pathways, such as non-homologous end joining (NHEJ) or microhomology-mediated end joining (MMEJ), to introduce precise edits at pre-programmed CRISPR DSB-induced loci. While this approach has potential to induce precise sequence-dependent insertions or deletions, its inherent stochasticity complicates the prediction of mutational outcomes, creating uncertainty during experimental design. To overcome this, we propose a model that **predicts CRISPR-induced mutational outcomes**. Moreover, generating the datasets to train such models is expensive, so currently these datasets only exist for a limited set of genomic contexts. Thus, we further demonstrate how our model can utilise **transfer learning** to improve generalisability to new genomic contexts, especially in data-scarce domains.

Next, we investigate the use of large-scale CRISPR datasets to **elucidate DNA DSB repair mechanisms**. We analyze frequency distributions of **CRISPR repair outcomes** collected under individual **gene knockout** conditions and describe our two main challenges. First, we develop a method to identify genes whose knockouts cause significant deviations from expected wild-type repair distributions estimated without the need for **experimental controls**. Second, we propose a clustering method to group gene knockouts based on their repair outcome profiles, revealing potential **gene functions** and **membership** of DNA repair pathways. Based on this analysis, we identify and recommend several high-impact candidate repair genes for further experimental validation.

Finally, we shift focus to **genome-wide CRISPR functional screens**, which evaluate the impact of gene knockouts on cell survival. Using these datasets, we propose a computational framework to predict **synthetic lethal** interactions (SL) between genes, a concept with therapeutic implications, particularly in cancer treatment. Unlike earlier sections focused on DNA repair pathways, this approach extends to broader cellular contexts. We address a major challenge in SL prediction – **selection bias** – by proposing strategies to enhance the reliability and applicability of predictive models, making them more effective for identifying therapeutic targets.

Through this research, we make novel contributions to multiple fields that utilize large-scale CRISPR datasets, showcasing the power of this technology to uncover new biological insights while addressing current key challenges, oversights, and limitations in its application.

Samenvatting

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technologie heeft de moleculaire biologie veranderd door een strategie mogelijk te maken voor precieze, efficiënte en relatief eenvoudige genoombewerking. Geleid door een kleine RNA-streng, lokaliseert CRISPR specifieke DNA-sequenties in het genoom en introduceert **double-strand breaks** (DSBs). Een typische cel kan de schade detecteren en herstellen door een beroep te doen op een van de verschillende **DNA-reparatie mechanisme**. Reparatie is echter niet foutloos en introduceert vaak mutaties. De mutagene aard van reparatie mechanisme kan gebruikt worden om genen of regulerende elementen zeer specifiek te verstoren, wat een krachtig hulpmiddel is om inzicht te krijgen in de functie van genen. Onderzoekers kunnen ook datasets van mutaties genereren die zijn achtergebleven na DSB inductie en reparatie binnen verschillende **genomische contexten** om meer te leren over de mutagene effecten van DNA reparatie. In dit proefschrift verkennen we uitdagingen en nieuwe benaderingen voor het analyseren van grootschalige datasets van mutaties en genessentialiteit gegenereerd met CRISPR-technologie.

We beginnen met het onderzoeken van manieren om de effectiviteit van **template-free CRISPR editing** te verbeteren, een gen-editing aanpak die gebruik maakt van de meer gangbare cellulaire reparatie mechanisme, zoals non-homologous end joining (NHEJ) of microhomology-mediated end joining (MMEJ), om precieze edits aan te brengen op voorgeprogrammeerde CRISPR DSB-geïnduceerde loci. Hoewel deze aanpak het potentieel heeft om precieze volgorde-afhankelijke inserties of deleties te induceren, bemoeilijkt de inherente stochasticiteit het voorspellen van mutatieresultaten, en leidt tot onzekerheid bij het experimentele ontwerp. Om dit te verhelpen stellen we een model voor dat CRISPR-geïnduceerde mutatieresultaten voorspelt. Bovendien is het genereren van datasets die nodig zijn om zulke modellen te trainen duur, dus momenteel bestaan deze datasets alleen voor een beperkte set van genomische contexten. Daarom laten we verder zien

hoe ons model **transfer learning** kan benutten om de generaliseerbaarheid naar nieuwe genomische contexten te verbeteren, vooral in domeinen met weinig gegevens.

Vervolgens onderzoeken we hoe grootschalige CRISPR datasets gebruikt kunnen worden om DNA DSB reparatiemechanismen te doorgronden. We analyseren frequentieverdelingen van **CRISPR reparatie uitkomsten** verzameld onder afzonderlijke **gen knock-out** condities en beschrijven onze twee belangrijkste uitdagingen. Ten eerste ontwikkelen we een methode om genen te identificeren waarvan de knock-outs significante afwijkingen veroorzaken ten opzichte van de verwachte wild-type reparatiedistributies, geschat zonder de behoefte aan **experimentele controles**. Ten tweede introduceren we een clustermethode om gen-knockouts te groeperen op basis van hun reparatie-uitkomstprofielen, waardoor potentiële functies van genen en betrokkenheid bij DNA-reparatie mechanismes aan het licht komen. Op basis van deze analyse identificeren en bevelen we meerdere cruciale potentiële herstelgenen aan voor verdere experimentele validatie.

Ten slotte richten we ons op genoombrede functionele CRISPR screening, die de impact van gen knock-outs op de overleving van cellen evalueren. Met behulp van deze datasets stellen we een computationeel raamwerk voor om **synthetic lethal** interacties (SL) tussen genen te voorspellen, een concept met therapeutische implicaties, vooral in de context van kankerbehandeling. In tegenstelling tot eerdere secties die zich richtten op DNA reparatie mechanismes, strekt deze aanpak zich uit tot bredere cellulaire contexten. We adresseren een belangrijke uitdaging in de voorspelling van SL-interacties door strategieën te introduceren om de betrouwbaarheid en toepasbaarheid van voorspellende modellen te verbeteren, zodat ze effectiever worden in het identificeren van therapeutische doelen.

Door dit onderzoek doen we nieuwe bijdragen aan meerdere vakgebieden die gebruik maken van grootschalige CRISPR datasets, waarbij we de kracht van deze technologie demonstreren om nieuwe biologische inzichten te ontdekken terwijl we de huidige belangrijkste uitdagingen, tekortkomingen en beperkingen in de toepassing ervan aanpakken.

Achoimre

Tá teicneolaíocht CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) tar éis réabhlóid a dhéanamh i mbitheolaíocht móilíneach trí stráitéis a chur ar fáil do **modhnú géanóim** atá cruinn, éifeachtúil agus measartha simplí. Treoraithe ag ribe beag RNA, aimsíonn CRISPR seiceamh shonrach DNA sa ghéanóim agus tugann sé isteach **briseadh déribe** (DSB). Is féidir le gnáth cheall an damáiste a aimsiú agus a chóiriú trí ceann de na **conair deisiúcháin DNA** éagsúla a chur i bhfeidhm. Mar sin féin, níl deisiúcháin saor ó earráidí agus is minic a chruthaíthear **socháin** dá bharr. Is féidir an nádúr só-ghineach a bhaineann leis na cosáin deisiúcháin a úsáid chun **géinte** nó eilimintí rialaitheacha le sainghnéithe arda, ag cur uirlis cumhachtach ar fáil do thaighdeoirí a thugann léargas níos fearr ar fheidhmiú géine. Is féidir le taighdeoirí **tacair sonraí ar shocháin** a fágadh ina diaidh tar éis indochtú agus deisiúcháin DSB laistigh de **chomhthéacsanna géanómaíoch** a chruthú chun níos mó a fhoghlaim faoi éifeachtaí só-ghineach a bhaineann le deisiúcháin DNA. Sa tráchtas seo, déantar plé ar dhúshláin agus ar cur chuige nua chun anailís a dhéanamh ar thacar sonraí **ar scála mór** ar shocháin agus ar nádúr géine a chruthaítear trí theicneolaíocht CRISPR.

Tosaímid trí aghaidh a thabhairt ar bealaí chun éifeachtúlacht **eagarthóireacht CRISPR gan teimpléad** a fheabhsú. Is cur chuige eagarthóireacht géine é seo a bhaineann leas as na conair deisiúcháin is láidre, ar nós ceangal críoch neamh-homalógach (NHEJ) nó ceangal críoch trí mheán na micrea-homalógachta (MMEJ), chun eagarthóireacht cruinn a thabhairt isteach ag láithreacha réamh-ríomhchláraithe. Cé go bhfuil póiteansail ag an gcur chuige seo ionsánna nó scriosanna seiceamh-spleach cruinn a chruthú, is féidir lena nádúr stocastach na tuartha ar thorthaí sóchánúla a dhéanamh níos casta, ag cruthú éiginnteacht le linn dearadh trialach. Chun é seo a shárú, molaimid múnla a **thuarann torthaí sochána ionduchtaithe ag CRISPR**. Ina theannta sin, tá sé costasach na tacair sonraí don cur chuige seo a thraenáil, mar sin, níl na tacair sonraí seo ar fáil ach do thacar géanómach teoranta faoi láthair.

Mar sin, léirímid conas gur féidir lenár gcur chuige **foghlaim thraschurtha** a úsáid chun inghinearálaitheacht do chomhthéacsanna géanómach a fheabhsú, go mór mór i réimsí atá gann ó thaobh sonraí de.

Ina dhiaidh sin, déanfaidh muid fiosrúchán ar úsáid thacair sonraí CRISPR ar scála mór chun **modhanna deisiúcháin DNA DSB a shoiléirú**. Déanfaidh muid anailís ar dháileacháin mhinicíochta **thorthaí deisiúcháin CRISPR** bailithe faoi choinníollacha **asleagan géine** aonaracha agus déanfaidh muid cur síos ar ár dhá phríomh dhúshláin. I dtosach báire, forbraímid modh chun géine a aithint a bhfuil a gcuid asleagan ag cruthú athruithe suntasacha ó dáileacháin fiáin a mheastar gan gá le **rialaitheoirí turgnamhacha**. Ar an dara dul síos, molaímid modh bailiúcháin chun asleagan géine a chur i ngrúpaí bunaithe ar a bpróifilí torthaí deisiúcháin, ag nochtadh **feidhmeanna géine** féideartha agus **ballraíochtaí** chonair deisiúcháin DNA. Bunaithe ar an anailís seo, aithnímid agus molaímid roinnt géinte deisiúcháin a d'fhéadfadh a bheith tábhachtach do bhailíochtú turgnamhach breise.

Ar deiridh, aistrímid ár bhfócas chuig **scagthástálacha feidhmiúla CRISPR ar fud an ghéanóim** a dhéanann measunú ar an tionchar atá ag asleagan géine ar mharthanacht cille. Ag baint úsáid as na tacair sonraí seo, molaímid creatlach ríomhaireachtúil chun caidrimh **sintéiseach marfach** (SL) idir géinte a mheas, coincheap le himpleachtaí teiripeacha, go mór mór ó thaobh cóir leighis le haghaidh ailse de. Difriúil ó na cuideanna roimhe seo a bhí dírithe ar chonair deisiúcháin DNA, clúdaíonn an cur chuige seo comhthéacsanna ceallach níos leathan. Tugann muid aghaidh ar dhúshlán mór i dtuar SL – **laofacht roghnúcháin** – trí stráitéisí a mholadh chun iontaofacht agus infheidhmeacht do mhúnlaí tuarthacha a fheabhsú, á dhéanamh níos éifeachtach chun spriocanna teiripeacha a aithint.

Tríd an taighde seo, tugaímid ionchur nua do réimse iomadúil a bhaineann úsáid as tacair sonraí CRISPR ar scála mór, ag taispeáint cumhacht an teicneolaíocht seo chun tuiscintí bitheolaíochta nua a aimsiú agus ag an am céanna, ag tabhairt faoi dúshláin, dearmaid agus srianta tábhachtacha reatha le linn a úsáid.

1

Introduction

> *The fourth R of gene physiology, essential to both the survival and mutability of organisms, might be "repair."*
>
> **— Siddhartha Mukherjee**
> (The Gene: An Intimate History)

Deoxyribonucleic acid (DNA) is a special molecule at the centre of all life as it exists today. DNA consists of four nucleotides – adenine, cytosine, guanine, and thymine – and from these four nucleotides stems all of the diversity and wonder we witness in the biological world, from flowers to trees, insects, birds, fish, and more. But how does DNA achieve this? DNA is often considered a sort of language, and when arranged into a full genome, it describes the entire set of instructions for cellular function and, thus, for life. These instructions are passed from parent to child, from old cells to new, from generation to generation. They may shift and change and mutate both within and between generations, but core genetic elements must remain to retain cell viability. The central dogma of molecular biology states that genetic information flows in one direction: DNA is transcribed into ribonucleic acids (RNA), some of which are translated into proteins, the machines that fuel life. This means that DNA is the only method by which the instructions for life are preserved. Therefore, it should be quite apparent to the reader that protecting this information is of utmost concern to maintaining the very existence of life itself.

There is much we understand about how DNA functions, "The Three R's" as described by Mukherjee: how it (r)eplicates itself to produce new copies for new cells; how it (r)ecombines to facilitate the exchange of genetic information, promoting diversity and providing an engine for evolution to occur; and how it (r)egulates itself within a cell to maintain homeostasis, to respond to its

environment, and to differentiate into various cell types. Mukherjee alludes to the presence of a fourth "R", upon which this dissertation is focused - (r)epair: the ability of the genome to self-correct damage to the DNA structure and maintain its integrity and stability, generation after generation. Specifically, this dissertation focuses on the repair of a particular type of damage to DNA where its sequence has been split into two - otherwise called a double-strand break (DSB).

This chapter aims to introduce the basic concepts of DNA, DSBs, and the need for repair, and to familiarise the reader with the current state of scientific knowledge surrounding the topic. We discuss some of the latest research developments and describe how they produce vast quantities of data for the community, what opportunities have arisen from these data, and the challenges alongside them. Finally, we present the contributions made by this thesis to advance research in the field.

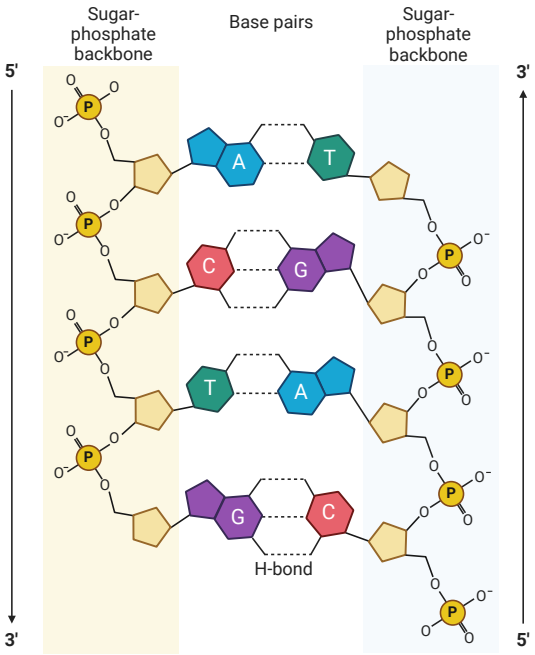


Fig. 1.1: Illustration of the building blocks of DNA. The DNA double-helix structure consists of two strands running in opposite directions. Each consists of a sugar-phosphate backbone and nucleotide bases that pair specifically across strands to form ladder-like rungs. Created with BioRender.com.

1.1 DNA - the language of life

DNA is the fundamental molecule that encodes the genetic instructions required for all living organisms to grow, develop, function, reproduce, and evolve. Structurally, DNA is a double-stranded helix comprised of four nucleotide bases: adenine (A), cytosine (C), guanine (G), and thymine (T) (Fig. 1.1). These bases form specific pairings (A with T and C with G) [26]. In natural language terminology, we can think of these bases as the "letters" of the language of DNA. These letters can be arranged in a sequence which constitutes the genetic code. It can be read in sets of three (called "codons") to synthesize proteins, the building blocks and functional units of cells [28]. A **gene** is a DNA segment containing instructions for building proteins. If the bases are the letters, we might think of the codons as words and genes as sentences. The full body of text is called the **genome** and represents the entirety of an organism's DNA, including all genes and non-coding regions.

1.2 Breaks in DNA - causes and consequences

Throughout the lifetime of the cell, the DNA stored within that cell can experience different forms of genetic damage. One of the most severe forms of damage to DNA is a **double-strand break** (DSB), which occurs when both strands of the DNA double helix are broken, either directly opposite each or within a few nucleotides apart. The average human somatic cell suffers roughly 10-50 DSB events a day [24]. DSBs can be endogenously induced during cellular processes such as meiosis, V(D)J recombination, or as a byproduct of processes like DNA replication, or caused by factors external to the cell like ionizing radiation or certain chemicals [7].

The presence of unrepaired DSBs can activate DNA damage response (DDR) pathways, resulting in cell cycle arrest or apoptosis (programmed cell death) in the case of excessive damage [7]. To return to our earlier language metaphor, if the genome is the full body of text representing all the instructions for life, then tearing the book into two severely hampers the ability of the cell to continue to read from this text. Cells do not have memory, so without instructions, they will struggle to produce the necessary components they need to survive (see Fig 1.2).

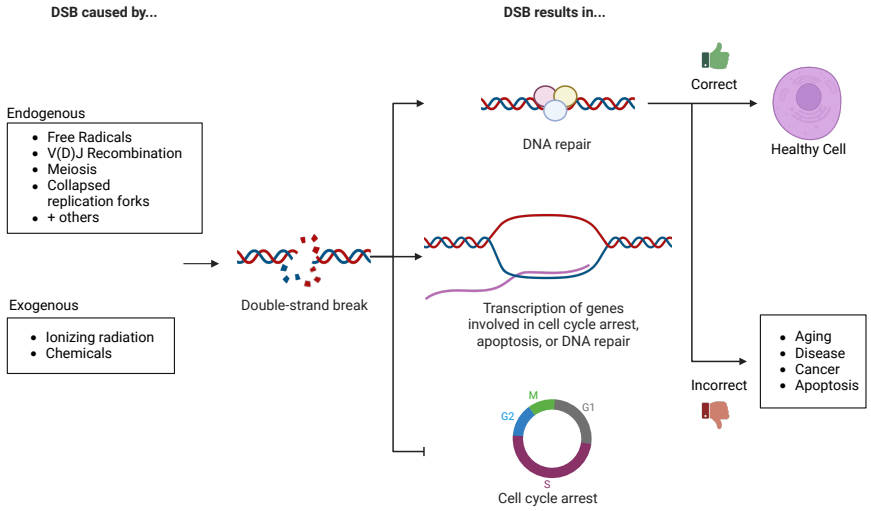


Fig. 1.2: Causes and consequences of double-strand breaks (DSBs). DSBs can be caused by various stressors, leading to cell cycle arrest, transcription and activation of the DNA damage response, and initiation of DNA repair mechanisms. If DSBs are not properly repaired or remain unresolved, they can trigger apoptosis, premature aging, genetic disorders, and potentially contribute to the development of cancer. Adapted from [15]. Created with BioRender.com.

Considering the dire consequences that DSBs can exert on cellular function, cells have evolved numerous mechanisms to repair and recover from such events. Yet, these repair mechanisms (described in more detail later) can also result in cellular abnormalities. Erroneous DSB repair can lead to chromosomal aberrations, such as translocations, inversions, or deletions, which are associated with various genetic diseases, including the development of cancer. These problems can become more pronounced if the repair machinery of the cell is defective. Thus, while DSBs are common, their uncontrolled occurrence or improper repair has profound implications for organismal health and development, and must be treated.

1.3 Repairing breaks - what we know and why we care

The initial discoveries that demonstrated that the genome had encoded within it the ability to recognise and repair damage to its own DNA were made by Evelyn Witkin and Steve Elledge. Witkin discovered the bacterial SOS response, a global regulatory network that activates DNA repair mechanisms

when cells experience extensive damage to their DNA [27]. Working independently, Steve Elledge focused on DNA damage response in eukaryotic cells, identified key protein interactions and revealed how cells detect, signal, and repair DNA damage [10]. In the time since these pivotal works, the scientific community has greatly expanded its understanding of the cell's capabilities and limitations for sensing and repairing DNA damage. We know now that the cell has multiple **pathways** – networks of proteins and other molecules that interact with one another in a cascading fashion – which can be called upon to sense and fix irregularities in the DNA. These pathways can be separated into two main modes of function: **homologous recombination** (HR) and **non-homologous end joining** (NHEJ) [19] (Fig. 1.3).

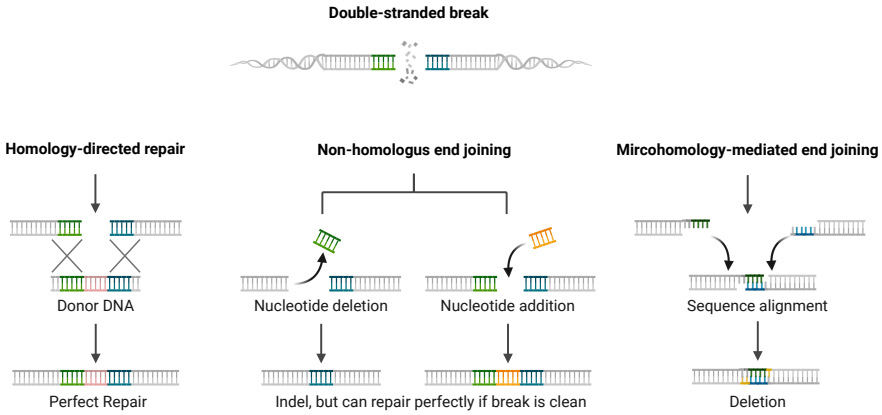


Fig. 1.3: DSB repair pathways. Homologous recombination (HR): a precise DNA repair process that uses a homologous template, usually the sister chromatid, to accurately restore the original DNA sequence. Non-homologous end joining (NHEJ): error-prone DNA repair mechanism that directly ligates broken DNA ends back together. Mirochomology-mediated end joining (MMEJ): DNA repair pathway that uses short homologous sequences near the break to join DNA ends, often resulting in small deletions. Created with BioRender.com.

Homologous recombination is a highly accurate repair mechanism that uses homologous sequences as templates. This process typically occurs during the S and G2 phases of the cell cycle, when the sister chromatid is readily available to serve as a repair template to restore the original DNA sequence. NHEJ is the preferred repair pathway during other phases and can be subdivided into two separate pathways: **classical-NHEJ** (cNHEJ) and **alternative-NHEJ** (alt-NHEJ). In cNHEJ, the broken ends are directly ligated together, often resulting in small indels (insertions or deletions of nucleotides) at the break site [22]. In alt-NHEJ, also called **microhomology-mediated end joining**

(MMEJ), opposing ends of the DSB are resected along opposite strands, after which small regions of homology on either side of the break are used to align and repair the broken DNA, resulting in a deletion of nucleotides [20]. Research has explored many corridors of the labyrinth of interactions that constitute the entire DSB repair process [19]. Yet, there are still avenues left to explore in order to fully map out these systems. One specific interest is to understand how the DNA sequence itself affects the choice of which pathways to use and how these pathways select for the mutations that they leave behind [14, 17].

Beyond the noble and ongoing pursuit of understanding the world around us, there are immediate and practical benefits to uncovering the mechanisms of DNA double-strand break (DSB) repair. As mentioned above, faulty DSB repair can lead to the accumulation of mutations within the genome, increasing the risk of cancer [4]. In fact, defective DSB repair is a hallmark of many cancer cells [23, 4]. While these faulty mechanisms contribute to cancer development, they can also be exploited for cancer therapy. Treatments like chemotherapy and radiation therapy work by inducing DSBs across the genome [23]. Healthy cells with intact repair systems can better recover from this damage. In contrast, cancer cells with impaired DSB repair are less likely to recover, leading to their selective elimination.

Additionally, drug therapies such as PARP inhibitors further weaken a cancer cell's ability to repair DNA damage, making them more vulnerable to chemotherapy or radiation [8]. Treatments like PARP inhibitors can reduce the need for, or exposure to, radiation therapies, which are notorious for their toxicity and the toll they take on a patient's overall health and quality of life [8]. By advancing our understanding of DNA repair, we can expand the range of personalized treatment options available to clinicians [11]. This could reduce the exposure of a patient to highly toxic therapies like chemotherapy, improving both recovery time and quality of life.

One method to further DNA repair research is by profiling the mutations left behind by the different repair pathways and reverse engineering their causal mechanisms. This can be done in two ways: by randomly inducing DNA breaks across the genome, such as with ionizing radiation, followed by whole genome sequencing to capture mutations; or by creating breaks at specific sites with precision tools, and then sequencing only the targeted regions. The random approach captures mutations at multiple sites but makes it difficult

to pinpoint break locations, making these approaches useful for studying what mutations the repair pathways leave behind across the genome if we are not interested in the local sequence context surrounding each break. The targeted approach allows precise identification of break sites, allowing the investigation of how local sequence context influences the mutations that occur, but introduces bias through non-random sequence selection. This dissertation utilises the second approach, employing **CRISPR** technology to induce DSBs at targeted locations.

1.4 CRISPR - Pushing the boundaries of DNA repair research

CRISPR-Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats) is one of the latest technological advances that has revolutionised modern molecular biology research efforts [13]. CRISPR offers a cost-effective, efficient, and precise method to induce double-strand breaks (DSBs) at specific, targeted locations within the genome [5, 16, 25, 12, 2]. The key to this targeting ability lies in the **single-guide RNA** (sgRNA), which contains a short programmable nucleotide sequence designed to be complementary to the region of DNA that researchers wish to cut. The sgRNA directs the Cas9 protein to this complementary sequence, where Cas9 functions like molecular scissors to create a DSB at the targeted site (Fig. 1.4). Once the break occurs, the cell's natural DNA repair mechanisms – such as HR or NHEJ – are triggered. This targeted approach allows researchers to study how cells respond to DNA damage, which repair pathways are activated, and how the chosen genomic location influences the outcome of DSB repair. CRISPR technology has brought upon us the advent of new data types we can use to advance our understanding of DNA repair and beyond.

1.4.1 CRISPR knockout screening

CRISPR enables the creation of gene **knockouts** with ease [5, 16, 25, 12, 2]. When the cell repairs the DSB, the genome may either be restored to its original sequence or modified with small insertions or deletions. If such alterations occur within a coding region, such as a gene, they can effectively "knockout" that gene, allowing researchers to generate cells with altered genetic backgrounds. This ability to create gene knockouts is especially valuable for studying the roles of genes in processes like DNA repair or for

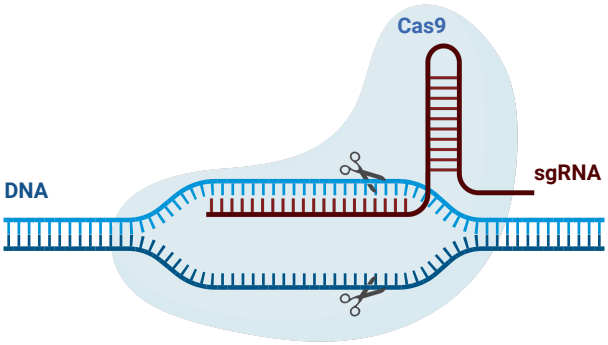


Fig. 1.4: The CRISPR-Cas9 genetic scissors. When researchers edit a genome using CRISPR-Cas9, they design a single guide RNA (sgRNA) that matches the specific DNA sequence where the cut is intended. The Cas9 protein then forms a complex with the sgRNA, guiding it to the exact location in the genome to cut the DNA at the targeted site. Adapted from [18]. Created with BioRender.com.

exploring how pathway dynamics shift when specific genes are altered or disabled. Furthermore, CRISPR knockout screens can be performed at a large scale, enabling the generation of thousands of single-gene knockouts across a population of cells to systematically assess gene function on a genome-wide scale. Compared to other technologies such as shRNA (short hairpin RNA) or siRNA (small interfering RNA), which temporarily reduce gene expression ("knockdown"), CRISPR creates permanent and complete knockouts, offering greater precision, cleaner results, and more consistent outcomes in functional studies [6].

1.4.2 CRISPR repair outcomes

With the advent of CRISPR-Cas9 technology and programmable target locations for the induction of DSBs has come a new type of data for researchers to exploit: CRISPR repair outcomes. As discussed, after the Cas9 protein cleaves the DNA in a cell, repair pathways kick into action and work to fix the lesion. Often, the post-repair DNA product is altered due to error-*prone* repair pathway activity, such as NHEJ or MMEJ. Small-scale research has shown that the DNA products left behind post-repair are non-random, as they are influenced by the state of the cellular repair mechanisms and by the sequence context surrounding the cleavage site [14, 17]. These findings have created new possibilities: by studying these repair outcomes, we get another

window into the world of interactions among DSB repair factors and how they influence DSB repair (Fig. 1.5).

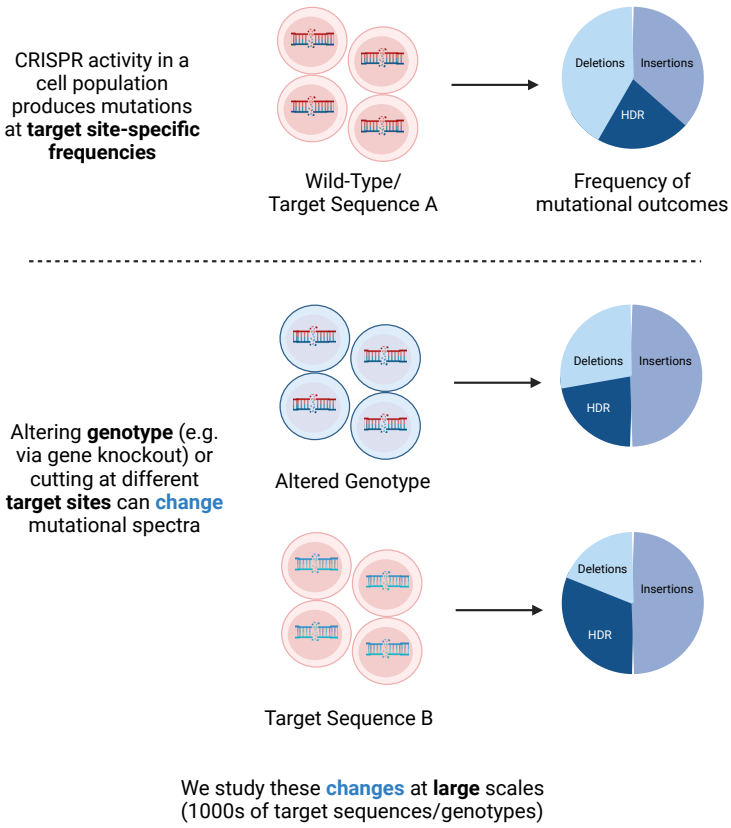


Fig. 1.5: Studying DSB Repair via CRISPR repair outcomes. CRISPR-Cas9 induces a DSB in DNA, which, when repaired, produces a sequence context-specific distribution of mutational outcomes across a population of cells. Altering the cell state or targeting a different sequence can change the distribution. Large-scale studies of these changes may inform us about the underlying DSB repair mechanisms. Created with BioRender.com.

Furthermore, CRISPR-Cas9 technology itself allows for the cost-effective and efficient induction of DSBs at thousands of target sites simultaneously. These advances have allowed researchers to produce large-scale datasets analysing how the repair outcomes are influenced by thousands of different sequence contexts [21, 3, 9], or by the knockout of hundreds of individual genes along one or multiple repair pathways [1] (Fig. 1.5).

A CRISPR repair outcome dataset typically has three main dimensions: the **mutational outcome**, the **target sequence**, and the **genotype** (Fig. 1.6). The mutational outcomes are the unique products that can be produced (or have been observed) post-repair. The outcomes are always highly dimensional since cleavage at a single target site can typically produce approximately 450 to 550 unique mutational outcomes. The target sequence describes the sequence context surrounding the DSB used to generate the mutational outcomes. Mutational outcomes are typically unique to a particular target site, so comparing across target sites may require some form of aggregation across outcomes to make them comparable between sites. The genotype describes the specific genetic makeup of the cell within which the outcome was observed (i.e. the organism, the tissue, whether the cell is wild-type or has any genes knocked out, and so on). Existing datasets are typically low-dimensional in either the gene knockout, target sequence, or both dimensions, and no datasets exist with high dimensionality along both the gene knockout and sequence context dimensions (see Table 1.1 for a summary of current datasets).

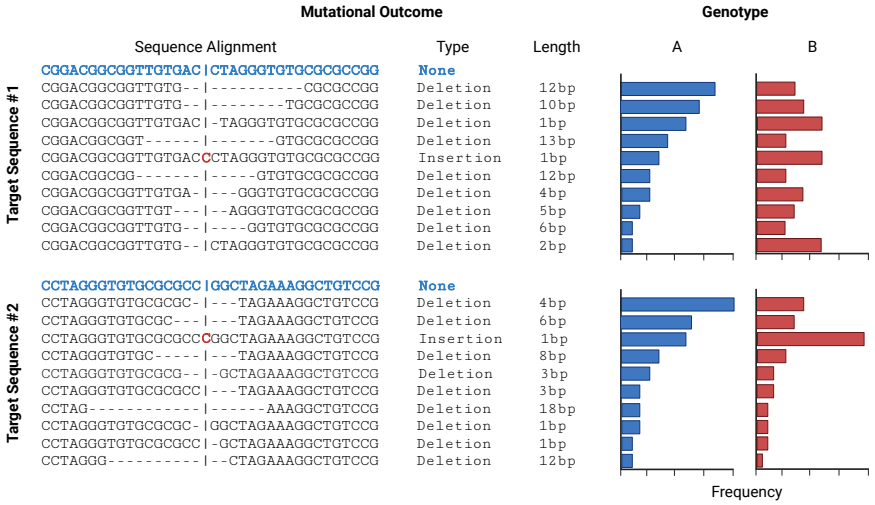


Fig. 1.6: Illustrative Example of CRISPR Repair Outcome Dataset. Each dataset usually has 3 dimensions: target sequence, mutational outcomes, and genotype. The mutational outcomes are dependent on the target sequence. A different genotype (different organism, tissue, or other modification such as a knockout) can produce different frequencies for the mutational outcomes.

Genotype Sequence Context	Low dimensionality	High dimensionality
Low dimensionality	[14, 17]	[17]
High dimensionality	[21, 3, 9]	None

Tab. 1.1: Breakdown of the types of CRISPR outcome datasets present in the current literature.

1.4.3 Computational challenges of using CRISPR data in research

While these new CRISPR-generated datasets offer exciting opportunities to explore previously ambitious questions, they also present new challenges, including:

Scarcity of data for most genomic contexts. While we see a growing number of available CRISPR datasets, it is still relatively tiny when compared to the vast number of possible genomic contexts. For instance, different organisms such as humans, mice, fish, plants, or different tissue types within an organism, such as lung, heart, liver, and so on. In reality, only a small number of cell line models with specific contexts will get large datasets created for them, as these require considerable resources to produce. Therefore, one of the challenges here is: how do we generalise the findings or results we identify in one dataset, in one specific genomic context, to be applicable in other contexts?

Hundreds of unique outcomes per target site. When analyzing CRISPR repair outcome data, each target site presents a partially unique set of hundreds of potential repair outcomes. Even for repair outcomes shared between different sites, their relative frequencies vary in a target site-dependent. This variability complicates cross-site comparisons. Simple approaches to this problem have employed some form of categorization that, in turn, challenges the interpretability of models and results. Therefore, the challenge here is to develop a strategy to handle the thousands of possible outcomes in a ubiquitous manner across target sites while maintaining interpretability.

Rarer outcomes are sparse. The frequency of CRISPR repair outcomes observed at any given target site is not uniformly distributed across all possible outcomes. In wild-type cells, the repair pathways often produce a small number of frequently occurring outcomes. Current studies tend to emphasise

these common outcomes. Yet, it also happens that these frequently occurring outcomes result from pathways or routes through these pathways that are already relatively well-characterised. Therefore, some of the potentially more interesting information or areas where discoveries may reside are within the rarely occurring outcomes. This introduces a few problems: sequencing must be performed with deep enough coverage to accurately capture the frequency distributions of rare outcomes. This coverage is limited by the resources of the labs conducting these experiments. Therefore, data for rarer outcomes can be sparse or noisy, introducing complexity in how to extract useful information from these sets of outcomes.

Pathway interactions confound results. When analysing repair outcome data, we are aware that multiple repair pathways are, in a sense, competing to perform repair. Furthermore, different repair pathways can also collaborate or independently produce the same post-repair mutational outcome. Therefore, when looking at the final distribution of observed outcomes, it is difficult to delineate what pathways, and to what extent, are responsible for producing each of the repair outcomes. Some genes also function in more than one pathway, complicating the manner of determining causality further. This can make interpreting CRISPR outcome data difficult along either the mutational outcome or genotype dimension, and analysing the marginal effects along one dimension only may result in missing important interactions. Therefore, the challenge here is to attempt to deconvolve these mixtures of signals coming from multiple pathways to allow for a better understanding of the genes and pathways responsible for producing these mutational patterns.

1.5 Thesis outline and contributions

In this dissertation, we begin by introducing a model capable of predicting the frequency of occurrence of repair outcomes for any given target sequence. We introduce new features which are ubiquitous to different categories of repair outcomes, simplifying the best-in-class architecture for this problem while allowing for better interpretability of the model. Furthermore, we demonstrate how transfer learning may be used to tackle the issue of data scarcity in the genomic contexts for the available CRISPR outcome datasets. Next, we introduce a ranking algorithm for mutational spectra that quantifies their deviation from the central distributional trend without the need for controls.. Then, we present an approach to analyse CRISPR repair outcome

data to identify and functionally characterize the "fingerprints" or "signatures" left behind by the DNA repair pathways at given target sites. We use these signatures to quantify repair pathway activity through association with known genes, intending to uncover novel functions for candidate genes. Finally, we shift focus from DSB repair-focused work to focus on improving synthetic lethality (SL) prediction. SL describes a relationship between genes where the simultaneous loss or mutation of two genes leads to cell death, while a defect in only one is not lethal. We develop models using features engineered from CRISPR gene dependency screens (among others) to predict SL interactions between genes, which can also be applied to DSB repair genes.

1.5.1 X-CRISP: Interpretable and Domain-Adaptable CRISPR Repair Outcome Prediction

Precise CRISPR-based gene editing requires control over repair outcomes. Since donor template-based editing is often inefficient, researchers have sought to use DSB repair pathways that do not rely on a template to achieve the desired results. Machine learning models have been developed to predict the distribution of repair outcomes based on target sequences, but generalizability remains a challenge—how well these models perform in genomic contexts beyond the original training cell line is unclear. Additionally, current top-performing models suffer from limited interpretability due to suboptimal feature representations and model architectures. Chapter 2 introduces X-CRISP (eXplainable CRISPR PRedictions), a more interpretable and adaptable machine learning model designed to predict CRISPR-based DNA repair outcomes. It uses a unified encoding of sequence features to improve accuracy, explainability, and transferability to new, data-scarce domains. The model is evaluated across multiple datasets and compared to other leading models, with results demonstrating the benefits of transfer learning in improving predictions in data-scarce domains.

1.5.2 MUSICiAN: Detecting Gene-DNA Repair Associations via Control-Free Mutational Spectra Analysis

Enhancing our understanding of DNA double-strand break (DSB) repair mechanisms is crucial for understanding and treating diseases like cancer. Attempting to identify gene-repair pathway interactions is time-consuming and expensive. Traditional approaches to speed up this process rely on gene

knockout screens and indirect measures of DSB repair involvement. For example, growth assays that measure sensitivity to ionizing radiation. A more promising approach is to analyse CRISPR repair outcome distributions resulting from genome-wide CRISPR-Cas9 knockouts to reveal novel genes involved in DSB repair. Chapter 3 introduces MUSICiAN (Mutational Signature Catalogue ANalysis), a multivariate algorithm to detect and rank mutational spectra based on how their behaviour deviates from the expected wild-type spectra across multiple target sites in genome-wide assays, without the need for traditional controls, facilitating the discovery of lesser-known DNA repair factors. We demonstrate and evaluate MUSICiAN on a published genome-wide CRISPR mutational spectra dataset against several sets of experimentally validated DSB repair genes.

1.5.3 Signatures in CRISPR Mutational Spectra Reveal Role and Interplay of Genes in DNA Repair

Understanding DSB repair is key to genomic instability in cancer and therapy. Genome-wide studies link many genes to DSB repair, but their roles remain unclear. Evidence from other studies shows that related genes similarly modulate the frequency of specific mutational outcomes following DSB repair but have largely ignored the fact that DSB repair pathways share genes, functions, and repair outcomes. In Chapter 4, we present a computational method to exploit this connection between mutational outcomes and repair pathways to link genes to DSB repair function. We use non-negative matrix factorization (NMF) to analyze CRISPR repair outcome screens conducted on both established and candidate repair genes, as identified in Chapter 3, and identify signatures of repair pathway activity. We further employ these signatures to characterize the shared roles of repair pathways in shaping mutational patterns and link candidate genes to potential functions within these pathways based on these shared responsibilities between known and candidate DSB repair genes.

1.5.4 Overcoming Selection Bias in Synthetic Lethality Prediction

Chapter 6 explores approaches to enhance our understanding of gene interactions in DNA repair pathways by stepping outside of CRISPR-based repair outcome analyses and examining gene-gene interactions through the lens of synthetic lethality (SL). SL occurs when the simultaneous loss of function in

two genes leads to cell death. This concept holds great promise for developing anti-cancer therapies, particularly in cases where defective DNA repair pathways are prevalent. The discovery of SL interactions within these pathways has paved the way for personalized cancer treatments. However, identifying new SL pairs — and thus new therapeutic opportunities — is both costly and time-consuming. As a result, computational methods are increasingly used to predict SL interactions and guide experimental validation. Despite this, current methods often suffer from selection bias as they tend to rely heavily on known SL interactions, which limits their generalizability and overall performance. In response, we introduce SBSL (Selection Bias-resilient Synthetic Lethality) prediction models, designed to enhance robustness and generalizability across various cancer types. By integrating molecular features from cancer cell lines, patient tumour samples, and healthy donor tissues, this approach aims to improve predictive accuracy while reducing the impact of selection bias present in existing SL data. Importantly, we also highlight a class of methods in the literature whose performance is often overestimated due to their lack of generalizability when applied to new data.

References

[1] Britt Adamson, Agata Smogorzewska, Frederic D Sigoillot, Randall W King, and Stephen J Elledge. “A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response”. In: *Nature cell biology* 14.3 (2012), pp. 318–328 (cit. on p. 9).

[2] Mazhar Adli. “The CRISPR tool kit for genome editing and beyond”. In: *Nature Communications* 9.1 (2018), pp. 1–13 (cit. on p. 7).

[3] Felicity Allen, Luca Crepaldi, Clara Alsinet, et al. “Predicting the mutations generated by repair of Cas9-induced double-strand breaks”. In: *Nature Biotechnology* 37.1 (2019), pp. 64–72 (cit. on pp. 9, 11).

[4] Tomas Aparicio, Richard Baer, and Jean Gautier. “DNA double-strand break repair pathway choice and cancer”. In: *DNA repair* 19 (2014), pp. 169–175 (cit. on p. 6).

[5] Dipankan Bhattacharya, Chris A Marfo, Davis Li, Maura Lane, and Mustafa K Khokha. “CRISPR/Cas9: An inexpensive, efficient loss of function tool to screen human disease genes in Xenopus”. In: *Developmental Biology* 408.2 (2015), pp. 196–204 (cit. on p. 7).

[6] Michael Boettcher and Michael T McManus. “Choosing the right tool for the job: RNAi, TALEN, or CRISPR”. In: *Molecular cell* 58.4 (2015), pp. 575–585 (cit. on p. 8).

[7] Wendy J Cannan and David S Pederson. “Mechanisms and consequences of double-strand DNA break formation in chromatin”. In: *Journal of Cellular Physiology* 231.1 (2016), pp. 3–14 (cit. on p. 3).

[8] Alice Chen. “PARP inhibitors: its role in treatment of cancer”. In: *Chinese journal of cancer* 30.7 (2011), p. 463 (cit. on p. 6).

[9] Wei Chen, Aaron McKenna, Jacob Schreiber, et al. “Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair”. In: *Nucleic Acids Research* 47.15 (2019), pp. 7989–8003 (cit. on pp. 9, 11).

[10] David Cortez, Zheng Zhou, and Yolanda Sanchez. “Stephen Elledge and the DNA damage response”. In: *DNA repair* 35 (2015), pp. 156–157 (cit. on p. 5).

[11] Navnath S Gavande, Pamela S VanderVere-Carozza, Hilary D Hinshaw, et al. “DNA repair targeted therapy: The past or future of cancer treatment?” In: *Pharmacology & therapeutics* 160 (2016), pp. 65–83 (cit. on p. 6).

[12] Patrick D Hsu, Eric S Lander, and Feng Zhang. “Development and applications of CRISPR-Cas9 for genome engineering”. In: *Cell* 157.6 (2014), pp. 1262–1278 (cit. on p. 7).

[13] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, et al. “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity”. In: *Science* 337.6096 (2012), pp. 816–821 (cit. on p. 7).

[14] Hiroko Koike-Yusa, Yilong Li, E-Pien Tan, Martin Del Castillo Velasco-Herrera, and Kosuke Yusa. “Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library”. In: *Nature Biotechnology* 32.3 (2014), pp. 267–273 (cit. on pp. 6, 8, 11).

[15] Huiming Lu and Anthony J Davis. “Human RecQ helicases in DNA double-strand break repair”. In: *Frontiers in Cell and Developmental Biology* 9 (2021), p. 640755 (cit. on p. 4).

[16] Amy Maxmen. “Faster, better, cheaper: the rise of CRISPR in disease detection”. In: *Nature* 566.7745 (2019), pp. 437–438 (cit. on p. 7).

[17] Megan van Overbeek, Daniel Capurso, Matthew M Carter, et al. “DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks”. In: *Molecular Cell* 63.4 (2016), pp. 633–646 (cit. on pp. 6, 8, 11).

[18] F Ann Ran, Patrick D Hsu, Jason Wright, et al. “Genome engineering using the CRISPR-Cas9 system”. In: *Nature protocols* 8.11 (2013), pp. 2281–2308 (cit. on p. 8).

[19] Ralph Scully, Arvind Panday, Rajula Elango, and Nicholas A Willis. “DNA double-strand break repair-pathway choice in somatic mammalian cells”. In: *Nature Reviews Molecular Cell Biology* 20.11 (2019), pp. 698–714 (cit. on pp. 5, 6).

[20] Agnel Sfeir and Lorraine S Symington. “Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway?” In: *Trends in biochemical sciences* 40.11 (2015), pp. 701–714 (cit. on p. 6).

[21] Max W Shen, Mandana Arbab, Jonathan Y Hsu, et al. “Predictable and precise template-free CRISPR editing of pathogenic variants”. In: *Nature* 563.7733 (2018), pp. 646–651 (cit. on pp. 9, 11).

[22] Benjamin M Stinson and Joseph J Loparo. “Repair of DNA double-strand breaks by the nonhomologous end joining pathway”. In: *Annual review of biochemistry* 90.1 (2021), pp. 137–164 (cit. on p. 5).

[23] Anika Trenner and Alessandro A Sartori. “Harnessing DNA double-strand break repair for cancer treatment”. In: *Frontiers in oncology* 9 (2019), p. 1388 (cit. on p. 6).

[24] Michael M Vilenchik and Alfred G Knudson. “Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer”. In: *Proceedings of the National Academy of Sciences* 100.22 (2003), pp. 12871–12876 (cit. on p. 3).

[25] Tim Wang, Jenny J Wei, David M Sabatini, and Eric S Lander. “Genetic screens in human cells using the CRISPR-Cas9 system”. In: *Science* 343.6166 (2014), pp. 80–84 (cit. on p. 7).

[26] James D Watson and Francis HC Crick. “Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid”. In: *Nature* 171.4356 (1953), pp. 737–738 (cit. on p. 3).

[27] Evelyn M Witkin. “Ultraviolet mutagenesis and inducible DNA repair in Escherichia coli”. In: *Bacteriological reviews* 40.4 (1976), pp. 869–907 (cit. on p. 5).

[28] Charles Yanofsky. “Establishing the triplet nature of the genetic code”. In: *Cell* 128.5 (2007), pp. 815–818 (cit. on p. 3).

2

X-CRISP: Interpretable and Domain-Adaptable CRISPR Repair Outcome Prediction

> *Have no doubt, this technology will — someday, somewhere — be used to change the genome of our own species in ways that are heritable, forever altering the genetic composition of human kind.*
>
> **— Jennifer A. Doudna**
> (A Crack In Creation: Gene Editing and the Unthinkable Power to Control Evolution)

Controlling the outcomes of CRISPR editing is crucial for the success of gene therapy. Since donor template-based editing is often inefficient, alternative strategies have emerged that leverage mutagenic end-joining repair instead. Existing machine learning models can accurately predict end-joining repair outcomes, however: generalisability beyond the specific cell line used for training remains a challenge, and interpretability is typically limited by suboptimal feature representation and model architecture. We propose X-CRISP, a flexible and interpretable neural network for predicting repair outcome frequencies based on a minimal set of outcome and sequence features, including microhomologies (MH). Outperforming prior models on detailed and aggregate outcome predictions, X-CRISP prioritised MH location over MH sequence properties such as GC content for deletion outcomes. Through transfer learning, we adapted X-CRISP

Colm Seale and Joana P. Gonçalves. "X-CRISP: Domain-Adaptable and Interpretable CRISPR Repair Outcome Prediction." bioRxiv, 10.1101/2025.02.06.636858

pre-trained on wild-type mESC data to target human cell lines K562, HAP1, U2OS, and mESC lines with altered DNA repair function. Adapted X-CRISP models improved over direct training on target data from as few as 50 samples, suggesting that this strategy could be leveraged to build models for new domains using a fraction of the data required to train models from scratch.

2.1 Introduction

Gene therapies that alter the DNA to treat diseases have been made widely accessible with the emergence of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technology [11], providing faster, cheaper, and more effective gene editing [5, 20, 31, 10, 3]. The CRISPR strategy to gene editing performs enzyme-based cleavage of the DNA at a programmed location, determined by a guide RNA sequence, and subsequently exploits endogenous mechanisms recruited by the cell to repair the DNA break and introduce the desired changes. The success of gene therapies relies on CRISPR editing to produce a precise outcome, regardless of whether aiming to inactivate a disease-causing gene, to replace such a gene with a healthy copy, or to introduce a new gene with therapeutic properties. In principle, homology-directed repair (HDR) offers the most control over the repair outcome, given that it can make use of a donor template. However, HDR is typically inefficient, as it is only available during the G2 and S phases. As an alternative, template-free editing can be performed throughout the cell cycle, leading to repair by HDR or one of the more error-prone non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ) pathways [26].

Template-free editing is appealing for its wide availability but presents challenges to ensure a precise post-repair outcome, given the stochasticity of the repair processes as a result of pathway choice and inaccuracies of the repair machinery. Notably, numerous studies have reported a strong dependence of CRISPR-induced repair outcomes on the DNA sequence surrounding the break site [13, 23]. This suggests that it might be possible to influence the post-repair outcome distribution by purposefully designing guide RNAs to target sequence contexts favouring desired outcomes.

Several machine learning models leverage the relationship with sequence context to predict the frequency distribution of CRISPR-induced DNA repair

outcomes, aimed at improving guide RNA design. We categorise these models regarding predicted outcomes into more or less granular. The more granular models estimate the distribution of individual repair outcomes, and include: inDelphi, using a dual neural network and k-nearest neighbours model [27]; and FORECasT [4] and Lindel [8], both multinomial logistic regression models. The less granular models predict the frequencies of aggregated or higher-level repair outcomes: SPROUT, using a gradient-boosted tree [15]; and CROTON, based on a convolutional neural network [18]. Less granular models lack detail to offer control over precise repair outcomes. For example, CROTON predicts the overall frequency of 1bp insertions, whereas inDelphi predicts a 1bp insertion frequency per nucleotide.

Model interpretability is another key aspect of repair outcome prediction that has been insufficiently explored. The ability to explain predictions for individual target sequences and delineate how features such as sequence properties influence changes in outcome frequency provides a means to scrutinise model output and gain insight into DNA repair processes, as well as to optimise gene editing. However, less granular models make it infeasible to explain individual outcomes, whereas more granular models show a tradeoff between interpretability and performance. Specifically, FORECasT and Lindel outperform inDelphi [8], but are difficult to interpret due to the use of a large number of features with suboptimal encoding. Despite their linear model architecture, FORECasT and Lindel rely on over 3000 binary features related to sequence context and repair outcome characteristics. Furthermore, features such as deletion or microhomology length are one-hot encoded, making it challenging to recover the inherent relationship between different values of the same feature. This also leads to significant sparsity, with many features showing marginal contributions to a large proportion of outcomes. The inDelphi model uses non-linearity to leverage a more compact feature set, but ignores microhomology location and increases model complexity by distributing the prediction of deletion outcomes across multiple models. We argue that improved interpretability could be achieved by pairing a compact set of interpretable features with a non-linear model, while keeping model complexity as low as possible.

Repair outcomes are further influenced by cellular and genomic context which may affect the reliability of model predictions in diverse genomic contexts. However, training models for each context requires considerably large data that might be unavailable or challenging to generate. Notably,

most pre-trained models focus on a few mice and human cell lines for which data either previously existed or was purposely generated [27, 4, 8, 15, 18]. The limited diversity, combined with the high cost of generating new data, impedes model adoption in more diverse or unique genomic contexts, such as those encountered when developing CRISPR therapeutics [7] for precision medicine or rare diseases. To overcome this challenge, we explore transfer learning (TL) as a means to reuse and adapt knowledge from repair outcome prediction in data-rich genomic contexts for prediction in new contexts with limited data availability [24, 32, 30]. The success of TL is influenced by the degree of "relatedness" between the source and target prediction domains, while evidence suggests that DNA repair mechanisms are highly conserved among eukaryotes [16, 6] and that models trained on mouse embryonic stem cell (mESC) data can reasonably predict outcomes in zebrafish and *Xenopus* embryos [21]. We hypothesise that TL could exploit the conservation of DNA repair mechanisms to facilitate adaptation of pre-trained CRISPR outcome prediction models to data-scarce genomic contexts.

Here, we introduce X-CRISP (eXplainable CRISPR Predictions), a repair outcome prediction model designed to be granular, interpretable, and sufficiently flexible to enable adaptation to new genomic contexts. X-CRISP integrates a neural network model based on five deletion-descriptive features for prediction of deletion outcomes, alongside two multinomial logistic regression models for prediction of insertion outcomes and deletion-insertion ratio. We employ Shapley values [19] to interpret the behaviour of X-CRISP for each outcome prediction. Finally, we demonstrate several transfer learning strategies, wherein X-CRISP models pre-trained on wild-type (WT) mESC cells are adapted to other domains encompassing different cell types, organisms, and genotypes with altered DNA repair function.

2.2 Methods

2.2.1 Data and preprocessing

We used sequence data from two template-free CRISPR targeting screens: FORECasT [1] and inDelphi [2]. Both studies employed thousands of designed gRNAs paired with a 55bp (inDelphi) or 79bp (FORECasT) DNA sequence containing a PAM-adjacent 20bp target, which were delivered to Cas9-expressing cells via lentiviral transduction. Following several days of cell

culture for genomic integration, DNA cleavage, and repair, DNA sequencing was performed to capture the CRISPR repair products.

Targets, samples, and outcome sequence data. From FORECasT, we analyzed the 11,058 "Explorative gRNA-Targets" in the "FORWARD" orientation ("NGG" PAM, not "CCN"), requiring a minimum of 30bp on both sides of the cut site. From inDelphi, we used the 1,996 "FORWARD" gRNA-target pairs in "Lib-A". Specifically, we examined data from mouse embryonic stem cells (mESCs), either in their wild-type form (WT) or upon double-knockout of Prkdc and Lig4 (denoted by $Prkdc^{-/-}Lig4^{-/-}$) resulting in NHEJ deficiency (hereafter denoted by $NHEJ^{-/-}$, [33]). We also included data from human leukemic near-haploid cells (HAP1), human osteosarcoma cells (U2OS), and modified human chronic myelogenous leukaemia cells (TREX2), where the TREX2 modification fuses the Cas9 protein to the three-prime repair exonuclease 2. All FASTQ files were obtained from the European Nucleotide Archive [17] (see Supplementary Table 2.S1 for study and accession numbers).

Sequence alignment, repair outcome calling. We relied on the same set of tools to process all datasets. For FORECasT data, we used PEAR v0.9.11 [34] to merge paired-end reads using parameters "-n 20 -p 0.1" (specifying a minimum combined sequence length of 20 and a probability of no overlap below 0.1) and the "indelmap" tool [4] to map the merged reads to target sequences, both parameterised and performed as in [4], also discarding reads mapped to multiple targets. For inDelphi, we reverse complemented the target-containing reverse reads before mapping, also using the same "indelmap" tool as described in [4] (Supplementary Fig. 2.S1-2.S6 show the distributions of counts of reads mapped to target sites per dataset). Finally, we used SIQ v4.3 [25] to call repair outcomes per read with options "-m 2 -c -e 0.05", specifying a minimum number of 2 reads for the event to be counted, the collapsing of identical events to a single record with the corresponding sum of counts, and a maximum permitted base error rate of 0.05.

Repair outcome profile generation. We calculated outcome frequency distributions (or repair outcome profiles) per gRNA-target and screen as follows. We considered all deletions up to 30bp long overlapping the cut site or adjacent to the upstream nucleotide neighbouring the cut site. This cut-off was selected because deletion lengths over 30 bp were rarely observed

(Supplementary Fig. 2.S7–2.S12), and also to maintain consistency with other studies on CRISPR repair outcome profile prediction. [4, 8]. We then determined unique deletion outcomes by grouping deletions that produced identical repair products. Only MH-*based* deletions presented such ambiguity, as a result of the loss of one of two microhomologous sequences flanking the deleted region (Fig. 2.1). All remaining unique deletions were categorised as "MH-*less* deletions". This yielded unique sets of approximately 330-480 deletion outcomes per target site. We also considered all unique single- and di-nucleotide insertions, as well as one single category for insertions of at least 3bp due to their rarity of occurrence (Supplementary Fig. 2.S13-2.S18), totalling 21 insertion outcomes. We mapped reads to outcomes per gRNA-target, and discarded gRNA-targets with less than 100 mutated reads. Outcome counts were divided by the sum of counts per gRNA-target to obtain the final outcome profiles. The target sequences surviving the filtering step (and their respective repair outcome profiles) were then randomly split into non-overlapping train and test sets per source study (i.e. the same sequences are used for training and testing across datasets sourced from the same study), with train sets later used for model hyperparameter optimisation and test sets held out for evaluation (Table 2.1). The only exception is the inDelphi WT mESC dataset, where all the points were used as an extra held-out dataset for testing. We calculated Needleman-Wunsch [22] and Smith-Waterman [29] alignment scores (scoring: match=1, mismatch=0.0, gap=0.0), and Hamming distances between sequences in the train and test sets to ensure there were no near-identical sequences between them (Supplementary Fig. 2.S19-2.S21).

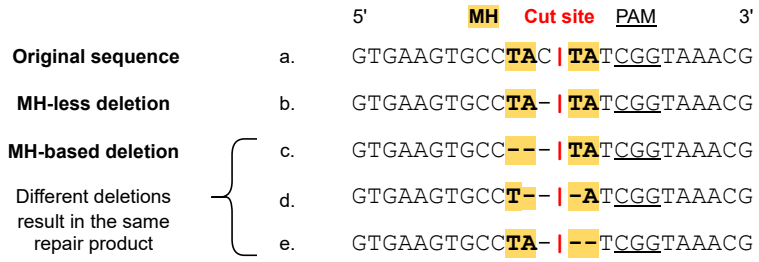


Fig. 2.1: Deletion type categorisation.

Cell line	Genotype	Study	Total	Train	Test
mESC	WT	FORECasT	9854	5900	3954
mESC	WT	inDelphi	1961	0	1961
mESC	*NHEJ*$^{-/-}$	inDelphi	1485	500	985
K562	TREX2	FORECasT	3855	500	3355
HAP1	WT	FORECasT	4450	500	3950
U2OS	WT	inDelphi	1462	500	962

Tab. 2.1: Counts of processed repair outcome profiles per cell line, genotype, and study. Total counts, as well as train and test set splits.

2.2.2 X-CRISP

The proposed model, X-CRISP, uses three sub-models to predict the repair outcome profile for a 60bp sequence centred at the cut site. The first two sub-models predict individual deletion and insertion outcomes, and the third predicts the overall frequency of deletions and insertions. The outputs of the first two sub-models are scaled by the output of the third and concatenated to construct the complete predicted repair outcome profile.

Deletion model. The deletion model predicts a frequency distribution per target over all considered deletion outcomes. Different from the other granular deletion models, X-CRISP introduces common features for MH-*based* and MH-*less* deletions, and avoids one-hot encoding by representing integer features as-is. Five features are reconciled and used ubiquitously by X-CRISP across deletion categories to consolidate deletion prediction into one single interpretable model: "Left edge" and "Right edge", representing the left and right deletion edges or the positions of the nucleotides closest to the cut site for the left and right MHs; "Gap", denoting the distance between the two edges; and MH length and MH GC fraction, which are both zero for MH-*less* deletions. These features are fed to a fully connected neural network that independently scores each outcome between 0 and 1.

The network contains two hidden layers of 16 nodes and one output node, using sigmoid activation at every layer. We trained two models, "X-CRISP KLD" and "X-CRISP MSE", respectively using the Kullback-Leibler divergence (KLD) [14] and mean-squared error (MSE) loss functions. Training was performed using PyTorch v.1.8.0 with the Adam optimiser [12] ($\beta_1 = 0.99$, $\beta_2 = 0.999$, and remaining default settings), a batch size of 200, and learning

Feature	Description
MH length	Length of MH, defined as $M \in [0, 30]$. Zero for MH-*less* deletions. For MH-*based* deletions, $M \in [1, 30]$, since the deletion contains one of the MHs (Fig. 2.1) and 30 is the maximum deletion length.
MH GC	Fraction of GC in MH, defined as $F \in [0, 1]$. Zero for MH-*less* deletions.
Gap	Length between MHs, defined as $G \in [1, 30]$. For MH-*less* deletions, it equals the deletion length.
Left edge	Deletion left edge position $L \in [-30, 0]$, where 0 indicates the edge is at the cut site. For MH-*based* deletions: L is the distance in bp from the cut site to the closest base pair of the PAM-distal MH.
Right edge	Deletion right edge position $R \in [0, 30]$, where 0 indicates the edge is at the cut site. For MH-*based* deletions: R is the distance in bp from the cut site to the closest base pair of the PAM-proximal MH.

Tab. 2.2: X-CRISP deletion outcome feature descriptions.

rates 0.05 and 0.01 for KLD and MSE, respectively. We applied an exponential learning rate decay with $\gamma = 0.999$ per epoch. We used L2 regularisation and optimised hyperparameters with 5-fold cross-validation (CV) on the train set. The final models were trained on the entire train set using the hyperparameter values yielding the lowest mean loss (Supplementary Table 2.S2 for tested and final hyperparameters).

Insertion and deletion-insertion models. The insertion model predicts frequencies for the 21 insertion outcomes, while the deletion-insertion model predicts the overall frequency of deletions and insertions. Both models use softmax regression and take a DNA sequence as input, represented by one-hot encodings of single nucleotides and dinucleotides at each position. The insertion model uses the six nucleotides directly upstream of the PAM, and the deletion-insertion model considers the 20bp target sequence. Both models were trained using the Adam optimiser and an exponential learning rate decay. We used L2 regularisation and optimised hyperparameters using 5-fold CV on the train set. The final models were trained on the entire train set using the hyperparameter values yielding the lowest mean MSE (Supplementary Table 2.S2 for tested and final hyperparameters). We trained for a maximum of 200 epochs, with early stopping if there was no improvement after two epochs.

Other prediction models. We compared X-CRISP against four other published models at the time of writing: inDelphi [27], FORECasT [4], Lindel [8], and CROTON [18] (Supplementary Table 2.S3). We excluded SPROUT [15], as it only predicts aggregate outcomes at a higher level, namely: "average insertion length", "average deletion length", "diversity", and "most likely inserted pair". These prediction tasks do not align well with the practical applications we envision for X-CRISP, which require more detailed outcomes. Every model was trained on the same data (Table 2.1), following the procedures outlined in the respective publication, with two exceptions: we trained inDelphi with a maximum deletion length of 30bp for consistency across models, and we trained CROTON using the architecture provided by the authors without redoing the architecture search.

Evaluation of prediction performance. We trained each model on 5900 FORECasT WT mESC target repair profiles and tested it against 3954 FORECasT and 1961 inDelphi target repair profiles, without overlap between train and test. Direct comparisons were challenging due to the different outcome categorisations used by each model (Supplementary Table 2.S3). To address this, we assessed each model on the outcomes described in its original publication, as well as on three sets of outcomes comparable across models: common MH-*based* deletions, common MH-*less* deletions, and 1bp insertions. Across models, predicted deletion outcomes were limited to 30bp in size. For comparison with FORECasT for 1bp insertions, non-repeat insertions of nucleotides neighbouring the cut site were grouped as one outcome. To measure the error between predicted and observed repair outcome probability distributions, we used the Jensen-Shannon distance (JSD) and Pearson's correlation coefficient. The JSD is designed to measure the divergence between probability distributions, and its symmetry and bounded range (0 to 1) make it respectively invariant to the order of the two distributions and more comparable across targets with varying numbers of possible repair outcomes than other relevant metrics such as KLD or cross-entropy. The Pearson's correlation quantifies the linear relationship between the predicted and observed frequency vectors of each target. It is less suitable for probability distributions because it assumes independence between the values in each vector, whereas the probabilities associated with the repair outcomes of any target sum to 1 and are necessarily dependent on one another. Nevertheless, Pearson's correlation is still commonly used and we include it for completeness. For the overall frequencies of deletions and insertions, we used MSE.

We also assessed the ability of each model to classify "precision-X%" target sequences, defined as those where a single outcome represented at least $X\%$ of the reads assigned to all considered repair outcomes, with $X \in \{20, 30, 40, 50, 60, 70\}$. Targets were classified as positive if they met the criteria and negative otherwise. We used precision, recall, and Matthew's correlation coefficient (MCC) to evaluate performance. Precision denotes the rate at which the model is expected to be correct when it predicts a target sequence as a precision-X% site. Recall expresses the fraction of all observed precision-X% targets of a set we can expect the model to identify. For CRISPR-based gene editing, the focus is on which targets can be more confidently used to produce the desired outcome with the highest possible fidelity for more precise gene editing. As a result, higher emphasis is placed on precision and reducing the risk of false precision-X% target predictions, compared to recall and recovering the most precision-X% targets. The MCC evaluates the quality of the binary precision-X% predictions made by the model, including but also beyond precision and recall, providing a value between -1 and +1, where +1 indicates perfect agreement, 0 indicates random prediction, and -1 indicates total disagreement between model predictions and observed outcomes. The MCC is generally reported as more informative and reliable to assess binary classification than other combined performance metrics, such as F1 and AUC. It is also more suitable for evaluation under class imbalance, which is present in the case of precision-X% prediction tasks, where most targets are non-precision-X%.

Lastly, we assessed six aggregate prediction tasks (deletion, 1bp insertion, 1bp deletion, 1bp frameshift, 2bp frameshift, and frameshift frequencies) using the MSE and Pearson's correlation. We emphasise the MSE to determine the best performing model for these prediction tasks because MSE takes the magnitude of the errors into account, while Pearson's correlation does not (it is both scale- and shift-invariant).

Explainability. To interpret X-CRISP predictions, we used SHapley Additive exPlanations (SHAP, python library v0.39.0 [19]). We calculated SHAP values to estimate the contribution of each feature to the change in outcome frequency predicted by the X-CRISP model for a given target, relative to the average frequency obtained for a background set of targets. For the deletion model, we used the "DeepExplainer" function with 10k randomly selected deletions from target sequences in the train set as background. For the insertion and deletion-insertion models, we used the "LinearExplainer" with

5k randomly selected targets as background. All other parameters were set as default. We generated SHAP values to explain 400 randomly selected target sequences from the test set per model. We aggregate the SHAP values for the nucleotide features by summing the SHAP values of their binary encodings. For example, the SHAP value for "A" at position 16 is the sum of the SHAP values for the feature|value pairs (A16|1, C16|0, G16|0, T16|0).

Code	Weight initialisation	Frozen layers	Trained & fine-tuned on target
SO	Random	No	No
TO	Random	No	Both
FT	Pre-trained	No	Fine-tuned
PF0	Pre-trained	No	Both
PF1	Pre-trained	First hidden	Both
PF2	Pre-trained	Both hidden	Both

Tab. 2.3: Baseline X-CRISP models and transfer learning strategies. Baselines: SO, source only; TO, target only. Transfer learning: FT, pre-trained on source and fine-tuned for target; PF0-2, pre-trained on source and retrained + fine-tuned on target with 0-2 frozen hidden layers.

Transfer learning. We investigated whether X-CRISP models trained on FORECasT WT mESC data as a source domain could be adapted using transfer learning (TL) to predict on the following different cell lines as target domains: mESC *NHEJ*$^{-/-}$, TREX2, HAP1, and U2OS. Our general TL approach was to first initialise a new model using the learned weights from the pre-trained X-CRISP KLD mESC model, and then either fine-tune (FT) or retrain and fine-tune (PF0-2) the initialised model using n samples from the other cell line of interest to adapt the model to the target domain. We used sets of samples of increasing size, with $n \in \{2, 5, 10, 20, 50, 200, 500\}$, where each subsequent set was a superset of the preceding one. Fine-tuning alone (FT) involved training the model on the target domain samples using a low learning rate. When retraining before fine-tuning (PF0-2), we controlled the flexibility of the model to adapt itself (i.e. the number of learnable model parameters or weights) by freezing the weights of none (PF0), the first (PF1), or both (PF2) of its hidden layers before retraining, allowing only unfrozen layers to change [30] (Supplementary Table 2.S2 for hyperparameter details). As baselines for comparison, we used the X-CRISP KLD model trained only on the WT mESC source domain data (SO) and another X-CRISP KLD model trained only on the target cell line data (TO) (Table 2.3).

2.3 Results and Discussion

2.3.1 X-CRISP accurately predicts detailed repair profiles

We first evaluated the ability of all models trained on the 5900 FORECasT mESC target sequences to predict repair outcome profiles for the 3954 and 1961 mESC WT target sequences in the FORECasT and inDelphi test sets, respectively.

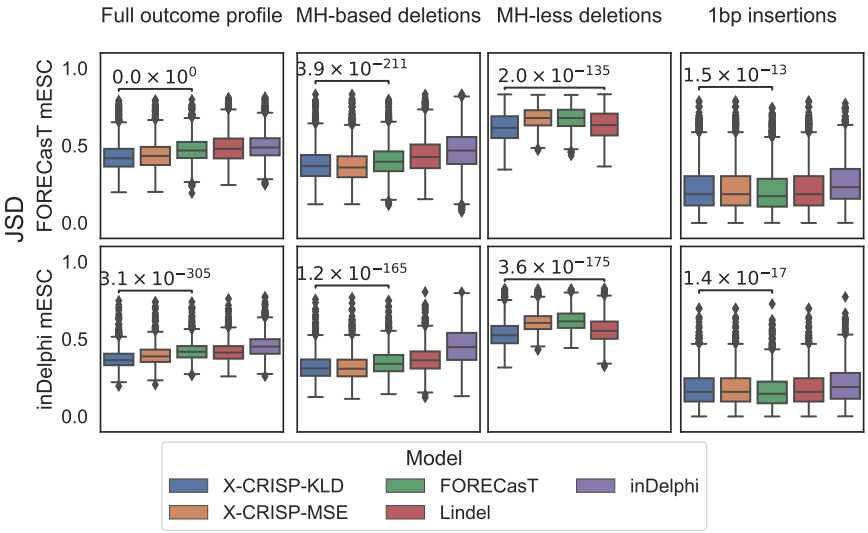


Fig. 2.2: Detailed repair outcome prediction performance. Jensen-Shannon distance (JSD) between predicted and observed outcomes for (top) 3954 FORECasT or (bottom) 1961 inDelphi mESC WT test target sites, considering: (left to right) original publication outcomes; common MH-*based* deletions; common MH-*less* deletions; 1bp insertions. Significance p-values calculated using Wilcoxon two-sided signed-rank tests, comparing X-CRISP KLD to the best of the non-X-CRISP models. For 1bp insertions, X-CRISP KLD and Lindel perform identically, given that they are based on the same model, so the comparison is then made with FORECasT, the next best of the non-X-CRISP models.

When predicting the profile as described in the original publication (Fig. 2.2, Original outcomes; Supplementary Fig. 2.S22 for Pearson's correlation), both X-CRISP KLD and X-CRISP MSE achieved the best performances with a significantly lower median JSD compared to FORECasT, the best non-X-CRISP model (FORECasT|inDelphi mESC data X-CRISP KLD: 0.43|0.37, FORECasT: 0.47|0.42; Wilcoxon two-sided signed rank test p-values < 0.05).

For deletion frequency prediction, the X-CRISP KLD model significantly outperformed all others, with FORECasT and Lindel as the best non-X-CRISP models for MH-*based* (FORECasT|inDelphi mESC data X-CRISP KLD: 0.38|0.32, FORECasT: 0.40|0.35) and MH-*less* deletions (FORECasT|inDelphi mESC data X-CRISP KLD: 0.61|0.53, Lindel: 0.62|0.56), respectively (Fig. 2.2). The FORECasT and Lindel models could be at a disadvantage compared to X-CRISP due to their linear architectures and binary feature representations, since the one-hot encoding of integer-valued features (such as deletion length) and assumption of feature independence impairs their ability to model relationships between values within and across features. In contrast, X-CRISP uses integer-valued features as-is to safeguard interpretation, and leverages non-linearity to learn feature interactions. The inDelphi model uses a similar strategy, however X-CRISP additionally considers deletion edge locations, which seem to confer a further boost in performance.

For insertion predictions, we had two considerations: (i) inDelphi only predicts 1bp insertions; (ii) FORECasT groups all 1bp insertions that do not repeat the nucleotides flanking the cut site into one outcome. Thus, we compared only 1bp insertion frequencies and aggregated non-cut site-repeating insertions into one outcome. Both X-CRISP models performed comparably to Lindel, which was expected given that the X-CRISP insertion model is based on Lindel (Fig. 2.2). Moreover, X-CRISP/Lindel achieved a lower JSDs than inDelphi, but higher than FORECasT (FORECasT|inDelphi mESC data X-CRISP KLD: 0.22|0.18, FORECasT: 0.21|0.17). We reason that given FORECasT is trained to predict the frequency of non-cut site-repeating insertions as a single group directly, it has a slight advantage in these comparisons. In addition, the advantage of X-CRISP/Lindel over inDelphi could possibly be explained by the use of a wider sequence context surrounding the cut site (6bp vs. 3bp), providing additional degrees of freedom to model insertion frequencies.

2.3.2 X-CRISP generalises well to frameshift prediction tasks

We further investigated if the detailed repair profiles predicted by the models could be useful to address higher-level prediction tasks, focusing on precision-$X\%$ targets and broader outcome categories.

First, we assessed the prediction of precision-$X\%$ targets, defined as target sequences for which a single outcome accounts for at least $X\%$ of all reads

assigned to any outcome [27]. The ability to predict the precision-X% property can help with selecting CRISPR targets that maximise the proportion of the desired outcome relative to all other outcomes, for more precise gene editing. Performance was measured using precision (Table 2.4), recall, and Matthew's correlation coefficient (Supplementary Table 2.S4). On the FORECasT mESC test set, X-CRISP excelled in precision-20% but came second to FORECAST in precision-30% through to precision-70% (Table 2.4). On the inDelphi mESC test set, FORECasT led in precision-20/30/40/50%, while X-CRISP MSE achieved top performance in precision-60%. Note that each model performed worse on inDelphi than on FORECasT data, highlighting the challenges of generalising to other datasets, even within the same cell type and genotype.

Test set	Model	Prediction of precision-X% (performance with $X = ..$)					
		20	30	40	50	60	70
FORECasT WT mESC	X-CRISP KLD	**0.75**	0.84	0.84	0.80	0.71	0.66
	X-CRISP MSE	**0.75**	0.83	0.83	0.77	0.69	0.65
	FORECasT	0.72	**0.85**	**0.86**	**0.83**	**0.76**	**0.74**
	Lindel	0.72	0.77	0.69	0.56	0.33	0.50
	inDelphi	0.40	0.45	0.24	0.09	0.06	0.03
inDelphi WT mESC	X-CRISP KLD	0.73	0.70	0.57	0.40	**0.28**	0.00
	X-CRISP MSE	0.73	0.70	0.54	0.34	0.21	0.00
	FORECasT	**0.77**	**0.83**	**0.71**	**0.46**	0.21	0.00
	Lindel	0.69	0.64	0.53	0.37	0.16	0.00
	inDelphi	0.36	0.31	0.07	0.02	0.00	0.00

Tab. 2.4: Performance of six precision-X% prediction tasks, with $X \in \{20, 30, 40, 50, 60, 70\}$, measured using the precision performance score. All five models were trained on the 5900 FORECasT WT mESC target train set from Table 2.1, and then tested separately on the 3954 FORECasT and 1961 inDelphi WT mESC target test sets from Table 2.1.

We further evaluated the performance of each model on six outcome profile aggregation tasks: deletion, 1bp insertion, 1bp deletion, 1bp frameshift, 2bp frameshift, and frameshift frequency prediction. We also included a CROTON model [18] in the evaluation, which was originally developed to predict these six broader outcomes. To ensure comparability, we retrained CROTON on the same data as the other models, and aggregated the predictions of the other models per broader outcome.

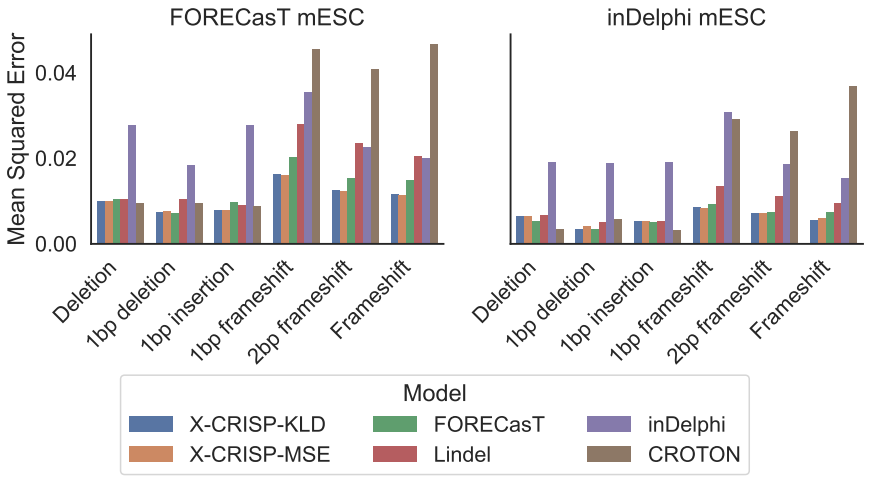


Fig. 2.3: Broader repair outcome prediction performance. Mean-squared error (MSE) for six outcomes: deletion, 1bp insertion, 1bp deletion, 1bp frameshift, 2bp frameshift, and frameshift frequency prediction. Models trained on FORECasT WT mESC and tested on 3954 FORECasT and 1961 inDelphi WT mESC target sites.

The X-CRISP model achieved top or near-top performance across all tasks (Fig. 2.3 for MSE, Supplementary Fig. 2.S23 for Pearson's correlation). In both datasets, X-CRISP competed for the best deletion frequency and 1bp insertion frequency prediction performance with Lindel and CROTON, respectively. FORECasT led in 1bp deletion frequency prediction for both datasets, while X-CRISP excelled on all frameshift prediction tasks. Frameshift prediction is an especially important task as frameshifts often result in gene knockouts, useful for studying gene function and developing therapeutics. We attribute the improved performance of X-CRISP on broader outcome prediction to the superiority already demonstrated when predicting detailed repair outcome profiles. In this case, the ability to predict more accurately across the entire frequency distribution seemed to translate into improved aggregate frequency predictions.

Overall, the results indicate that repair outcome profiles predicted by X-CRISP generalise better than those predicted by other models to higher-level prediction tasks, such as precision-X% and broader outcome prediction. However, the loss function used to achieve optimal performance could be task-specific.

2.3.3 Deletion prediction is most influenced by MH location

Compared to models like FORECasT and Lindel that break down (combinations of) outcome attributes such as MH length and position into thousands of categorical binary feature encodings for specific values or bins, X-CRISP preserves each of the five attributes it uses as a single feature in its original integer or real-valued range. For instance, Lindel splits MH length into five categorical bins (0, 1, 2, 3, and 4+) and couples them with positional and deletion length categories, resulting in a total of 2649 features where the effect of MH length alone cannot be easily discerned from the effects of other attributes. Similarly, FORECasT bins MH length into seven categories ("No MH", 1, 2, 3, 4–6, 7–10, 11–15) and pairs those with additional attributes like deletion length or microhomology location, both also binned with varying ranges, yielding 525 MH-related features out of a total of 3633 used by the FORECasT model. In contrast, X-CRISP represents the MH length attribute using a single integer-valued feature and relies on the neural network to learn eventual interactions with other attributes or features in a data-driven manner. Similar observations can be made for the remaining attributes: where X-CRISP uses a single feature per attribute, FORECasT and Lindel typically use hundreds of features combining specific values or bins from multiple attributes.

While the encoding employed by FORECasT and Lindel is not necessarily limiting in terms of performance, given that the large numbers of features provide sufficient degrees of freedom to learn good prediction models, the dispersion and combination of attribute values across features make it challenging to isolate and interpret the contribution of each attribute on its own. On the other hand, the X-CRISP approach allows the learning of feature interactions and thus introduces black-box characteristics to the model. Nevertheless, post-hoc explainability tools such as SHAP are precisely designed to summarise and quantify how input attributes influence the predictions of (black-box) models, which we can readily use to gain insight into the impact of sequence characteristics on CRISPR outcome prediction across target sites and outcomes. To explore this, we obtained SHAP values for 400 randomly selected targets from the FORECasT test data to elucidate the influence of each feature on the predicted score of each X-CRISP submodel.

For MH-*based* deletions (Fig. 2.4A), the influence (SHAP value) of the left and right edges increased as they got closer to the cut site (feature values near

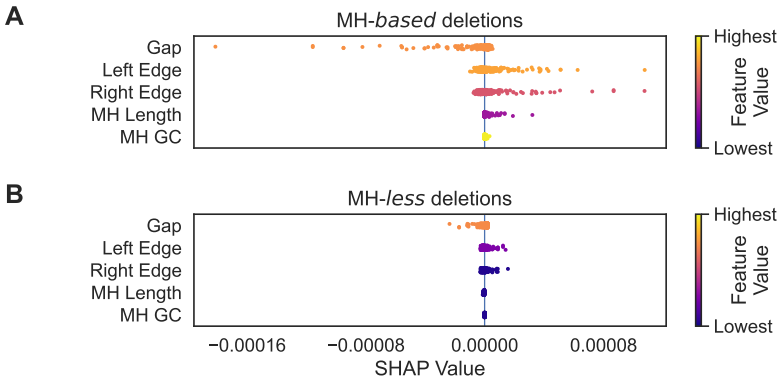


Fig. 2.4: Influence of X-CRISP model features on repair outcome prediction. Feature effects expressed by SHAP values, calculated for 400 targets from the FORECasT WT mESC test set. Strip plots show SHAP values per feature for 10k randomly selected (A) MH-*based* and (B) MH-*less* deletions.

zero), meaning that an MH-*based* deletion became more likely. We consider this intuitive, since the edge features also denote MH location, and MHs closer to the cut site should be easier to select and anneal during repair due to their physical proximity. Counter-intuitively, as the gap between MHs decreased, the influence (SHAP value) of the gap feature decreased. We reason that this could be a correction applied by the model when both edges are near zero, to prevent the deletion frequency from growing disproportionately large. Longer MHs also led to larger MH-*based* deletion frequencies, yet were considerably less influential than MH location (left/right edge), while GC content exhibited minimal contribution. This could indicate that the selection of an MH during repair might be more influenced by MH position than by sequence content. For MH-*less* deletions, the gap and edge features showed similar behaviour to that described above for MH-*based* deletions, but with smaller SHAP value ranges.

For insertions, we focused on the most prominent group: 1bp insertions (Fig. 2.5A). Cut site-proximal positions showed more influence than others, where an A or T immediately upstream of the DSB (at the 17th position/index position 16 of the 20nt gRNA sequence, see Fig. 2.5C for an illustration of target sequence indexing) promoted the insertion of an A or T, respectively. Insertions of C were positively influenced by the presence of CC immediately upstream of, or CG centred at, the cut site. We did not observe strong

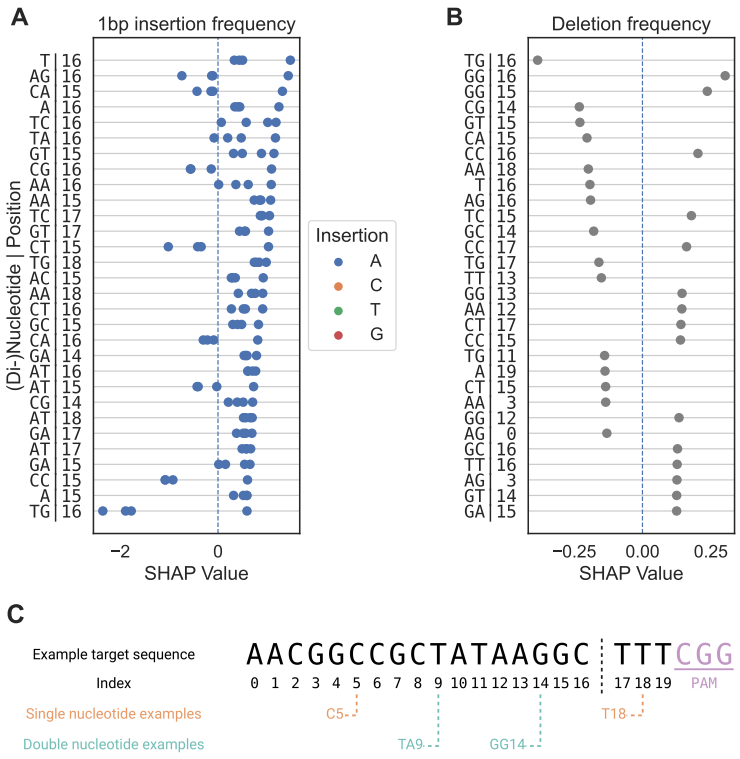


Fig. 2.5: Influence of X-CRISP model features on repair insertion prediction. Feature effects expressed by SHAP values, calculated for 400 targets from the FORECasT WT mESC test set. (A) Top 30 features for all 1bp insertions, ranked by maximum SHAP value. (B) Top 30 features for the deletion-insertion model, with SHAP values denoting impact on deletion frequency and ranked by absolute SHAP value. (C) Indexing of 1- and 2-nucleotide features for an example target DNA sequence.

associations between single or dinucleotide sequence content and G insertions. These findings align with existing literature [23, 27, 4, 8, 28, 9].

For the deletion-insertion model, DSB-proximal nucleotides were strong influencers as well (Fig. 2.5B), with C or G at position 16 promoting deletions. However, A or T at position 16 (e.g. T|16, GT|15, CA|15, or AG|16) promoted insertions. A standout observation was that dinucleotide repeats centred at the cut site (except AA|16) favoured deletions of DSB-adjacent nucleotides, as previously observed elsewhere [4].

2.3.4 Transfer learning greatly reduces data required for new domains

We explored if transfer learning could adapt pre-trained X-CRISP models to new domains, like other cell types or genomic contexts, and reduce the requirements of domain-specific training data.

We first examined changes in the distributions of repair outcomes between WT mESC cells and each of the four different target domains: human U2OS, human HAP1, NHEJ-deficient mESC, and human TREX2 cells (Fig. 2.6). The HAP1 cells revealed high similarity with mESC, exhibiting similar MH-*based* and MH-*less* deletion frequency distributions, average frequencies per deletion length, and insertion frequency distributions. The U2OS cells showed a larger variation in overall deletion frequency, a higher proportion of single A insertions, and a lower proportion of $\geq$3bp insertions. Cell lines with modified repair function deviated the most from the others: mESC *NHEJ*$^{-/-}$ cells favoured MH-*based* deletions and led to less frequent insertions, especially 1bp insertions; TREX2 cells preferred deletions of 10-16bp over deletions of 3-9bp, and MH-*less* over MH-*based* deletions.

For the transfer learning task, we adapted the X-CRISP models pre-trained on WT mESC data to each target domain using several techniques (Fig. 2.7 for JSD, Supplementary Fig. 2.S24 for Pearson's correlation). Each adapted model was further tested against heldout unseen data from the corresponding target domain. Pre-training on mESC data alone (SO, source only) generalised well to HAP1 cells (full repair profile mean JSD: mESC 0.428, HAP1 0.432), and achieved comparable performance to training directly on HAP1 data using at least 500 samples (TO, target only).

The TL strategies showed little benefit here, likely due to the similarity between mESC and HAP1 (Fig. 2.6), requiring at least 500 target HAP1 samples before a consistent gain in performance could be observed across all TL methods (TL mean JSD: 0.420). For the transfer to U2OS cells, all TL methods significantly improved the full repair profile performance over the SO and TO baselines after retraining and fine-tuning on 50 target U2OS samples, with PF0 (retrained and fine-tuned on target data without layer freezing) achieving the best results (Wilcoxon two-tailed signed-rank test *p*-values PF0 vs. SO: 6×10^{-92}; and vs. TO: 1×10^{-137}). The improvement was also seen for both U2OS MH- and MH-*less* deletion prediction. In addition, fine-tuning (FT) on U2OS cells improved deletion-insertion ratio prediction

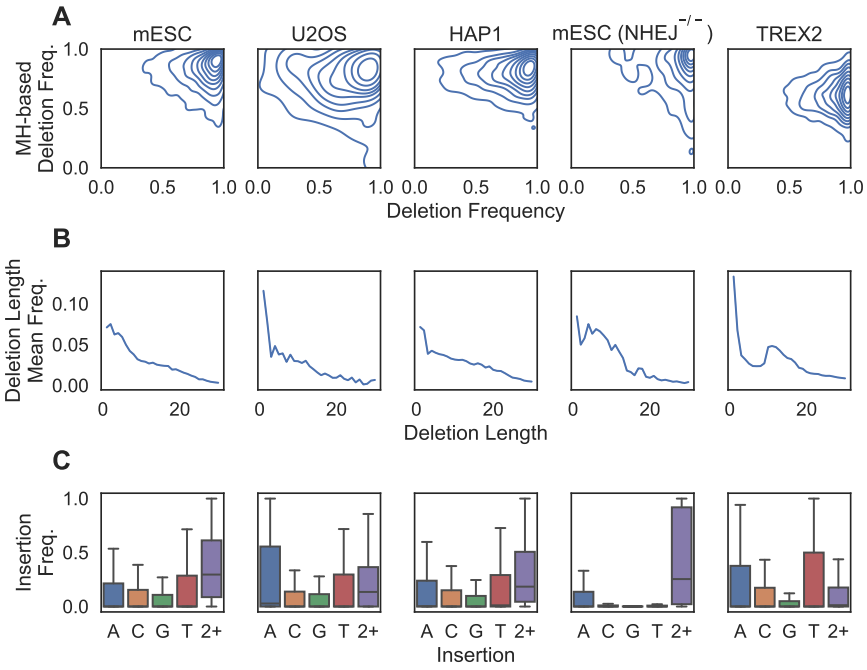


Fig. 2.6: Repair outcome distributions across all X-CRISP processed repair outcome profiles observed for the mouse and human cell lines in Table 2.1: (A) densities for overall deletion frequency (horizontal) vs. MH-*based* deletion frequency (vertical), with each contour line denoting 10% of the data; (B) trend line of average frequency per deletion length; (C) frequency distributions of 1bp and $\geq$ 2bp insertions, outliers excluded.

over SO and TO from only two target samples. These results indicate that TL could be used to reduce the number of target samples required to train new models for some target domains.

Lastly, we examined the results for cell lines with modified cellular DNA repair function. For mESC *NHEJ*$^{-/-}$ cells, all TL methods showed small improvements in full repair profile performance over SO and TO after training on 50 target samples. Using 500 target samples, only PF0 outperformed SO and TO (mean JSD; PF0: 0.505, SO: 0.533), driven by small improvements in MH-*based* deletion prediction, consistent with the fact that impairment of NHEJ visibly altered MH-*based* deletion activity (Fig. 2.6). For TREX2 cells, PF0 was the only effective TL method, showing gains from just five target samples. Here, the insertion model displayed a significant performance benefit over both SO and TO baselines up to 500 target samples, while the deletion model did

not benefit from TL, likely due to the large distribution shift towards longer deletions driven by the Cas9-fused three-prime exonuclease 2. The TO model achieved comparable performance to TL using 500 samples.

Overall, the most flexible transfer learning strategy (PF0, no layer freezing) showed the most effective adaptation to new domains, requiring only 50 target domain samples to consistently achieve results comparable or superior to training directly on a larger number of samples for the 4 different target domains. The gap between the least and most flexible TL strategies was especially apparent when adapting to larger changes between source and target context distributions (Fig.2.7, TREX2). These changes seemed largest when the underlying biological mechanisms for cutting or repairing the DNA were modified than across wild-type cell lines (Fig. 2.6, NHEJ$^{-/-}$ and TREX2 vs. WT mESC, U2OS, and HAP1). This suggested a higher conservation of repair mechanisms between wild-type cells, even across mouse and human. Importantly, our results also showed that TL strategies provided the most benefit for genotypes affecting CRISPR-Cas9 function and DNA damage response (Fig. 2.7), creating opportunities to better understand the associated biological mechanisms and to improve the control over CRISPR-induced outcomes for more precise gene editing across fundamental and translational applications. Increasing the number of learnable parameters allowed the X-CRISP model to better realign to the larger changes in repair outcome distributions exhibited by genetically modified cells, highlighting that the effectiveness of TL is domain-dependent and considerations such as further tweaking of the models could be necessary for successful adaptation to more challenging contexts. We also analysed the impact of TL on the prediction performance of broader repair outcomes, where TL consistently improved the MSE over the baselines on the frameshift frequency tasks after 50 target domain samples (Supplementary Fig. 2.S25, see Supplementary Fig. 2.S26for Pearson's correlation). We envision that similar TL strategies could be used to adapt models for precise CRISPR therapeutic interventions considering the unique genetic landscape of specific patient cohorts or individual patients. To determine how successful such strategies could be would need extensive investigation and validation across a wide variety or at least a representative selection of human donors.

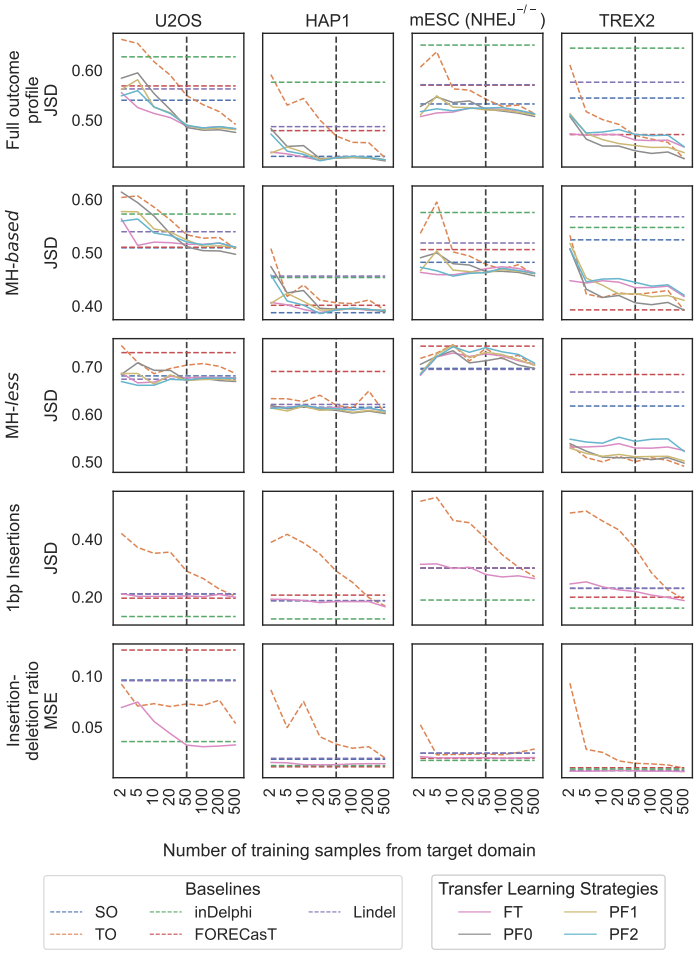


Fig. 2.7: X-CRISP model adaptation to new domains or cell lines using transfer learning (TL). Prediction performance of baseline X-CRISP models and TL strategies, as the average JSD or the MSE between predicted and observed frequencies per model and number of training samples from the target domain. (Baseline models) TO, X-CRISP trained on target only; SO, X-CRISP trained on source only; FORECasT, Lindel, inDelphi. (Transfer learning) FT, X-CRISP pre-trained on source and fine-tuned for target; PF0-2, X-CRISP pre-trained on source and retrained + fine-tuned on target using 0-2 frozen hidden layers. (Top to bottom) Prediction models for full repair profile, MH-*based* deletions, MH-*less* deletions, 1bp insertions, and deletion-insertion ratios. Note: horizontal axis is categorical, thus it is not to scale; and the SO, FORECasT, inDelphi, and Lindel baselines do not use any samples from the target domain, so their performance remains constant along the horizontal axis. (Left to right) Models tested against U2OS (962), HAP1 (3950), TREX2 (3355), and mESC *NHEJ*$^{-/-}$ (985) test target sites.

2.4 Conclusion

We introduced X-CRISP, an interpretable and domain-adaptable model to predict the frequency of CRISPR repair outcomes. The X-CRISP model exhibited superior accuracy and generalisation in both detailed and broader repair outcome frequency prediction, especially when predicting frameshift mutations – a crucial task in experimental and therapeutic genome editing applications.

Top performance was achieved while retaining interpretability and flexibility to adapt X-CRISP models to additional genomic contexts. Contributing to this was the inclusion of informative features like deletion edges, which also function as indicators of MH location, and showed a prominent influence on deletion frequency prediction of X-CRISP models. Additionally, the ubiquitous representation of features for both MH-*based* and MH-*less* deletions, coupled with a non-linear neural network architecture, enabled X-CRISP to leverage a compact set of 5 features for improved and interpretable deletion prediction, compared to the next best models using $\sim$3k features.

Finally, we showed that pre-trained X-CRISP models could be successfully adapted using transfer learning for prediction in additional domains, spanning different organisms, cell lines, and DNA repair function characteristics. Transfer learning was effective from 50 target domain samples, suggesting that the typical range of thousands of domain-specific samples required to train a repair outcome prediction model for a new domain could be reduced by adapting existing pre-trained models, potentially using orders of magnitude less target domain samples.

More flexible TL approaches, with freedom to adjust all weights of the prediction model, generally adapted better and were especially effective in transferring to more distant contexts characterised by larger changes in the repair outcome frequency distribution. These results highlight the potential of transfer learning to expedite the development of CRISPR repair outcome prediction models for contexts where generating extensive data may not be feasible and thus facilitate research involving CRISPR assays. Specifically, these models could help improve the design of guide RNAs to efficiently knock out specific genes with the aim of studying their function, or contribute to the development of CRISPR therapeutics by improving the design of guide RNAs to correct pathogenic variants across various genomic contexts.

2.5 Supplementary Tables

Source	Cell Type	Genotype	Sample Accession(s)
FORECasT [4]	mESC	Wild-Type	SAMEA5093999, SAMEA5094000
	K562	TREX2	SAMEA104549464
	HAP1	Wild-Type	SAMEA5094017, SAMEA5094018
inDelphi [27]	mESC	Wild-Type	SAMN08955971
	U2OS	Wild-Type	SAMN09689449
	mESC	$Prkdc^{-/-}Lig4^{-/-}$	SAMN08955971

Tab. 2.S1: CRISPR repair outcome sequencing data accession numbers.

Model	β_1	β_2	Learning rate (LR)	LR decay (γ)	Penalty (optimised via 5-fold CV)	Penalty (final)
Training on source domain						
Deletion	.99	.999	0.05	0.999	Tested L2 regularisation with weights in $\{0.1, 0.05, 0.01, 0.001\}$	0.00001 (L2)
Insertion	.99	.999	0.001	0.99	Tested L1 and L2 regularisation independently with weights in the range of between 10^{-10} and 10^{-1}	0.0005011872 (L1)

Continued on next page

– continued from previous page

Model	β_1	β_2	Learning rate (LR)	LR decay (γ)	Penalty (optimised via 5-fold CV)	Penalty (final)
Deletion-insertion	.99	.999	0.001	0.99	Tested L1 and L2 regularisation independently with weights in the range of between 10^{-10} and 10^{-1}	0.00025118864 (L1)
Transfer Learning: Further training on target domain						
Deletion	.99	.999	0.05	0.999	NA	0.0
Insertion	.99	.999	0.001	0.99	NA	0.0005011872 (L1)
Deletion-insertion	.99	.999	0.001	0.99	NA	0.00025118864 (L1)
Transfer Learning: Fine-tuning on target domain						
Deletion	.99	.999	0.0005	0.999	NA	0.0
Insertion	.99	.999	0.0001	0.99	NA	0.0005011872 (L1)
Deletion-insertion	.99	.999	0.0001	0.99	NA	0.00025118864 (L1)

Continued on next page

– continued from previous page

Model	β_1	β_2	Learning rate (LR)	LR decay (γ)	Penalty (optimised via 5-fold CV)	Penalty (final)

Tab. 2.S2: Hyperparameters for model training on source domain and transfer to target domains.

Model	Input	Model architecture	Predicts frequency of...
FORECasT [4]	3,633 binary features describing each potential indel	Logistic regression model	All deletion outcomes touching the nucleotide directly downstream of the cut site, up a deletion length of 30nt, and all insertions of up to 2nt in the -3/0 window upstream of the cut site.

Continued on next page

– continued from previous page

Model	Input	Model architecture	Predicts ratio of...
Lindel [8]	2,649 binary features describing MH locations within the target sequence and 384 binary features for the one-hot encoded target sequence	Three multioutput logistic regression models	All 536 of 550 possible deletion outcomes of length $<$30nt that overlap with the -3/+2 window around the cut site, 21 insertion outcomes including all single and dinucleotide insertions and insertions of length $\geq$ 3bp.
inDelphi [27]	3 features describing each MH deletion, one-hot encodings of the -3, -4, and -5 nucleotides (upstream of the PAM), and 2 features describing the predicted distribution of deletions	Combined two neural networks and k-nearest neighbour model	All MH-based deletions and all non-MH-based deletions grouped by deletion length, with a deletion length $<$ 60nt, and all single nucleotide insertions.

Continued on next page

– continued from previous page

Model	Input	Model architecture	Predicts ratio of...
CROTON [18]	One-hot encoded 60nt target sequence	Convolutional neural network	Aggregate categories of outcomes against all others. Six separate CROTON models trained for predictions of: deletions, 1nt insertions, 1nt deletions, 1nt frameshift mutations, 2nt frameshift mutations, frameshift mutations.

Tab. 2.S3: Repair outcome prediction model comparison. FORECasT, Lindel, and inDelphi predict repair outcome profiles. CROTON predicts the ratio of six aggregate categories of repair outcomes against all others.

Test set	Model	Performance of precision-X% prediction tasks (task: is the target precision-X%? yes/no)					
		X =20	30	40	50	60	70
		Precision performance score					
FORECasT WT mESC [4]	X-CRISP KLD	**0.746**	0.841	0.844	0.795	0.702	0.655
	X-CRISP MSE	**0.751**	0.831	0.828	0.766	0.695	0.653
	FORECasT	0.723	**0.853**	**0.858**	**0.836**	**0.755**	**0.739**
	Lindel	0.714	0.769	0.693	0.596	0.333	0.333
	inDelphi	0.400	0.447	0.238	0.086	0.059	0.029
inDelphi WT mESC [27]	X-CRISP KLD	0.726	0.702	0.569	0.396	**0.278**	0.000
	X-CRISP MSE	0.728	0.698	0.537	0.340	0.211	0.000
	FORECasT	**0.771**	**0.827**	**0.704**	**0.464**	0.211	0.000

Continued on next page

– continued from previous page

Test set	Model	Performance of precision-X% prediction tasks (task: is the target precision-X%? yes/no)					
		X =20	30	40	50	60	70
		Precision performance score					
	Lindel	0.682	0.617	0.512	0.3326	0.115	0.000
	inDelphi	0.363	0.313	0.071	0.022	0.000	0.000
		Recall performance score					
FORECasT	X-CRISP KLD	**0.643**	0.597	0.474	0.412	0.301	0.191
WT mESC	X-CRISP MSE	0.652	**0.617**	**0.518**	**0.458**	**0.378**	**0.250**
[4]	FORECasT	0.522	0.428	0.323	0.244	0.157	0.080
	Lindel	0.520	0.329	0.172	0.075	0.011	0.004
	inDelphi	0.559	0.368	0.280	0.173	0.154	0.093
inDelphi	X-CRISP KLD	0.712	0.771	0.669	0.462	0.333	0.000
WT mESC	X-CRISP MSE	**0.726**	**0.780**	**0.706**	**0.555**	**0.400**	0.000
[27]	FORECasT	0.613	0.572	0.460	0.359	0.216	0.000
	Lindel	0.648	0.551	0.412	0.271	0.162	0.000
	inDelphi	0.608	0.316	0.075	0.020	0.000	0.000
		Matthew's correlation coefficient (MCC)					
FORECasT	X-CRISP KLD	0.283	0.498	0.515	0.504	0.418	0.337
WT mESC	X-CRISP MSE	**0.298**	**0.501**	**0.534**	**0.519**	**0.468**	**0.386**
[4]	FORECasT	0.172	0.384	0.409	0.388	0.311	0.232
	Lindel	0.150	0.267	0.218	0.143	0.039	0.042
	inDelphi	-0.016	0.143	0.095	0.020	0.036	0.027
inDelphi	X-CRISP KLD	0.370	0.589	**0.535**	0.387	**0.292**	**-0.001**
WT mESC	X-CRISP MSE	**0.381**	**0.591**	0.528	**0.388**	0.275	-0.002
[27]	FORECasT	0.376	0.561	0.496	0.367	0.198	-0.003
	Lindel	0.253	0.387	0.365	0.268	0.146	-0.004
	inDelphi	0.055	0.166	0.010	-0.003	-0.006	0.000

Continued on next page

– continued from previous page

Test set	Model	Performance of precision-X% prediction tasks (task: is the target precision-X%? yes/no)					
		X =20	30	40	50	60	70
		Precision performance score					

Tab. 2.S4: Performance of five models for six different precision-X% prediction tasks and two test sets, according to three performance metrics. The goal of each prediction task was to predict if a given target had a precision-X% outcome or not for a specific value of X. We evaluated six precision-X% prediction tasks, each for a different value of $X \in \{20, 30, 40, 50, 60, 70\}$, using three performance metrics: precision, recall, and Matthew's correlation coefficient. All models were trained on the FORECasT wild-type mESC train set and were then tested separately on the FORECast and the inDelphi wild-type mESC test sets from Table 1 of the main article, as specified in the "Test set" column. Note that precision-X% refers to a target sequence for which a single CRISPR-induced repair outcome accounts for at least $X\%$ of all reads assigned to any outcome observed for that target sequence, as defined by Shen et al. in [27].

2.6 Supplementary Figures

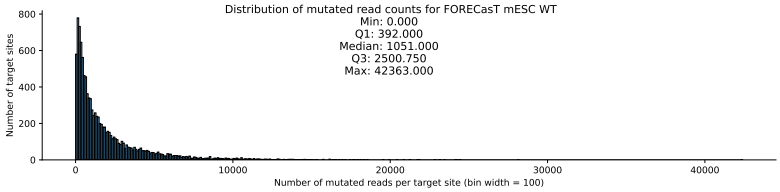


Fig. 2.S1: Distribution of mutated read counts per target site for the FORECasT mESC WT data. The horizontal axis shows the number of mutated reads assigned to each target site in bins of 100 in width. The vertical axis shows the number of target sites within each bin.

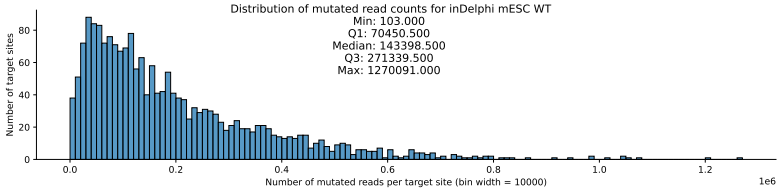


Fig. 2.S2: Distribution of mutated read counts per target site for the inDelphi mESC WT data. The horizontal axis shows the number of mutated reads assigned to each target site in bins of 10000 in width. The vertical axis shows the number of target sites within each bin.

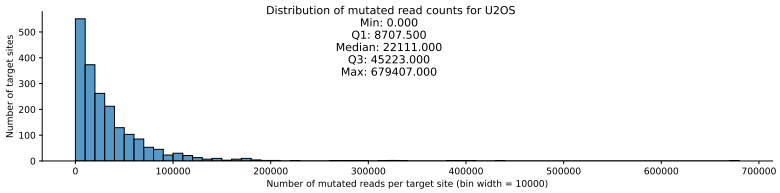


Fig. 2.S3: Distribution of mutated read counts per target site for the U2OS data. The horizontal axis shows the number of mutated reads assigned to each target site in bins of 10000 in width. The vertical axis shows the number of target sites within each bin.

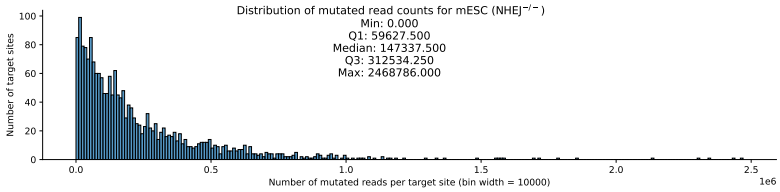


Fig. 2.S4: Distribution of mutated read counts per target site for the NHEJ-deficient mESC data. The horizontal axis shows the number of mutated reads assigned to each target site in bins of 10000 in width. The vertical axis shows the number of target sites within each bin.

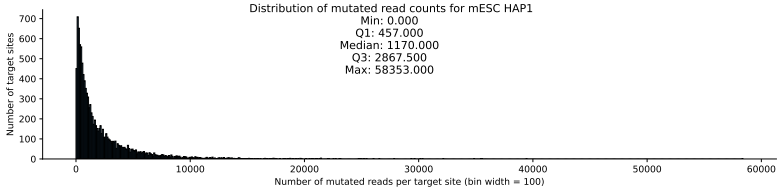


Fig. 2.S5: Distribution of mutated read counts per target site for the HAP1 data. The horizontal axis shows the number of mutated reads assigned to each target site in bins of 100 in width. The vertical axis shows the number of target sites within each bin.

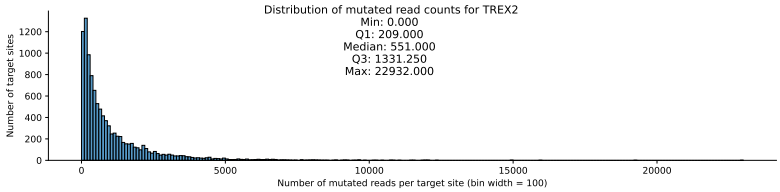


Fig. 2.S6: Distribution of mutated read counts per target site for the TREX2 data. The horizontal axis shows the number of mutated reads assigned to each target site in bins of 100 in width. The vertical axis shows the number of target sites within each bin.

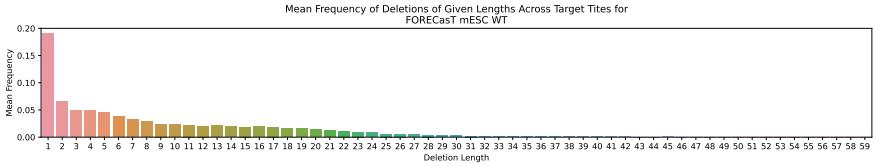


Fig. 2.S7: Mean frequency of deletions of given lengths across target sites for the FORECasT mESC WT data.

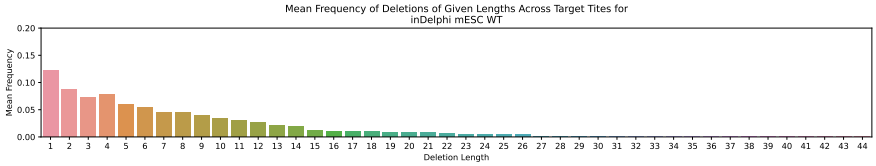


Fig. 2.S8: Mean frequency of deletions of given lengths across target sites for the inDelphi mESC WT data.

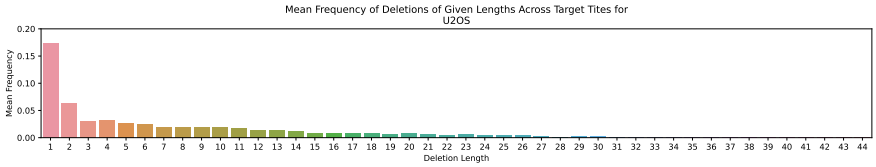


Fig. 2.S9: Mean frequency of deletions of given lengths across target sites for the U2OS data.

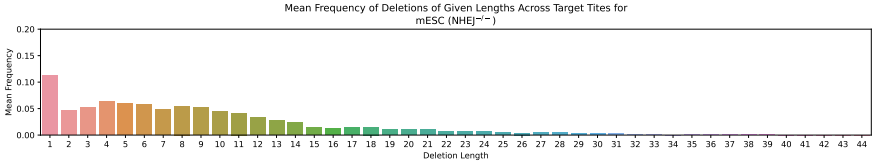


Fig. 2.S10: Mean frequency of deletions of given lengths across target sites for the NHEJ-deficient mESC data.

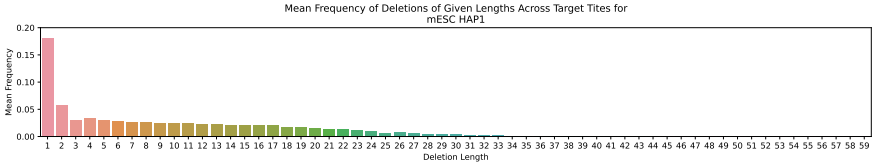


Fig. 2.S11: Mean frequency of deletions of given lengths across target sites for the HAP1 data.

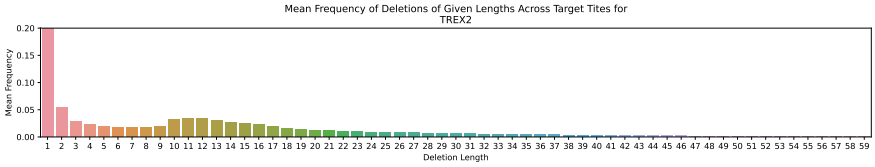


Fig. 2.S12: Mean frequency of deletions of given lengths across target sites for the TREX2 data.

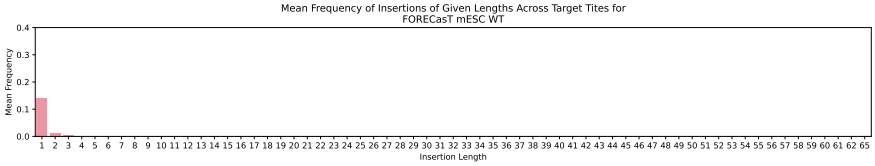


Fig. 2.S13: Mean frequency of insertions of given lengths across target sites for the FORECasT mESC WT data.

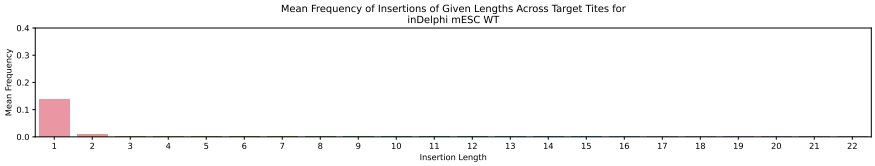


Fig. 2.S14: Mean frequency of insertions of given lengths across target sites for the inDelphi mESC WT data.

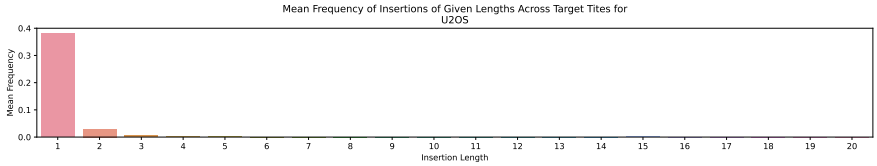


Fig. 2.S15: Mean frequency of insertions of given lengths across target sites for the U2OS data.

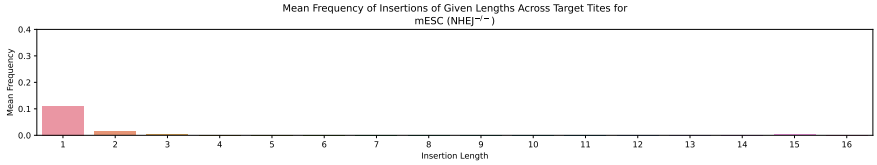


Fig. 2.S16: Mean frequency of insertions of given lengths across target sites for the NHEJ-deficient mESC data.

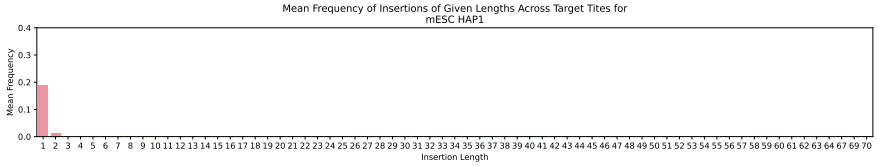


Fig. 2.S17: Mean frequency of insertions of given lengths across target sites for the HAP1 data.

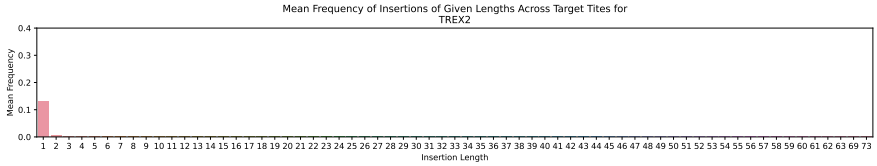


Fig. 2.S18: Mean frequency of insertions of given lengths across target sites for the TREX2 data.

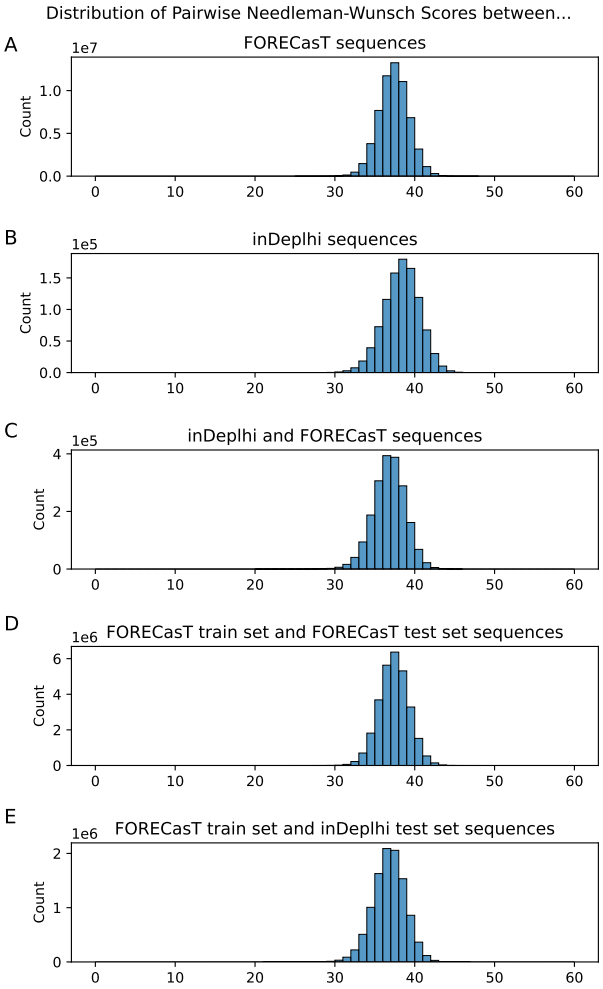


Fig. 2.S19: Pairwise Needleman-Wunsch scores between (A) sequences in the FORECasT dataset (B) sequences from the inDelphi dataset (C) sequences from the inDelphi dataset and the FORECasT dataset (D) sequences from the inDelphi dataset and the FORECasT dataset (E) sequences from the FORECasT train set and the inDelphi test set. All sequences are 60bp in length, centred at the cut site. Scoring parameters used: match: 1.0, mismatch: 0.0, gap: 0.0.

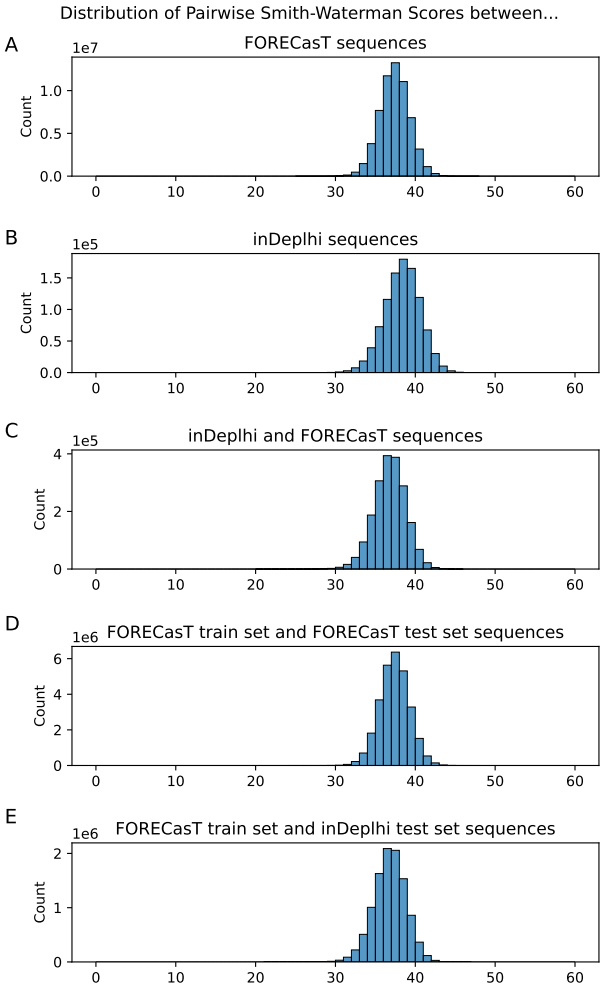


Fig. 2.S20: Pairwise Smith-Waterman scores between (A) sequences in the FORECasT dataset (B) sequences from the inDelphi dataset (C) sequences from the inDelphi dataset and the FORECasT dataset (D) sequences from the inDelphi dataset and the FORECasT dataset (E) sequences from the FORECasT train set and the inDelphi test set. All sequences are 60bp in length, centred at the cut site. Scoring parameters used: match: 1.0, mismatch: 0.0, gap: 0.0.

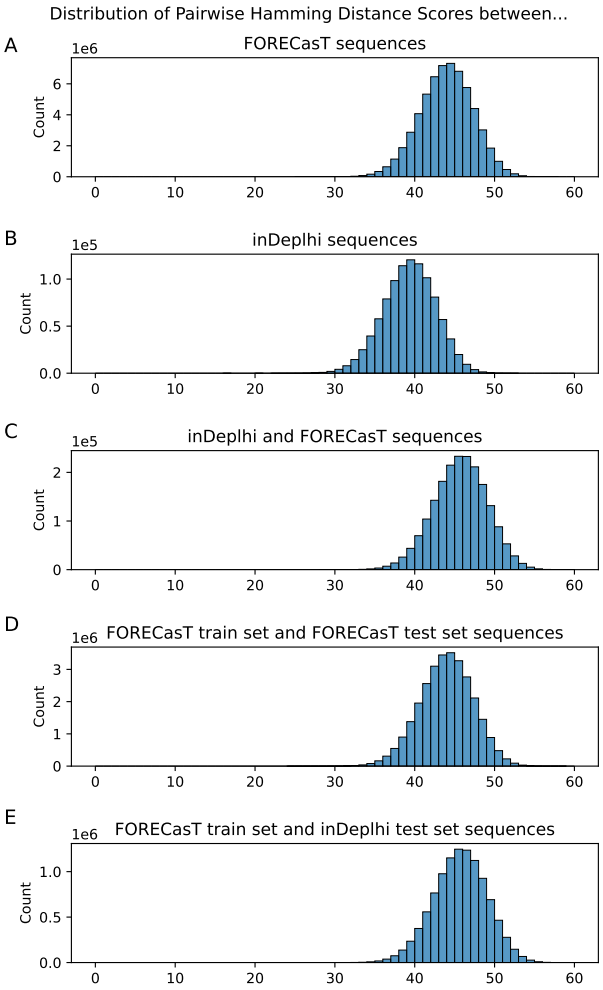


Fig. 2.S21: Pairwise Hamming distances between (A) sequences in the FORECasT dataset (B) sequences from the inDelphi dataset (C) sequences from the inDelphi dataset and the FORECasT dataset (D) sequences from the inDelphi dataset and the FORECasT dataset (E) sequences from the FORECasT train set and the inDelphi test set. All sequences are 60bp in length, centred at the cut site.

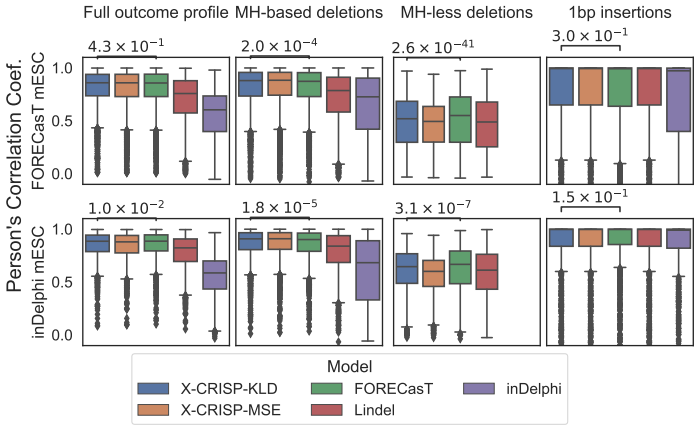


Fig. 2.S22: Detailed repair outcome prediction performance. Pearson's correlation coefficient between predicted and observed outcomes for (top) FORECasT or (bottom) inDelphi test data, considering: (left to right) original publication outcomes; common MH-*based* deletions; common MH-*less* deletions; 1bp insertions. Significance p-values calculated using Wilcoxon signed-rank tests, comparing X-CRISP KLD to the best of the non-X-CRISP models. For 1bp insertions, X-CRISP KLD and Lindel perform identically, given that they are based on the same model, so the comparison is then made with FORECasT, the next best of the non-X-CRISP models.

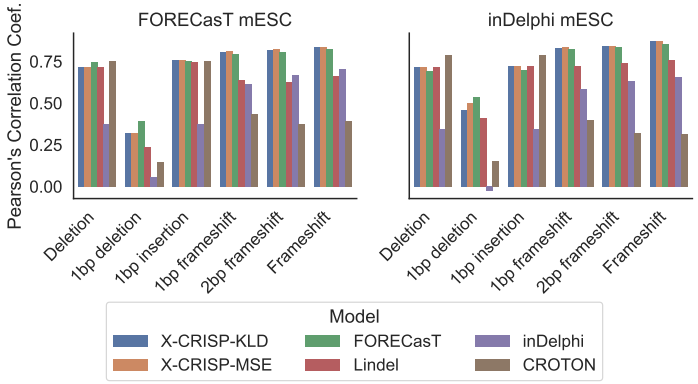


Fig. 2.S23: Broader repair outcome prediction performance. Pearson's correlation coefficient for six outcomes: deletion, 1bp insertion, 1bp deletion, 1bp frameshift, 2bp frameshift, and frameshift frequency prediction. Models trained on FORECasT WT mESC and tested on 3954 FORECasT/1961 inDelphi WT mESC target sites.

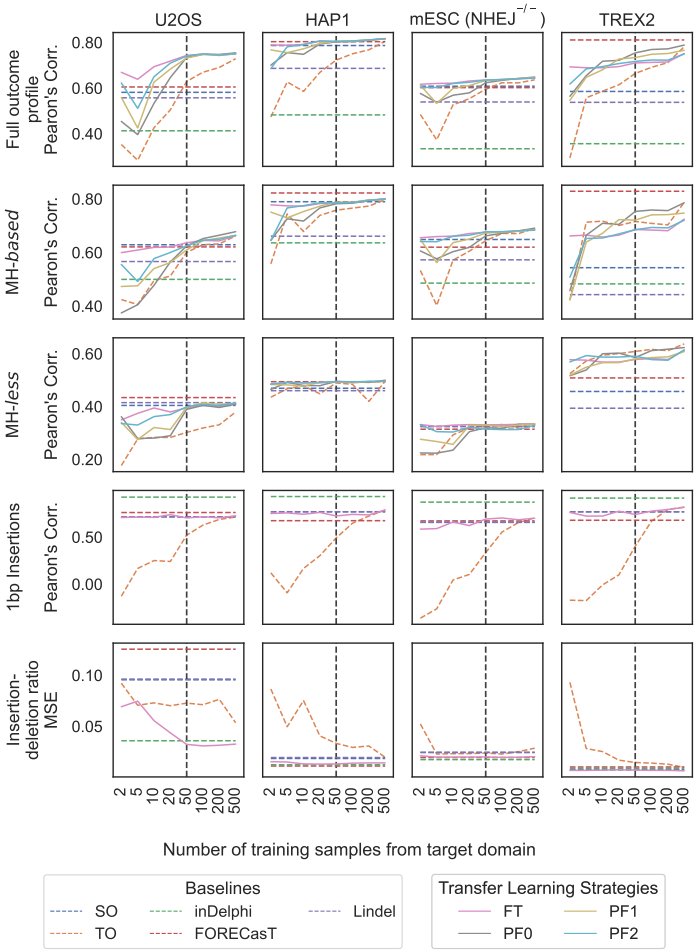


Fig. 2.S24: X-CRISP model adaptation to new domains or cell lines using transfer learning (TL). Prediction performance of baseline models and TL strategies, as the average Pearson's correlation or the MSE between predicted and observed frequencies per model and number of training samples. Baseline models: TO, X-CRISP trained on target only; SO, X-CRISP trained on source only; FORECasT, Lindel, inDelphi. Transfer learning: FT, pre-trained on source and fine-tuned for target; PF0-2, pre-trained on source and retrained + fine-tuned on target using 0-2 frozen hidden layers. (Top to bottom) Prediction models for full repair profile, MH-*based* deletions, MH-*less* deletions, insertions, and deletion-insertion ratios. Note: horizontal axis is not to scale; and the SO baseline does not use any samples from the target domain, so its performance remains constant along the horizontal axis.

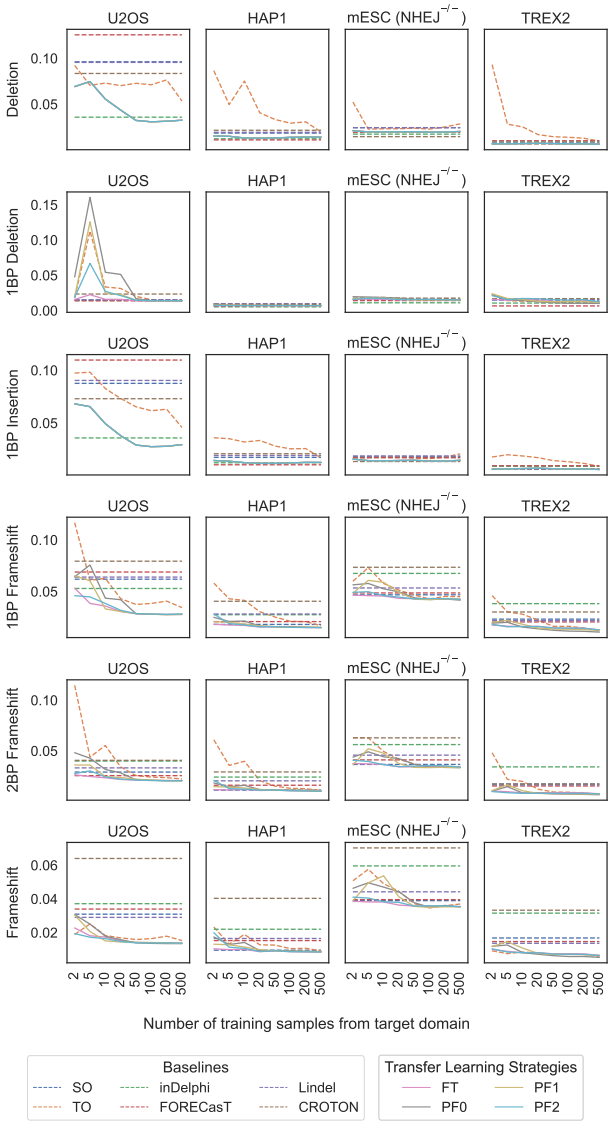


Fig. 2.S25: X-CRISP model adaptation via Transfer Learning on aggregate tasks evaluated via MSE. Prediction performance of baseline models and TL strategies, as the MSE between predicted and observed frequencies per model and number of training samples. Baseline models: TO, X-CRISP trained on target only; SO, X-CRISP trained on source only; FORECasT, Lindel, inDelphi, CROTON. Transfer learning: FT, pre-trained on source and fine-tuned for target; PF0-2, pre-trained on source and retrained + fine-tuned on target using 0-2 frozen hidden layers. (Top to bottom) Prediction models for frequency of deletions, 1bp deletions, 1bp insertions, 1bp frameshift, 2bp frameshift, and any frameshift. Note: The horizontal axis is not to scale, and the SO, FORECasT, inDelphi, Lindel, and CROTON baselines do not use any samples from the target domain, so its performance remains constant along the horizontal axis.

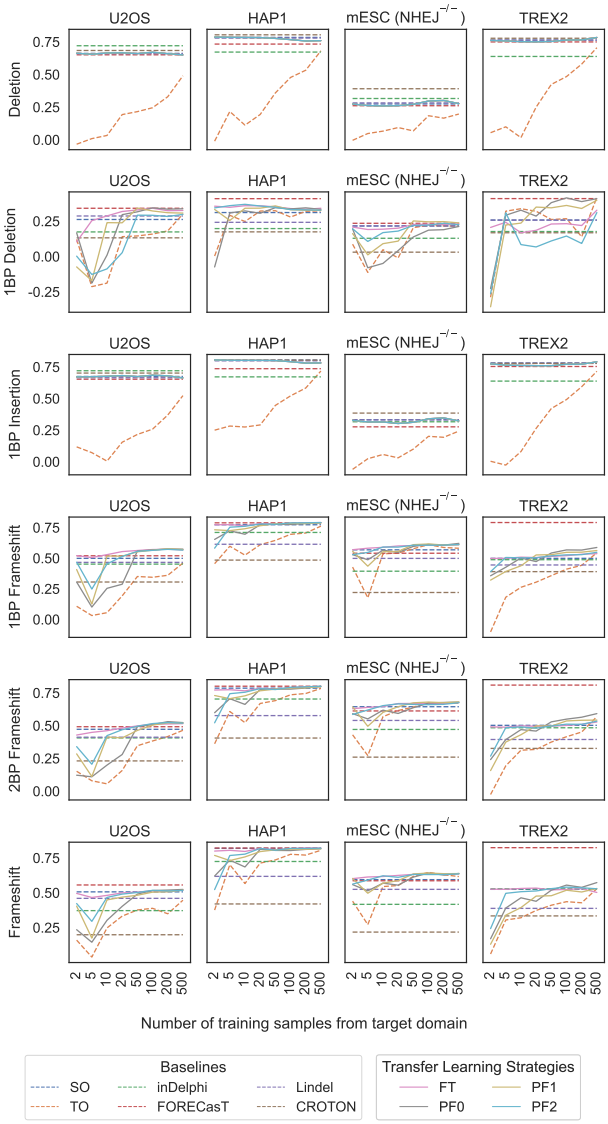


Fig. 2.S26: X-CRISP model adaptation via Transfer Learning on aggregate tasks evaluated via Pearson's correlation. Prediction performance of baseline models and TL strategies, as the Pearson's correlation between predicted and observed frequencies per model and number of training samples. Baseline models: TO, X-CRISP trained on target only; SO, X-CRISP trained on source only; FORECasT, Lindel, inDelphi, CROTON. Transfer learning: FT, pre-trained on source and fine-tuned for target; PF0-2, pre-trained on source and retrained + fine-tuned on target using 0-2 frozen hidden layers. (Top to bottom) Prediction models for frequency of deletions, 1bp deletions, 1bp insertions, 1bp frameshift, 2bp frameshift, and any frameshift. Note: The horizontal axis is not to scale, and the SO, FORECasT, inDelphi, Lindel, and CROTON baselines do not use any samples from the target domain, so its performance remains constant along the horizontal axis.

References

[1] Felicity [dataset]* Allen, Luca Crepaldi, Clara Alsinet, et al. “Predicting the mutations generated by repair of Cas9-induced double-strand breaks”. In: *European Nucleotide Archive. PRJEB29746* (2018) (cit. on p. 22).

[2] Max W [dataset]* Shen, Mandana Arbab, Jonathan Y Hsu, et al. “Deep sequencing of Cas9 editing outcomes in mouse cells”. In: *NCBI Sequence Read Archive. SRP141144* (2018) (cit. on p. 22).

[3] Mazhar Adli. “The CRISPR tool kit for genome editing and beyond”. In: *Nature Communications* 9.1 (2018), pp. 1–13 (cit. on p. 20).

[4] Felicity Allen, Luca Crepaldi, Clara Alsinet, et al. “Predicting the mutations generated by repair of Cas9-induced double-strand breaks”. In: *Nature Biotechnology* 37.1 (2019), pp. 64–72 (cit. on pp. 21–24, 27, 36, 42, 44, 46, 47).

[5] Dipankan Bhattacharya, Chris A Marfo, Davis Li, Maura Lane, and Mustafa K Khokha. “CRISPR/Cas9: An inexpensive, efficient loss of function tool to screen human disease genes in Xenopus”. In: *Developmental Biology* 408.2 (2015), pp. 196–204 (cit. on p. 20).

[6] Dana Cahill, Brian Connor, and James P Carney. “Mechanisms of eukaryotic DNA double-strand break repair”. In: *Frontiers in Bioscience (Landmark Ed)* 11.2 (2006), pp. 1958–1976 (cit. on p. 22).

[7] Mohammad Chehelgerdi, Matin Chehelgerdi, Milad Khorramian-Ghahfarokhi, et al. “Comprehensive review of CRISPR-based gene editing: mechanisms, challenges, and applications in cancer therapy”. In: *Molecular cancer* 23.1 (2024), p. 9 (cit. on p. 22).

[8] Wei Chen, Aaron McKenna, Jacob Schreiber, et al. “Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair”. In: *Nucleic Acids Research* 47.15 (2019), pp. 7989–8003 (cit. on pp. 21, 22, 24, 27, 36, 45).

[9] Santiago Gisler, Joana P. Gonçalves, Waseem Akhtar, et al. “Multiplexed Cas9 targeting reveals genomic location effects and gRNA-based staggered breaks influencing mutation efficiency”. In: *Nature Communications* 10.1 (Dec. 2019), p. 1598 (cit. on p. 36).

[10] Patrick D Hsu, Eric S Lander, and Feng Zhang. “Development and applications of CRISPR-Cas9 for genome engineering”. In: *Cell* 157.6 (2014), pp. 1262–1278 (cit. on p. 20).

[11] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, et al. “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity”. In: *Science* 337.6096 (2012), pp. 816–821 (cit. on p. 20).

[12] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 25).

[13] Hiroko Koike-Yusa, Yilong Li, E-Pien Tan, Martin Del Castillo Velasco-Herrera, and Kosuke Yusa. “Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library”. In: *Nature Biotechnology* 32.3 (2014), pp. 267–273 (cit. on p. 20).

[14] Solomon Kullback and Richard A Leibler. “On information and sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86 (cit. on p. 25).

[15] Ryan T Leenay, Amirali Aghazadeh, Joseph Hiatt, et al. “Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells”. In: *Nature Biotechnology* 37.9 (2019), pp. 1034–1037 (cit. on pp. 21, 22, 27).

[16] Alan R Lehmann and Elaine M Taylor. “Conservation of eukaryotic DNA repair mechanisms”. In: *DNA Damage and Repair*. Springer, 2001, pp. 377–401 (cit. on p. 22).

[17] Rasko Leinonen, Ruth Akhtar, Ewan Birney, et al. “The European Nucleotide Archive”. In: *Nucleic Acids Research* 39.suppl_1 (2010), pp. D28–D31 (cit. on p. 23).

[18] Victoria R Li, Zijun Zhang, and Olga G Troyanskaya. “CROTON: an automated and variant-aware deep learning framework for predicting CRISPR/Cas9 editing outcomes”. In: *Bioinformatics* 37.Supplement_1 (2021), pp. i342–i348 (cit. on pp. 21, 22, 27, 32, 46).

[19] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: (2017). Ed. by I. Guyon, U. V. Luxburg, S. Bengio, et al., pp. 4765–4774 (cit. on pp. 22, 28).

[20] Amy Maxmen. “Faster, better, cheaper: the rise of CRISPR in disease detection”. In: *Nature* 566.7745 (2019), pp. 437–438 (cit. on p. 20).

[21] Thomas Naert, Dieter Tulkens, Nicole A Edwards, et al. “Maximizing CRISPR/Cas9 phenotype penetrance applying predictive modeling of editing outcomes in Xenopus and zebrafish embryos”. In: *Scientific Reports* 10.1 (2020), pp. 1–12 (cit. on p. 22).

[22] Saul B Needleman and Christian D Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of molecular biology* 48.3 (1970), pp. 443–453 (cit. on p. 24).

[23] Megan van Overbeek, Daniel Capurso, Matthew M Carter, et al. “DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks”. In: *Molecular Cell* 63.4 (2016), pp. 633–646 (cit. on pp. 20, 36).

[24] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2009), pp. 1345–1359 (cit. on p. 22).

[25] Robin van Schendel, Joost Schimmel, and Marcel Tijsterman. “SIQ: easy quantitative measurement of mutation profiles in sequencing data”. In: *NAR Genomics and Bioinformatics* 4.3 (2022), lqac063 (cit. on p. 23).

[26] Ralph Scully, Arvind Panday, Rajula Elango, and Nicholas A Willis. “DNA double-strand break repair-pathway choice in somatic mammalian cells”. In: *Nature Reviews Molecular Cell Biology* 20.11 (2019), pp. 698–714 (cit. on p. 20).

[27] Max W Shen, Mandana Arbab, Jonathan Y Hsu, et al. “Predictable and precise template-free CRISPR editing of pathogenic variants”. In: *Nature* 563.7733 (2018), pp. 646–651 (cit. on pp. 21, 22, 27, 32, 36, 42, 45–48).

[28] Xin Shi, Jia Shou, Mohammadreza M Mehryar, et al. “Cas9 has no exonuclease activity resulting in staggered cleavage with overhangs and predictable di-and tri-nucleotide CRISPR insertions without template donor”. In: *Cell Discovery* 5.1 (2019), pp. 1–4 (cit. on p. 36).

[29] Temple F Smith, Michael S Waterman, et al. “Identification of common molecular subsequences”. In: *Journal of molecular biology* 147.1 (1981), pp. 195–197 (cit. on p. 24).

[30] Chuanqi Tan, Fuchun Sun, Tao Kong, et al. “A survey on deep transfer learning”. In: *International Conference on Artificial Neural Networks*. Springer. 2018, pp. 270–279 (cit. on pp. 22, 29).

[31] Tim Wang, Jenny J Wei, David M Sabatini, and Eric S Lander. “Genetic screens in human cells using the CRISPR-Cas9 system”. In: *Science* 343.6166 (2014), pp. 80–84 (cit. on p. 20).

[32] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big Data* 3.1 (2016), pp. 1–40 (cit. on p. 22).

[33] Xiaoqiao Yue, Chenjun Bai, Dafei Xie, Teng Ma, and Ping-Kun Zhou. “DNA-PKcs: A multi-faceted player in DNA damage response”. In: *Frontiers in Genetics* (2020), p. 1692 (cit. on p. 23).

[34] Jiajie Zhang, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. “PEAR: a fast and accurate Illumina Paired-End reAd mergeR”. In: *Bioinformatics* 30.5 (2014), pp. 614–620 (cit. on p. 23).

3

MUSICiAn: Genome-wide Identification of Genes Involved in DNA Repair via Control-Free Mutational Spectra Analysis

> *The intellectual question has remained the same over the last 20-something years: How does DNA repair work? The complexity is quite astounding.*
>
> — **Patrick Sung**
> (Interview with American Society for Biochemistry and Molecular Biology Today)

Understanding the factors involved in DNA double-strand break (DSB) repair is crucial for the development of targeted anti-cancer therapies, yet the roles of many genes remain unclear. Recent studies show that perturbations of certain genes can alter the distribution of sequence-specific mutations left behind after DSB repair. This suggests that genome-wide screening could reveal novel DSB repair factors by identifying genes whose perturbation causes the mutational distribution spectra observed at a given DSB site to deviate significantly from the wild-type. However, designing proper controls for a genome-wide perturbation screen could be challenging. We explore the idea that a genome-wide screen might allow us to forgo the use of traditional non-targeting controls by reframing the analysis as an outlier detection problem, assuming that most genes have

Colm Seale, Marco Barazas, Robin van Schendel, Marcel Tijsterman, and Joana P. Gonçalves. "MUSICiAn: Detecting Gene-DNA Repair Associations via Control-Free Mutational Spectra Analysis." bioRxiv, 10.1101/2025.01.27.635038

minimal influence on DSB repair. We propose MUSICiAn (Mutational Signature Catalogue Analysis), a compositional data analysis method that ranks gene perturbation-specific mutational spectra without controls by measuring deviations from the central tendency in the distributions of all spectra. We show that MUSICiAn can effectively estimate pseudo-controls for the existing Repair-seq dataset, screening 476 genes and 60 non-targeting controls. We further apply MUSICiAn to a genome-wide dataset profiling mutational outcomes induced by CRISPR-Cas9 at three target sites across cells with individual perturbations of 18,406 genes. MUSICiAn successfully recovers known genes, highlights the spliceosome as a lesser-appreciated player in DSB repair, and reveals candidates for further investigation.

3.1 Introduction

Double-strand breaks (DSBs) in DNA are critical cellular events that occur spontaneously due to endogenous processes like replication or external agents like ionizing radiation. Left unaddressed, DSBs can lead to genomic instability and eventually cell death or cancer [9]. As a result, cells have evolved a suite of mechanisms to repair DSBs, including the non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ, also called alt-NHEJ), and homology-directed repair (HDR) pathways [38]. Understanding the roles that genes play in DSB repair can importantly contribute to the development of targeted therapies for diseases such as cancer [45, 5]. For example, PARP inhibitor drugs are indicated to treat cancers with impaired HDR or BRCA gene function, whose synthetic lethality with PARP is leveraged to block DSB repair and cause a fatal accumulation of DNA damage in HDR- or BRCA-deficient cancer cells [10]. The ability to discover further opportunities for targeted therapy requires deeper insight into gene function, yet for many genes the link with DSB repair remains unclear.

In searching for these links, research has turned to large-scale gene functional screens enabled by CRISPR technology [5]. Originally, functional screens for DSB repair focused on the effect of gene silencing or knockout on readouts such as cell growth and proliferation to identify repair factors [18, 42, 50, 30, 32]. While valuable to characterize gene essentiality, inhibition of cellular growth is only indirectly related to DSB repair and could lead to results confounded by other mechanisms of cellular activity. For more precise readouts and biological insights, recent advances use CRISPR targeting to induce DSBs

and deep sequencing to analyze how disruption of gene function alters the mutational spectra, or the frequency distributions of mutations arising at DSB sites following repair [33, 47, 7, 40, 41, 19]. Multiple studies have demonstrated that knockouts of certain genes yield distinct, sequence-specific mutational spectra [33, 40, 19], but focused on the screening of known DSB repair genes. Notably, the first genome-wide study characterizing the effect of gene perturbation on mutational spectra will soon be released. We obtained early access to the data from this study, termed Mutational Signature Catalogue (MUSIC), to be made available upon publication.

Using CRISPR targeting with mutational spectra as readout, the primary approach to link genes to DSB repair is to quantify how much the mutational spectrum deviates from the expected wild-type distribution following the knockout of each individual gene. The larger the deviation, the higher the confidence that the perturbed gene has an effect on the outcomes and could thus be involved in DSB repair. Recent work by [19] defined this deviation as the "overall outcome redistribution activity", quantified by a chi-squared-like statistic relying on non-targeting controls to determine the expected wild-type spectrum.

For genome-wide screens, a limited set of non-targeting controls might not be suitable. While the majority of genes is not expected to produce an effect on the mutational spectra, it is unclear if targeting such genes could indirectly or mildly influence the outcomes, an effect which would not be appropriately captured by non-targeting controls. At the same time, it would be challenging to design realistic controls for all levels of variation at play in a genome-wide screen, while trying to maximize the coverage per mutational spectra and mitigate batch effects. We explore an alternative approach leveraging the assumption that most genes in a genome-wide perturbation screen have minimal impact on the mutational spectra to frame the identification of DSB repair genes as an outlier detection problem [2], and investigate if it can forgo the need for conventional controls.

When analyzing mutational spectra, it is also important to consider their compositional nature. In other words, each mutational spectrum is a distribution of relative frequencies over a collection of mutation categories whose overall sum is one. This composition property introduces a negative correlation bias caused by dependencies between the different frequencies, where an increase for one mutation type necessarily causes a reduction in others. Ignoring the

dependencies in compositional data using standard data analysis techniques can lead to misleading results and interpretation [3]. Additionally, the covariance structure of mutational spectra is likely to be skewed by the outlier gene knockouts that significantly affect DSB repair, emphasizing the need for methods tailored for compositional data analysis.

We introduce MUSICiAn (Mutational Signature Catalogue Analysis), a computational approach to score gene associations with DSB repair via genome-wide mutational spectra analysis. MUSICiAn operates without non-targeting controls, framing the task as an outlier detection problem under the assumption that most genes do not influence DSB repair. MUSICiAn uses the compositional data analysis (CoDA) framework to address dependencies and outliers in genome-wide mutational spectra data, for an improved estimation of pseudo-controls. By ranking gene knockouts based on their robust deviation from the overall mutational spectra distribution, MUSICiAn provides a control-free approach for genome-wide discovery of DSB repair-related genes.

We evaluate the MUSICiAn estimation of pseudo-controls on the Repair-seq dataset, screening 476 DSB genes and 60 non-targeting controls [19]. We further apply MUSICiAn to the genome-wide MUSIC mutational spectra dataset, covering 18,406 genes, to investigate the ability of this control-free method to recover established repair genes and suggest new candidates for experimental validation.

3.2 Methods

We introduce the MUSICiAn method using outlier detection to identify DSB repair genes from genome-wide CRISPR mutational spectra without traditional controls. The aim is to quantify the effect that each gene knockout produces on the mutational spectra relative to the expected wild-type or control spectra. In the absence of controls, MUSICiAn leverages the assumption that most genes are not involved in DSB repair to estimate the center of the mutational spectra distribution as a representative point, close to which the spectra will be most alike the expected wild-type. To quantify the deviation, MUSICiAn calculates a distance between each spectra and the estimated center also taking the covariance of the spectra distribution into account. This is done using a combination of data transformation and robust covariance estimation

designed to address dependencies and outliers in the mutational spectra data. Finally, MUSICiAn creates a unified gene outlier score based on the distances obtained across target sites.

3.2.1 Data and preprocessing

Mutational outcome data. We analyzed data from two gene perturbation screens with CRISPR-induced mutational outcome readout, Repair-seq and MUSIC (Supplementary Fig. 3.S1 for an illustration of the experimental setup). The Repair-seq screen used CRISPR interference with each of 1,573 single-guide RNAs (sgRNAs) to individually silence each of 476 DSB repair genes, and 60 non-targeting control sgRNAs [19]. To generate mutational outcomes, Repair-seq used CRISPR-Cas9 to create DSBs for a single target site across the population of cells with and without silenced genes, in two biological replicates. The genome-wide MUSIC screen was similarly set up, but used CRISPR knockouts rather than interference, with 89,571 sgRNAs spanning 18,406 genes, and generated outcomes for three target sites in two biological replicates each. We downloaded the raw Repair-seq sequence data [19] from the NCBI Sequence Read Archive, Bioproject PRJNA746980, runs SRR15164738 and SRR15164739. We also obtained early access to the MUSIC sequence data, to be made available upon publication.

We called mutations from the sequence data using the Sequence Interrogation and Quantification (SIQ) tool [35] v4.3 with parameters "-m 2 -c -e 0.05", specifying a minimum number of 2 reads for counting an event, the collapsing of identical events to a single record with the sum of counts, and a maximum permitted base error rate of 0.05. The SIQ tool mapped the reads to the sgRNAs used for gene perturbation and identified mutations observed at the CRISPR-Cas9-induced DSB sites.

Mutation aggregation and categorization. To reduce sparsity and improve statistical power, the fine-grained mutational outcomes output by the SIQ tool were aggregated into 8 higher-level categories: *wild-type*, denoting a sequence without mutations; *deletion with 1\2\3+bp\no microhomology*, for a deletion overlapping the cut site with a microhomology (MH) of length 1bp to 3+bp or no MH at all, where an MH is a short homologous sequence on both sides of the DSB and used for repair by MMEJ [24]; *insertion*, for a new sequence added at the cut site; *deletion with insertion*, for a combination of deletion and insertion; and *homology-directed repair*, for any insertion

matching the donor template DNA (Supplementary Table 3.S3 for SIQ vs. MUSICiAn categories). Any other rare mutation types, such as single-base substitutions, were excluded, since they are not typical outcomes of DSB repair. Wild-type reads were also excluded to avoid confounding by gene essentiality, as a decrease in wild-type read abundance could indicate that the gene was essential for survival, but not if it was relevant for DSB repair. We thus considered a final set of 7 mutation categories.

Quality analysis and filtering of perturbation sgRNAs. For each replicate, we filtered out perturbation sgRNAs yielding a total read count below 700 across the 7 mutation categories, and excluded genes with less than 3 associated perturbation sgRNAs. Additionally, we controlled for inconsistencies in the effect of the different sgRNAs used for perturbation of the same gene, which could be indicative of sgRNA off-target effects, less effective gene perturbation, or any other undesirable effect. We excluded sgRNAs whose count profiles over the 7 mutation categories showed a median pairwise Pearson's correlation below 0.6 with the profiles for other sgRNAs perturbing the same gene within the same replicate (and target site), or a median pairwise Pearson's correlation below 0.6 with replicate profiles for the same sgRNA and target site (Supplementary Tables 3.S1 and 3.S2). To avoid numerical issues with the data transformation applied by MUSICiAn later on, in the rare cases where some mutation categories had zero counts, we imputed real values drawn independently from a uniform distribution between the detection limit DL and $0.1 \times DL$, where $DL = 1$. [26].

Generating mutational spectra per gene. We first computed mutational spectra by dividing the count of each of the 7 mutation categories by the total per sgRNA and replicate. Then we aggregated across sgRNAs by calculating the geometric mean of the sgRNAs-associated spectra per gene and replicate, producing replicate spectra per gene (two for each target site). Finally, we computed the geometric mean between replicate spectra per target site, resulting in one mutational spectrum per gene and target site. After every aggregation step, the frequencies in each mutational spectrum were divided by the sum to ensure they summed to one.

3.2.2 MUSICiAn scoring of gene effect on mutational spectra

The MUSICiAn method scores genes for DSB repair association by computing the distance between the mutational spectrum of each perturbed gene and

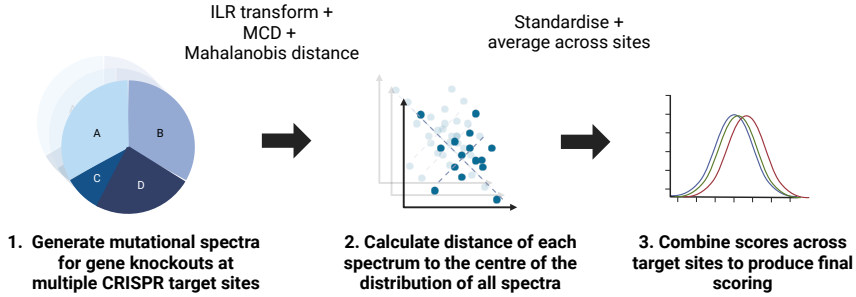


Fig. 3.1: Overview of MUSICiAn scoring of gene effect on mutational spectra. The method quantifies the effect of gene perturbation using the Mahalanobis distance of the gene mutational spectrum to the estimated center of the spectra distribution of all genes, under the assumption that most genes have a negligible effect. Estimation is improved using ILR-transformed spectra and robust covariance (MCD) to mitigate biases from data closure and outliers. Distances are normalized and averaged across target sites to produce a unified gene effect score.

the estimated center of the distribution of all spectra obtained for a given target site (Fig. 3.1). For experiments with multiple targets, target-specific scores can be normalized and averaged to produce one single gene score. Genes with larger scores have a more prominent effect on the mutational spectra, thus also a higher likelihood of being involved in DSB repair. Genes with the lowest scores are assumed to approximate the wild-type or control distribution. We are interested in both the most outlying and the most central spectra for downstream analysis.

Gene scoring. To calculate gene effect scores, MUSICiAn computes the Mahalanobis distance [27] per gene spectrum relative to the overall spectra distribution per target site, assuming that most genes are not directly involved in DSB repair and have negligible effect (Fig. 3.1). We chose the Mahalanobis distance as it takes the distribution and covariation of the data into account, unlike the Euclidean distance. Informative Mahalanobis distances require reliable covariance estimation, which is affected by: data closure, where dependencies between spectra categories summing to one introduces a negative correlation bias [3]; and outlier genes with a significant impact on mutational spectra and therefore also on the distribution.

To mitigate data closure, MUSICiAn applies an isometric log-ratio (ILR) transformation [15] to the mutational spectra using the defaults for scikit-bio 0.5.4. The ILR transformation maps the data from a constrained simplex space

to an unconstrained Euclidean space, allowing for independent statistical analysis of components. To mitigate outlier effects, MUSICiAn uses the minimum covariance determinant (MCD) as a robust covariance matrix estimator [16], using scikit-learn 1.2.1 defaults. The MUSICiAn method calculates the robust Mahalanobis distances for each ILR-transformed spectra, and unifies the individual distances into gene scores across target sites by: *(i)* selecting the common genes with mutational spectra in all target sites to act as a reference, *(ii)* calculating the mean and standard deviation of the distances of the reference genes per target site, *(iii)* normalizing all gene distances per site by subtracting the means and dividing standard deviations to place them on a common scale, and *(iv)* averaging the normalized gene distances across sites, ignoring missing values, to produce a final unified gene score.

Pseudo-control selection. The target-specific distances and unified gene scores enable the selection of "pseudo-controls" as the lowest-scoring genes per target or common across target sites. These pseudo-controls enable comparative analyses by estimating the central tendency of traditional controls, but may not recapitulate their natural variation.

3.2.3 Evaluation

Before applying MUSICiAn to the genome-wide MUSIC dataset, we assessed the outlier detection and pseudo-control selection on the Repair-seq dataset, the only other dataset available of CRISPR mutational spectra for multiple individual gene knockouts. While not a genome-wide dataset, Repair-seq included non-targeting controls, allowing us to assess if and how well MUSICiAn could estimate the wild-type distribution center. Furthermore, Repair-seq data focused on genes involved in DNA repair, so we also checked if MUSICiAn could recover similar mutational patterns for the genes screened in both studies. We preprocessed the Repair-seq data as described and held out the non-targeting controls from the scoring for later validation.

Estimation of pseudo-controls. We used PCA to visualize the effect of ILR transformation and robust MCD covariance, proposed to mitigate compositional data closure and outlier spectra, on the estimation of the mutational spectra distribution center and selection of pseudo-controls for the Repair-seq dataset. We applied PCA in four scenarios: *Classical Covariance*, using the original mutational spectra with the classical covariance estimation;

MCD Covariance, using the original spectra with the outlier-robust MCD covariance estimation; *ILR & Classical Covariance*, using ILR-transformed spectra with classical covariance estimation; and *ILR & MCD Covariance*, using ILR-transformed spectra with MCD covariance estimation. After ILR transformation, location and covariance estimation, we back-transformed the data to centered log-ratio (CLR) space to analyze the relation between PCA components and mutationcategories [16].

To evaluate pseudo-control selection, we identified 60 pseudo-controls for each scenario and calculated the Jensen-Shannon distance (JSD) between the geometric means of the non-targeting control and the pseudo-control spectra. As a baseline, we also calculated the distance from the non-targeting control spectra to the geometric mean across all gene-targeting sgRNAs, without pseudo-control selection. The JSD quantifies the distance between distributions, where a lower distance indicates greater similarity between distributions.

Cross-dataset estimation of pseudo-controls. To further assess the selection of pseudo-controls, we analyzed the consistency in mutational patterns retrieved for the MUSIC and Repair-seq datasets, using pseudo-controls estimated by MUSICiAn jointly from the two datasets. Specifically, we applied MUSICiAn to select 60 pseudo-controls for the set of all mutational spectra associated with the 434 genes shared across both datasets, with four target sites in total (three for MUSIC, one for Repair-seq). We then calculated the difference in mutation frequency per category between each gene-related mutational spectra and the geometric mean of the pseudo-controls. Finally, we performed hierarchical clustering [46] on the resulting difference matrix, using Ward cluster linkage and distance between samples based on Pearson's correlation.

Gene scoring and ranking performance. To evaluate the quality of the MUSICiAn-derived gene effect scores for the genome-wide MUSIC dataset, we examined if MUSICiAn could effectively recover genes with known links to DSB repair by scoring or ranking them higher than other genes based on their effect on the mutational spectra. We assessed performance separately using precision-recall (PR) curves against known DSB repair genes from two sources: 476 experimentally validated genes curated by Repair-seq for their AX227 CRISPRi library [19], and 295 genes whose annotations matched the regex "double-strand break repair|interstrand cross-link repair" (interstrand

cross-link repair genes often crosstalk with DSB repair pathways such as HR, [28]) in any field in the Gene Ontology [4, 12]. For baseline comparison, we calculated PR curves after randomly ranking all genes in the MUSIC dataset. We preferred PR rather than ROC curves, given that the dataset is highly imbalanced, where most genes have no known association with DSB repair and are therefore considered negative for the purpose of the evaluation.

Functional enrichment for top 500 ranked genes. We performed enrichment analysis for the top 500 genes ranked by MUSICiAn against the background of all genes in the MUSIC dataset, using the "gseapy" python package 1.0.4. We employed four sets of annotations, including KEGG pathways "KEGG_2019_Mouse" [21], and Gene Ontology terms across the three ontologies "GO_Biological_Process_2023", "GO_Molecular_Function_2023", "GO_Cellular_Component_2023". We performed a hypergeometric test per term within each gene set, and the resulting *p*-values were FDR corrected using the Benjamini-Hochberg method. [6]. We further estimated the effect of the genes annotated with each of the top 10 enriched terms or pathways on the mutation frequencies separately for the 4 annotation sets. To do this, we fitted an ordinary least squares (OLS) regression model per term t and mutation outcome category o to explain the variation in mutation frequency ($Frequency$) based on term or pathway membership ($Group$), according to the following R-style formula

$$Frequency_{g,o} \sim Group_{g,t}, \tag{3.1}$$

where each sample is a mutational spectra for a given gene knockout g. The $Frequency_{g,o}$ variable denotes the frequency of the given mutation outcome o for gene g, and $Group_{g,t}$ is a binary variable indicating if gene g is a member of term or pathway t. As case samples, we took the mutational spectra of all genes annotated with the enriched term in question. As control samples, we used the set of 100 pseudo-controls or lowest scoring genes, with valid mutational spectra across all target sites, and that were not members of any of the enriched pathways. We used the same control samples for the regression analysis of every annotation set, and report the regression coefficients and Benjamini-Hochberg corrected p-values for the $Group$ variable.

3.3 Results

3.3.1 MUSICiAn can estimate absent control mutational spectra

We first assessed the ability of MUSICiAn to estimate pseudo-control mutational spectra in the absence of actual controls. To do this, we applied MUSICiAn to the Repair-seq dataset, containing mutational spectra for one target site across knockouts of 476 different genes and 60 actual non-targeting controls. The actual controls were left out to be able to quantify how well they could be recovered by MUSICiAn. We also isolated the contributions of the ILR transformation and robust covariance (MCD) used by MUSICiAn to investigate if they improved the estimation of the distribution center location and covariance, and ultimately the selection of pseudo-controls, in the presence of outlier spectra and negative correlation bias. To visually examine the effect of ILR transformation and MCD on the distribution, we applied PCA to the original and ILR-transformed mutational spectra separately using classical PCA and a robust variant of PCA based on the MCD.

The estimated center of the distribution appeared to align the best with the center determined based on the actual controls (geometric mean of the 60 non-targeting controls) when both ILR and MCD were used to respectively address data closure and outliers in the mutational spectra data (Fig. 3.2A, "ILR & MCD Covariance" vs. others). We further observed that the pseudo-controls selected as the 60 mutational spectra closest to the center of the distribution estimated by MUSICiAn, using any of the four combinations of spectra and covariance types, were far more similar to the actual non-targeting controls than the average across all spectra. Specifically, the Jensen-Shannon (JS) distances between the geometric means of pseudo-controls and non-targeting controls were one order of magnitude smaller than those between the geometric means of all spectra and non-targeting controls (respectively < 0.005 and 0.012, Fig. 3.2B). Moreover, the selection of pseudo-controls using the preferred combination of techniques in MUSICiAn, ILR transformation and robust MCD covariance, produced the closest match with the actual non-targeting controls than the other three (JS distances 3.73×10^{-3} against 3.77×10^{-3}, 4.42×10^{-3}, and 4.69×10^{-3}; Fig. 3.2B). This result supported our choice to place ILR transformation and MCD at the core of the MUSICiAn outlier detection algorithm.

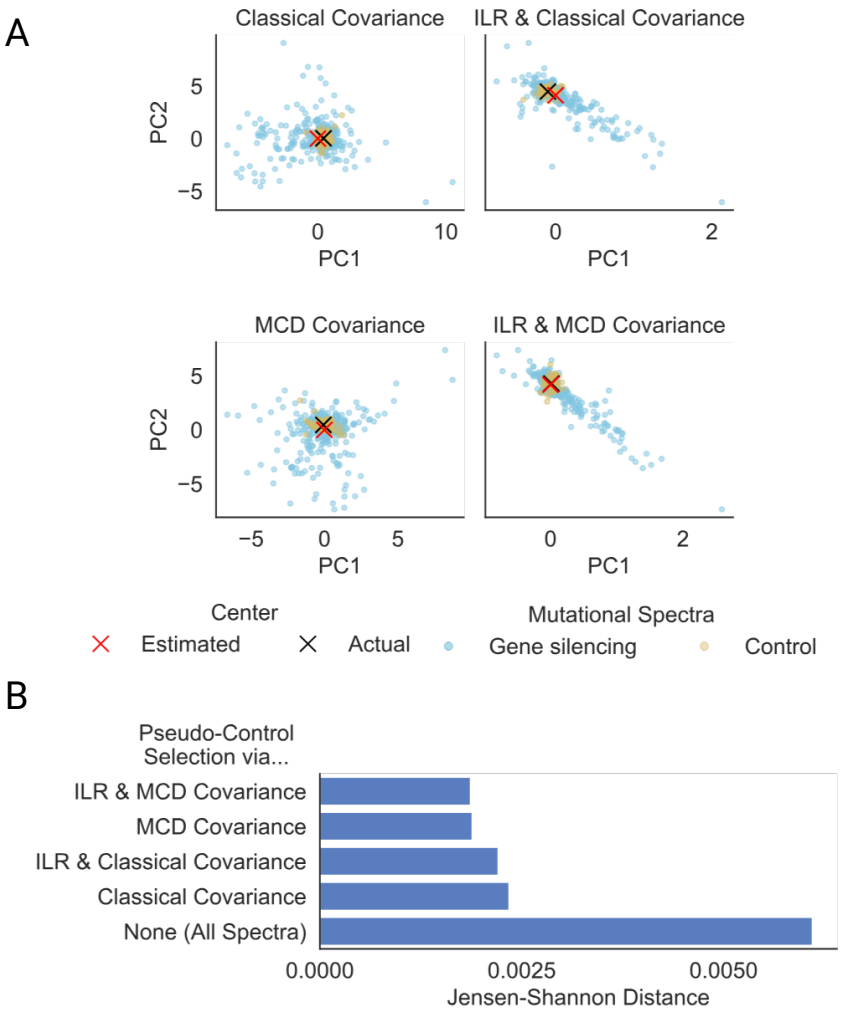


Fig. 3.2: Evaluation of MUSICiAn-selected pseudo-controls based on the estimated mutational spectra distribution center for Repair-seq. Effect of ILR transformation and MCD covariance on (A) the estimated center of the mutational spectra distribution and (B) the selected pseudo-controls, using the original or ILR transformed spectra with classical or MCD covariance. For (A), actual center (black cross) of the mutational spectra distribution as the geometric mean of the 60 actual controls (yellow points), and center estimated by MUSICiAn (red cross) based on the mutational spectra under gene-silencing (blue points), projected onto the two axes of largest variation in the data (first two principal components). For (B), Jensen-Shannon distances between the geometric means of the 60 actual non-targeting controls and either all mutational spectra or the 60 pseudo-controls closest to the center, estimated using each of the four combinations of spectra and covariance types.

We note that the majority of the 476 genes characterized in the Repair-seq screen are known to be involved in DNA repair, and therefore the assumption that most genes should not have an impact on the mutational spectra was in theory not necessarily met for this dataset. However, in practice, a large proportion of DNA repair genes still showed little effect on mutation frequencies (Fig. 3.2). The fact that MUSICiAn was able to recover controls in this scenario highlights that it could be applicable to more focused studies beyond genome-wide screens whenever a similar reasonable assumption can be made, for instance based on prior knowledge or the actual distribution of the data.

3.3.2 MUSICiAn controls reveal known repair patterns across studies

We further questioned if MUSICiAn could estimate pseudo-controls for mutational spectra aggregated from different studies, such that consistent mutational repair patterns would be revealed when applying the same controls as a baseline across the studies. To address this, we jointly analyzed the mutational spectra for knockouts of the 434 genes screened in both the genome-wide MUSIC and the focused Repair-seq studies. After selecting pseudo-controls, we calculated thedifferences between the frequencies in each mutational spectrum, obtained under silencing or knockout of a specific gene, and the geometric mean of the pseudo-controls (Fig. 3.3). We also performed hierarchical clustering of genes and mutation categories based on those differences (Fig. 3.3). The results revealed consistency in how HDR and insertion events were influenced by silencing of specific genes across targets and studies, as well as broadly consistent patterns for other mutation types with larger variations that could be attributed to differing target site-specific characteristics within and between studies.

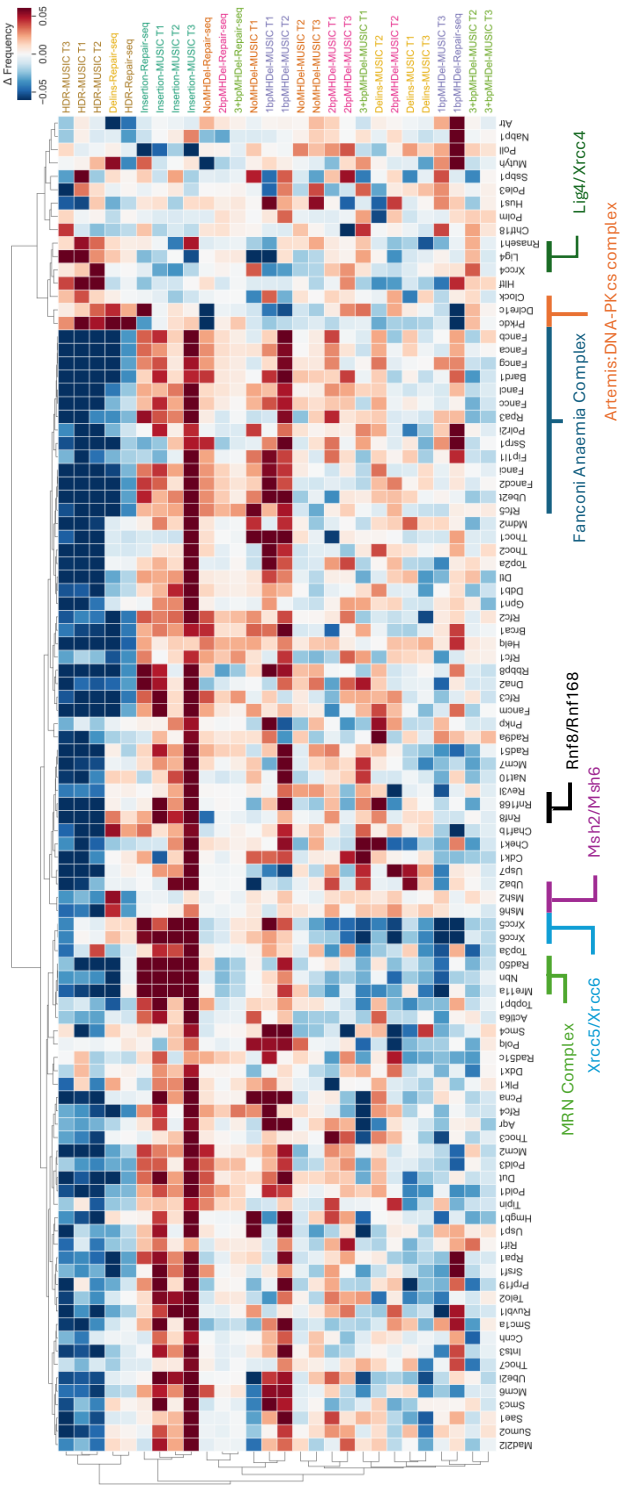

Δ Frequency
−0.05
0.00
0.05
HDR-MUSIC T3
HDR-MUSIC T1
HDR-MUSIC T2
Delins-Repair-seq
HDR-Repair-seq
Insertion-Repair-seq
Insertion-MUSIC T1
Insertion-MUSIC T2
Insertion-MUSIC T3
NoMHDel-Repair-seq
2bpMHDel-Repair-seq
3+bpMHDel-Repair-seq
NoMHDel-MUSIC T1
1bpMHDel-MUSIC T1
1bpMHDel-MUSIC T2
NoMHDel-MUSIC T2
NoMHDel-MUSIC T3
2bpMHDel-MUSIC T1
2bpMHDel-MUSIC T3
3+bpMHDel-MUSIC T1
Delins-MUSIC T2
2bpMHDel-MUSIC T2
Delins-MUSIC T1
Delins-MUSIC T3
1bpMHDel-MUSIC T3
1bpMHDel-Repair-seq
3+bpMHDel-MUSIC T2
3+bpMHDel-MUSIC T3
Madd2
Sumo2
Sae1
Smc3
Mcm6
Ube2l
Trioc7
Ints3
Ccnh
Smc1a
Ruvbl1
Telo2
Prpf19
Srsf1
Rpa1
Rfl1
Usp1
Hmgb1
Tipin
Pold1
Dut
Pold3
Mcm2
Trioc3
Agr
Rfc4
Pcna
Pik1
Dkn1
Rad51c
Pola
Smc4
Acd6a
Topbp1
Mre11a
Nbn
Rad50
Top3a
Xrcc5
Xrcc6
Msh6
Msh2
Ube2
Usp7
Chk1
Chek1
Chaf1b
Rnf8
Rnf168
Rev3l
Nat10
Mcm7
Rad51
Rad54l
Pnkp
Fancm
Rfc3
Dna2
Rpbp8
Rfc1
Helq
Brca1
Rfc2
Gpn1
Ddb1
Dll
Top2a
Trioc2
Trioc1
Mdm2
Rfc5
Ube2t
Fancd2
Fanci
Fip1l1
Ssrp1
Polr2i
Rpa3
Fanc
Fand1
Banf1
Fancg
Fanca
Fancb
Prkdc
Dclre1c
Clock
Hltf
Xrcc4
Lig4
Rnaseh1
Chtf18
Polm
Hus1
Pole3
Sbp1
Mutyh
Poll
Nabp1
Alr
MRN Complex
Xrcc5/Xrcc6
Msh2/Msh6
Rnf8/Rnf168
Fanconi Anaemia Complex
Artemis:DNA-PKcs complex
Lig4/Xrcc4

Fig. 3.3: Heatmap of the difference in mutation frequencies between each spectrum obtained for the knockout of a specific gene and the geometric mean of the pseudo-control spectra selected by MUSICiAn, per mutation category and target site. Shown are the top 100genes with the highest MUSICiAn outlier score across target sites (3 for MUSIC, denoted T1-T3, and 1 for Repair-seq). The horizontal axis represents genes. The vertical axis represents mutational outcomes, coloured by target site. Data was clustered on both dimensions, genes and mutation categories, using hierarchical clustering with Ward cluster linkage and distance between spectra based on Pearson's correlation coefficient.

Gene clustering also identified meaningful groups, including the Fanconi anemia core complex and related genes, whose silencing suppressed HDR events (Fig. 3.3). Interestingly, *Helq* displayed a mutational pattern similar to these genes, suggesting a potential association with FA and HDR, a topic of ongoing debate [20, 44]. Other notable clusters included: mismatch repair *MutS* homolog genes (*Msh2*, *Msh6*); ring finger protein genes with roles in DNA damage sensing and repair (*Rnf8*, *Rnf168*); NHEJ genes involved in early recognition of DNA damage and recruitment of additional repair factors (*Xrcc5*, *Xrcc6*), and in the processing of DNA ends (Artemis complex *Prkdc* and *Dclre1c*); and the MRN complex with roles in ATM checkpoint activation in response to DNA damage and also the tethering of broken DNA ends for further processing by NHEJ and HDR (*Mre11a*, *Rad50*, *Nbn*). The consistency in gene silencing effects on mutational spectra across the MUSIC and Repair-seq datasets, along with the identification of groups of genes with related function in DNA damage response, provided support for the effectiveness of the MUSICiAn control-free analysis in estimating pseudo-controls, quantifying effects, and ultimately generating meaningful insights from CRISPR targeting under gene silencing screens with mutational spectra readout.

3.3.3 MUSICiAn recovers known gene-DSB repair associations

In addition to estimating pseudo-controls, MUSICiAn attributes an outlier score to each gene, which determines the multivariate effect of gene silencing on mutational spectra to suggest (novel) associations between the gene and DNA damage response. In this context, we first applied MUSICiAn to the genome-wide MUSIC dataset to assess if it could recover known repair genes. Genes were ranked by their MUSICiAn outlier score, and the ranking

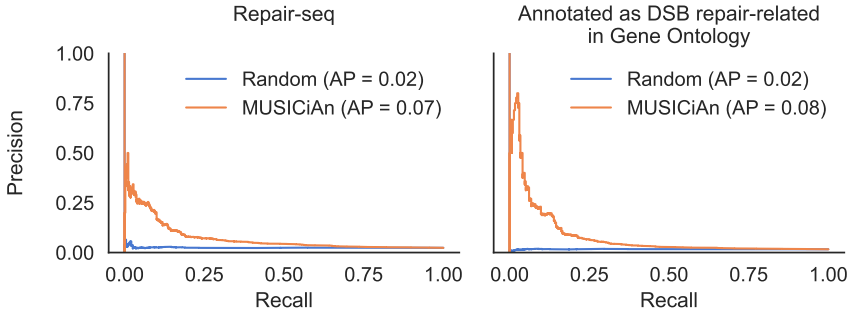


Fig. 3.4: Performance of MUSICiAn, recovery of known DNA repair genes. Precision-recall curves using MUSICiAn ranking (orange) or random ranking (blue) against the following gold standards for evaluation: (left) curated genes in Repair-seq and (right) genes annotated with DSB repair-related in Gene Ontology.

was evaluated against the set of 476 genes curated by Repair-seq and an alternative set of 295 genes retrieved from the Gene Ontology (GO). The closer to the top of the ranking these genes appeared, the better the results. We also performed the same evaluation on a randomly shuffled ranking as a baseline for comparison. The MUSICiAn method showed superior rankings for known associations with area under the precision-recall curve (AP) of 0.07 and 0.08 for the Repair-seq and GO gene sets, respectively, compared to an AP of 0.02 for the random baseline (Fig. 3.4).

Pathway enrichment analysis of the top 500 genes using KEGG annotations revealed significant associations with the "Fanconi anaemia" and "Homologous recombination" pathways (Fig. 3.5). A link with "Nucleotide excision repair" was also identified, supporting the idea that single and double-strand repair mechanisms are functionally intertwined [48]. Another enriched pathway, "Cell Cycle", is known to influence DNA repair pathway choice [48, 11]. Many DSB repair genes were also implicated in the "DNA replication" pathway [8, 13].

Functional enrichment analysis of the top 500 genes using GO annotations revealed links with repair-related biological processes (Fig. 3.5), including "DNA repair", "double-strand break repair", "double-strand break repair via homologous recombination", and "interstrand cross-link repair", further reinforcing the ability of MUSICiAn scores to capture and prioritize effects of genes on mutational spectra following the repair of CRISPR-induced DSB sites. Regarding molecular function, various binding activities, including

DNA, damaged DNA, and ubiquitin-like protein ligase binding, as well as single-strand DNA helicase activity were identified, all functions required for DNA damage signalling and repair [49, 36, 34] (Fig. 3.5).

Overall, MUSICiAn recovered known patterns and associations relevant to the repair of double-strand DNA breaks. While the AP performance may appear modest, it is significantly better than random. Nevertheless, mutational spectra exhibited relatively low coverage per sgRNA (median: MUSIC 2361.08 vs. Repair-seq 565201.97), leading to noisier mutational spectra that posed additional challenges in differentiating between true repair factors and noisy samples. Moreover, the assumption that mutational spectra deviating from the expected wild-type arise upon silencing of genes associated with DNA repair does not preclude the existence of other genes involved in DNA repair that do not affect mutational spectra. Such genes may not play a central role in the pathway, or their loss of function may be compensated by other genes, resulting in smaller effects and appropriately lower MUSICiAn rankings, while negatively biasing the AP.

3.3.4 MUSICiAn identifies lesser-appreciated players in DSB repair

After analyzing established genes and pathways, we also examined several lesser-recognized pathways and processes emerging from the MUSICiAn analysis of the MUSIC dataset. Intriguingly, "Ribosome biogenesis in eukaryotes" was the top enriched KEGG pathway (Fig. 3.5), aligning with emerging literature from the last decade suggesting a potential cross-talk between ribosome biogenesis and DNA repair pathways [31]. Recent studies have also implicated the nucleolus, a major site of ribosome synthesis and the top enriched cellular component, in the regulation of cellular processes, including DNA repair [25, 37, 22].

The proteasome and spliceosome were additionally identified as enriched pathways. The proteasome plays a role in the regulation of the *Rnf8-Rnf168* pathway, which itself works to recruit repair factors to DSB sites [36, 23], and the inhibition of which has been previously shown to reduce HDR events [14]. As for the spliceosome, there is growing evidence of a role in DNA repair, with studies suggesting that splicing regulates the expression of *Rnf8*, further controlling ubiquitin-signaling at DSBs [34].

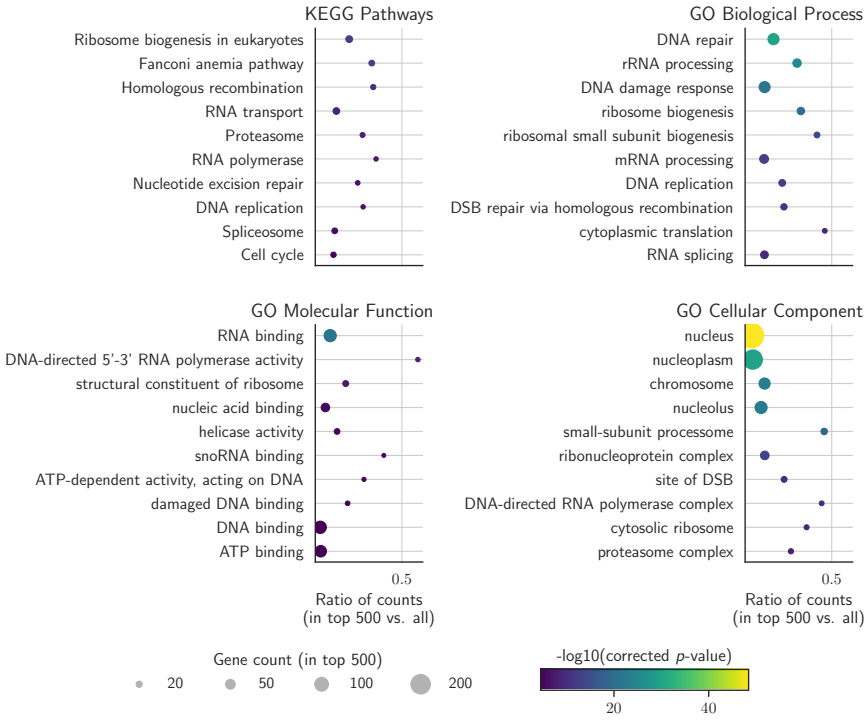


Fig. 3.5: Top 10 enriched KEGG pathways and GO terms among the top 500 genes ranked by MUSICiAn across targets for the genome-wide MUSIC dataset. Top left to bottom right - KEGG pathways, GO biological processes, GO molecular functions, and GO cellular components. The horizontal axis shows the ratio between the numbers of genes annotated with the pathway or GO term among the top 500 ranked genes vs. all genes. Circle color denotes the negative log10 of the FDR-corrected p-value, and circle size indicates the number of genes annotated with a pathway or GO term among the top 500 ranked genes.

3.3.5 Enriched pathways promote homology-directed repair

We analyzed how the genes in the identified pathways influenced the frequencies of different mutation types by fitting a linear regression model per pathway, mutation type, and target site, and using the mutation type frequency per gene knockout and target site as response variable. Some pathways lacked sufficient gene representation to fit a reliable regression model (< 3 samples) and were excluded on a per-analysis basis (Fig. 3.6).

Based on the fitted models, we observed that the genes in each of the enriched pathways promoted HDR events and repressed insertion events across the

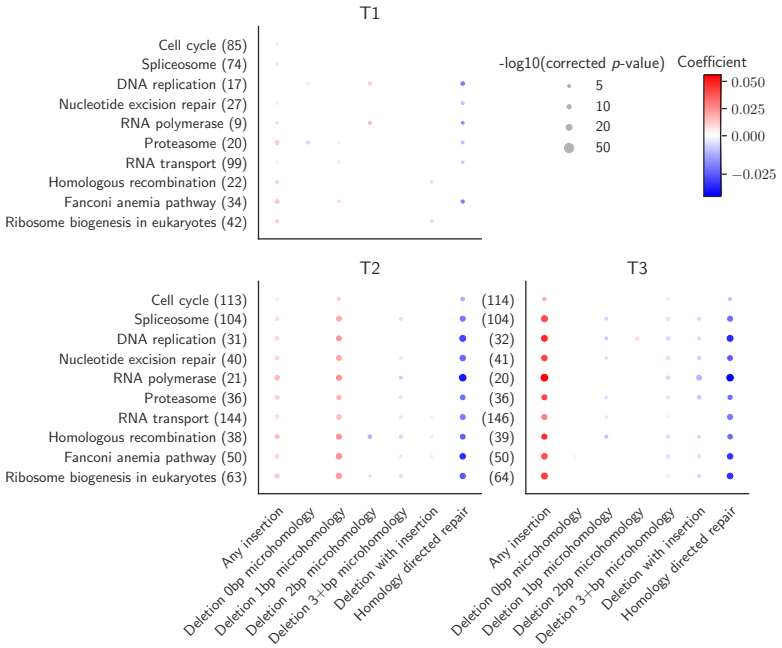


Fig. 3.6: Effect of genes annotated with the top 10 enriched pathways on mutation frequency. We considered all genes in the top 10 pathways enriched amongst the top 500 genes ranked by MUSICiAn based on the genome-wide MUSIC spectra, with the final number of genes available per target dependent on the quality of the obtained mutational spectra. Each dot denotes a linear regression analysis of gene effect on mutation frequency per term or pathway (vertical axis, with gene count), mutation category (horizontal axis), and target site (panels for targets T1-T3 in MUSIC). Dot color denotes the regression coefficient, and dot size indicates the negative log10 of the FDR-corrected p-value. Points with non-significant corrected p-values (> 0.05) were excluded.

target sites in the MUSIC genome-wide screen (Fig. 3.6, T1, T2, T3). Since NHEJ has been associated with introducing insertions at CRISPR-induced DSB sites [24, 29], we suggest that the rise and fall in the frequency of insertion and HDR events could reflect a change in the fraction of DSBs repaired via the NHEJ and HDR pathways. On the other hand, patterns pertaining to the promotion or inhibition of deletion events with or without MH were more sequence-context dependent, making it difficult to associate an inhibited pathway with how it might influence NHEJ and MMEJ. We note that the additional variation exhibited by MUSIC target site T1 could be an artifact of the noisier mutational profiles obtained for that target.

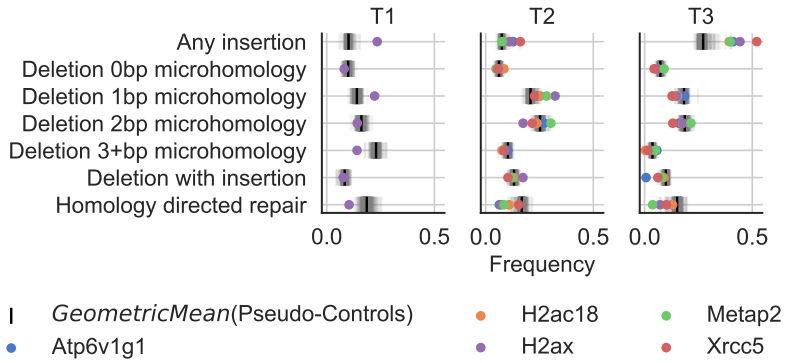


Fig. 3.7: Mutational spectra of the top 5 genes ranked by MUSICiAn based on the genome-wide MUSIC screen. Colored dots denote the frequency obtained under knockout of the indicated gene The vertical axis shows mutation types. The horizontal axis shows frequency. The grey lines represent 300 randomly sampled genes. The black lines show the geometric mean of the pseudo-controls. The colored dots show the top genes. Some dots are not shown for T1, as the sgRNAs were filtered out during quality analysis.

Analysis of the top 5 genes ranked by MUSICiAn for the genome-wide MUSIC dataset (Fig. 3.7) revealed two well-known DSB repair genes, *H2ax* [39] and *Xrcc5* [43]. The others three genes, *Atp6v1g1*, *Metap2*, and *H2ac18*, were not annotated with the "double-strand break repair" GO term. The top-ranked gene was *Atp6v1g1*, for which one other study has reported an effect on HDR repair frequency after knockdown of *Atp6v1g1* via RNA interference [1]. The MUSIC spectra for target sites T2 and T3 showed a relative decrease in the frequency of HDR events after CRISPR knockout of *Atp6v1g1* compared to the geometric mean of the pseudo-controls. A similar tendency was observed for *Metap2*, a gene associated with ribosomal activity, and for *H2ac18*, a histone gene. Identifying histones is not surprising, as the chromatin state regulates DNA damage response by modulating accessibility to DNA damage sites by repair factors [17]. However, to our knowledge, no previous studies have identified an influence of *Metap2* or *H2ac18* on DNA repair pathways or HDR in particular. Further experimentation will be required to validate the impact of these top-ranking genes on mutational spectra and to investigate their role within the DSB repair process.

3.4 Conclusion

In this work, we introduced MUSICiAn, a control-free method to identify genes involved in DSB repair from gene perturbation screens with mutational spectra readout. MUSICiAn is developed for genome-wide perturbation screens, and leverages the fact that most genes have negligible influence on DSB repair and mutational spectra to frame the discovery as an outlier detection task. The goal of MUSICiAn is to both estimate the central tendency and identify genes with outlying spectra by analyzing the distribution of all mutational spectra.

Pseudo-controls estimated by MUSICiAn provided a good approximation of the actual non-targeting controls available for the Repair-seq dataset, showing that MUSICiAn could also be effective at sub-genome scale, provided the assumption that most genes have minimal effect on the spectra can reasonably be made. Notably, the combination of ILR transformation and robust covariance used by MUSICiAn contributed to an improved estimation of the central tendency and pseudo-controls.

Further MUSICiAn analysis of the genome-wide MUSIC data demonstrated an ability to recover known DSB repair genes and suggest candidates for further investigation, including *Atp6v1g1*, *Metap2*, and *H2ac18*. Our findings indicated that genes involved in ribosome biogenesis, the proteasome, and the spliceosome could play a significant role in modulating the frequency of HDR events, suggesting their involvement in DSB repair.

Obtaining sufficient coverage in genome-wide perturbation studies with sequence-based output remains a challenge that has also been noted in prior studies [19]. Low coverage could limit the ability to detect subtle changes in mutagenic activity for rarer outcomes as the data becomes too sparse. To address this, we chose to aggregate mutational outcomes into broader categories. However, MUSICiAn could be applied with any collection of outcomes, as fine-grained as desired, and as the resolution across the different outcomes allows.

Overall, the results of MUSICiAn on the Repair-seq and the genome-wide MUSIC datasets highlighted that the method can effectively estimate pseudo-controls and identify genes with an impact on mutational spectra, enabling

analyses of large-scale screens where designing realistic controls may be challenging.

3.5 Supplementary Tables

Target	Sample	**Before**	Step *i*	Step *ii*	Step *iii*	**Step *iv***
T1	MB01	**89414**	75218	74654	64700	**60975**
	MB02	**89423**	76352	75789	67200	**63347**
T2	MB03	**89492**	80058	79887	79606	**79401**
	MB04	**89481**	82096	81926	80934	**80646**
T3	MB05	**89477**	79296	79042	78796	**78675**
	MB06	**89478**	78492	78237	78112	**78037**

Tab. 3.S1: Breakdown of sgRNA counts after each QA filtering step as described in "Quality analysis and sgRNA filtering" from the main article. Briefly, we filtered out sgRNAs per replicate by applying the following criteria in order: *(i)* a total mutated read count below 700; *(ii)* only two sgRNA representations for the gene; *(iii)* a median pairwise Pearson correlation coefficient below 0.6 compared to other sgRNAs for the same gene within the same replicate; or *(iv)* a median pairwise Pearson correlation below 0.6 compared to paired replicates at the same target site.

Target	Sample	**Before**	Step *i*	Step *ii*	Step *iii*	**Step *iv***
T1	MB01	**18406**	18023	17887	17130	**17004**
	MB02	**18405**	18025	17889	17251	**17135**
T2	MB03	**18406**	18167	18116	18090	**18089**
	MB04	**18406**	18264	18213	18144	**18140**
T3	MB05	**18406**	18224	18149	18134	**18133**
	MB06	**18406**	18178	18103	18095	**18092**

Tab. 3.S2: Breakdown of gene counts after each QA filtering step per the steps outlined in Table S1 above.

MUSICiAn Category	SIQ Categories	Description
Wild-type	WT	No mutation.
Deletion with insertion	DELINS	Deletion with insertion.
	TINS	Deletion with an insertion where the insert is copied from the flank.
	TANDEMDUPLICATION	Duplication of sequence immediately flanking the cut-site.
	TANDEMDUPLICATION_COMPOUND	A tandem duplication with some additional inserted sequence.
Insertion	INSERTION	Any simple insertion event.
Homology-directed repair	HDR	Homology-directed repair event.
Deletion with no/1bp/2bp/3+bp microhomology	DELETION	Any simple deletion event. Additional details recorded include the presence and length of any homology between one side of the deleted sequence and the opposing flank of the cut site.

Tab. 3.S3: Mapping of MUSICiAn mutation categories to SIQ mutation types and details.

3.6 Supplementary Figures

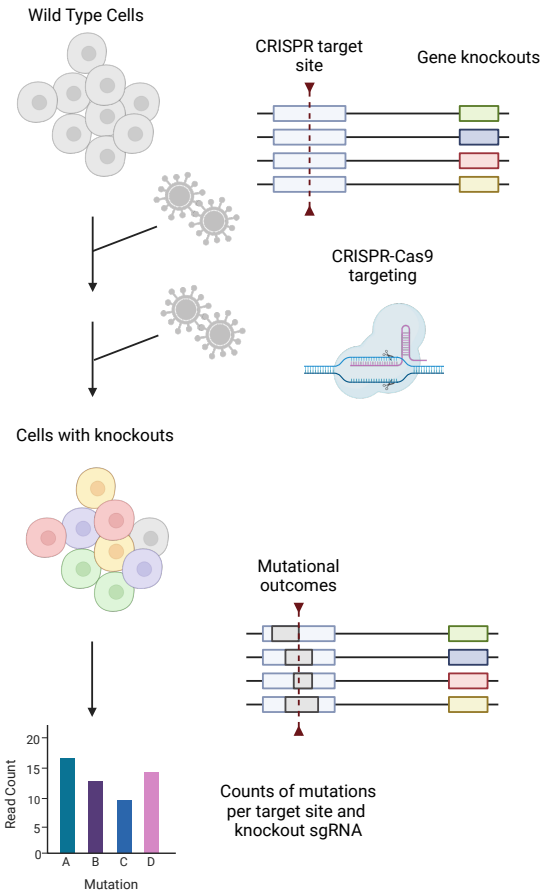


Fig. 3.S1: Illustration of CRISPR gene perturbation screens with mutational spectra readout. First, sequences are integrated into the genomes of cells via lentiviral transduction. Each sequence contains two elements: *(i)* a sgRNA-encoding region to knockout (MUSIC) or silence (Repair-seq) a single gene, and *(ii)* a region common to all integrated sequences to be targeted with CRISPR to produce the mutational spectra. After genomic integration, several days of cell culture are allowed for genes to be knocked out. Following this, MUSIC again uses lentiviral transduction to introduce sgRNAs targeting the common region to the Cas9-expressing cells. Repair-seq uses electroporation to introduce Cas9 RNP complexes to the cells to induce DSBs at the target site. After allowing time for cell culture for DNA cleavage and repair, DNA sequencing was performed to capture the final CRISPR repair products.

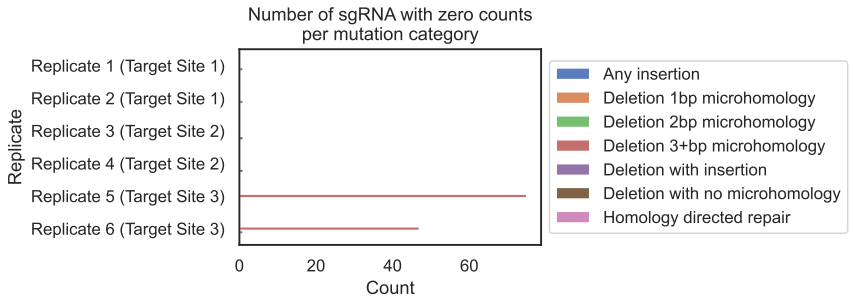


Fig. 3.S2: Counts of sgRNAs with zero values per mutation category and replicate.

References

[1] Britt Adamson, Agata Smogorzewska, Frederic D Sigoillot, Randall W King, and Stephen J Elledge. “A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response”. In: *Nature cell biology* 14.3 (2012), pp. 318–328 (cit. on p. 84).

[2] Charu C Aggarwal. *Outlier Analysis Second Edition*. 2016 (cit. on p. 67).

[3] John Aitchison. “Principal component analysis of compositional data”. In: *Biometrika* 70.1 (1983), pp. 57–65 (cit. on pp. 68, 71).

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, et al. “Gene Ontology: tool for the unification of biology”. In: *Nature genetics* 25.1 (2000), pp. 25–29 (cit. on p. 74).

[5] Samah W Awwad, Almudena Serrano-Benitez, John C Thomas, Vipul Gupta, and Stephen P Jackson. “Revolutionizing DNA repair research and cancer therapy with CRISPR–Cas screens”. In: *Nature Reviews Molecular Cell Biology* (2023), pp. 1–18 (cit. on p. 66).

[6] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300 (cit. on p. 74).

[7] Anne Bothmer, Tanushree Phadke, Luis A Barrera, et al. “Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus”. In: *Nature communications* 8.1 (2017), p. 13905 (cit. on p. 67).

[8] Peter MJ Burgers. “Eukaryotic DNA polymerases in DNA replication and DNA repair”. In: *Chromosoma* 107.4 (1998), pp. 218–227 (cit. on p. 80).

[9] Raphael Ceccaldi, Beatrice Rondinelli, and Alan D D’Andrea. “Repair pathway choices and consequences at the double-strand break”. In: *Trends in Cell Biology* 26.1 (2016), pp. 52–64 (cit. on p. 66).

[10] Alice Chen. “PARP inhibitors: its role in treatment of cancer”. In: *Chinese journal of cancer* 30.7 (2011), p. 463 (cit. on p. 66).

[11] Delisa E Clay and Donald T Fox. “DNA damage responses during the cell cycle: insights from model organisms and beyond”. In: *Genes* 12.12 (2021), p. 1882 (cit. on p. 80).

[12] The Gene Ontology Consortium. “The Gene Ontology resource: enriching a GOld mine”. In: *Nucleic acids research* 49.D1 (2021), pp. D325–D334 (cit. on p. 74).

[13] David Cortez. “Replication-coupled DNA repair”. In: *Molecular cell* 74.5 (2019), pp. 866–876 (cit. on p. 80).

[14] Kyle R Cron, Kaya Zhu, Deepa S Kushwaha, et al. “Proteasome inhibitors block DNA repair and radiosensitize non-small cell lung cancer”. In: *PloS one* 8.9 (2013), e73710 (cit. on p. 81).

[15] Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. “Isometric Logratio Transformations for Compositional Data Analysis”. In: *Mathematical geology* 35.3 (2003), pp. 279–300 (cit. on p. 71).

[16] Peter Filzmoser, Karel Hron, and Clemens Reimann. “Principal component analysis for compositional data with outliers”. In: *Environmetrics: The Official Journal of the International Environmetrics Society* 20.6 (2009), pp. 621–632 (cit. on pp. 72, 73).

[17] Clayton R Hunt, Deepti Ramnarain, Nobuo Horikoshi, et al. “Histone modifications and DNA double-strand break repair after exposure to ionizing radiations”. In: *Radiation research* 179.4 (2013), pp. 383–392 (cit. on p. 84).

[18] Kristen E Hurov, Cecilia Cotta-Ramusino, and Stephen J Elledge. “A genetic screen identifies the Triple T complex required for DNA damage signaling and ATM and ATR stability”. In: *Genes & development* 24.17 (2010), pp. 1939–1950 (cit. on p. 66).

[19] Jeffrey A Hussmann, Jia Ling, Purnima Ravisankar, et al. “Mapping the genetic landscape of DNA double-strand break repair”. In: *Cell* 184.22 (2021), pp. 5653–5669 (cit. on pp. 67–69, 73, 85).

[20] Tabitha Jenkins, Sarah J Northall, Denis Ptchelkine, et al. “The HelQ human DNA repair helicase utilizes a PWI-like domain for DNA loading through interaction with RPA, triggering DNA unwinding by the HelQ helicase core”. In: *NAR cancer* 3.1 (2021), zcaa043 (cit. on p. 79).

[21] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. “KEGG for taxonomy-based analysis of pathways and genomes”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D587–D592 (cit. on p. 74).

[22] Lea Milling Korsholm, Zita Gál, Blanca Nieto, et al. “Recent advances in the nucleolar responses to DNA double-strand breaks”. In: *Nucleic acids research* 48.17 (2020), pp. 9449–9461 (cit. on p. 81).

[23] Nevan J Krogan, Mandy HY Lam, Jeffrey Fillingham, et al. “Proteasome involvement in the repair of DNA double-strand breaks”. In: *Molecular cell* 16.6 (2004), pp. 1027–1034 (cit. on p. 81).

[24] Michael R Lieber. “The mechanism of human nonhomologous DNA end joining”. In: *Journal of Biological Chemistry* 283.1 (2008), pp. 1–5 (cit. on pp. 69, 83).

[25] Mikael S Lindström, Deana Jurada, Sladana Bursac, et al. “Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis”. In: *Oncogene* 37.18 (2018), pp. 2351–2366 (cit. on p. 81).

[26] Sugnet Lubbe, Peter Filzmoser, and Matthias Templ. “Comparison of zero replacement strategies for compositional data with large numbers of zeros”. In: *Chemometrics and Intelligent Laboratory Systems* 210 (2021), p. 104248 (cit. on p. 70).

[27] Prasanta Chandra Mahalanobis. “On the generalized distance in statistics”. In: *Sankhyā: The Indian Journal of Statistics, Series A (2008-)* 80 (2018), S1–S7 (cit. on p. 71).

[28] Johanna Michl, Jutta Zimmer, and Madalena Tarsounas. “Interplay between Fanconi anemia and homologous recombination pathways in genome integrity”. In: *The EMBO journal* 35.9 (2016), pp. 909–923 (cit. on p. 74).

[29] Kutubuddin A Molla and Yinong Yang. “Predicting CRISPR/Cas9-induced mutations for precise genome editing”. In: *Trends in biotechnology* 38.2 (2020), pp. 136–141 (cit. on p. 83).

[30] Brenda C O’Connell, Britt Adamson, John R Lydeard, et al. “A genome-wide camptothecin sensitivity screen identifies a mammalian MMS22L-NFKBIL2 complex required for genomic stability”. In: *Molecular cell* 40.4 (2010), pp. 645–657 (cit. on p. 66).

[31] LM Ogawa and SJ Baserga. “Crosstalk between the nucleolus and the DNA damage response”. In: *Molecular bioSystems* 13.3 (2017), pp. 443–455 (cit. on p. 81).

[32] Michele Olivieri, Tiffany Cho, Alejandro Álvarez-Quilón, et al. “A genetic map of the response to DNA damage in human cells”. In: *Cell* 182.2 (2020), pp. 481–496 (cit. on p. 66).

[33] Megan van Overbeek, Daniel Capurso, Matthew M Carter, et al. “DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks”. In: *Molecular Cell* 63.4 (2016), pp. 633–646 (cit. on p. 67).

[34] C Pederiva, S Böhm, A Julner, and M Farnebo. “Splicing controls the ubiquitin response during DNA double-strand break repair”. In: *Cell Death & Differentiation* 23.10 (2016), pp. 1648–1657 (cit. on p. 81).

[35] Robin van Schendel, Joost Schimmel, and Marcel Tijsterman. “SIQ: easy quantitative measurement of mutation profiles in sequencing data”. In: *NAR Genomics and Bioinformatics* 4.3 (2022), lqac063 (cit. on p. 69).

[36] Petra Schwertman, Simon Bekker-Jensen, and Niels Mailand. “Regulation of DNA double-strand break repair by ubiquitin and ubiquitin-like modifiers”. In: *Nature reviews Molecular cell biology* 17.6 (2016), pp. 379–394 (cit. on p. 81).

[37] Daniel D Scott and Marlene Oeffinger. “Nucleolin and nucleophosmin: nucleolar proteins with multiple functions in DNA repair”. In: *Biochemistry and cell biology* 94.5 (2016), pp. 419–432 (cit. on p. 81).

[38] Ralph Scully, Arvind Panday, Rajula Elango, and Nicholas A Willis. “DNA double-strand break repair-pathway choice in somatic mammalian cells”. In: *Nature Reviews Molecular Cell Biology* 20.11 (2019), pp. 698–714 (cit. on p. 66).

[39] Ralph Scully and Anyong Xie. “Double strand break repair functions of histone H2AX”. In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 750.1-2 (2013), pp. 5–14 (cit. on p. 84).

[40] Max W Shen, Mandana Arbab, Jonathan Y Hsu, et al. “Predictable and precise template-free CRISPR editing of pathogenic variants”. In: *Nature* 563.7733 (2018), pp. 646–651 (cit. on p. 67).

[41] Jia Shou, Jinhuan Li, Yingbin Liu, and Qiang Wu. “Precise and predictable CRISPR chromosomal rearrangements reveal principles of Cas9-mediated nucleotide insertion”. In: *Molecular cell* 71.4 (2018), pp. 498–509 (cit. on p. 67).

[42] Agata Smogorzewska, Rohini Desetty, Takamune T Saito, et al. “A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair”. In: *Molecular cell* 39.1 (2010), pp. 36–47 (cit. on p. 66).

[43] Guillermo E Taccioli, Tanya M Gottlieb, Tracy Blunt, et al. “Ku80: product of the XRCC5 gene and its role in DNA repair and V (D) J recombination”. In: *Science* (1994), pp. 1442–1445 (cit. on p. 84).

[44] Adam Thomas, Julie Cox, Kelly B Wolfe, et al. “Division of Labor by the HELQ, BLM, and FANCM Helicases during Homologous Recombination Repair in Drosophila melanogaster”. In: *Genes* 13.3 (2022), p. 474 (cit. on p. 79).

[45] Anika Trenner and Alessandro A Sartori. “Harnessing DNA double-strand break repair for cancer treatment”. In: *Frontiers in oncology* 9 (2019), p. 1388 (cit. on p. 66).

[46] Joe H Ward Jr. “Hierarchical Grouping to Optimize an Objective Function”. In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244 (cit. on p. 73).

[47] David W Wyatt, Wanjuan Feng, Michael P Conlin, et al. “Essential roles for polymerase θ-mediated end joining in the repair of chromosome breaks”. In: *Molecular cell* 63.4 (2016), pp. 662–673 (cit. on p. 67).

[48] Ye Zhang, Larry H Rohde, and Honglu Wu. “Involvement of nucleotide excision and mismatch repair mechanisms in double-strand break repair”. In: *Current genomics* 10.4 (2009), pp. 250–258 (cit. on p. 80).

[49] Fei Zhao, Wootae Kim, Jake A Kloeber, and Zhenkun Lou. “DNA end resection and its role in DNA replication and DSB repair choice in mammalian cells”. In: *Experimental & Molecular Medicine* 52.10 (2020), pp. 1705–1714 (cit. on p. 81).

[50] Michal Zimmermann, Olga Murina, Martin AM Reijns, et al. “CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions”. In: *Nature* 559.7713 (2018), pp. 285–289 (cit. on p. 66).

4

Signatures in CRISPR Mutational Spectra Reveal Role and Interplay of Genes in DNA Repair

> *If we understand how mutations are produced and how they're repaired and how they can be manipulated, we are really dealing with a fundamental aspect of cancer research.*
>
> — **Evelyn Witkin**
> (2015 Lasker Basic Medical Research Award Acceptance Speech)

Understanding double-strand break (DSB) repair and its disruption is key to decipher genomic instability driving diseases such as cancer and reveal therapeutic avenues. Numerous genes have been linked to DSB repair for the first time in recent genome-wide perturbation studies assessing effects on mutational outcomes. However, the functional roles of most such genes remain poorly understood. Evidence from other studies shows that related genes similarly modulate the frequency of specific mutational outcomes following DSB repair, but analysis has largely ignored the multiplicity of gene functions and cross-talk between pathways. Here, we infer functional roles for candidate genes based on knockout mutational spectra by identifying mutational signatures shared with known genes and then mapping them to DSB repair functions. Signatures are identified using non-negative matrix factorization (NMF) to reflect functional multiplicity and cross-talk. We generated mutational spectra for mouse embryonic stem

Colm Seale, Marco Barazas, Robin van Schendel, Marcel Tijsterman, and Joana P. Gonçalves.

(mES) cells at three target sites across CRISPR knockouts of 307 known and 459 candidate DSB repair genes. Analysis using NMF revealed four mutational signatures associated with homology-directed repair (HDR), microhomology-mediated end-joining (MMEJ), and the initiation and ligation components of non-homologous end-joining (NHEJ). Signatures suggested that candidate genes Dbr1 and Hnrnpk could influence MMEJ and Fanconi anaemia (FA), and that loss of core NHEJ components (e.g. MRN complex or Ku proteins) could shift repair preference towards Ku-independent NHEJ. These findings demonstrate the utility of NMF for characterizing the contribution of genes to repair pathways and provide a foundation to discover new gene functions in DSB repair.

4.1 Introduction

Effective cellular repair of DNA double-strand breaks (DSBs) is essential to prevent genomic instability leading to cell death or the development of diseases such as cancer [6]. Deficiencies in DNA damage response do not only drive tumor progression, but also expose vulnerabilities of tumor cells that can be leveraged for treatment. Conventional chemo and radiation therapies work by inducing substantial DNA damage, from which tumor cells with compromised repair struggle to recover. However, such therapies also affect healthy cells, causing debilitating side effects and possibly recurrence. Targeted treatments mitigate this toxicity by design by selectively exploiting unique handicaps of tumor cells. For instance, PARP inhibitors trigger the joint essentiality between PARP and BRCA genes to treat tumors with deficient BRCA function [8]. While the benefit of targeted therapies can be great, development requires in-depth knowledge of DSB repair mechanisms and the functions of individual genes. Advancing the understanding of DSB repair is therefore essential to reveal mechanistic relationships offering new therapeutic possibilities to improve patient care and quality of life [1].

The study of DSB repair has been accelerated by the availability of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technology to induce DSBs at predefined genomic loci [15]. A growing number of studies combine CRISPR targeting with DNA sequencing readout to analyze mutational spectra resulting from DSB repair activity and gain new insights into repair mechanisms at scale. Specifically, several studies have used CRISPR targeting across collections of loci to investigate the influence of sequence context around the DSB within a few different cell types or genomic

contexts [28, 36, 37]. Others have paired systematic knockouts of individual DSB repair genes with CRISPR targeting across loci to tease out the function of each gene and its specific contribution to DSB repair mechanisms [13]. For example, cells harboring knockouts of *Polq* show significantly fewer long deletions with microhomologies (MHs) at DSBs post-repair than wild-type cells, consistent with the key role of the gene in the microhomology-mediated end-joining repair pathway [41, 13]. Such insights have also recently inspired genome-wide and follow-up focused screens of gene knockout effects on mutational spectra aimed at uncovering new gene associations with DSB repair [3, 35].

Existing studies to elucidate the role of genes in DSB repair take different approaches to analyze the effects of the gene knockouts on mutational spectra relative to controls. One strategy is to separately examine the impact on individual outcomes of interest, such as 1bp insertions or deletions [3]. Such univariate analysis ignores covariation of outcomes produced by shared underlying mutational processes, whereas repair mechanisms and genes do not function in isolation. Multivariate methods are also applied to reveal genes involved in DSB repair through outlier analysis of spectra from genome-wide knockout screens, resulting in candidate associations of genes with DSB repair but no further insight into possible functions [35]. Alternative multivariate strategies are further used to discover relationships between known DSB repair genes based on similarity of their mutational spectra, relying on conventional clustering or manifold learning. Conventional clustering techniques such as hierarchical clustering [31] focus on global similarities across the entire mutational spectra and thus ignore that the same gene or mechanism can contribute to multiple outcomes, whereas manifold learning methods like UMAP [23] can capture local relationships between genes based on subsets of outcomes but are not straightforward to interpret due to their non-linear nature [29, 19].

Here, we propose using non-negative matrix factorization (NMF) to analyze mutational spectra of known and candidate DSB repair genes, and identify relationships offering insight into the functional role of such candidates. The advantage of NMF is that it captures local patterns while being interpretable and producing clusters or factors with "soft" gene memberships, reflecting the fact that genes can operate in multiple DSB repair pathways and that distinct pathways can produce identical outcomes at varying rates (also co-occurring with different sets of other outcomes). This is achieved by decomposing the

mutational spectra of all gene knockouts into mutational signatures capturing patterns across subsets of knockouts, and signature exposures quantifying the contribution of each signature to the spectrum of each gene knockout [22, 20].

In this work, we generate mutational spectra for three target sites under individual knockouts of 766 genes, including 307 known DSB repair genes and 459 candidates selected based on outlying spectra from a previous genome-wide knockout screen [35]. We investigate the ability of NMF to identify signatures and exposures that recover known mutational patterns and responsible genes linking signatures to DSB repair pathways. We also leverage signature depletions to suggest DSB repair mechanisms for candidate genes, and explore how joint depletion patterns may provide further functional granularity for genes of the same pathway (Fig 4.1).

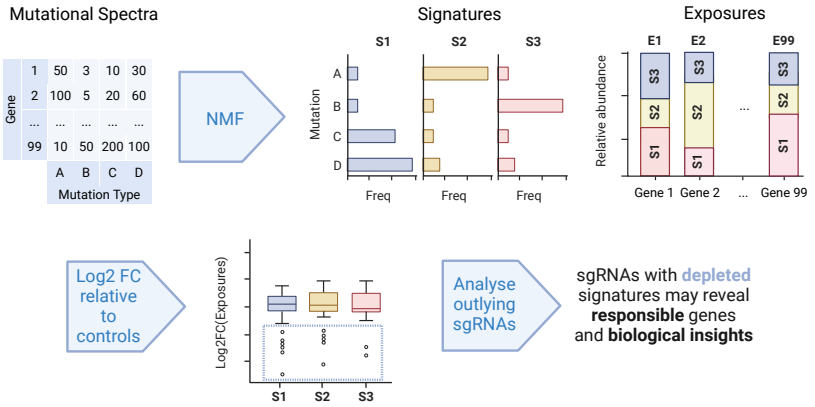


Fig. 4.1: Identifying DSB repair gene function via NMF analysis of mutational spectra. Mutational spectra, defined as counts over a set of unique mutational outcomes (columns A–D), for cell populations carrying knockouts of individual genes (rows E1-E99). Estimation of mutational signatures (S1-S3) and associated exposures (E1-E99) using NMF. Analysis of signature exposure ratio relative to controls (log2 fold-change) identifies gene knockouts depleted in signature-specific mutational outcomes, which can be further linked to other genes and functional mechanisms with impact on the same signatures.

4.2 Methods

We analyzed two datasets of mutational spectra for mouse embryonic stem (mES) cells carrying knockouts of individual genes: the primary dataset,

generated for this study, comprising knockouts of 766 genes and three target sites; and a published dataset, which we name Barazas2025 (*not publicly available at time of writing, access may be granted upon reasonable request to the corresponding author*), comprising knockouts of 742 genes and one target site [3]. There were 273 genes in common between the two screens. We included the Barazas2025 dataset mainly to assess the robustness of the mutational signatures identified from the primary dataset.

4.2.1 Generating mutational spectra

Selecting genes for the primary screen. The set of 766 genes for the primary screen consisted of 459 candidate genes and 307 known DSB repair genes. The candidate genes were based on the analysis of a genome-wide study quantifying the effect of individual gene knockouts on mutational spectra [35]. While the genome-wide study traded gene coverage for mutational spectra resolution, here we focused on profiling the more promising genes at higher resolution to increase the robustness of potential findings. Specifically, we selected the 459 genes with the largest impact on their knockout mutational spectra according to a multivariate outlier score measuring deviation relative to the center of the distribution of all mutational spectra.

To be able to relate mutational signatures and candidate genes to established DSB repair mechanisms, we additionally selected 307 DSB repair genes as the union of the following three sets of genes: *(i)* annotated with Gene Ontology (GO) terms "double-strand break repair" or "interstrand cross-link repair" [9]; *(ii)* annotated with KEGG pathways "Homologous recombination" (ID: mmu03440), "Non-homologous end joining" (ID: mmu03450), or "Fanconi anaemia pathway - Mus musculus (house mouse)" (ID: mmu03460) [16]; and *(iii)* curated as the 118 DSB repair genes with the largest effect in the Repair-seq study [13].

Screening DSB repair outcomes. To generate mutational spectra, we followed the experimental protocol in [3] (Supplementary Fig. 4.S1). Briefly, paired guide-target DNA sequences were integrated into the genomes of endogenously Cas9-expressing mouse embryonic stem (mES) cells via lentiviral transduction. Each integrated sequence contained a single-guide RNA (sgRNA) designed to knock out a specific gene or as a non-targeting control, and one of three 20bp target sequences with protospacer adjacent motif

(denoted T1-T3). For the primary screen, we used a total of 2,321 knockout sgRNAs (∼3 per gene), and 50 additional non-targeting control sgRNAs. The Barazas2025 screen included a total of 2,400 knockout sgRNAs for 742 genes, and 170 non-targeting control sgRNAs. After allowing 5 days for genomic integration and gene knockout, we split the bulk cell population into nine samples for the primary screen (3 targets $\times$ 3 replicates) or 3 samples for the Barazas2025 screen (1 target, 3 replicates). Per sample, we performed a second round of lentiviral transduction to express sgRNAs and induce DSBs at the integrated target sites. Following $\sim$80h of cell culture for cleavage and repair, we used paired-end DNA sequencing to characterize repair outcomes at the integrated sites.

Processing outcome sequences into mutational spectra. We used the SIQ v4.3 tool [33] to call mutational outcomes from the sequencing data, with parameters "-m 2 -c -e 0.05" (requiring a minimum number of 2 reads to count an event, collapsing identical events to a single record with the sum of counts, and allowing a maximum base error rate of 0.05). The outcomes were split into four categories: deletion; insertion; templated insertion, denoting a deletion with an insertion where the inserted sequence matches a region flanking the cut site; and HDR event, for any insertion matching a provided donor template DNA. Additionally, we recorded length and location for both insertions and deletions, as well as microhomology (MH) length when present for deletions. All templated insertions were collapsed into a single category, with the respective sum of counts over all observed events. Wild-type sequences were excluded to produce a final collection of mutational outcomes with the corresponding counts per sgRNA.

To filter rare outcomes, we excluded outcomes with a geometric mean frequency below 0.002 across the non-targeting controls. This resulted in a final set of mutational spectra containing 28-44 outcomes per target site and replicate for the primary screen and 32 outcomes per replicate for the Barazas2025 screen (Supplementary Table S1 for outcome counts, and Supplementary Tables S2-S5 for mutation details).

Replicate quality analysis and selection. We assessed data quality by calculating pairwise Pearson's correlations between replicates of the same target site, using the sgRNAs and outcomes common to each pair. Average

correlations were above 0.98, indicating high quality throughout (Supplementary Table S5). We selected the replicate with the largest number of recovered gene knockouts per target site (T1: 2,291, T2: 2,291, T3: 2,304, Barazas2025: 2,358; Supplementary Table S1) for downstream analysis.

4.2.2 Identifying co-occurring mutational patterns

Analyzing patterns across target sites. Since mutational spectra are target site-specific, to characterize patterns across sites for the primary dataset, we concatenated the mutational spectra of the three target sites per sgRNA into a single mutational spectrum covering all three sites. We excluded sgRNAs missing from any of the target sites. To mitigate target-specific batch effects, we scaled the counts per sgRNA to equalize the ranges of counts across the different targets. This resulted in a three-target mutational spectra count matrix of 2,288 sgRNAs $\times$ 112 mutational outcomes.

Identifying signatures and exposures. Each mutational spectrum reflects the aggregated contribution of all mutational processes, including DSB repair mechanisms, active in the cell population from which the spectrum was derived. Each mutational process tends to produce specific mutational outcomes at specific rates, resulting in a fingerprint or signature. Our goal is to decipher the mixture of signatures underlying the collection of mutational spectra using NMF, and later attempt to link each signature to responsible genes and pathways.

Formally, the input is a matrix $V \in \mathbb{N}^{m \times n}$ of n mutational spectra defined as count distributions over a set M of m mutational outcomes. We aim to decompose V into the product of two matrices $V \approx S \times E$: the signature matrix $S \in \mathbb{R}^{m \times k}$, representing a collection of k signatures, each defined as a frequency distribution over the set of outcomes M; and the exposure matrix $E \in \mathbb{N}^{k \times n}$ denoting the contribution of each of the k signatures to every one of the n mutational spectra.

Signatures S and exposures E can be estimated from V, given a fixed number of signatures k, using NMF [22]. Here, we used the SigProfilerExtractor [14] v1.1.23 framework, which includes additional bootstrapping steps to identify more robust or stable signatures for a fixed k, and an evaluation procedure to optimize the number of signatures k. We optimized the number of signatures

between a minimum of 1 and a maximum of 10, and used 30 bootstraps for signature robustness or stability.

SigProfilerExtractor selects the solution S and E for the number of signatures k that produces the closest reconstruction of the original mutational spectra V, with an average stability across bootstraps above a threshold (we used 0.8). We identified 4 signatures for the primary dataset and 5 for the Barazas2025 dataset (Supplementary Figs. 4.S2 and 4.S3). SigProfilerExtractor assigns a string identifier to each signature, such as 'CH112A', where 'CH' indicates the signature is extracted over a set of custom outcomes, '112' is the number of outcomes in the set, and 'A' is a letter of the alphabet identifying a specific signature.

4.2.3 Elucidating responsible genes and pathways

Identifying genes responsible for signatures. Given that exposures indicate the contribution of the signatures to the mutational spectra, changes in exposures can be used to detect genes with an effect on those signatures. If a gene is responsible for a signature, knocking it out should weaken the strength of that pattern and therefore lower the signature exposure compared to control cells, which we term as signature depletion. To detect this type of effect, we calculate the change in exposure $\delta_{t,s}$ for a gene knockout spectrum t and signature s as:

$$\delta_{t,s} = \log_2 \left(\frac{e_t^s}{\left(\prod_{c \in C} e_c^s\right)^{\frac{1}{|C|}}} \right) \tag{4.1}$$

where e_t^s and e_c^s denote exposures of signature s for gene knockout spectrum t and non-targeting control spectrum c, respectively, and C refers to the set of all non-targeting control spectra of size $|C|$. Note that the denominator corresponds to the geometric mean of the exposures of the non-targeting controls. For the log2 fold change calculations, we handled zeros by applying multiplicative replacement, using defaults as implemented in scikit-bio v0.5.9 [32].

Linking signatures to DSB repair functions. To associate signatures with biological functions, we performed GO enrichment analysis. First, we quan-

tified the effect of each gene knockout t on the exposures of the different signatures relative to the non-targeting controls using the change in exposure formula $\delta_{t,s}$ (Equation 4.1). Second, we singled out genes "responsible" for a given signature as those genes whose knockout spectra showed outlying low change in exposure for that signature (below Q1 - 1.5×IQR, where Q1 is the first quartile and IQR is the inter-quartile range). Finally, we performed enrichment analysis for the selected "responsible" genes against the full gene set as the background, using the *GOEnrichmentStudyNS* function in "goatools" v1.2.3 [17]. We corrected the resulting *p*-values for multiple testing using the Benjamini-Hochberg method [4].

Visualizing pathway-signature activity. To visualize the influence of gene-pathway members on exposure profiles, we obtained labels from public databases [2, 24] and lists curated by major studies [40, 18, 27]. We first compiled gene sets involved in HDR, NHEJ, and Fanconi anaemia (FA) per source. Then we used majority voting to assign one single label per gene, excluding ties to reduce ambiguity.

4.3 Results and Discussion

4.3.1 NMF identifies mutational processes and shared outcomes

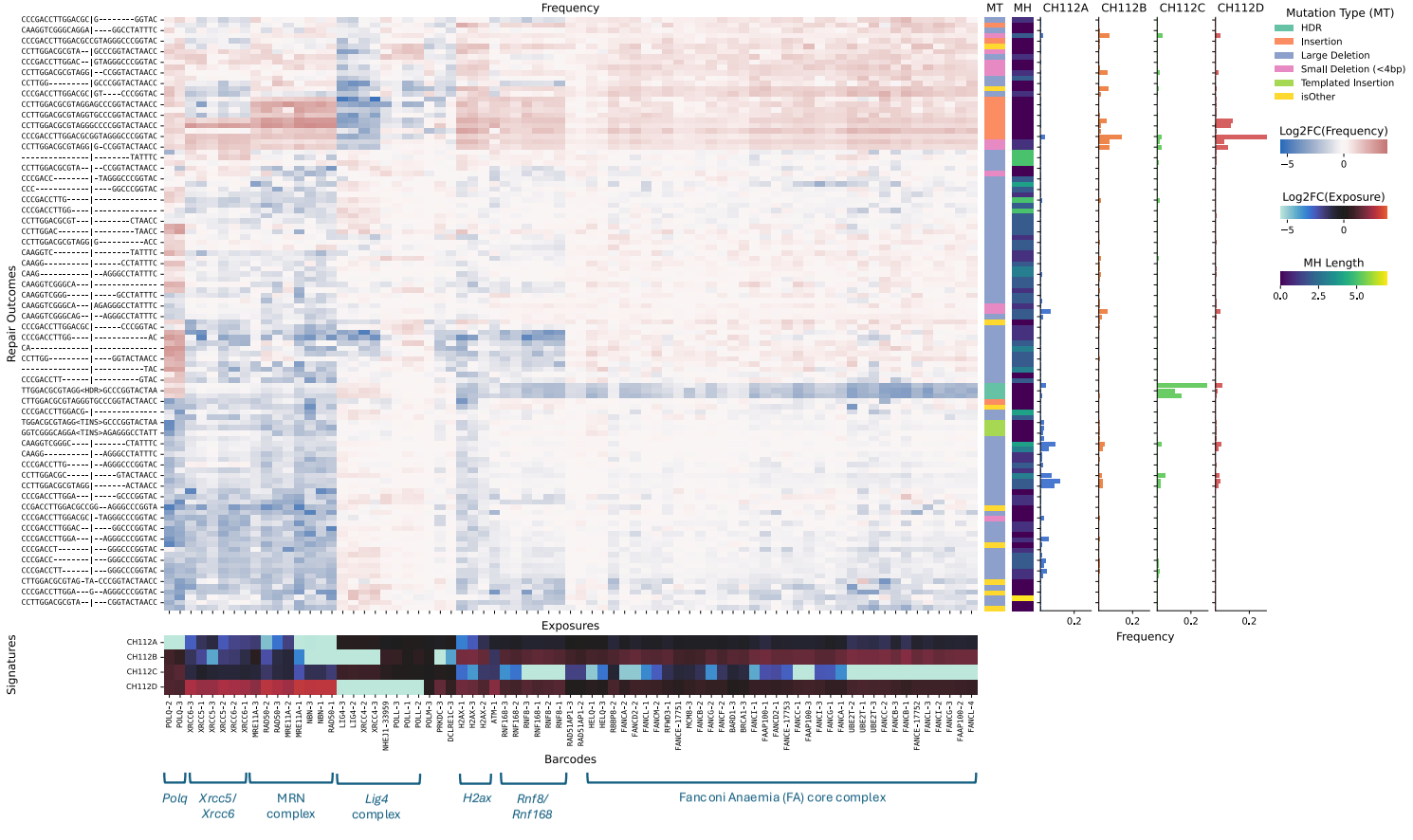
Our primary aim with this study was to uncover new functional roles for genes in DNA repair by leveraging their similarity to known DSB repair genes, based on the effect of gene knockouts on mutational patterns. To assess the potential of this approach, we first investigated if we could recover associations between genes with previously validated DSB repair functions. In addition, we sought to characterize relationships between mutational outcomes and describe how related genes might share responsibility for such outcomes via repair pathway co-membership or upstream co-regulation. We analyzed global associations between genes or outcomes using hierarchical clustering and also identified more localized co-occurring mutational patterns involving subsets of genes and mutational outcomes using NMF.

Hierarchical clustering of the gene knockouts based on changes in the frequency of mutational outcomes relative to controls across the three target sites revealed functionally related groups (Fig. 4.2, central heatmap), in-

cluding; MRN complex genes *Mre11a/Nbn/Rad50* involved in DNA damage sensing and repair, with influence on pathway choice; Lig4 complex genes involved in DNA ligation *Lig4/Xrcc4/Nhej1/Poll*, NHEJ-initiating Ku complex genes *Xrcc5/Xrcc6*; RING-type E3 ubiquitin ligases *Rnf8/Rnf168* involved in response to DNA damage and recruitment of repair factors with influence on pathway choice; and Fanconi anemia (FA) core complex genes. These associations confirmed that knockouts of genes with related roles tended to cluster together based on changes in the mutational spectra.

Clustering of mutational outcomes based on the same mutational frequency changes showed consistent effects across the three target sequence contexts for certain categories of mutations (Fig. 4.2 central heatmap and MT bar): shorter deletions (<4 bp) frequently co-occurred with insertions, while longer deletions with microhomologies correlated with templated insertion (TINS) events. Such co-occurrences suggest that the mutational outcomes could be influenced by similar repair processes and genes, independent of target site-specific effects.

Fig. 4.2: Impact of gene knockouts on mutational spectra and identified mutational signatures (overleaf). (Central heatmap) Effect of the 75 gene knockouts with the largest outcome redistribution (columns) on every mutational outcome across three target sites (rows), with outcomes further characterized by mutation type (MT) and microhomology length (MH). Effect quantified using the log2 fold-change in outcome frequency between the gene knockout spectrum and the geometric mean of the non-targeting control spectra. Gene knockouts and mutational outcomes ordered based on hierarchical clustering with Ward linkage and Euclidean distance. (Signature bar plots) Mutational profiles of the identified signatures, with bars denoting the frequency of each mutational outcome. (Bottom heatmap) Depletion in exposure of each identified signature per gene knockout, as the log2 fold-change between the exposure of the gene knockout sample and the geometric mean exposure of the non-targeting controls. Key DSB repair genes and complexes highlighted.


Frequency
MT
MH
CH112A
CH112B
CH112C
CH112D
Mutation Type (MT)
HDR
Insertion
Large Deletion
Small Deletion (<4bp)
Templated Insertion
isOther
Log2FC(Frequency)
Log2FC(Exposure)
MH Length
Repair Outcomes
Exposures
Signatures
Barcodes
Polq
Xrcc5/ Xrcc6
MRN complex
Lig4 complex
H2ax
Rnf8/ Rnf168
Fanconi Anaemia (FA) core complex

Nevertheless, clustering analysis ignores that genes can be involved in multiple repair pathways, and pathways may share responsibility for producing specific mutational outcomes. To capture such relationships, we analyzed the mutational spectra using the NMF-based SigProfiler framework (see Methods). We identified four distinct and reproducible mutational signatures, CH112A-CH112D, each with unique mutational profiles (Fig. 4.2 bar plots). Signature CH112A showed the highest proportion of longer deletions with larger MH lengths and the highest frequency of templated insertions (Fig. 4.2 bar plots and Fig. 4.3). Signature CH112B showed a larger fraction of small deletions and insertions, while CH112C and CH112D were dominated respectively by HDR events and insertions (Fig. 4.3). We also observed that outcomes such as HDR events were exclusively linked to one signature (CH112D), whereas other outcomes like insertions appeared in more than one signature (CH112B and CH112D) due to co-occurrence with different mutational outcomes in distinct sets of gene knockout spectra. These results show that NMF has the ability to capture more granular relationships between mutational outcomes possibly associated with multiple mutational processes, which could otherwise be missed using conventional clustering focusing on global similarities across mutational spectra.

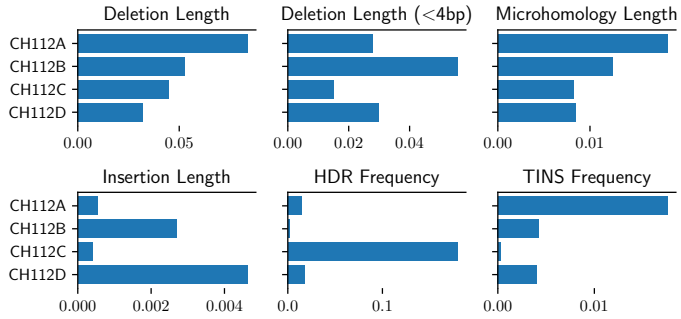


Fig. 4.3: Signature deletion and insertion properties. Per signature average of deletion and insertion outcome properties. Outcome properties: deletion length, small deletion length (< 4bp), deletion microhomology (MH) length, insertion length, homology-directed repair (HDR) frequency, and templated insertion (TINS) frequency. The first four length properties are weighted by the signature frequency of the respective outcomes.

We also analyzed the effect of each gene knockout on the contribution of the different signatures to the observed mutational spectrum, by looking at changes in the NMF-estimated signature exposures between gene knock-

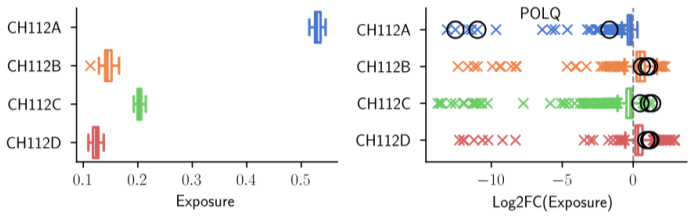


Fig. 4.4: Exposure baseline in controls and change in knockouts. Subfigures: (left) signature exposures of non-targeting controls; (right) impact of every gene knockout on signature exposures, as the log2 fold change between the exposure of the knockout and the geometric mean exposure of non-targeting controls. Black circles highlight *Polq* knockouts. Boxplot: box delimits the interquartile range between 1st and 3rd quartiles ($IQR = Q3 - Q1$), with a line across the box denoting the median; whiskers indicate the smallest and largest values within $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$, and points beyond them are considered outliers.

out and controls (see Methods). The diverse impact across genes reflected individual and shared responsibilities for distinct mutational patterns (Fig. 4.2 bottom heatmap). For example, *Polq* gene knockouts singularly reduced CH112A exposures, while FA core complex gene knockouts lowered CH112C exposures. In contrast, members of the MRN complex influenced multiple signatures (CH112A-B), reflecting their broader role as DNA damage checkpoint genes with an impact on DSB repair pathway choice [21].

Together, this NMF-based analysis of mutational spectra dissected important biological realities that would be missed by conventional techniques, including the multiplicity of roles played by genes and the complexity of shared mutational outcomes arising via alternative repair mechanisms.

4.3.2 Signature exposures reveal drivers of mutational patterns

We sought to identify mechanisms underlying each mutational signature by examining the biological functions of genes with a larger impact on signature exposures. Controls (Fig. 4.4, left) showed a prominent contribution from signature CH112A (median 52.9%) linked to deletions with longer MH, followed by CH112C (median 20.3%) associated with HDR events, CH112B (median 14.4%) producing insertions and small deletions, and CH112D (median 12.3%) dominated by insertions. As for the gene knockouts, the largest effects on exposures systematically pointed to a reduction relative to

controls (Fig. 4.4, right), suggesting that those genes tended to promote the mutational signature of interest.

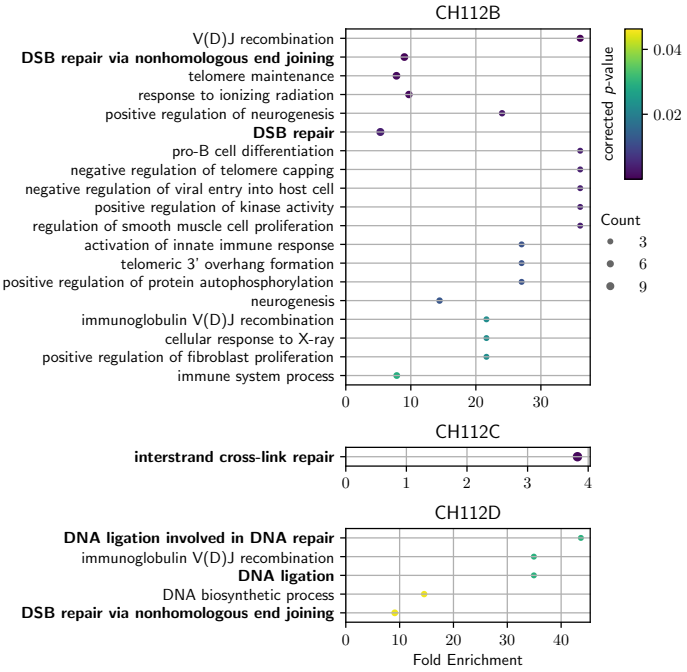


Fig. 4.5: Enriched biological processes for genes involved in signature depletions. Enriched Gene Ontology terms for genes whose knockouts promoted outlying depletion in the exposures of signatures CH112B-D (top to bottom). We analyzed the genes corresponding to outliers below $Q1 - 1.5 \times IQR$ in Fig.4.4. Signature CH112A did not yield significant results. Each plot shows, for each enriched process (vertical axis): number of genes annotated with the term among the outliers (circle size), fold enrichment as the ratio between the proportions of term-annotated genes among the outliers and among the full set of 766 knocked out genes (horizontal axis) and FDR-corrected p-value (color gradient). Terms directly related to DSB repair are highlighted in boldface.

To associate signatures with potential biological functions, we performed a functional enrichment analysis of Gene Ontology terms focusing on the genes causing the largest outlying depletion for each signature (see Methods). We found no significantly enriched terms for signature CH112A; however, its association with long MH deletions and the large depletion in exposure caused by *Polq* knockouts suggested a link with the MMEJ pathway (Figs. 4.3-4.4). The remaining three signatures, CH112B-D, were enriched for genes involved in DSB-related processes (Fig. 4.5). Specifically, CH112C was associated with

interstrand cross-link repair, a function of the FA core complex. Signatures CH112B and CH112D were both linked to non-homologous end joining (NHEJ), with CH112D specifically involved in DNA ligation via the ligase IV complex, and a higher frequency of insertions suggesting a role of the ligase IV complex in producing these mutations.

We recovered signatures linked to similar functions from the Barazas2025 screen of X genes on one target site (Supplementary Figs. 4.S4-4.S7), confirming the robustness of the NMF-based analysis. Naturally, the signatures were also similar but not identical across the datasets, as expected given the differences in gene sets, target sites, and sequencing depth. For instance, signature CH112C responsible for HDR events and related to the FA core complex and cross-link repair (Figs. 4.2/4.5), seemed to branch into two Barazas2025 signatures (Supplementary Fig. 4.S4): CH32A, producing HDR events and MH deletions, and influenced by FA core complex genes; and CH32E, linked to HDR events and longer MH deletions, influenced by members of the *Trp53bp1* pathway only present in the Barazas2025 dataset (Shieldin complex genes *Shld1* and *Shld2*; CST complex genes *Ctc1*, *Stn1*, and *Ten1*; and genes *Rif1* and *Mad2l2*, [25]) . Interestingly, the knockout of genes *Rnf8/Rnf168* led to depletion of both signatures CH32A and CH32E, with larger impact on CH32E. This effect was consistent with the role of the *Rnf8/Rnf168* genes as regulators of pathway choice and recruitment of repair factors, including *Trp53bp1* [26].

Overall, the analysis of exposure profiles enabled us to identify key drivers of mutational signatures, which we further linked to biological functions to obtain fingerprints of DSB repair activity at CRISPR-induced DSBs.

4.3.3 Exposures suggest DSB repair role for *Dbr1* and *Hnrnpk* genes

We analyzed genes with large knockout signature depletions and no direct annotations with DSB repair terms to infer their potential roles. In both datasets, *Dbr1* (Debranching RNA Lariats 1) emerged as the top unannotated gene. Knockouts of *Dbr1* resulted in a significant depletion of signatures CH112A and CH112C, respectively linked to MMEJ and FA activity. Similarly, knockouts of *Hnrnpk* (Heterogeneous Nuclear Ribonucleoprotein K), present only in the primary dataset, also showed depletion of CH112A and CH112C. It has been suggested that *Hnrnpk* could be a cofactor of p53-mediated DNA damage response, possibly including activation of *Rrm2* as a downstream

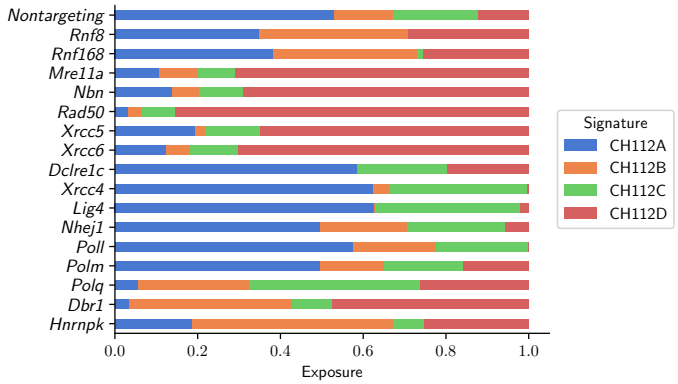


Fig. 4.6: Signature exposure profiles of selected genes. Each signature exposure value was divided by the sum of exposures per knockout and summarized per gene by taking the geometric mean value across knockouts of that gene.

target involved in nucleotide metabolism [12, 39, 5]. Our findings based on signature exposures indicated that *Dbr1* and *Hnrnpk* might specifically influence MMEJ and FA, but further experimental validation would be necessary to elucidate their precise roles in DSB repair.

4.3.4 Exposure analysis challenges existing repair models

Finally, we investigated if the ability of NMF to capture local co-occurrence patterns of subsets of mutational outcomes across subsets of gene knockouts could reveal genes involved in multiple DSB repair pathways. To achieve this, we analyzed the impact of gene knockouts simultaneously across every pair of signatures, focusing our interpretation on genes with known DSB repair pathway associations (Fig. 4.7). We observed that the knockout of key NHEJ genes and core members of the *Lig4* complex (*Xrcc4* or *Lig4*) promoted a large depletion of signatures CH112B and CH112D, respectively comprising shorter deletions and primarily 1bp insertions. This finding was consistent with the role of the *Lig4* complex in the final ligation step of NHEJ, whose disruption would be expected to hamper the overall NHEJ pathway function.

Interestingly, knockout of other NHEJ genes known as members of the Ku complex and involved in the initiation of repair by NHEJ, *Xrcc5* or *Xrcc6*, led to the depletion of CH112B but not CH112D. Given the association of CH112D with the *Lig4* complex, these results suggested a shift towards Ku-*independent*

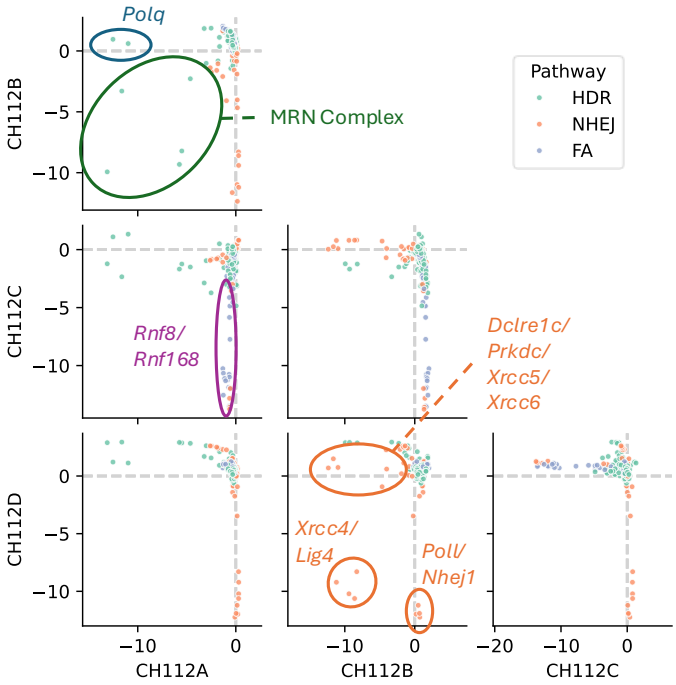


Fig. 4.7: Interplay of signatures based on shared depletion patterns. Pairwise signature depletion scatterplots, with every data point denoting a gene knockout coloured by the DSB repair pathway. The two axes represent signature depletion for the pair of signatures of interest, defined as the log2 fold change in exposure for the gene knockout sample relative to the geometric mean of the non-targeting controls. Grey dashed lines denote zero change. Notable gene knockouts are highlighted using ellipses and labelled.

NHEJ as the dominant repair mechanism in the absence of *Xrcc5* or *Xrcc6*. While classical NHEJ models consider *Xrcc5/Xrcc6/Dclre1c* required for *Lig4* activity, there is growing evidence for alternative NHEJ circumventing this requirement, though this remains poorly characterized [34, 11, 10, 38].

We also observed some similarity in the exposure profiles of the core *Lig4-Xrcc4* heterodimer and its accessory factors *Poll* and *Nhej1*, with all four genes causing a depletion of signature CH112D (Fig. 4.6), suggesting they could be responsible for NHEJ-related insertions [7]. Other candidate polymerases typically involved in NHEJ insertions, such as *Polm* [30], did not show the same effect on signature CH112D, highlighting the potential of NMF signature analysis to refine DSB repair function.

Knockouts of MRN complex genes reduced CH112A and CH112B activity, indicating the suppression of classical NHEJ and MMEJ pathways. In the absence of MRN complex genes, the Ku-*independent* NHEJ signature CH112D emerged as the predominant repair pathway. Overall, this multi-signature analysis of depletion patterns suggested a high degree of adaptability and cross-talk among DSB repair mechanisms. In particular, elements of how NHEJ compensates for inter- and intra-pathway deficiencies may deserve greater experimental attention.

4.4 Conclusion

In this work, we proposed a computational strategy to infer functions for genes in DSB repair based on high-resolution mutational spectra obtained following gene perturbation screens with CRISPR targeting. Specifically, we identify signatures of mutational activity influenced by both known and candidate DSB repair genes using NMF, attribute these signatures to established DSB repair mechanisms, and link these functions to new genes.

Our NMF analysis of the mutational spectra we generated for a combined set of 307 known and 459 candidate genes revealed an influence of *Dbr1* and *Hnrnpk* knockouts on MMEJ and FA activity, indicating potential roles for these genes in DSB repair. Additionally, signature contributions to mutational spectra revealed two signatures with varying dependencies on members of the *Lig4* complex, representing distinct branches of the NHEJ pathway. This provided evidence of *Lig4* complex activity in the absence of upstream NHEJ factors like *Xrcc5* or *Xrcc6*, suggesting that NHEJ could function independently of *Ku* factors.

Overall, our study highlights the potential of computational approaches for dissecting mutational patterns in CRISPR screens and exploring the genetic landscape of DSB repair. In particular, NMF enables us to identify mutational signatures that serve as fingerprints of repair pathway activity, powering an unsupervised approach to exploring large-scale CRISPR mutational data with the potential for nuanced interpretations and discovery of functional relationships between genes and DSB repair pathways.

4.5 Supplementary Tables

Dataset	Target	Replicate	Num. sgRNAs	Num. Frequent Outcomes
Primary	T1	1	2156	29
		2	2235	28
		3*	**2291**	28
	T2	1	2240	38
		2	2249	39
		3*	**2291**	40
	T3	1	2304	44
		2	2291	44
		3*	**2304**	44
Barazas2025	T1	1	2358	32
		2*	**2366**	32
		3	2350	32

Tab. 4.S1: Breakdown of the number of sgRNAs for which mutational spectra were recovered and the number of frequently occurring outcomes per dataset, target site, and replicate. * denotes the replicate that was selected for downstream analysis.

	Type	Start	End	Ins	MH	Repair Product
1	D	-10.0	2.0		1.0	CAAGG----------|--AGGGCCTATTTC
2	D	-10.0	5.0		3.0	CAAGG----------|-----GCCTATTTC
3	D	-10.0	6.0		2.0	CAAGG----------|------CCTATTTC
4	D	-11.0	2.0		3.0	CAAG-----------|--AGGGCCTATTTC
5	D	-13.0	17.0		3.0	CA-------------|--------------
6	D	-14.0	8.0		2.0	C--------------|--------TATTTC
7	D	-1.0	5.0		3.0	CAAGGTCGGGCAGG-|-----GCCTATTTC
8	D	-1.0	6.0		2.0	CAAGGTCGGGCAGG-|------CCTATTTC
9	D	-24.0	8.0		5.0	---------------|--------TATTTC
10	D	-2.0	2.0		2.0	CAAGGTCGGGCAG--|--AGGGCCTATTTC
11	D	-3.0	0.0		1.0	CAAGGTCGGGCA---|AGAGGGCCTATTTC
12	D	-3.0	17.0		2.0	CAAGGTCGGGCA---|--------------
13	D	-4.0	7.0		4.0	CAAGGTCGGGC----|-------CTATTTC
14	D	-5.0	5.0		2.0	CAAGGTCGGG-----|-----GCCTATTTC
15	D	-6.0	6.0		2.0	CAAGGTCGG------|------CCTATTTC
16	D	-8.0	7.0		1.0	CAAGGTC--------|-------CTATTTC
17	D	-8.0	8.0		1.0	CAAGGTC--------|--------TATTTC

Continued on next page

– continued from previous page

	Type	Start	End	Ins	MH	Repair Product
18	D	0.0	10.0		1.0	CAAGGTCGGGCAGGA|----------TTTC
19	D	0.0	1.0		1.0	CAAGGTCGGGCAGGA|-GAGGGCCTATTTC
20	D	0.0	3.0		2.0	CAAGGTCGGGCAGGA|---GGGCCTATTTC
21	D	0.0	4.0		0.0	CAAGGTCGGGCAGGA|----GGCCTATTTC
22	D	0.0	6.0		0.0	CAAGGTCGGGCAGGA|------CCTATTTC
23	D/L	0.0	3.0	G	0.0	CAAGGTCGGGCAGGAG---GGGCCTATTTC
24	HDR					AGGTCGGGCAGGAXAGAGGGCCTATT
25	I	0.0	0.0	GG		AAGGTCGGGCAGGAGGAGAGGGCCTATTTC
26	I	0.0	0.0	G		CAAGGTCGGGCAGGAGAGAGGGCCTATTTC
27	I	0.0	0.0	T		CAAGGTCGGGCAGGATAGAGGGCCTATTTC
28	I	1.0	1.0	A		CAAGGTCGGGCAGGAAAGAGGGCCTATTTC
29	TINS					GGTCGGGCAGGAXAGAGGGCCTATT

Tab. 4.S2: T1 frequently occurring repair outcomes across replicates. Type can be one of DELETION (D), INSERTION (I), DELINS (D/L), HDR, or TINS. Start describes the nucleotide position that the deletion begins relative to the cut site, where the cut site itself is position 0. Negative integer values indicate the position is upstream of the cut site. End describes the nucleotide position that the deletion/insertion ends relative to the cut site. InsSeq describes the inserted sequence. MH describes the microhomology length for deletions that feature a microhomology, and is 0 otherwise.

	Type	Start	End	Ins	MH	Repair Product
1	D	-1.0	2.0		0.0	CCTTGGACGCGTAG-|--CCGGTACTAACC
2	D	-1.0	8.0		1.0	CCTTGGACGCGTAG-|--------CTAACC
3	D	-2.0	0.0		0.0	CCTTGGACGCGTA--|GCCCGGTACTAACC
4	D	-2.0	2.0		0.0	CCTTGGACGCGTA--|--CCGGTACTAACC
5	D	-2.0	3.0		0.0	CCTTGGACGCGTA--|---CGGTACTAACC
6	D	-3.0	6.0		2.0	CCTTGGACGCGT---|------TACTAACC
7	D	-3.0	8.0		2.0	CCTTGGACGCGT---|--------CTAACC
8	D	-4.0	11.0		2.0	CCTTGGACGCG----|----------ACC
9	D	-4.0	1.0		2.0	CCTTGGACGCG----|-CCCGGTACTAACC
10	D	-4.0	2.0		2.0	CCTTGGACGCG----|--CCGGTACTAACC
11	D	-5.0	129.0		7.0	CCTTGGACGC-----|--------------

Continued on next page

– continued from previous page

	Type	Start	End	Ins	MH	Repair Product
12	D	-5.0	2.0		1.0	CCTTGGACGC-----|--CCGGTACTAACC
13	D	-5.0	5.0		3.0	CCTTGGACGC-----|-----GTACTAACC
14	D	-7.0	9.0		2.0	CCTTGGAC-------|---------TAACC
15	D	-9.0	0.0		0.0	CCTTGG---------|GCCCGGTACTAACC
16	D	-9.0	1.0		0.0	CCTTGG---------|-CCCGGTACTAACC
17	D	-9.0	4.0		2.0	CCTTGG---------|----GGTACTAACC
18	D	0.0	11.0		1.0	CCTTGGACGCGTAGG|----------ACC
19	D	0.0	21.0		5.0	CCTTGGACGCGTAGG|----------
20	D	0.0	2.0		1.0	CCTTGGACGCGTAGG|--CCGGTACTAACC
21	D	0.0	3.0		0.0	CCTTGGACGCGTAGG|---CGGTACTAACC
22	D	0.0	5.0		1.0	CCTTGGACGCGTAGG|-----GTACTAACC
23	D	0.0	7.0		2.0	CCTTGGACGCGTAGG|-------ACTAACC
24	D	1.0	11.0		2.0	CCTTGGACGCGTAGG|G----------ACC
25	D	1.0	2.0		1.0	CCTTGGACGCGTAGG|G-CCGGTACTAACC
26	D	1.0	33.0		5.0	CCTTGGACGCGTAGG|G----------
27	D	1.0	5.0		1.0	CCTTGGACGCGTAGG|G----GTACTAACC
28	D	2.0	11.0		2.0	CCTTGGACGCGTAGG|GC---------ACC
29	D/L	-1.0	1.0	TA		CTTGGACGCGTAG-TA-CCCGGTACTAACC
30	D/L	-2.0	1.0	TA		CTTGGACGCGTA--TA-CCCGGTACTAACC
31	D/L	0.0	1.0	GTA		CTTGGACGCGTAGGGTA-CCCGGTACTAAC
32	D/L	0.0	2.0	G		CCTTGGACGCGTAGGG--CCGGTACTAACC
33	D/L	0.0	2.0	T		CCTTGGACGCGTAGGT--CCGGTACTAACC
34	HDR					TTGGACGCGTAGGXGCCCGGTACTAA
35	I	0.0	0.0	A		CCTTGGACGCGTAGGAGCCCGGTACTAACC
36	I	0.0	0.0	GT		CTTGGACGCGTAGGGTGCCCGGTACTAACC
37	I	0.0	0.0	G		CCTTGGACGCGTAGGGGCCCGGTACTAACC
38	I	0.0	0.0	T		CCTTGGACGCGTAGGTGCCCGGTACTAACC
39	I	2.0	2.0	C		CCTTGGACGCGTAGGGCCCCGGTACTAACC
40	TINS					TGGACGCGTAGGXGCCCGGTACTAA

Continued on next page

– continued from previous page

	Type	Start	End	Ins	MH	Repair Product

Tab. 4.S3: T2 frequently occurring repair outcomes across replicates. Type can be one of DELETION (D), INSERTION (I), DELINS (D/L), HDR, or TINS. Start describes the nucleotide position that the deletion begins relative to the cut site, where the cut site itself is position 0. Negative integer values indicate the position is upstream of the cut site. End describes the nucleotide position that the deletion/insertion ends relative to the cut site. InsSeq describes the inserted sequence. MH describes the microhomology length for deletions that feature a microhomology, and is 0 otherwise.

	Type	Start	End	Ins	MH	Repair Product
1	D	-12.0	4.0		2.0	CCC----------|----GGCCCGGTAC
2	D	-13.0	9.0		2.0	CC----------|---------GGTAC
3	D	-16.0	11.0		3.0	----------|----------TAC
4	D	-1.0	30.0		4.0	CCCGACCTTGGACG-|----------
5	D	-1.0	7.0		1.0	CCCGACCTTGGACG-|-------CCGGTAC
6	D	-1.0	8.0		1.0	CCCGACCTTGGACG-|--------CGGTAC
7	D	-23.0	7.0		4.0	----------|-------CCGGTAC
8	D	-2.0	0.0		0.0	CCCGACCTTGGAC--|GTAGGGCCCGGTAC
9	D	-2.0	2.0		0.0	CCCGACCTTGGAC--|--AGGGCCCGGTAC
10	D	-2.0	4.0		1.0	CCCGACCTTGGAC--|----GGCCCGGTAC
11	D	-3.0	1.0		0.0	CCCGACCTTGGA---|-TAGGGCCCGGTAC
12	D	-3.0	2.0		1.0	CCCGACCTTGGA---|--AGGGCCCGGTAC
13	D	-3.0	5.0		1.0	CCCGACCTTGGA---|-----GCCCGGTAC
14	D	-3.0	6.0		2.0	CCCGACCTTGGA---|------CCCGGTAC
15	D	-3.0	8.0		0.0	CCCGACCTTGGA---|--------CGGTAC
16	D	-4.0	12.0		1.0	CCCGACCTTGG----|----------AC
17	D	-4.0	28.0		5.0	CCCGACCTTGG----|----------
18	D	-4.0	2.0		1.0	CCCGACCTTGG----|--AGGGCCCGGTAC
19	D	-4.0	6.0		2.0	CCCGACCTTGG----|------CCCGGTAC
20	D	-5.0	16.0		5.0	CCCGACCTTG-----|----------
21	D	-5.0	2.0		2.0	CCCGACCTTG-----|--AGGGCCCGGTAC
22	D	-6.0	10.0		2.0	CCCGACCTT------|----------GTAC
23	D	-6.0	3.0		1.0	CCCGACCTT------|---GGGCCCGGTAC
24	D	-7.0	10.0		2.0	CCCGACCT-------|----------GTAC

Continued on next page

– continued from previous page

	Type	Start	End	Ins	MH	Repair Product
25	D	-7.0	3.0		1.0	CCCGACCT-------|---GGGCCCGGTAC
26	D	-8.0	1.0		2.0	CCCGACC--------|-TAGGGCCCGGTAC
27	D	-8.0	3.0		2.0	CCCGACC--------|---GGGCCCGGTAC
28	D	-9.0	1.0		0.0	CCCGAC---------|-TAGGGCCCGGTAC
29	D	-9.0	6.0		2.0	CCCGAC---------|------CCCGGTAC
30	D	0.0	16.0		1.0	CCCGACCTTGGACGC|--------------
31	D	0.0	1.0		0.0	CCCGACCTTGGACGC|-TAGGGCCCGGTAC
32	D	0.0	5.0		1.0	CCCGACCTTGGACGC|-----GCCCGGTAC
33	D	0.0	6.0		1.0	CCCGACCTTGGACGC|------CCCGGTAC
34	D	1.0	10.0		1.0	CCCGACCTTGGACGC|G---------GTAC
35	D	1.0	13.0		2.0	CCCGACCTTGGACGC|G----------C
36	D	1.0	3.0		2.0	CCCGACCTTGGACGC|G--GGGCCCGGTAC
37	D	1.0	9.0		1.0	CCCGACCTTGGACGC|G--------GGTAC
38	D	2.0	6.0		1.0	CCCGACCTTGGACGC|GT----CCCGGTAC
39	D/L	-1.0	2.0	G		CCCGACCTTGGACG-G--AGGGCCCGGTAC
40	D/L	-3.0	2.0	G		CCCGACCTTGGA---G--AGGGCCCGGTAC
41	D/L	0.0	2.0	CGG		CCGACCTTGGACGCCGG--AGGGCCCGGTA
42	HDR					CGACCTTGGACGCXGTAGGGCCCGGT
43	I	0.0	0.0	C		CCCGACCTTGGACGCCGTAGGGCCCGGTAC
44	I	1.0	1.0	G		CCCGACCTTGGACGCGGTAGGGCCCGGTAC
45	TINS					GACCTTGGACGCXGTAGGGCCCGGT

Tab. 4.S4: T3 frequently occurring repair outcomes across replicates. Type can be one of DELETION (D), INSERTION (I), DELINS (D/L), HDR, or TINS. Start describes the nucleotide position that the deletion begins relative to the cut site, where the cut site itself is position 0. Negative integer values indicate the position is upstream of the cut site. End describes the nucleotide position that the deletion/insertion ends relative to the cut site. InsSeq describes the inserted sequence. MH describes the microhomology length for deletions that feature a microhomology, and is 0 otherwise.

	Type	Start	End	Ins	MH	Repair Product
1	D	-10.0	2.0		1.0	CAAGG----------|--AGGGCCTATTTC
2	D	-10.0	5.0		3.0	CAAGG----------|-----GCCTATTTC

Continued on next page

– continued from previous page

	Type	Start	End	Ins	MH	Repair Product
3	D	-10.0	6.0		2.0	CAAGG----------|------CCTATTTC
4	D	-11.0	2.0		3.0	CAAG----------|--AGGGCCTATTTC
5	D	-13.0	17.0		3.0	CA----------|----------
6	D	-14.0	8.0		2.0	C----------|--------TATTTC
7	D	-1.0	5.0		3.0	CAAGGTCGGGCAGG-|-----GCCTATTTC
8	D	-1.0	6.0		2.0	CAAGGTCGGGCAGG-|------CCTATTTC
9	D	-24.0	8.0		5.0	----------|--------TATTTC
10	D	-2.0	2.0		2.0	CAAGGTCGGGCAG--|--AGGGCCTATTTC
11	D	-3.0	0.0		1.0	CAAGGTCGGGCA---|AGAGGGCCTATTTC
12	D	-3.0	17.0		2.0	CAAGGTCGGGCA---|----------
13	D	-4.0	7.0		4.0	CAAGGTCGGGC----|-------CTATTTC
14	D	-5.0	5.0		2.0	CAAGGTCGGG-----|-----GCCTATTTC
15	D	-6.0	6.0		2.0	CAAGGTCGG------|------CCTATTTC
16	D	-8.0	7.0		1.0	CAAGGTC--------|-------CTATTTC
17	D	-8.0	8.0		1.0	CAAGGTC--------|--------TATTTC
18	D	-9.0	2.0		0.0	CAAGGT---------|--AGGGCCTATTTC
19	D	0.0	10.0		1.0	CAAGGTCGGGCAGGA|----------TTTC
20	D	0.0	1.0		1.0	CAAGGTCGGGCAGGA|-GAGGGCCTATTTC
21	D	0.0	3.0		2.0	CAAGGTCGGGCAGGA|---GGGCCTATTTC
22	D	0.0	4.0		0.0	CAAGGTCGGGCAGGA|----GGCCTATTTC
23	D	0.0	6.0		0.0	CAAGGTCGGGCAGGA|------CCTATTTC
24	D	0.0	7.0		0.0	CAAGGTCGGGCAGGA|-------CTATTTC
25	D/L	-3.0	1.0	AG	0.0	AAGGTCGGGCA---AG-GAGGGCCTATTTC
26	D/L	0.0	3.0	G		CAAGGTCGGGCAGGAG---GGGCCTATTTC
27	HDR					AGGTCGGGCAGGAXAGAGGGCCTATT
28	I	0.0	0.0	GG		AAGGTCGGGCAGGAGGAGAGGGCCTATTTC
29	I	0.0	0.0	G		CAAGGTCGGGCAGGAGAGAGGGCCTATTTC
30	I	0.0	0.0	T		CAAGGTCGGGCAGGATAGAGGGCCTATTTC
31	I	1.0	1.0	ATA		AAGGTCGGGCAGGAAATAGAGGGCCTATTT
32	I	1.0	1.0	A		CAAGGTCGGGCAGGAAAGAGGGCCTATTTC
33	TINS					GGTCGGGCAGGAXAGAGGGCCTATT

Continued on next page

– continued from previous page

	Type	Start	End	Ins	MH	Repair Product

Tab. 4.S5: Barazas2025 frequently occurring repair outcomes across replicates. Type can be one of DELETION (D), INSERTION (I), DELINS (D/L), HDR, or TINS. Start describes the nucleotide position that the deletion begins relative to the cut site, where the cut site itself is position 0. Negative integer values indicate the position is upstream of the cut site. End describes the nucleotide position that the deletion/insertion ends relative to the cut site. InsSeq describes the inserted sequence. MH describes the microhomology length for deletions that feature a microhomology, and is 0 otherwise.

Dataset	Target	Replicate	1	2	3
Primary	T1	1	1.000	0.982	0.983
		2	0.982	1.000	0.984
		3	0.983	0.984	1.000
	T2	1	1.000	0.981	0.983
		2	0.981	1.000	0.982
		3	0.982	0.983	1.000
	T3	1	1.000	0.990	0.990
		2	0.990	1.000	0.990
		3	0.990	0.990	1.000
Barazas2025	T1	1	1.000	0.993	0.992
		2	0.993	1.000	0.992
		3	0.992	0.992	1.000

Tab. 4.S5: Mean Pearson's correlation coefficient across common genes and outcomes between mutational spectra of replicates for the same target site.

4.6 Supplementary Figures

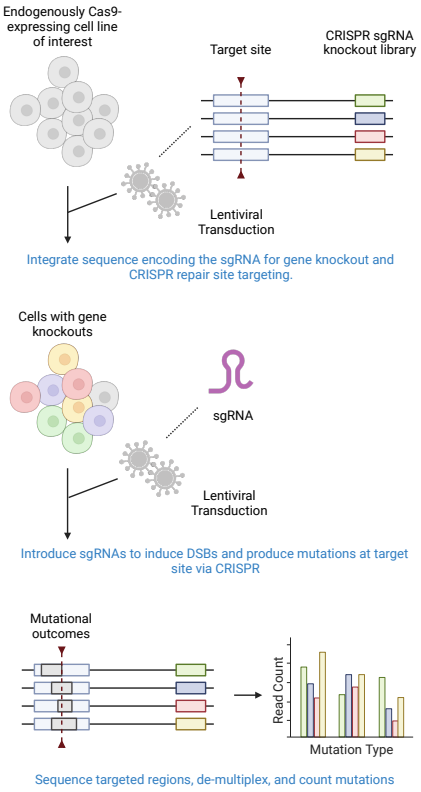


Fig. 4.S1: Illustration of CRISPR gene knockout screens with mutational spectra readout. First, sequences are integrated into the genomes of cells via lentiviral transduction. Each sequence contains two elements: *(i)* a sgRNA-encoding region to knockout a single gene, and *(ii)* a region common to all integrated sequences to be targeted with CRISPR to produce the mutational spectra. After genomic integration, several days of cell culture are allowed for genes to be knocked out. Following this, lentiviral transduction is used to introduce sgRNAs targeting the common region to the Cas9-expressing cells. After allowing time for cell culture for DNA cleavage and repair, DNA sequencing was performed to capture the final CRISPR repair products.

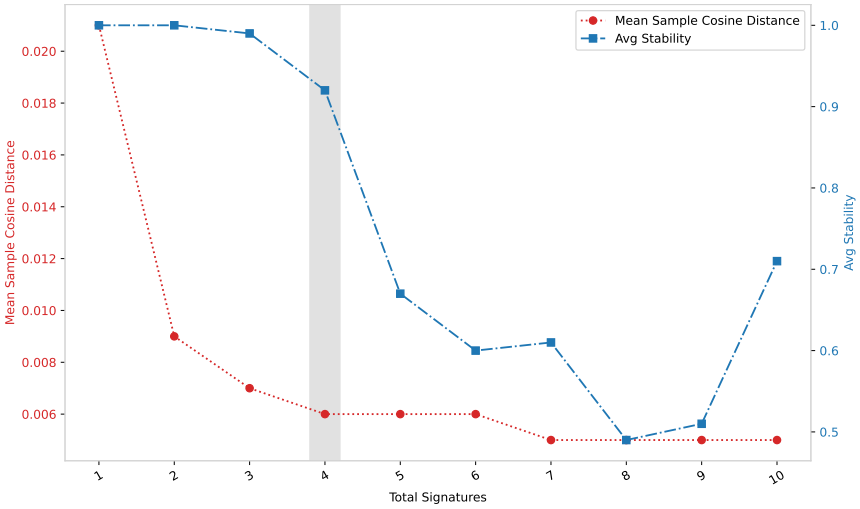


Fig. 4.S2: Comparison of NMF solutions for different k values for primary dataset. The left axis represents the reconstruction error as "Mean Sample Cosine Distance". Lower error means better reconstruction. The right axis represents the stability of the solution as the "average silhouette similarity". A higher stability means the solution does not vary a lot between successive runs of the SigProfiler algorithm. The horizontal axis represents the different values tested under for k.

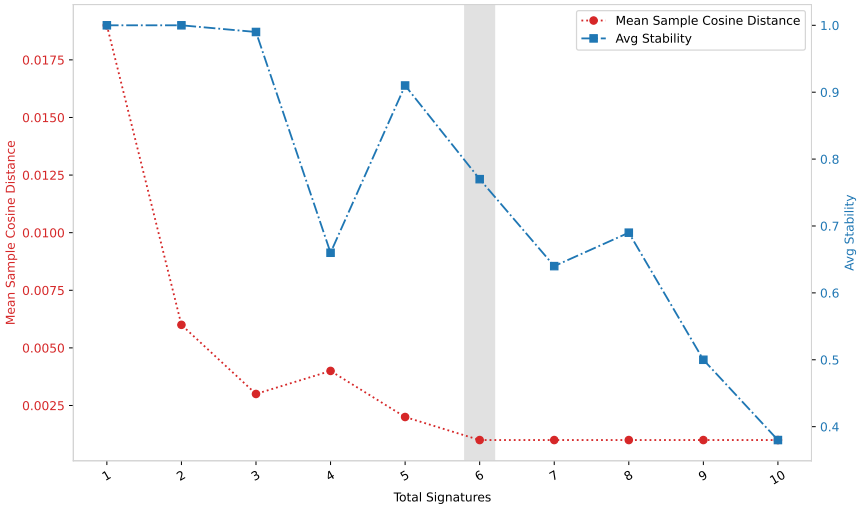


Fig. 4.S3: Comparison of NMF solutions for different k values for Barazas2025 dataset. The left axis represents the reconstruction error as "Mean Sample Cosine Distance". Lower error means better reconstruction. The right axis represents the stability of the solution as the "average silhouette similarity". A higher stability means the solution does not vary a lot between successive runs of the SigProfiler algorithm. The horizontal axis represents the different values tested under for k.

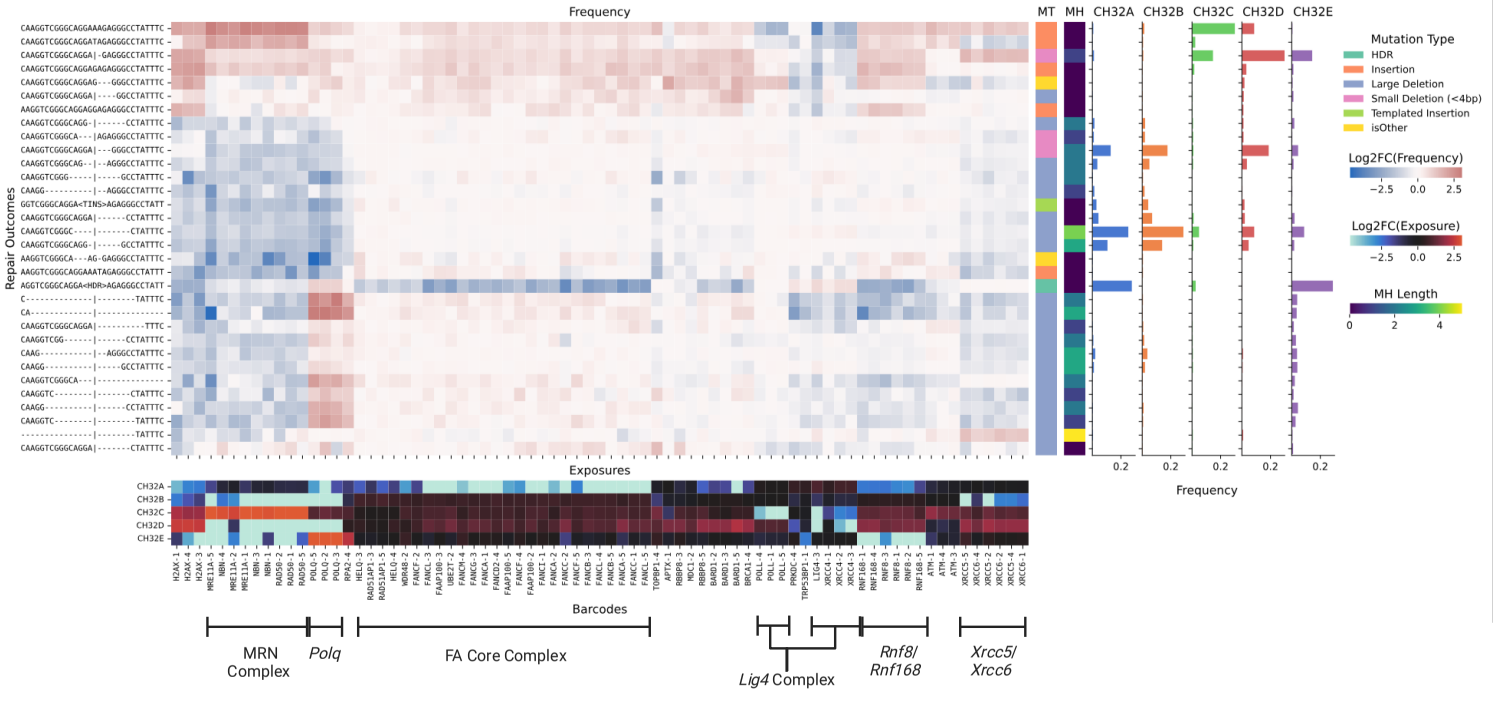
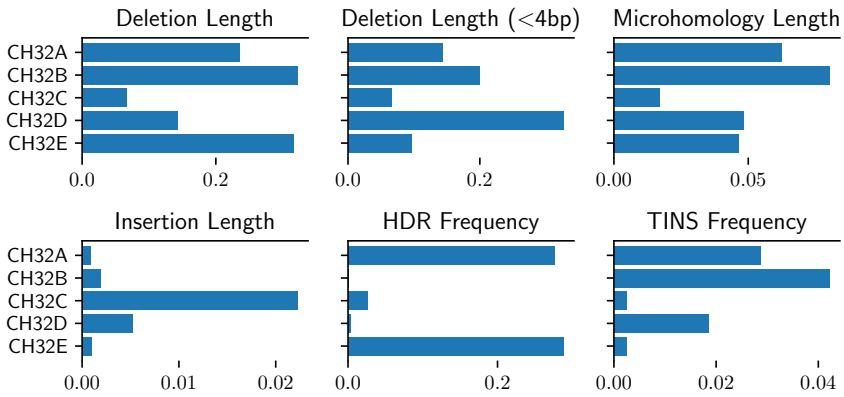

Frequency
Repair Outcomes
MT
MH
CH32A
CH32B
CH32C
CH32D
CH32E
Mutation Type
HDR
Insertion
Large Deletion
Small Deletion (<4bp)
Templated Insertion
isOther
Log2FC(Frequency)
Log2FC(Exposure)
MH Length
Exposures
Barcodes
MRN Complex
Polq
FA Core Complex
Lig4 Complex
Rnf8/ Rnf168
Xrcc5/ Xrcc6

Fig. 4.S4: Impact of gene knockouts on mutational spectra and identified mutational signatures for the Barazas2025 dataset. (Central heatmap) Effect of the 75 gene knockouts with the largest outcome redistribution (columns) on every mutational outcome across three target sites (rows), with outcomes further characterized by mutation type (MT) and microhomology length (MH). Effect quantified using the log2 fold-change in outcome frequency between the gene knockout spectrum and the geometric mean of the non-targeting control spectra. Gene knockouts and mutational outcomes ordered based on hierarchical clustering with Ward linkage and Euclidean distance. (Signature bar plots) Mutational profiles of the identified signatures, with bars denoting the frequency of each mutational outcome. (Bottom heatmap) Depletion in exposure of each identified signature per gene knockout, as the log2 fold-change between the exposure of the gene knockout sample and the geometric mean exposure of the non-targeting controls. Key DSB repair genes and complexes highlighted.



Average repair outcome characteristics (weighted by outcome frequency).

Fig. 4.S5: Barazas2025 signature deletion and insertion properties. Per signature average of deletion and insertion outcome properties. Outcome properties: deletion length, small deletion length (< 4bp), deletion microhomology (MH) length, insertion length, homology-directed repair (HDR) frequency, and templated insertion (TINS) frequency. The first four length properties are weighted by the signature frequency of the respective outcomes.

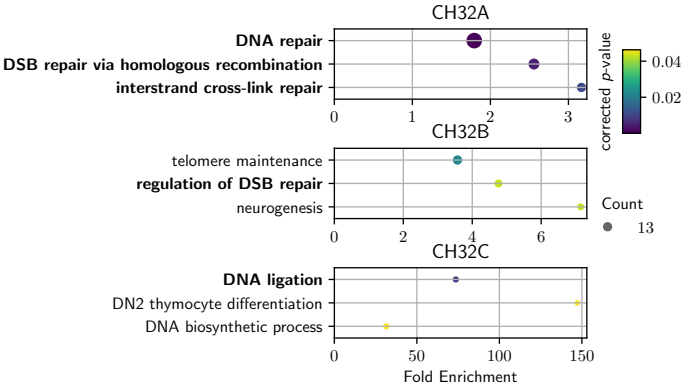


Fig. 4.S6: Enriched biological processes for genes involved in Barazas2025 signature depletions. Enriched Gene Ontology terms for genes whose knockouts promoted outlying depletion in the exposures of signatures CH112B-D (top to bottom). We analyzed the genes corresponding to outliers below $Q1 - 1.5 \times IQR$ in Fig.4.4. Signature CH112A did not yield significant results. Each plot shows, for each enriched process (vertical axis): number of genes annotated with the term among the outliers (circle size), fold enrichment as the ratio between the proportions of term-annotated genes among the outliers and among the full set of 766 knocked out genes (horizontal axis) and FDR-corrected p-value (color gradient). Terms directly related to DSB repair are highlighted in boldface.

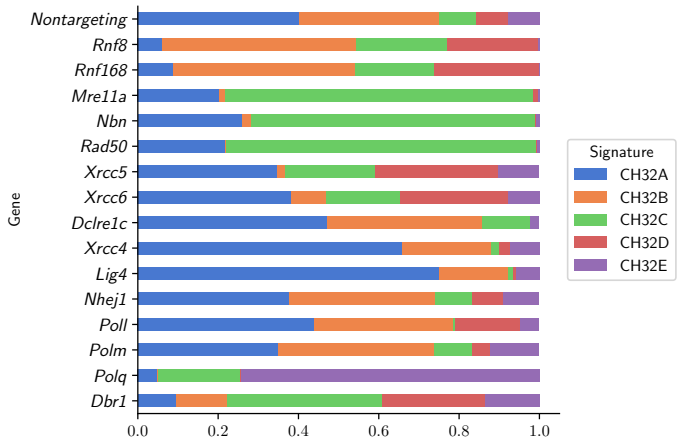


Fig. 4.S7: Barazas2025 signature exposure profiles of selected genes. Each signature exposure value was divided by the sum of exposures per knockout, and summarized per gene by taking the geometric mean value. across knockouts of that gene.

References

[1] Samah W Awwad, Almudena Serrano-Benitez, John C Thomas, Vipul Gupta, and Stephen P Jackson. "Revolutionizing DNA repair research and cancer therapy with CRISPR–Cas screens". In: *Nature Reviews Molecular Cell Biology* (2023), pp. 1–18 (cit. on p. 98).

[2] Richard M Baldarelli, Cynthia L Smith, Martin Ringwald, Joel E Richardson, and Carol J Bult. "Mouse Genome Informatics: an integrated knowledgebase system for the laboratory mouse". In: *Genetics* 227.1 (2024), iyae031 (cit. on p. 105).

[3] Marco Barazas, Robin van Schendel, and Marcel Tijsterman. "Mutational signature catalogue (MUSIC) of DNA double-strand break repair". In: *bioRxiv* (2025). eprint: `https://www.biorxiv.org/content/early/2025/04/29/2025.04.29.650617.full.pdf` (cit. on pp. 99, 101).

[4] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300 (cit. on p. 105).

[5] Elke Braems, Valérie Bercier, Evelien Van Schoor, et al. "HNRNPK alleviates RNA toxicity by counteracting DNA damage in C9orf72 ALS". In: *Acta Neuropathologica* 144.3 (2022), pp. 465–488 (cit. on p. 112).

[6] Wendy J Cannan and David S Pederson. "Mechanisms and consequences of double-strand DNA break formation in chromatin". In: *Journal of Cellular Physiology* 231.1 (2016), pp. 3–14 (cit. on p. 98).

[7] Howard HY Chang, Nicholas R Pannunzio, Noritaka Adachi, and Michael R Lieber. "Non-homologous DNA end joining and alternative pathways to double-strand break repair". In: *Nature reviews Molecular cell biology* 18.8 (2017), pp. 495–506 (cit. on p. 113).

[8] Alice Chen. "PARP inhibitors: its role in treatment of cancer". In: *Chinese journal of cancer* 30.7 (2011), p. 463 (cit. on p. 98).

[9] The Gene Ontology Consortium. "The Gene Ontology resource: enriching a GOld mine". In: *Nucleic acids research* 49.D1 (2021), pp. D325–D334 (cit. on p. 101).

[10] Farjana Fattah, Eu Han Lee, Natalie Weisensel, et al. “Ku regulates the non-homologous end joining pathway choice of DNA double-strand break repair in human somatic cells”. In: *PLoS Genetics* 6.2 (2010), e1000855 (cit. on p. 113).

[11] Josée Guirouilh-Barbat, Emilie Rass, Isabelle Plo, Pascale Bertrand, and Bernard S Lopez. “Defects in XRCC4 and KU80 differentially affect the joining of distal nonhomologous ends”. In: *Proceedings of the National Academy of Sciences* 104.52 (2007), pp. 20902–20907 (cit. on p. 113).

[12] Maite Huarte, Mitchell Guttman, David Feldser, et al. “A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response”. In: *Cell* 142.3 (2010), pp. 409–419 (cit. on p. 112).

[13] Jeffrey A Hussmann, Jia Ling, Purnima Ravisankar, et al. “Mapping the genetic landscape of DNA double-strand break repair”. In: *Cell* 184.22 (2021), pp. 5653–5669 (cit. on pp. 99, 101).

[14] SM Ashiqul Islam, Marcos Díaz-Gay, Yang Wu, et al. “Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor”. In: *Cell Genomics* 2.11 (2022) (cit. on p. 103).

[15] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, et al. “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity”. In: *Science* 337.6096 (2012), pp. 816–821 (cit. on p. 98).

[16] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. “KEGG for taxonomy-based analysis of pathways and genomes”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D587–D592 (cit. on p. 101).

[17] DV Klopfenstein, Liangsheng Zhang, Brent S Pedersen, et al. “GOATOOLS: A Python library for Gene Ontology analyses”. In: *Scientific reports* 8.1 (2018), pp. 1–17 (cit. on p. 105).

[18] Theo A Knijnenburg, Linghua Wang, Michael T Zimmermann, et al. “Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome atlas”. In: *Cell Reports* 23.1 (2018), pp. 239–254 (cit. on p. 105).

[19] Dmitry Kobak and George C Linderman. “Initialization is critical for preserving global data structure in both t-SNE and UMAP”. In: *Nature Biotechnology* 39.2 (2021), pp. 156–157 (cit. on p. 99).

[20] Gene Koh, Andrea Degasperi, Xueqing Zou, Sophie Momen, and Serena Nik-Zainal. “Mutational signatures: emerging concepts, caveats and clinical applications”. In: *Nature Reviews Cancer* 21.10 (2021), pp. 619–637 (cit. on p. 100).

[21] Brandon J Lamarche, Nicole I Orazio, and Matthew D Weitzman. “The MRN complex in double-strand break repair and telomere maintenance”. In: *FEBS letters* 584.17 (2010), pp. 3682–3695 (cit. on p. 109).

[22] Daniel D Lee and H Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (1999), pp. 788–791 (cit. on pp. 100, 103).

[23] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. “UMAP: Uniform Manifold Approximation and Projection”. In: *Journal of Open Source Software* 3.29 (2018), p. 861 (cit. on p. 99).

[24] Marija Milacic, Deidre Beavers, Patrick Conley, et al. “The Reactome Pathway Knowledgebase 2024”. In: *Nucleic Acids Research* 52.D1 (2024), pp. D672–D678 (cit. on p. 105).

[25] Zachary Mirman, Francisca Lottersberger, Hiroyuki Takai, et al. “53BP1–RIF1–shieldin counteracts DSB resection through CST-and Polα-dependent fill-in”. In: *Nature* 560.7716 (2018), pp. 112–116 (cit. on p. 111).

[26] Shinichiro Nakada. “Opposing roles of RNF8/RNF168 and deubiquitinating enzymes in ubiquitination-dependent DNA double-strand break response signaling and DNA-repair pathway choice”. In: *Journal of Radiation Research* 57.S1 (2016), pp. i33–i40 (cit. on p. 111).

[27] Michele Olivieri, Tiffany Cho, Alejandro Álvarez-Quilón, et al. “A genetic map of the response to DNA damage in human cells”. In: *Cell* 182.2 (2020), pp. 481–496 (cit. on p. 105).

[28] Megan van Overbeek, Daniel Capurso, Matthew M Carter, et al. “DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks”. In: *Molecular Cell* 63.4 (2016), pp. 633–646 (cit. on p. 99).

[29] Krishan Pal and Mayank Sharma. “Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space”. In: *2020 fourth international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC)*. IEEE. 2020, pp. 1106–1110 (cit. on p. 99).

[30] Dale A Ramsden. “Polymerases in nonhomologous end joining: building a bridge over broken chromosomes”. In: *Antioxidants & redox signaling* 14.12 (2011), pp. 2509–2519 (cit. on p. 113).

[31] Xingcheng Ran, Yue Xi, Yonggang Lu, Xiangwen Wang, and Zhenyu Lu. "Comprehensive survey on hierarchical clustering algorithms and the recent developments". In: *Artificial Intelligence Review* 56.8 (2023), pp. 8219–8264 (cit. on p. 99).

[32] JR Rideout, JG Caporaso, E Bolyen, et al. "scikit-bio/scikit-bio: scikit-bio 0.6. 1". In: *Zenodo. doi* 10 (2024) (cit. on p. 104).

[33] Robin van Schendel, Joost Schimmel, and Marcel Tijsterman. "SIQ: easy quantitative measurement of mutation profiles in sequencing data". In: *NAR Genomics and Bioinformatics* 4.3 (2022), lqac063 (cit. on p. 102).

[34] Leonie Schulte-Uentrop, Raafat A El-Awady, Lena Schliecker, Henning Willers, and Jochen Dahm-Daphi. "Distinct roles of XRCC4 and Ku80 in non-homologous end-joining of endonuclease-and ionizing radiation-induced DNA double-strand breaks". In: *Nucleic Acids Research* 36.8 (2008), pp. 2561–2569 (cit. on p. 113).

[35] Colm Seale, Marco Barazas, Robin van Schendel, Marcel Tijsterman, and Joana P. Gonçalves. "MUSICiAn: Genome-wide Identification of Genes Involved in DNA Repair via Control-Free Mutational Spectra Analysis". In: *bioRxiv* (2025). eprint: `https://www.biorxiv.org/content/early/2025/01/28/2025.01.27.635038.full.pdf` (cit. on pp. 99–101).

[36] Max W Shen, Mandana Arbab, Jonathan Y Hsu, et al. "Predictable and precise template-free CRISPR editing of pathogenic variants". In: *Nature* 563.7733 (2018), pp. 646–651 (cit. on p. 99).

[37] Jia Shou, Jinhuan Li, Yingbin Liu, and Qiang Wu. "Precise and predictable CRISPR chromosomal rearrangements reveal principles of Cas9-mediated nucleotide insertion". In: *Molecular cell* 71.4 (2018), pp. 498–509 (cit. on p. 99).

[38] Benjamin M Stinson, Sean M Carney, Johannes C Walter, and Joseph J Loparo. "Structural role for DNA Ligase IV in promoting the fidelity of non-homologous end joining". In: *Nature Communications* 15.1 (2024), p. 1250 (cit. on p. 113).

[39] Nadine Wiesmann, Judith Strozynski, Carina Beck, et al. "Knockdown of hnRNPK leads to increased DNA damage after irradiation and reduces survival of tumor cells". In: *Carcinogenesis* 38.3 (2017), pp. 321–328 (cit. on p. 112).

[40] Richard D Wood, Michael Mitchell, John Sgouros, and Tomas Lindahl. “Human DNA repair genes”. In: *Science* 291.5507 (2001), pp. 1284–1289 (cit. on p. 105).

[41] David W Wyatt, Wanjuan Feng, Michael P Conlin, et al. “Essential roles for polymerase θ-mediated end joining in the repair of chromosome breaks”. In: *Molecular cell* 63.4 (2016), pp. 662–673 (cit. on p. 99).

5

Overcoming Selection Bias in Synthetic Lethality Prediction

> *The discovery of many synthetic lethal interactions between proteins involved in DNA repair allowed the development of personalized therapeutic treatments that target specific DNA repair enzymes to kill cancer cells.*
>
> **— Lodovichi et al. 2020**
> (Inhibition of DNA Repair in Cancer Therapy: Toward a Multi-Target Approach)

Synthetic lethality (SL) between two genes occurs when simultaneous loss-of-function leads to cell death. This holds great promise for developing anti-cancer therapeutics that target synthetic lethal pairs of endogenously disrupted genes. Identifying novel SL relationships through exhaustive experimental screens is challenging, due to the vast number of candidate pairs. Computational SL prediction is therefore sought to identify promising SL gene pairs for further experimentation. However, current SL prediction methods lack consideration for generalisability in the presence of selection bias in SL data. We show that SL data exhibit considerable gene selection bias. Our experiments designed to assess robustness of SL prediction reveal that models driven by the topology of known SL interactions (e.g. graph, matrix factorisation) are especially sensitive to selection bias. We introduce selection bias-resilient synthetic lethality (SBSL) prediction using regularised logistic regression or random forests. Each gene

Colm Seale, Yasin Tepeli, & Joana P. Gonçalves. (2022). "Overcoming selection bias in synthetic lethality prediction." Bioinformatics, 38(18), 4360-4368, 10.1093/bioinformatics/btac523.

pair is described by 27 molecular features derived from cancer cell line, cancer patient tissue, and healthy donor tissue samples. SBSL models are built and tested using ∼8000 experimentally derived SL pairs across breast, colon, lung, and ovarian cancers. Compared to other SL prediction methods, SBSL showed higher predictive performance, better generalisability and robustness to selection bias. Gene dependency, quantifying the essentiality of a gene for cell survival, contributed most to SBSL predictions. Random forests were superior to linear models in the absence of dependency features, highlighting the relevance of mutual exclusivity of somatic mutations, co-expression in healthy tissue, and differential expression in tumour samples.

5.1 Introduction

Synthetic lethality (SL) describes a relationship between two genes where simultaneous loss-of-function in both genes causes cell death, but independent disruption of either gene does not affect cell viability. An SL relationship can be exploited for precision anti-cancer treatment by targeting a gene known to be synthetic lethal with another gene that is deleteriously mutated in the tumour cells. This targeted gene disruption not only induces the death of tumour cells, it is also unlikely to affect healthy cells if they do not carry the mutation. For instance, PARP inhibitor drugs are preferentially lethal towards tumour cells with BRCA1 or BRCA2 mutations and were the first SL-based therapy approved for use in the clinic [40]. Developing SL-based therapies requires the identification of novel SL interactions through SL loss-of-function screens, which silence gene pairs of interest and measure the respective effect on cell viability [47]. However, exhaustive screening is expensive and becomes impractical due to the vast number of possible gene pairs. This is where computational SL prediction comes into play to guide experimental follow-up and reduce screening to only promising SL pairs.

Previous computational approaches have derived SL through the analysis of gene mutation or expression [27, 63, 62], patient survival [33, 17], metabolic networks [50, 19], protein-protein interactions [30, 24], signalling pathways [66], existing SL networks [37, 23, 7], evolutionary conservation within and between species [9, 64, 41, 13], among others. We categorise existing SL prediction approaches as either SL topology-based or SL feature-based methods. Informally, SL topology prediction methods: (i) consider a limited prediction universe based on a predefined set of genes, usually induced

by the availability of SL labels; (ii) are explicitly aware of the gene-gene SL label graph structure, where nodes denote genes and edges denote SL relationships between pairs of genes. SL topology methods can be further categorised into matrix factorisation techniques like pca-gCMF [34], SL2MF [37], GRSMF [23], and graph-based methods like SLant [5], DDGCN [7], and GCATSL [38]. Conversely, SL feature methods are unaware of gene identity or the structure defined by the SL relationship labels and rely exclusively on molecular features of genes for the prediction task. For this reason, feature models can be used to predict an SL relationship for any pair of genes with a corresponding feature-based representation. SL feature methods include statistical techniques like DAISY [27] and BiSEp [63], and supervised learning models such as [41], DiscoverSL [12], and EXP2SL [62].

Significant challenges remain before existing SL prediction methods can be routinely used to guide experimental screening. To be effective, they must rank positive SL gene pairs consistently high across multiple datasets, be able to make predictions for unseen genes, and generalise to unseen gene pairs. However, most studies assess prediction performance under limited scenarios, for instance focusing on a single cancer type and testing of gene pairs whose genes individually appear in the training set. We hypothesise that some genes may be overrepresented in existing SL labels while others remain understudied for historical or academic reasons [56]. The extreme case, where SL labels are available for many pairs but involving only a few genes, is also likely to induce SL relationship biases because pairs involving the same gene are not independent from each other. We argue that the presence of strong biases in SL labels can lead to performance overestimation, particularly of SL topology models which are explicitly designed to exploit them.

In this work, we propose different experiments to assess sensitivity of SL prediction methods to selection biases. We also introduce SBSL (Selection Bias-resilient Synthetic Lethality) prediction models, with two main goals in mind: (i) improving model resilience to biases in SL prediction; and (ii) bridging the performance gap between SL topology and SL feature methods currently perceived in the SL prediction literature. To improve bias-resilience, we propose SL feature models based on supervised ML that explicitly ignore the structure of the SL label graph. To improve performance, we define novel features based on molecular data that could be relevant for SL prediction but remains underexplored in the SL prediction context. Specifically, these are:

the interaction between gene dependency scores (measuring cell viability upon gene silencing) and mutations in cancer cell lines [14, 46, 44, 4], as increased dependency on one gene in cell lines harbouring a deleterious mutation in another gene may indicate SL between the two; gene expression from healthy donor tissue, in addition to expression from patient tumour tissue, which could help identify tumour-specific changes in the relationship between the pair of genes; measures of mutual exclusivity, quantifying the non-co-occurrence of mutations in a pair of genes [2, 8]; change in survival time between cancer patients with and without mutations or aberrant expression in the pair of genes, both of which may be associated with SL [55, 33].

5.2 Methods

Our proposed models aim to predict if a given pair of genes is synthetically lethal for a specific cancer type, where the pair is described by a collection of molecular features. We approach it as a binary classification problem.

5.2.1 Data

Synthetic lethality labels. We obtained cancer-specific SL labels from two studies, ISLE and DiscoverSL [33, 12]. Together, they included thousands of SL relationships experimentally-derived by 21 other studies using double gene knockdown/knockout experiments or targeting of one gene using CRISPR or RNAi in contexts where the other gene is either endogenously inactive or rendered inactive through the use of drug compounds. We removed duplicate gene pair entries from ISLE and DiscoverSL separately by retaining a single entry if all entries agreed, or removing all duplicate entries if any of them disagreed on the label. To combine the two datasets, we reduced 63 gene pairs with duplicate entries across the datasets to a single entry per pair. In case of disagreement, we chose the label from DiscoverSL, since there was a lower level of disagreement within DiscoverSL than within ISLE. We ended up with 7962 labelled gene pairs distributed over the four cancer types that had at least 200 positive and negative labels after preprocessing, namely breast (BRCA), colon (COAD), lung (LUAD), and ovarian (OV) (Table 5.1). The ISLE, DiscoverSL, and combined SL gold standards had differing cancer type representations and class imbalances. We used the combined SL gold standard in our experiments except where otherwise specified.

	ISLE		DiscoverSL		Combined		Num.	
	+	-	+	-	+	-	Genes	Labelled %
BRCA	713	1168	835	72	1548	1240	1072	.39
COAD	859	806	0	0	859	806	1560	.14
LUAD	202	5155	347	312	549	5467	804	1.66
OV	223	449	0	0	223	449	86	18.14
All	1997	7578	1182	384	3179	7962	3072	.05

Tab. 5.1: SL gold standard statistics. Breakdown of labels into positives and negatives, unique gene count, and percentage of labelled pairs. Columns + and - show number of positive and negative labels for each dataset. Num. Genes and Labelled % denote the number of unique genes and percentage of labelled pairs (of all possible pairs involving genes from the combined dataset).

Cancer cell line data. We used cancer cell line gene dependency scores based on CRISPR (CERES) [14, 46] and RNA interference (DEMETER2) [44, 4] screens from the 19Q3 DepMap and DEMETER2 Data v6 public releases, respectively. We also obtained functionally categorised mutation data per gene [21].

Patient tumour and clinical data. We collected the following patient tumour sample data from the Broad GDAC Firehose pipeline run *stddata__2016_01_28* [59]: mutation data, discrete copy-number variation (CNV) scores from GISTIC [45], patient race, age, sex, and survival time (days). We also obtained gene expression data from the GEO (accession GSM1536837) as aggregated read counts [49].

Healthy tissue data. We collected expression data from GTEx for breast, lung, colon, and ovarian tissue of healthy donors, provided as gene-aggregated TPM values (dbGaP accession phs000424.v8.p2) [39]. We also included expression data of TCGA matched normal BRCA and LUAD samples from GEO, as described for patient tumours.

Biological pathway data. We downloaded KEGG [28], PID [53] and Reactome [26] pathway gene sets from the molecular signatures database v7.1 [57, 35].

Protein-protein interaction and Gene Ontology data. Protein-protein interaction data was downloaded from STRINGdb, version 11 [58]. We selected only interactions supported by curated experimental evidence. GO

biological process and cellular process data were downloaded from GO on March 18, 2021 [1, 10].

5.2.2 Features

Every example denotes a tissue type-specific relationship between a pair of genes (A, B), characterized by the following 27 molecular features. (see Supplementary Table 5.S1 for a summary of all individual features).

Gene dependencies. We calculated five features for each type of gene dependency, CRISPR or RNAi (10 in total). We performed two two-tailed Wilcoxon rank-sum tests [42], one for (A, B) and another for the same pair in reverse order (B, A). Each test quantifies the change in dependency on the first gene between cell lines with and without a non-silent mutation in the second gene. We chose as features the test statistic and *p*-value for the tested pair (A, B) or (B, A) that yielded the smallest *p*-value. We defined two additional features as the Pearson's correlation coefficient and corresponding two-tailed *t*-test *p*-value between the dependency scores of A and B. The fifth feature was the average of the means of the dependency scores for genes A and B. Respectively, the features are termed *CRISPR/RNAi_dep_stat*, *CRISPR/RNAi_dep_pvalue*, *CRISPR/RNAi_cor_stat*, *CRISPR/RNAi_cor_pvalue*, *CRISPR/RNAi_avg*.

Mutual exclusivity. We calculated seven mutual exclusivity features based on tumour mutation data using three methods: DiscoverSL (four features) [12], DISCOVER [8], and MUTEX [2]. These features are termed *discoversl_mutex_amp*, *discoversl_mutex_del*, *discoversl_mutex_mut*, *discoversl_mutex*, *discover_mutex*, and *MUTEX*. We calculated an additional mutual exclusivity *p*-value, *mutex_alt*, by treating every non-silent mutation, amplification (CNV = 2), and deletion (CNV = -2) as an "alteration" event. We used a hypergeometric test:

$$p = 1 - \sum_{j=n_{A,B}}^{min(n_A,n_B)} \frac{\binom{n_A}{j}\binom{n_T-n_A}{n_B-j}}{\binom{n_T}{n_B}}, \tag{5.1}$$

where n_A and n_B are the numbers of tumour samples with an alteration in A and B, respectively, $n_{A,B}$ is the number of samples with alterations in both, and n_T is the total number of samples.

Survival. We modelled patient survival time using Cox proportional hazard models accounting for the alteration status of gene pair (A, B) in patient tumours. We defined the status as "altered" if any of the following alterations occur in both A and B (unaltered otherwise): copy-number amplifications (CNV = 2) or deletions (CNV = -2), non-silent mutations, or aberrant expression. We defined aberrant expression as having a gene expression level in the upper or lower fifth percentile across all patient samples. We also controlled for age, race, and sex as follows:

$$ln\ h(t) \sim ln\ h_0 + \beta_1 s(A,B)\ + \beta_2 sex + \beta_3 age + \beta_4 race \tag{5.2}$$

where $h(t)$ is the hazard function defined as the conditional probability of a patient dying at time t given that the patient has survived to time t [6]. The indicator variable $s(A,B)$ denotes the alteration status of gene pair (A, B) in a patient tumour sample. The β values are the regression coefficients. One feature, *logrank_pval*, was defined as the two-tailed *p*-value of $\beta_1 \neq 0$ using the Wald statistic [3].

Co-expression. We determined co-expression between a gene pair for three types of biological samples: tumour and normal TCGA samples (for BRCA and LUAD), and healthy donor GTEx samples. We used pairwise Pearson's correlations and two-tailed *t*-test *p*-values, yielding four to six features: *tumour_corr/pvalue*, *normal_corr/pvalue*, *gtex_corr/pvalue*.

Differential expression. We calculated differential expression using tumour samples to quantify the variation in expression of one gene given the presence or absence of non-silent mutations in the other gene. We performed two differential expression tests per gene pair (A, B), for gene A based on the mutation status of gene B and vice versa, and used the minimum of the two *p*-values and the corresponding log fold-change as features for the gene pair. These were calculated using edgeR based on the read count data [52]. Using the edgeR default parameter values, we performed Trimmed Mean of M-values normalisation (TMM), and calculated gene-wise log2 fold-changes and *p*-values as features, respectively termed *diff_exp_logFC* and *diff_exp_pvalue*.

Pathway co-participation. We calculated a *pathway_coparticipation* *p*-value denoting the significance of co-occurrence of a pair of genes in a set of pathways using a hypergeometric test as defined in Eq. 5.1. Here, n_A and n_B

are the number of occurrences of genes A and B in all pathways, respectively, $n_{A,B}$ is the number of occurrences of both genes in the same pathway, and n_T is the total number of pathways. The set of pathways was defined as the union of the KEGG, PID, and Reactome gene sets.

5.2.3 Synthetic Lethality Prediction Models

SBSL prediction models. We trained logistic regression and random forest models with regularisation, as representatives of linear and non-linear models. For logistic regression, we used L0 and L2 (L0L2), or L1 and L2 (Elastic Net) regularisation, as implemented respectively in the L0Learn and glmnet packages [22, 20]. We also tried two regularised random forest implementations: Multivariate random forests with Unbiased Variable selection in R (MUVR, Shi et al. [54]), and Regularised Random Forests (RRF, Deng and Runger [15]). MUVR combines a random forest model with feature selection through repeated, nested, cross-fold validation and backward feature elimination on the train set. RRF is a random forest variant that uses two parameters to control model complexity: *mtry* determining how many features are randomly sampled at each new node; and *coefReg* to control the penalisation of the information gained when adding a new feature to the model to split at a given node.

Other SL prediction models for comparison. We compared the SBSL models against five other published methods: statistical approach DAISY [27], supervised model DiscoverSL [12], graph-based GCATSL [38] and GRSMF [23], and matrix factorisation pca-gCMF [34].

5.2.4 Training and Evaluation

For each experiment, we created 10 different train/test set splits of the available dataset(s), so that we could better assess the robustness of the models. For each pair of train and test sets, which we term run for short, we performed the following steps: hyperparameter search on the train set using cross-validation, learning of a final model on the entire train set using the best parameters, and assessing the final model on the corresponding disjoint test set. We report the averages and standard deviations of our performance evaluation metrics across the 10 runs. These steps are further detailed below.

Train and test sets. All pairs of train and test sets were created as follows, unless otherwise specified. To handle class imbalances (Table 5.1), we uniformly downsampled the dataset to ensure an equal number of SL and non-SL pairs. We then divided it into train and test sets with a 70/30 split via uniform sampling. We standardised every feature in both sets by subtracting the mean and dividing by the standard deviation calculated from the train set. We also excluded any feature for which at least 95% of the values in its feature vector were constant.

Hyperparameter tuning. To select model hyperparameters for Elastic Net and RRF, we defined a search space per model as follows; Elastic Net: *lambda* = [0,1], *alpha* = [0,1]; RRF: *mtry* = [4,8], *coefReg* = [.5,1]. For L0L2, the search space hyperparameters were set to *nGamma* = 20 and *nLambda* = 50. For the Elastic Net, RRF and L0L2 models, we conducted 10-fold cross-validation on the train set with 5 repeats using the area under the ROC (AUROC) as performance metric. Results of hyperparameter search for these three models can be found in Supplementary Fig. 5.S1-5.S3. The hyperparameters used for the MUVR backwards feature elimination algorithm were *nRep* = 5, *nOuter* = 10, and *varRatio* = 0.8.

Evaluation. Following hyperparameter tuning, SBSL models were trained on the entire train set using the best hyperparameters. Performance was then assessed on the corresponding disjoint test set, using receiver operating characteristic (ROC) and precision-recall (PR) curves. The curves were summarised by area under the ROC or PR curve metrics (AUROC, AUPRC). We report averages and standard deviations of the AUROC and AUPRC across the 10 runs.

Comparison with other SL prediction methods. We calculated DAISY scores for all gene pairs, and predicted DiscoverSL scores for test set pairs using the package provided by the authors. GCATSL, GRSMF, and pca-gCMF models were trained on the train set using their default parameter settings (see Supplementary Methods). Scores obtained by all methods for gene pairs in the test set were used for comparison during evaluation.

Feature importance. We calculated permutation feature importance (FI) values for SBSL models based on the test set to determine which features contributed most to the predictions [18]. Interpreting FI scores can be confounded by multicollinearity, as importance may spread over correlated

features. For this reason, we assessed multicollinearity using variance inflation factors (VIF) [25].

5.3 Results and Discussion

5.3.1 SBSL and SL topology methods are the top performers

We first evaluated the performance of the SL prediction models separately within each cancer type (BRCA, COAD, LUAD, and OV). We evaluated the predictive performance of the SBSL logistic regression (L0L2, Elastic Net) and random forest (MUVR, RRF) models against published methods DAISY, DiscoverSL, GCATSL, GRSMF and pca-gCMF.

On BRCA and LUAD, the SBSL models and the matrix factorisation methods GRSMF and pca-gCMF performed most consistently considering the two metrics, with average AUROC and AUPRC above 0.80 (Tables 5.2 and 5.3, Supplementary Fig. 5.S4 for ROC and PR curves). SBSL models did better at predicting true SL pairs for BRCA and LUAD than the other approaches with the exception of pca-gCMF on BRCA (Table 5.3). GRSMF performed reasonably with average AUPRC above 0.80, but GCATSL performed poorly on BRCA (average AUPRC of .55) while scoring highest among the SL Topology methods on LUAD (average AUPRC of .85). On COAD, AUROC performances were very modest across the board, with SBSL models featuring on the higher end (.38 < average AUROC < .64).

On OV, the SBSL models predicted poorly whereas GCATSL, GRSMF, and pca-gCMF showed high AUROC and AUPRC scores above 0.90. We hypothesise that the low performance of SBSL models in OV could be due to the modest mutational burden typically observed for this cancer type [61], which could affect the resolution and informativeness of features relying on mutation data. We confirmed that OV cell lines contained a much lower average number of mutations per gene pair than the other cancer types (OV: 1.6, BRCA: 4.33, LUAD: 11.35, COAD: 5.97). As for the high performance of SL topology methods on OV, we reasoned that it could be due to selection bias, which we investigate in a later section.

DAISY and DiscoverSL performed poorly overall and were excluded from subsequent experiments. We note that DAISY is not an ML approach, and

does not involve separate training and prediction. For fairness, DAISY was applied to the entire dataset, per cancer type, and then evaluated on the same test sets as the other models (see Methods).

Our results suggest that SBSL models and pca-gCMF are the most consistent and thus may be better suited for pre-selecting SL pairs for experimental follow-up in BRCA and LUAD. Most methods struggled to predict SL for COAD according to one or both performance metrics.

We advance that low mutational burden could negatively affect the performance of SBSL models on OV, and go on to further investigate a possible link between selection bias and the high performance of SL topology methods.

5.3.2 Selection bias drives SL topology method predictions

Since SL topology methods are driven by the structure of the SL label graph, we hypothesised that their predictive performance could be affected by selection bias in SL screens. We sought to assess the impact of this bias.

Selection bias in SL labels. We examined the coverage and structure of SL labelled gene pairs. The OV set of labelled pairs stood out from the other cancer types for three reasons. First, it had limited gene coverage, comprising only 86 unique genes whereas the other cancer types included 804 to 1560 labelled genes. Second, 18.14% of all possible pairs formed by these 86 genes were labelled in OV, compared to a maximum of 1.66% for the other cancer types (Table 5.1). Third, the structure of the labels was quite striking: many rows were nearly identical to one another, showing very consistent patterns of SL and non-SL relationships with the same genes. These formed visibly distinct groups, indicative of heavy gene selection bias (Fig. 5.1). For example, the genes highlighted with red labels in Fig. 5.1 are functionally related: they mostly consist of tyrosine kinases, which are all reported targets of the same drug dasatinib [29].

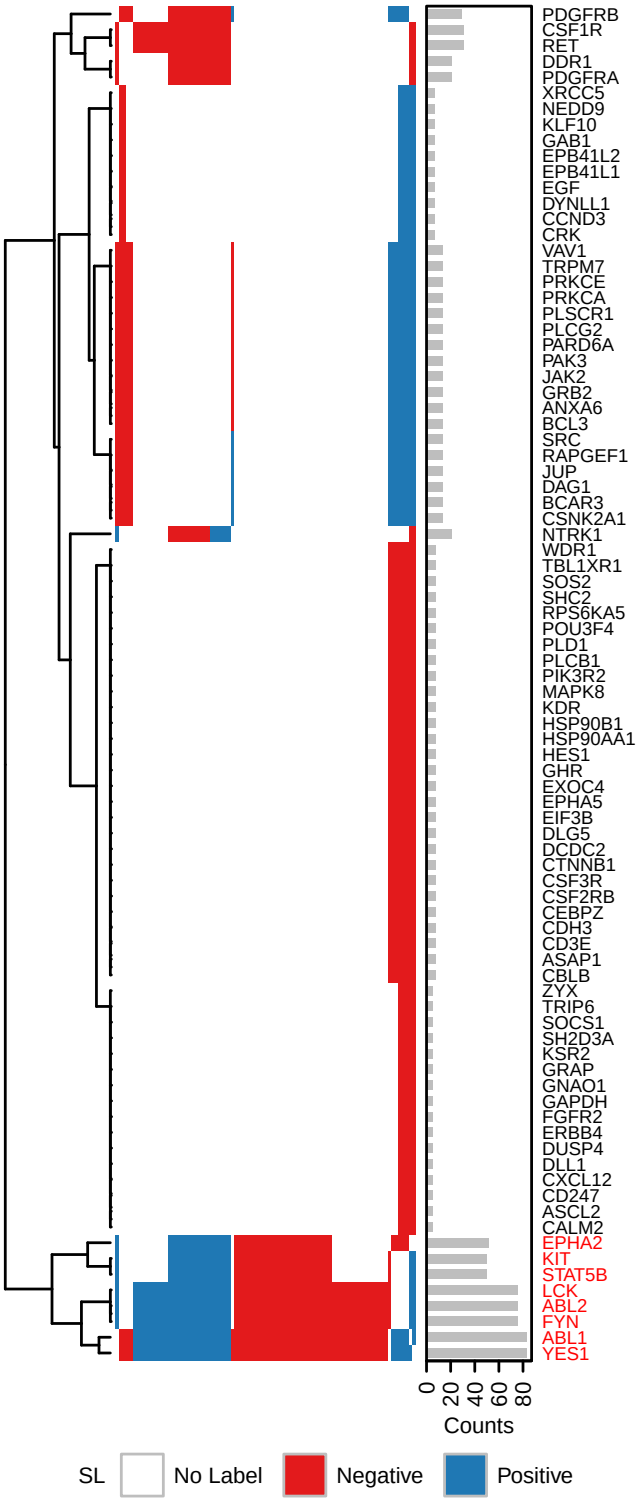

PDGFRB
CSF1R
RET
DDR1
PDGFRA
XRCC5
NEDD9
KLF10
GAB1
EPB41L2
EPB41L1
EGF
DYNLL1
CCND3
CRK
VAV1
TRPM7
PRKCE
PRKCA
PLSCR1
PLCG2
PARD6A
PAK3
JAK2
GRB2
ANXA6
BCL3
SRC
RAPGEF1
JUP
DAG1
BCAR3
CSNK2A1
NTRK1
WDR1
TBL1XR1
SOS2
SHC2
RPS6KA5
POU3F4
PLD1
PLCB1
PIK3R2
MAPK8
KDR
HSP90B1
HSP90AA1
HES1
GHR
EXOC4
EPHA5
EIF3B
DLG5
DCDC2
CTNNB1
CSF3R
CSF2RB
CEBPZ
CDH3
CD3E
ASAP1
CBLB
ZYX
TRIP6
SOCS1
SH2D3A
KSR2
GRAP
GNAO1
GAPDH
FGFR2
ERBB4
DUSP4
DLL1
CXCL12
CD247
ASCL2
CALM2
EPHA2
KIT
STAT5B
LCK
ABL2
FYN
ABL1
YES1
0
20
40
60
80
Counts
SL
No Label
Negative
Positive

Fig. 5.1: Structure of SL labels. Adjacency plot showing OV gene pairs. Elements along horizontal and vertical axes represent unique genes. Each coloured cell denotes a negative (red) or positive (blue) SL pair. White cells denote pairs with no label. Rows are ordered according to hierarchical clustering with complete linkage and Euclidean distance. Columns follow the ordering of rows. The barplot to the right shows the number of pairs each gene is involve in. The group of genes highlighted in red consists mostly of tyrosine kinases.

The high performances of matrix factorisation and graph-based methods on OV data could be expected, given that they are designed to exploit this structure. However, the consistency of patterns seen in these OV labels will not likely generalise well to most randomly selected pairs of genes. The SL labels for the other cancer types exhibited similar bias, albeit less pronounced given the larger sample size and gene coverage (Table 5.1, Supplementary Figs. 5.S5-5.S8). As an example, for BRCA the 5 most frequently occurring genes were involved in 52% of all SL labelled gene pairs (PARP1 18%, BRCA1 12%, PTEN 11%, TP53 7%, BRCA2 4%). We also identified two distinct groups of genes with visibly coordinated patterns, which also happened to be functionally related: one group comprised members of cell proliferation pathways (JAK2, GATA3, PIK3C3, FLI1, MAP2K4, PPARA, BIRC3, CREBBP, KRAS, MAP3K1, and others), and the other group contained genes involved in DNA damage response (CHD1, USP6, CANT1, ERCC4, MAML2, DHRS13, FHIT).

Method	BRCA	COAD	LUAD	OV
Elastic Net	$.84 \pm .01$	$.60 \pm .02$	$.85 \pm .02$	$.59 \pm .03$
L0L2	$.84 \pm .01$	$.60 \pm .02$	$.85 \pm .02$	$.59 \pm .03$
MUVR	$.86 \pm .01$	$\mathbf{.64 \pm .01}$	$\mathbf{.87 \pm .01}$	$.56 \pm .07$
RRF	$.86 \pm .01$	$.63 \pm .02$	$.87 \pm .02$	$.57 \pm .07$
DAISY	$.61 \pm .02$	$.38 \pm .02$	$.44 \pm .03$	$.41 \pm .04$
DiscoverSL	$.54 \pm .02$	$.54 \pm .02$	$.54 \pm .03$	$.45 \pm .04$
GCATSL	$.59 \pm .04$	$.51 \pm .01$	$.86 \pm .03$	$\mathbf{.99 \pm .02}$
GRSMF	$.82 \pm .01$	$.57 \pm .02$	$.87 \pm .02$	$\mathbf{.99 \pm .01}$
pca-gCMF	$\mathbf{.90 \pm .02}$	$.54 \pm .03$	$.87 \pm .02$	$.94 \pm .05$

Tab. 5.2: Classification performance of SL prediction models within a cancer type, denoted by the area under the ROC curve (AUROC). Mean and standard deviations over 10 runs.

Method	BRCA	COAD	LUAD	OV
Elastic Net	$.87 \pm .01$	$.59 \pm .01$	$\mathbf{.87 \pm .02}$	$.58 \pm .03$
L0L2	$.88 \pm .01$	$.59 \pm .02$	$\mathbf{.87 \pm .02}$	$.58 \pm .04$
MUVR	$\mathbf{.89 \pm .01}$	$.62 \pm .02$	$\mathbf{.87 \pm .02}$	$.54 \pm .06$
RRF	$\mathbf{.89 \pm .01}$	$\mathbf{.63 \pm .02}$	$\mathbf{.87 \pm .02}$	$.55 \pm .05$
DAISY	$.58 \pm .02$	$.43 \pm .01$	$.47 \pm .02$	$.48 \pm .04$
DiscoverSL	$.55 \pm .02$	$.53 \pm .02$	$.55 \pm .03$	$.49 \pm .04$
GCATSL	$.55 \pm .02$	$.50 \pm .01$	$.82 \pm .04$	$\mathbf{.98 \pm .03}$
GRSMF	$.81 \pm .01$	$.59 \pm .02$	$.85 \pm .02$	$.97 \pm .04$
pca-gCMF	$\mathbf{.89 \pm .04}$	$.56 \pm .03$	$.83 \pm .03$	$.93 \pm .06$

Tab. 5.3: Classification performance of SL prediction models within a cancer type, denoted by the area under the precision-recall curve (AUPRC). Mean and standard deviations over 10 runs.

Cross-SL-dataset generalisation. We assessed the impact of selection bias on the ability of SL prediction methods to generalise across the two datasets of SL labels. We trained BRCA models on gene pairs from ISLE and tested them against DiscoverSL. We also trained LUAD models on DiscoverSL and tested them against ISLE. We focused on these specific combinations, since the number of available SL pairs was insufficient for the reverse combinations. Any labelled pairs present in both datasets were removed from the train set. Our results showed that SBSL models generalised better across gold standards (Fig. 5.2, Supplementary Fig. 5.S9), with linear models performing best overall. The SL topology approaches (GCATSL, GRSMF, and pca-gCMF) struggled to generalise, and their performances decreased to nearly random on LUAD data. We found that pca-gCMF did only marginally better than the graph-based methods on LUAD, but was comparable to our SBSL models on BRCA.

Contributing to the poor performance of SL topology models is the fact that these techniques have difficulty making connections to genes that are not involved in pairs in the train set, an issue that is most prevalent in LUAD SL data. Specifically, for BRCA, 522 of the 907 pairs in DiscoverSL contained genes that also appeared in ISLE. However, for LUAD, only 19 out of 659 DiscoverSL pairs shared a gene with ISLE (Supplementary Fig. 5.S10-5.S11). SL topology methods would be more affected by this than SBSL models due to missing prior SL information for the genes in the test set.

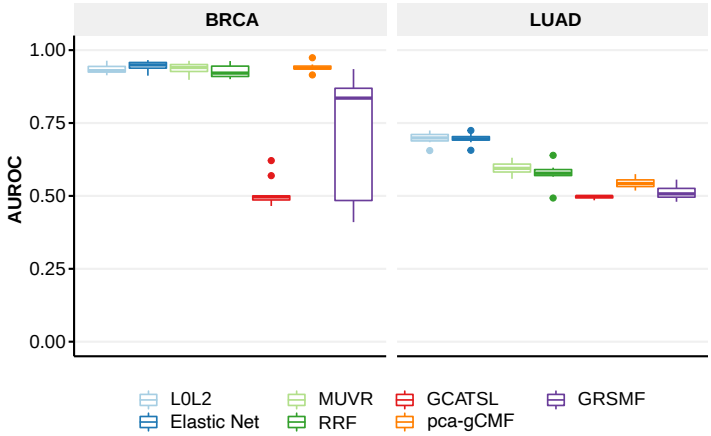


Fig. 5.2: Cross-SL gold standard performances. AUROC values averaged over 10 runs for: (left) BRCA models trained on ISLE and tested on DiscoverSL; (right) LUAD models were trained on DiscoverSL and tested on ISLE.

Gene holdout experiments. We further investigated the impact of selection bias on SL prediction using *gene holdout* experiments, where train/test sets were constructed in three different ways, seeking to control the number of genes shared between the two sets. In our original baseline scenario, also termed *None*, we only ensured that there was no overlap in gene pairs between train and test sets. For *Single* holdout, we constructed the train and test sets such that for every gene pair in the test set, only one of the genes from the pair could be present in the train set. For *Double* holdout, we created the train and test sets such that they did not share any genes. Note that there was not enough OV data to conduct the *Double* experiment.

The SBSL models were more robust to gene holdouts than SL topology models on BRCA, COAD, and LUAD. For these cancers, SBSL models showed negligible decrease between baseline and *Single* holdout, and a more pronounced drop to mean AUROC values between 0.60 and 0.75 using *Double* holdout. Comparatively, the performance of SL topology models varied more and became approximately random with *Double* holdout (Fig. 5.3, Supplementary Fig. 5.S12). OV was the exception, where SL topology methods seemed to do better, possibly due to the previously described bias in SL labels. We note that our results are confounded by shrinking of the train set size as we move through the scenarios from *None* to *Double*, and that OV is the smallest of the datasets.

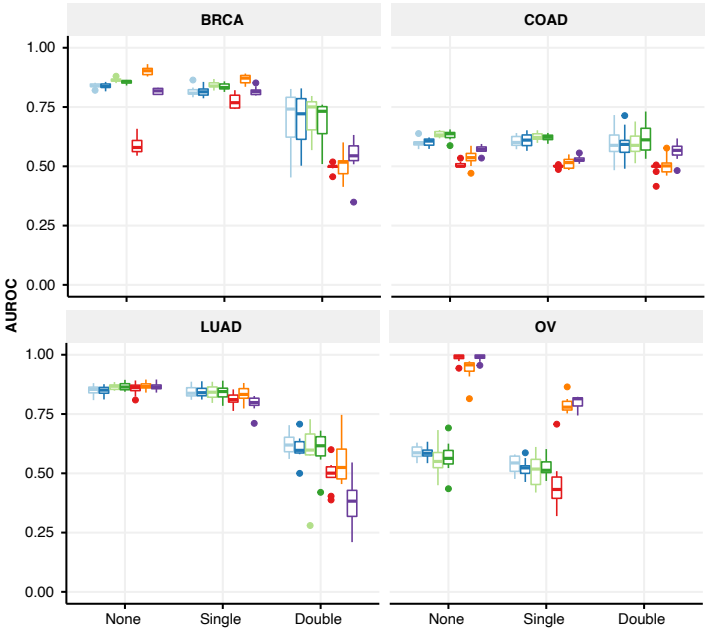


Fig. 5.3: Performances of gene holdout experiments, where bias is controlled by ensuring that none, one, or both genes of pairs in the test set are excluded from the train set. Shown are AUROC values for each gene-holdout experiment per cancer type (10 runs). For "None", we only guarantee that train and test sets are disjoint in terms of gene pairs, not individual genes; for "Single", only one gene from a gene pair in the test set can be present in the train set; for "Double" neither gene of a pair in the test set appears in the train set. The results for "None" correspond to those also reported in Table 5.2. Note: there was insufficient data to conduct the OV "Double" experiment.

5.3.3 Not all cancers are equal in SL prediction

We wondered whether the underlying molecular patterns that allow us to recognize when two genes are synthetically lethal could be independent of cancer type and thus generalisable across cancers. To answer this question, we assessed the potential benefits of training pan-cancer L0L2 and MUVR models, which could also help mitigate the sparsity and selection bias affecting some of the cancer types (Supplementary Tables 5.S2-5.S3 and Fig. 5.S13-5.S14 for results including all SBSL models).

First, we trained two pan-cancer models on data from all four cancer types (Table 5.4). One was an unbalanced model, with gene pairs uniformly selected from the combined dataset to keep the original cancer type ratios.

The other model was trained with balanced proportions of cancer types and class labels by undersampling. Both models were evaluated against held out data from every cancer type. Model performances improved in almost every case when training on balanced compared to unbalanced data. However, training on multiple cancers resulted in a degradation of overall performance relative to the cancer-specific models. We note that balanced models typically had less gene pairs to train on.

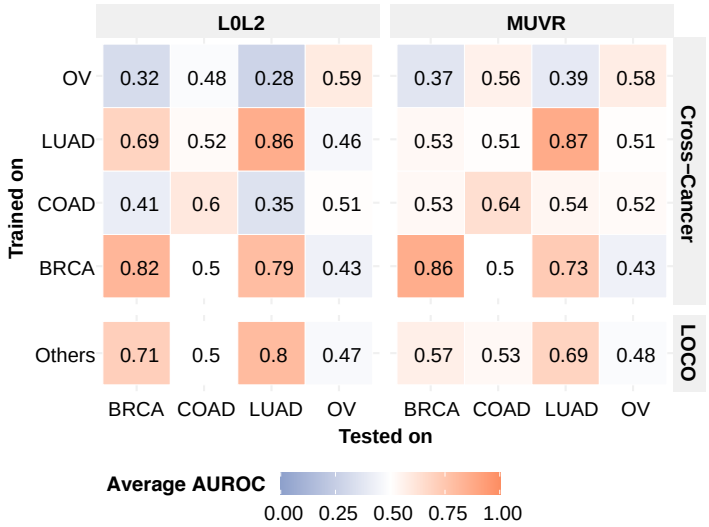


Fig. 5.4: Cross-cancer and leave-one-cancer-out (LOCO) performances. Average AUROC for L0L2 and MUVR models over 10 runs. *Cross-cancer:* Vertical and horizontal axes denote the cancer types used to train and test, respectively. *LOCO:* Horizontal axis denotes the cancer type held out for testing. Models trained on balanced data from all other cancers.

We then assessed the ability of SBSL cancer-specific models to make SL predictions for other cancer types (Fig. 5.4). As expected, models that performed poorly within the same cancer type, like COAD and OV, could not generalise to other cancer types either. The better performing models, BRCA and LUAD, could not predict well on COAD and OV either. Notably, the L0L2 linear models generalised reasonably both when trained on BRCA and tested against LUAD (mean AUROC 0.79) and vice-versa (0.69). The MUVR random forests could generalise when trained on BRCA and tested against LUAD (0.73), but not vice-versa (0.53), showing they were more prone to overfit.

Method	Cancer	Pan-cancer		One-cancer
		Unbalanced	Balanced	
L0L2	BRCA	$.64 \pm .02$	$.75 \pm .01$	$\mathbf{.83 \pm .01}$
	COAD	$.52 \pm .02$	$.51 \pm .02$	$\mathbf{.60 \pm .02}$
	LUAD	$.73 \pm .03$	$.79 \pm .02$	$\mathbf{.83 \pm .02}$
	OV	$.40 \pm .04$	$.53 \pm .04$	$\mathbf{.58 \pm .03}$
MUVR	BRCA	$.76 \pm .01$	$.82 \pm .02$	$\mathbf{.86 \pm .01}$
	COAD	$.62 \pm .02$	$.60 \pm .01$	$\mathbf{.64 \pm .01}$
	LUAD	$.81 \pm .02$	$.83 \pm .02$	$\mathbf{.86 \pm .01}$
	OV	$\mathbf{.55 \pm .06}$	$.52 \pm .04$	$.54 \pm .07$

Tab. 5.4: Performance of one-cancer and pan-cancer models (AUROC). Mean and standard deviation calculated over 10 runs. One-cancer and pan-cancer models trained with unbalanced or balanced cancer representation, tested on held-out gene pairs.

We also investigated the ability of models trained on multiple cancers to make SL predictions for unseen cancer types. In this leave-one-cancer-out experiment (LOCO), we held out one cancer type for testing and trained models using samples from the other three, with balanced class labels and cancer types (Fig. 5.4, bottom row). The results were consistent with those of the cross-cancer experiment (Fig. 5.4). For example, the three-cancer models trained on COAD, LUAD, and OV generalised to BRCA as well as the LUAD-specific models. This indicates that training on multiple cancer types is not necessarily detrimental to cross-cancer generalisation.

5.3.4 Gene dependency-based features are most important

We used permutation feature importance scores (PFI) to quantify the contribution of the 27 features to the predictions of our SBSL models. To obtain meaningful PFI scores, the models should be reasonably accurate, thus we excluded the lower performing OV and COAD models (Table 5.2).

Gene dependency-based features were the largest contributors to the performance of BRCA and LUAD models (Supplementary Fig. 5.S15-5.S16). The highest ranked feature overall was *CRISPR_dep_stat*, which quantifies the change in dependency score of one gene between cell lines with and without a non-silent mutation in the other gene. Specifically, for linear models, the importance of *CRISPR_dep_stat* was nearly 2-fold greater than the importance of the second ranked feature. Ranking second were features denoting the

average of means of gene dependency scores across all cell lines. For all LUAD models and the BRCA random forest models, the choice went to the CRISPR-based feature (*CRISPR_avg*), while BRCA logistic regression models picked the RNAi-based feature (*RNAi_avg*). Even though CRISPR and RNAi-based dependency scores exhibit some differences, they are still moderately to highly correlated (multicollinearity VIF >2, Supplementary Table 5.S4), and thus fairly equivalent in contribution to SL prediction.

To further assess the reliance of our SBSL models on dependency-based features, we retrained and tested BRCA and LUAD models without these features. This led to a significant decrease in mean AUROC across all models, from between 0.83 and 0.85 to between 0.64 and 0.76, for both cancer types (Fig. 5.5). We also calculated PFI values for these models, which showed higher variability but also a few clear patterns. The DISCOVER mutual exclusivity score [8], *discover_mutex*, ranked first across all BRCA models (Supplementary Fig. 5.S17). Gene co-expression in healthy tissue samples (GTEx), *gtex_corr*, and co-expression in matched normal tissue samples from cancer patients (TCGA), *normal_corr*, respectively ranked second and third for all BRCA models (Supplementary Fig. 5.S17). Differential expression features, *diff_exp_logFC* and *diff_exp_pvalue*, were most important for LUAD random forest models (Supplementary Fig. 5.S18). These results indicate that features other than those based on gene dependency could also be informative for SL prediction.

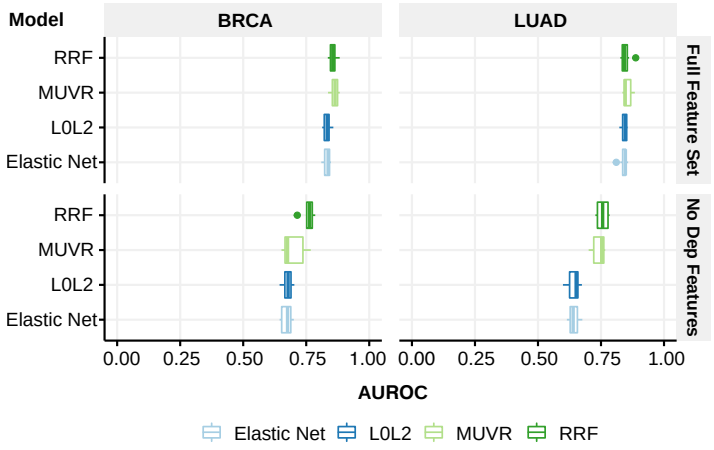


Fig. 5.5: Performance of SBSL models with and without gene dependency-based features (AUROC over 10 runs), respectively labelled "Full Feature Set" and "No Dep Features".

5.4 Conclusion

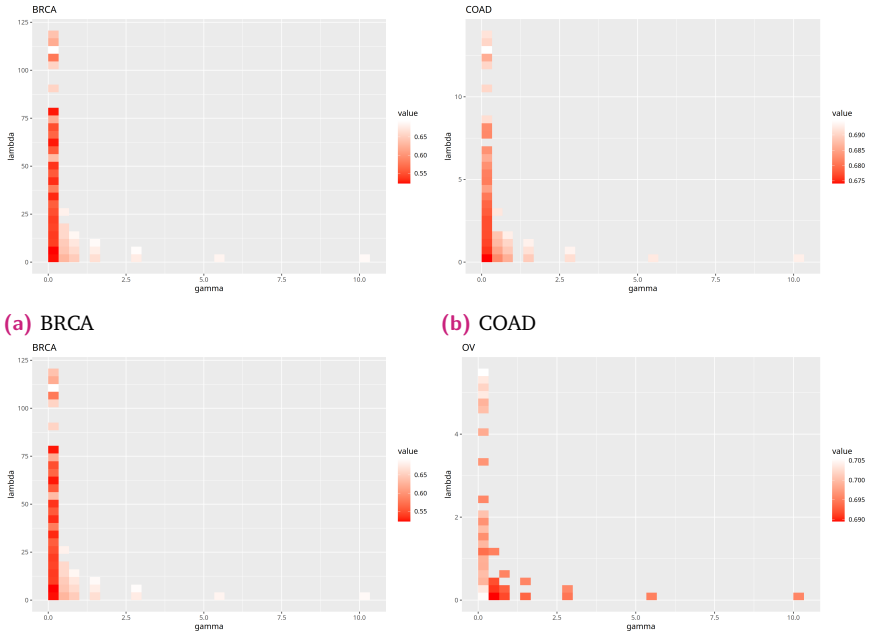
We proposed synthetic lethality (SL) prediction models with increased resilience to selection bias. We used logistic regression and random forest models based on molecular features characterising genes and gene pair relationships. Without explicit knowledge of gene or pair identity, SBSL models generalised better across SL label datasets, and were more robust to gene holdout compared to methods that predict based on the structure of SL labels. In addition, SBSL models improved over existing feature-based SL prediction approaches by focusing on underexplored data such as cancer cell line gene dependencies with mutation data, gene expression from healthy donors, mutual exclusivity of somatic mutations in patient tumours, and cancer patient survival. One limitation of our SBSL models is that they rely heavily on gene dependency scores, which are not available for rarer cancer types. We showed that other features could partially compensate for the absence of gene dependency scores, but led to a significant decrease in performance. In addition, we also note that some of the most relevant features in SBSL models are less effective for cancer types typically characterised by low mutational burden. Further research is therefore needed into alternative data and strategies to improve SL prediction. Since the ultimate goal of SL prediction models is to identify SL partners for drug target genes, systematic validation of SBSL models should be conducted to assess the therapeutic potential of predicted pairs.

Analysis of SL label data revealed the presence of strong gene selection bias. Further experiments showed that SL prediction methods relying on the structure of SL labels were more sensitive to such bias. This vulnerability persisted even when the methods incorporated additional data sources. Our observations align with a study on prediction of protein-protein interactions by Richoux et al. [51], which showed that including the same proteins in the train and test set led to performance overestimation. We believe that performances reported for SL topology methods under these conditions could be optimistic and should be viewed with caution.

We put forward two recommendations for the evaluation of SL prediction models. First, inspecting performance across cancer types, SL datasets, and other variables of interest is crucial to ensure that results are consistent and reproducible. Second, we advocate that gene selection biases are considered to avoid that performance metrics report on ability to exploit selection bias

rather than predict SL interactions. We show that plotting SL label adjacencies and conducting gene holdout experiments are effective ways to assess selection bias and its impact on SL prediction.

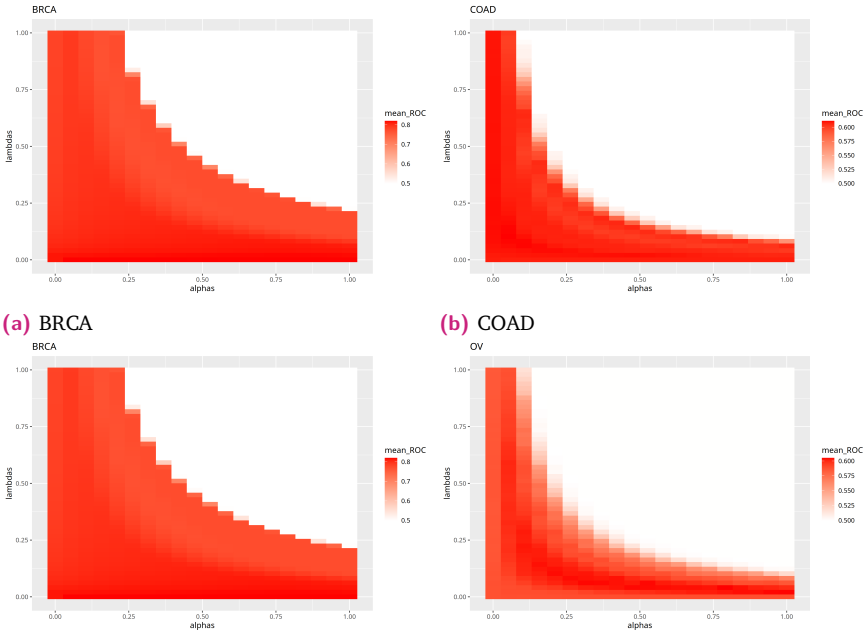
5.5 Supplementary Figures



(a) BRCA **(b)** COAD

(c) LUAD **(d)** OV

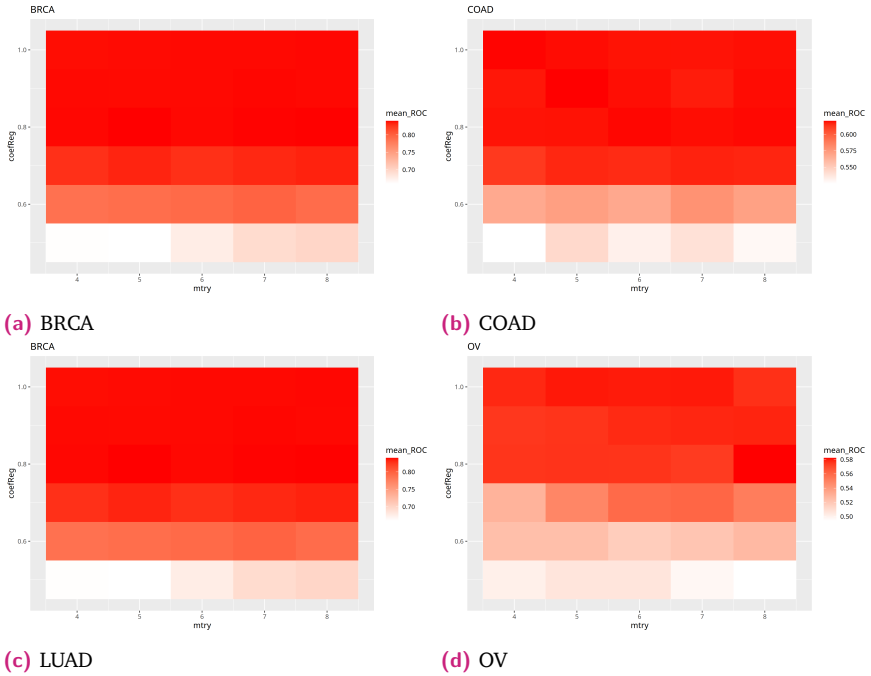
Fig. 5.S1: Hyperparameter search for the L0L2 model [22]. Mean logistic loss values of the optimised local search cross-validation results across each of the 10 folds across all 10 cross-validation runs for each cancer type. Deeper red values indicate lower mean logistic loss for that combination of *gamma* and *lambda*. L0Learn uses a custom local search algorithm to find the optimal values for *lambda*.



(a) BRCA **(b)** COAD

(c) LUAD **(d)** OV

Fig. 5.S2: Hyperparameter search for the ElasticNet model [20]. Mean AUROC grid search cross-validation results across each of the 10 folds across all 10 cross-validation runs for each cancer type. Deeper red values indicate higher mean AUROC for that combination of *alpha* and *lambda*.



(a) BRCA (b) COAD

(c) LUAD (d) OV

Fig. 5.S3: Hyperparameter search for the RRF model [15]. Mean AUROC grid search cross-validation results across each of the 10 folds across all 10 cross-validation runs for each cancer type. Deeper red values indicate higher mean AUROC for that combination of *coefReg* and *mtry*.

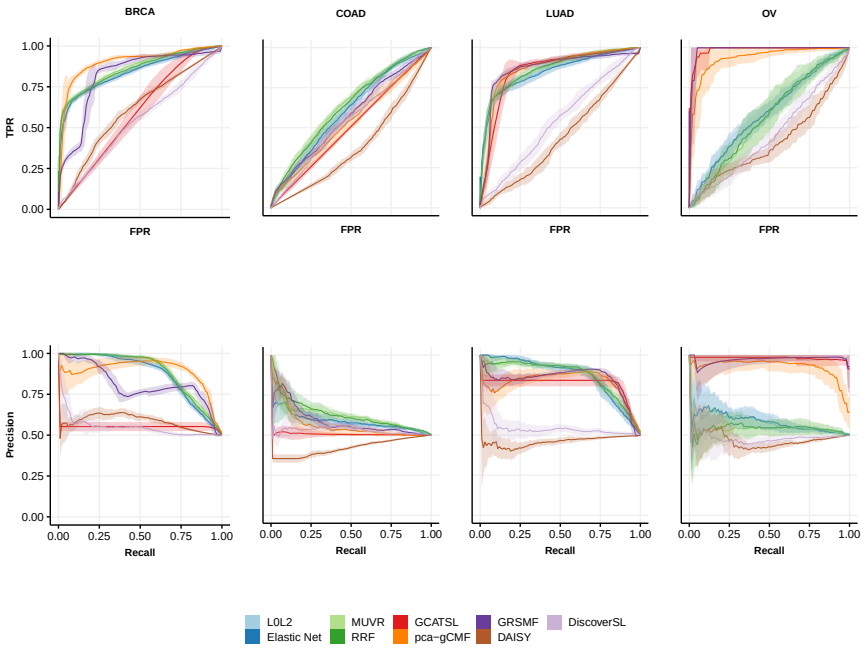


Fig. 5.S4: Average receiver-operating characteristic (ROC) curves and precision-recall (PR) curves for each cancer-specific model tested against that same cancer type. The top plots show ROC curves, the bottom plots show PR curves, and each column corresponds to a different cancer type. The ROC and PR curves were averaged across 10 runs using the vertical-averaging method [16]. The shaded regions represent the standard deviation.

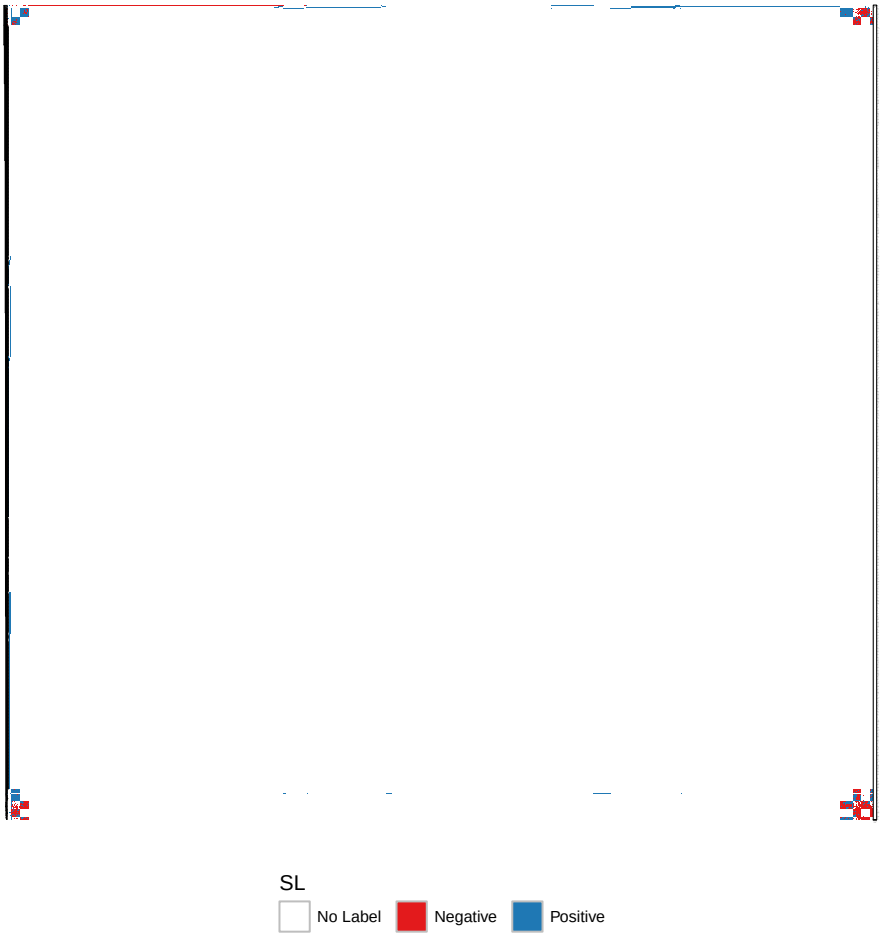


Fig. 5.S5: Adjacency matrix plot showing clusters of labelled genes in BRCA. Elements along the horizontal and vertical axes represent unique genes. Each coloured dot corresponds to a negatively (red) or positively (blue) labelled gene pair. Whitespace denotes a gene pair with no label. Rows are clustered using complete linkage and Euclidean distance with "No Label", "Negative", and "Positive" encoded as 0.5, 0 and 1, respectively. Both the rows and columns are ordered based on these clusters. The barplot to the right shows the number of occurrences of each gene in the BRCA dataset.

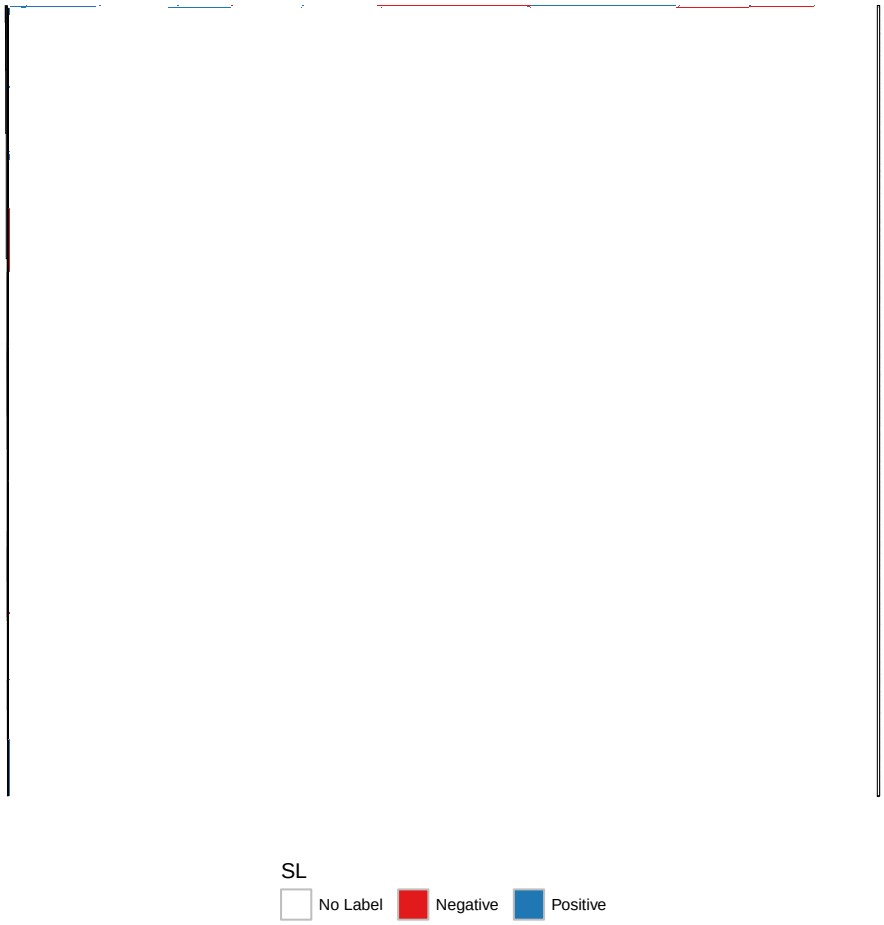


Fig. 5.S6: Adjacency matrix plot showing clusters of labelled genes in COAD. Elements along the horizontal and vertical axes represent unique genes. Each coloured dot corresponds to a negatively (red) or positively (blue) labelled gene pair. Whitespace denotes a gene pair with no label. Rows are clustered using complete linkage and Euclidean distance with "No Label", "Negative", and "Positive" encoded as 0.5, 0 and 1, respectively. Both the rows and columns are ordered based on these clusters. The barplot to the right shows the number of occurrences of each gene in the COAD dataset.

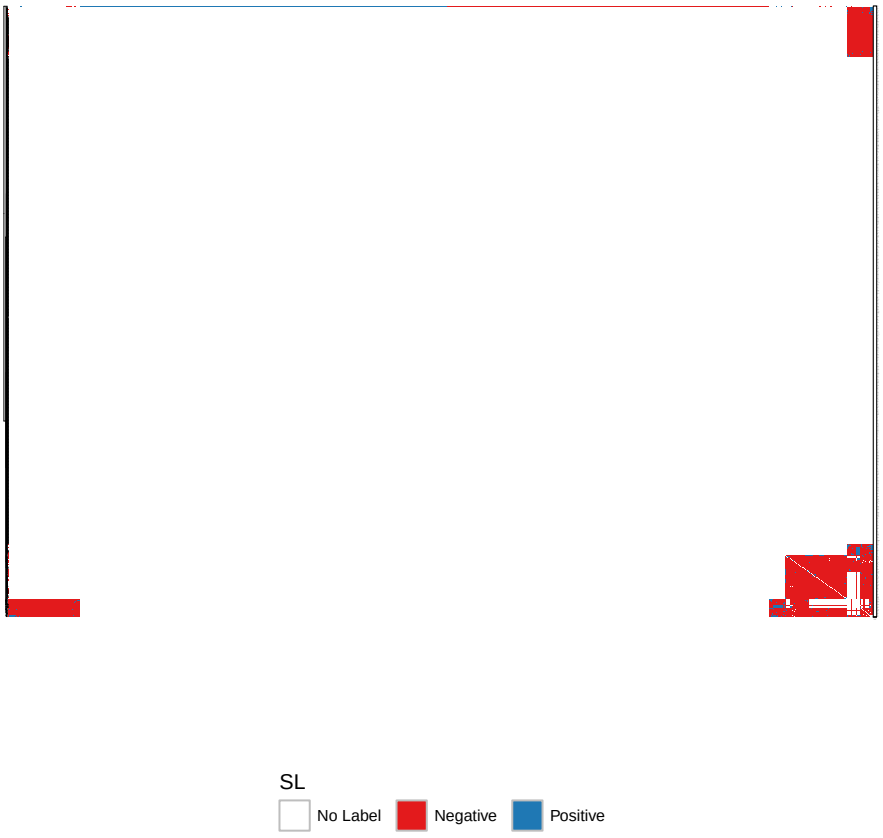


Fig. 5.S7: Adjacency matrix plot showing clusters of labelled genes in LUAD. Elements along the horizontal and vertical axes represent unique genes. Each coloured dot corresponds to a negatively (red) or positively (blue) labelled gene pair. Whitespace denotes a gene pair with no label. Rows are clustered using complete linkage and Euclidean distance with "No Label", "Negative", and "Positive" encoded as 0.5, 0 and 1, respectively. Both the rows and columns are ordered based on these clusters. The barplot to the right shows the number of occurrences of each gene in the LUAD dataset.

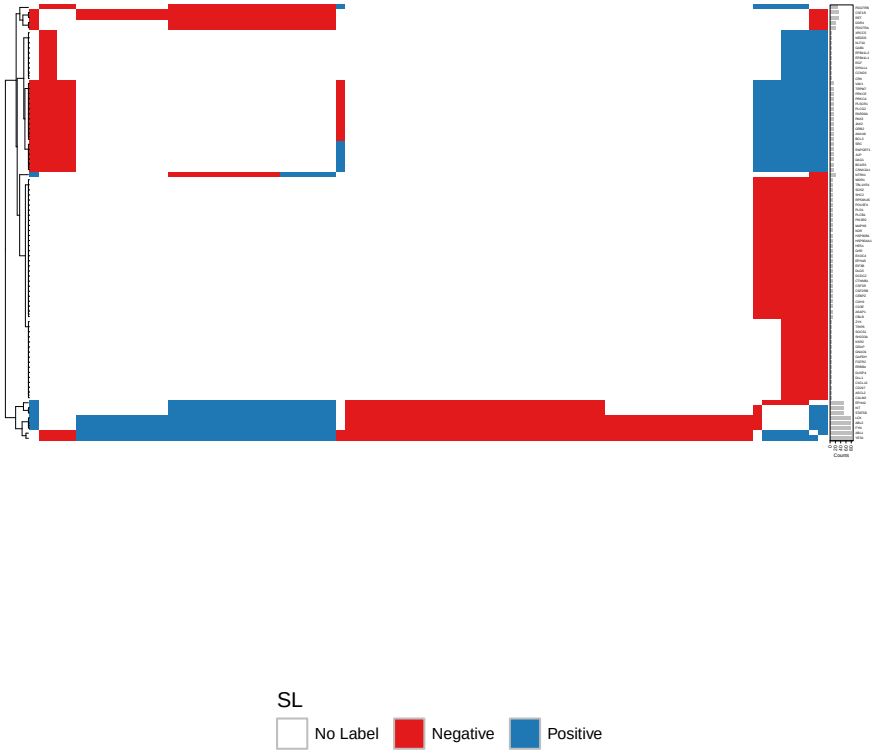


Fig. 5.S8: Adjacency matrix plot showing labelled gene pairs in OV. Elements along the horizontal and vertical axes represent unique genes. Each coloured dot corresponds to a negatively (red) or positively (blue) labelled gene pair. Whitespace denotes a gene pair with no label. Rows are clustered using complete linkage and Euclidean distance with "No Label", "Negative", and "Positive" encoded as 0.5, 0 and 1, respectively. Both the rows and columns are ordered based on these clusters. The barplot to the right shows the number of occurrences of each gene in the OV dataset.

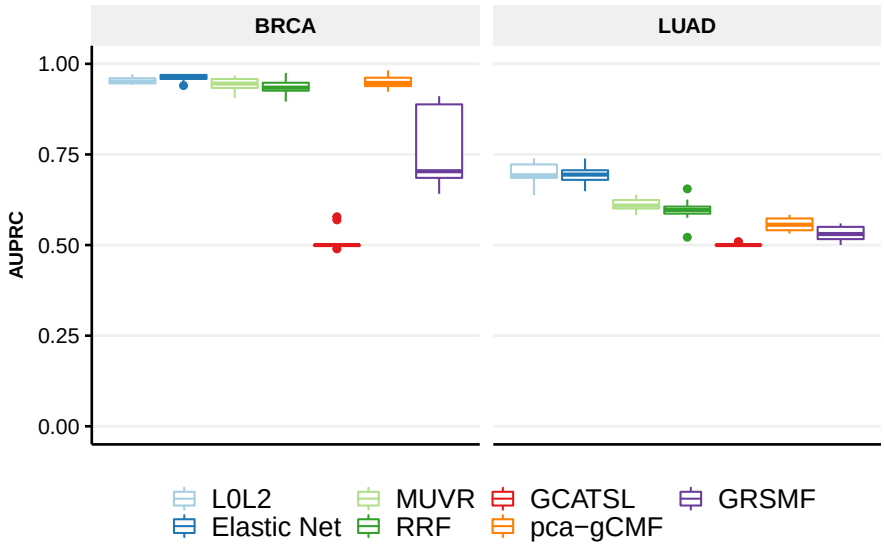


Fig. 5.S9: AUPRC values averaged over 10 runs for: (left) BRCA models trained on ISLE and tested on DiscoverSL; (right) LUAD models were trained on DiscoverSL and tested on ISLE.

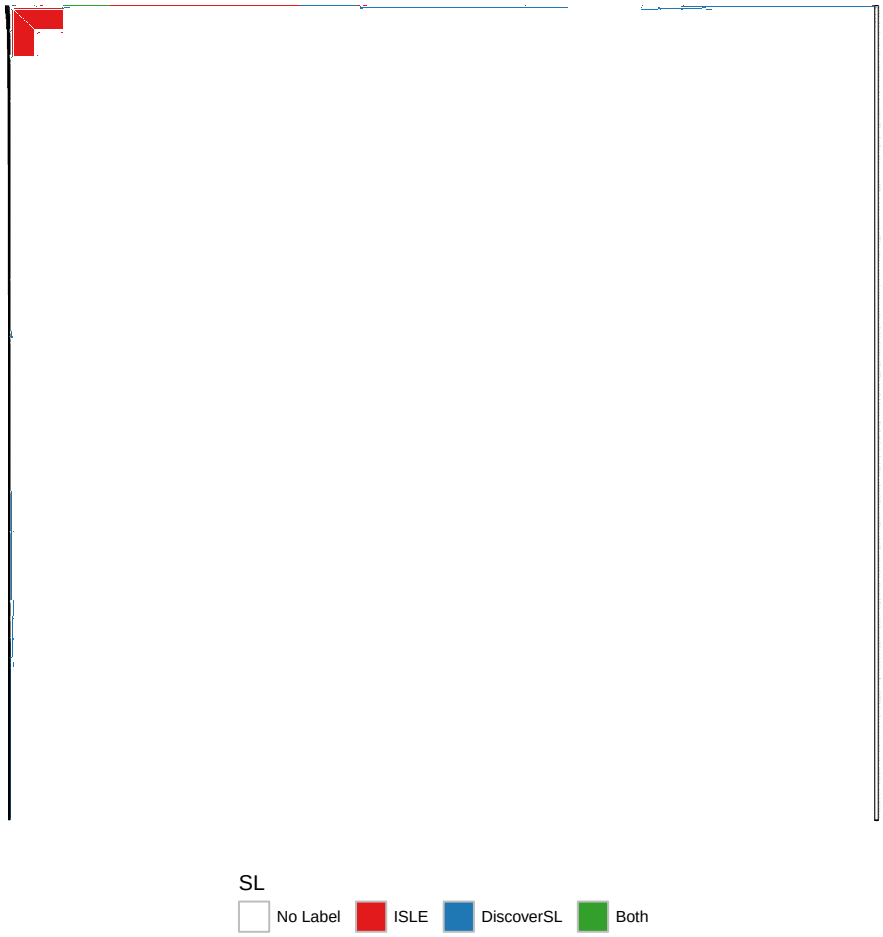


Fig. 5.S10: Structure of BRCA SL labels in the ISLE [33] and DiscoverSL datasets [12]. Heatmap showing labelled gene pairs from BRCA data from both the ISLE and DiscoverSL dataset. Elements along the horizontal and vertical axes represent unique genes. Each coloured dot represents a gene pair where a label exists in either the ISLE (red), DiscoverSL (blue), or both (green) datasets. Whitespace denotes a gene pair with no label. Rows are clustered using complete linkage and Euclidean distance with "No Label", "ISLE", "DiscoverSL", and "Both" encoded as -1, 0, 1, and 2, respectively. Both the rows and columns are ordered based on these clusters. The barplot to the right shows the number of occurrences of each gene in the BRCA dataset.

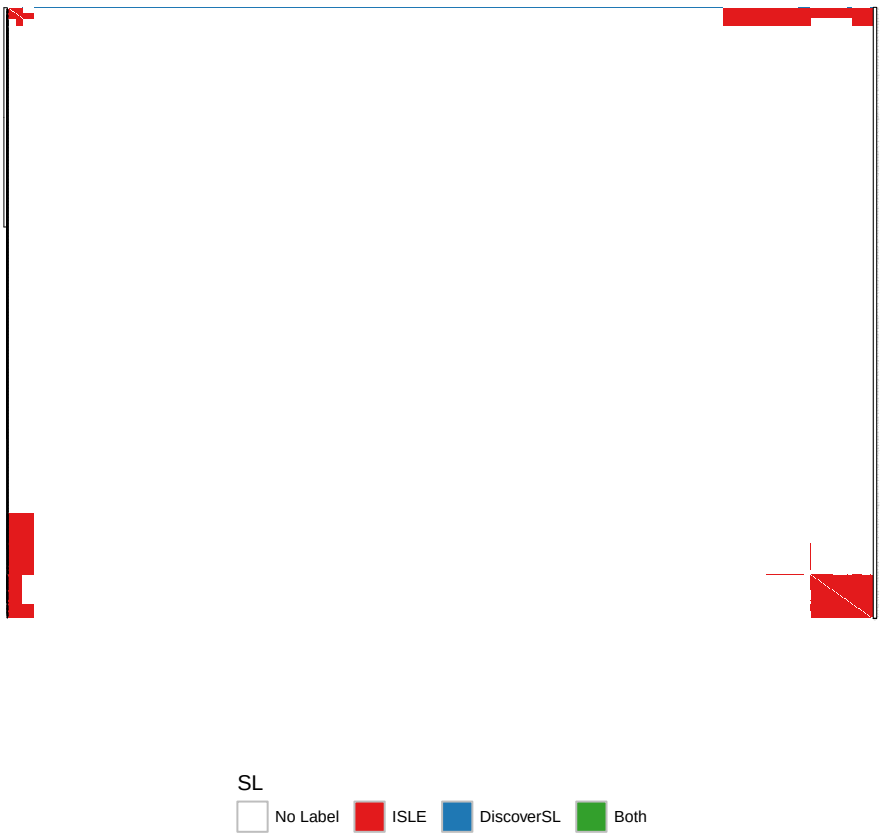


Fig. 5.S11: Structure of LUAD SL labels in the ISLE [33] and DiscoverSL datasets [12]. Heatmap showing labelled gene pairs from LUAD data from both the ISLE and DiscoverSL dataset. Elements along the horizontal and vertical axes represent unique genes. Each coloured dot represents a gene pair where a label exists in either the ISLE (red), DiscoverSL (blue), or both (green) datasets. Whitespace denotes a gene pair with no label. Rows are clustered using complete linkage and Euclidean distance with "No Label", "ISLE", "DiscoverSL", and "Both" encoded as -1, 0, 1, and 2, respectively. Both the rows and columns are ordered based on these clusters. The barplot to the right shows the number of occurrences of each gene in the LUAD dataset.

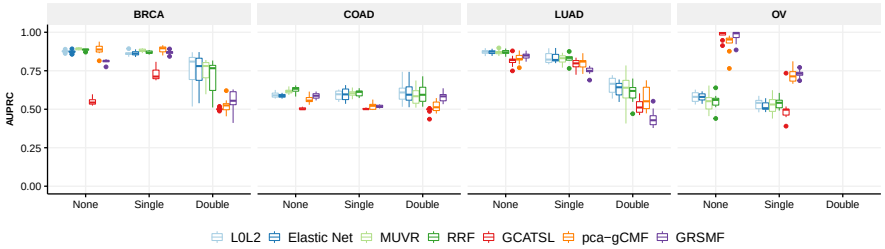


Fig. 5.S12: Performances of gene holdout experiments, where bias is controlled by ensuring that none, one, or both genes of pairs in the test set are excluded from the train set. Shown are AUPRC values for each gene-holdout experiment per cancer type (10 runs). For *None*, we only guarantee that train and test sets are disjoint in terms of gene pairs, not individual genes; for *Single*, only one gene from a gene pair in the test set can be present in the train set; for *Double* neither gene of a pair in the test set appears in the train set. The results for "None" are calculated from the same experiment as Tables 2-3 in the main text. Note: there was insufficient data to conduct the OV *Double* experiment.

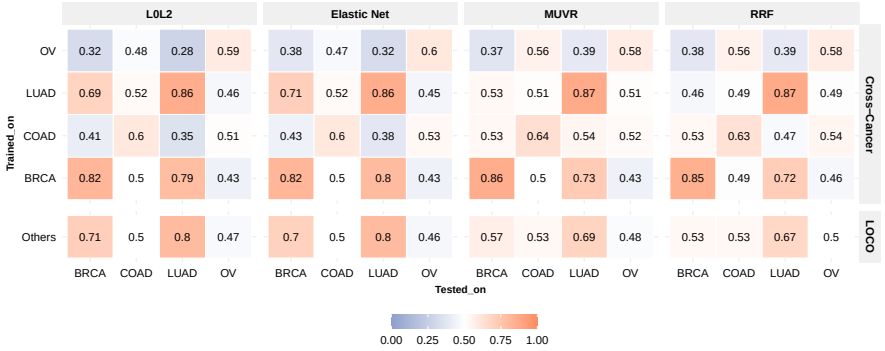


Fig. 5.S13: Heatmaps of average AUROC performances for the L0L2 [22], Elastic Net [20], MUVR [54], and RRF [15] models over 10 runs. *Cross-cancer:* Vertical and horizontal axes denote the cancer types used to train and test, respectively. *LOCO:* Horizontal axis denotes the cancer type held out for testing. Models trained on balanced data from all other cancers.

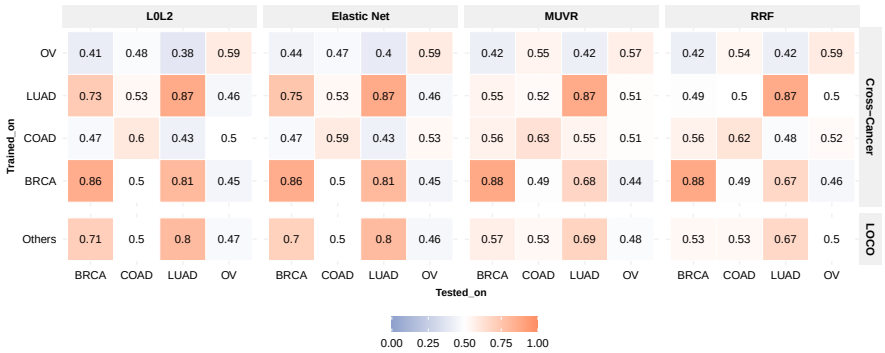


Fig. 5.S14: Heatmaps of average AUPRC performances for the L0L2 [22], Elastic Net [20], MUVR [54], and RRF [15] models over 10 runs. *Cross-cancer:* Vertical and horizontal axes denote the cancer types used to train and test, respectively. *LOCO:* Horizontal axis denotes the cancer type held out for testing. Models trained on balanced data from all other cancers.

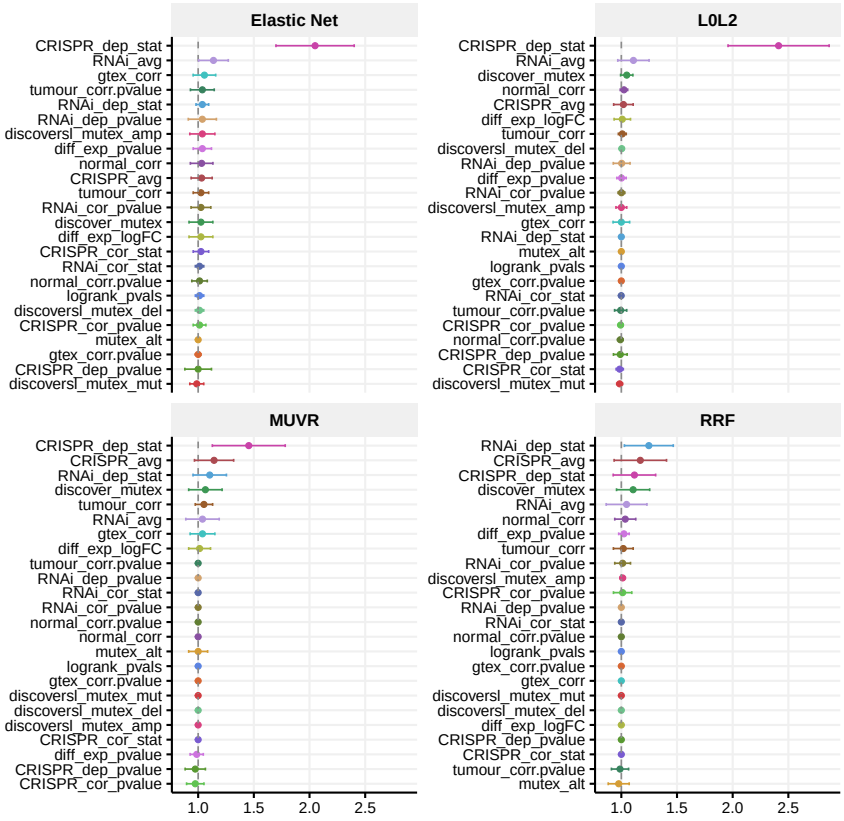


Fig. 5.S15: Median feature importance scores for the BRCA one-cancer models.These were scored using 100 repetitions of the model agnostic permutation feature importance algorithm [18]. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions.

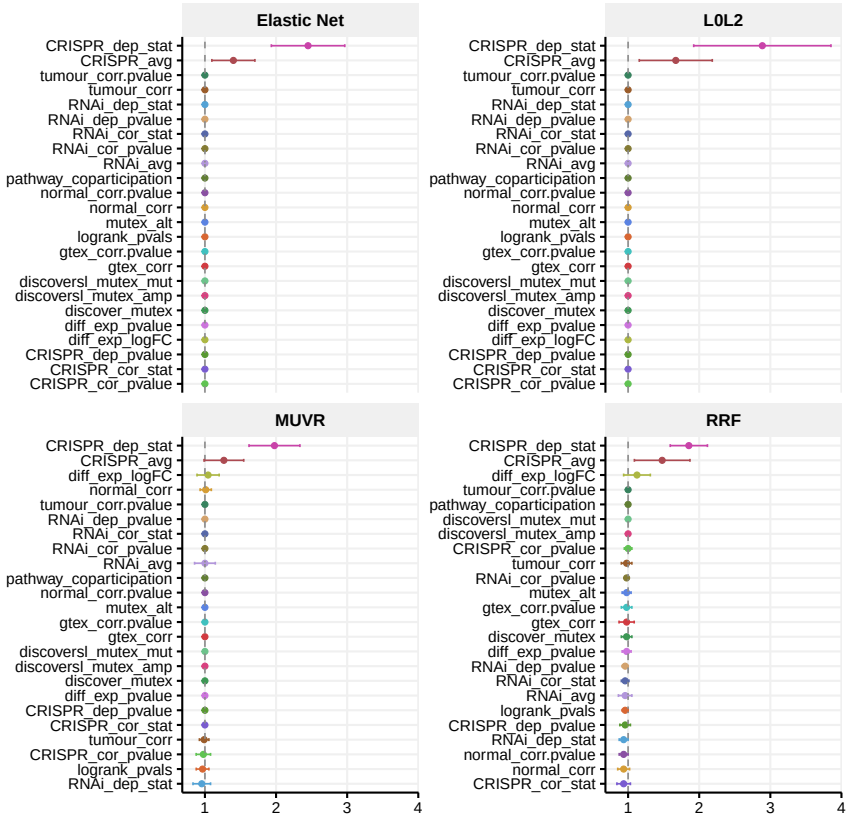


Fig. 5.S16: Median feature importance scores for the LUAD one-cancer models.These were scored using 100 repetitions of the model agnostic permutation feature importance algorithm [18]. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions.

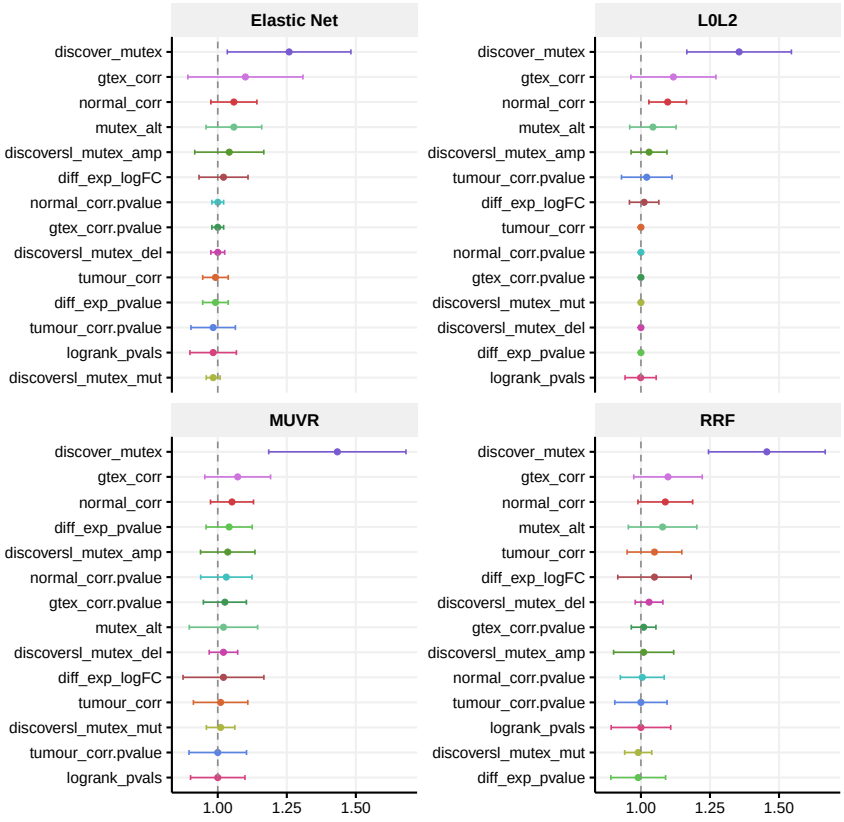


Fig. 5.S17: Median feature importance scores for the BRCA models trained without gene dependency-based features. These were scored using 100 repetitions of the model agnostic permutation feature importance algorithm [18]. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions..

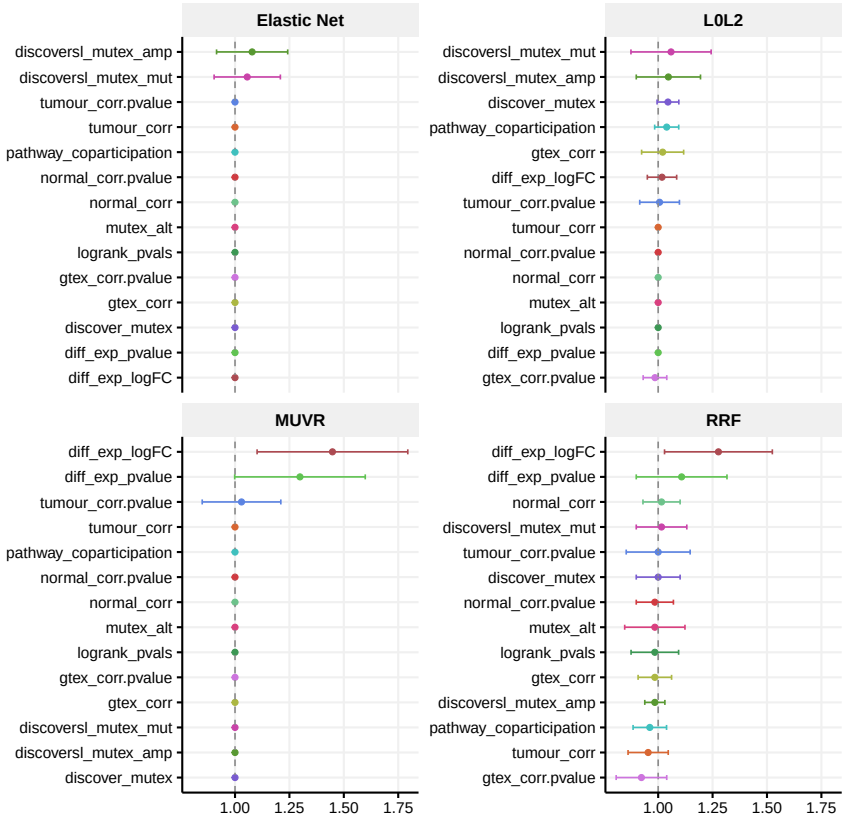


Fig. 5.S18: Median feature importance scores for the LUAD models trained without gene dependency-based features. These were scored using 100 repetitions of the model agnostic permutation feature importance algorithm [18]. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions.

5.6 Supplementary Tables

Symbol	Description	Biological sample	Data type	Used
CRISPR_dep_stat	Change in CRISPR dependency score of one gene based on non-silent mutations in the other (Wilcoxon)	Cancer cell lines	CRISPR dependency [46, 14] Mutation [21]	Here
CRISPR_dep_pvalue	Significance of change in dependency of one gene based on non-silent mutations in the other (Wilcoxon)	Cancer cell lines	CRISPR dependency [46, 14] Mutation [21]	Here
CRISPR_cor_stat	Correlation of gene-wise CRISPR dependency scores (Pearson's)	Cancer cell lines	CRISPR dependency [46, 14]	Here

Continued on next page

– continued from previous page

Symbol	Description	Biological sample	Data type	Used
CRISPR_cor_pvalue	Significance of correlation of gene-wise CRISPR dependencies (*t*-test)	Cancer cell lines	CRISPR dependency [46, 14]	Here
CRISPR_avg	Average of gene-wise means of CRISPR dependency scores	Cancer cell lines	CRISPR dependency [46, 14]	Here
RNAi_dep_stat	See CRISPR equivalent	Cancer cell lines	RNAi dependency [60, 44] Mutation [21]	Here
RNAi_dep_pvalue	See CRISPR equivalent	Cancer cell lines	RNAi dependency [60, 44] Mutation [21]	[27]
RNAi_cor_stat	See CRISPR equivalent	Cancer cell lines	RNAi dependency [60, 44]	Here
RNAi_cor_pvalue	See CRISPR equivalent	Cancer cell lines	RNAi dependency [60, 44]	Here
RNAi_avg	See CRISPR equivalent	Cancer cell lines	RNAi dependency [60, 44]	Here

Continued on next page

– continued from previous page

Symbol	Description	Biological sample	Data type	Used
discover_mutex	Mutual exclusivity score [8]	Patient tumour	CNV, mutation [45, 32]	Here
discoversl_mutex_amp	Significance of non-co-occurrence of amplifications (hyper-geom.)	Patient tumour	CNV [45, 32]	[12]
discoversl_mutex_del	Significance of non-co-occurrence of deletions (hyper-geom.)	Patient tumour	CNV [45, 32]	[12]
discoversl_mutex_mut	Significance of non-co-occurrence of non-silent mutations (hyper-geom.)	Patient tumour	Mutation [32]	[12]
discoversl_mutex	Combined *p*-value of previous three scores using Fisher's method	Patient tumour	CNV, mutation [45, 32]	[12]

Continued on next page

– continued from previous page

Symbol	Description	Biological sample	Data type	Used
mutex_alt	Significance of non-co-occurrence of amplifications, deletions or non-silent mutations	Patient tumour	CNV, mutation [45, 32]	Here
MUTEX	Mutual exclusivity score [2] (failed to identify mutually exclusive pairs in our datasets and was thus not used in training)	Patient tumour	Mutation [32]	Here
logrank_pval	Significance of change in survival time [6] between patients with(out) aberrant expression or CNV in both genes	Patient tumour	CNV, mutation, expression, [45, 32, 49] patient clinical data [49]	Here

Continued on next page

– continued from previous page

Symbol	Description	Biological sample	Data type	Used
diff_exp_logFC	Differential expression of a gene based on mutations in other (log fold-change)	Patient tumour	Mutation, expression [32, 49]	[12]
diff_exp_pvalue	Significance of differential expression of a gene based on mutations in the other (edgeR test p-value, [52])	Patient tumour	Mutation, expression [32, 49]	[12]
gtex_corr	Co-expression (Pearson's correlation)	Healthy donor	Expression [39]	Here
gtex_corr._pvalue	Significance of co-expression (t-test)	Healthy donor	Expression [39]	Here

Continued on next page

– continued from previous page

Symbol	Description	Biological sample	Data type	Used
tumour_corr	Co-expression (Pearson's correlation)	Patient tumour	Expression [49]	[12]
tumour_corr._pvalue	Significance of co-expression (*t*-test)	Patient tumour	Expression [49]	[12]
normal_corr	Co-expression (Pearson's correlation)	Patient normal	Expression [49]	Here
normal_corr._pvalue	Significance of co-expression (*t*-test)	Patient normal	Expression [49]	Here
pathway_coparticipation	Significance of co-occurrence in pathways (hypergeom.)	Pathway databases	Pathway gene sets [35]	[12]

Tab. 5.S1: Features used in SBSL prediction models. Columns "Biological sample" and "Data type" indicate the type of biological samples and the kind of data (e.g. molecular profiles) acquired for those samples, respectively. Descriptions of how each feature was calculated, and sources of the data, are provided in the main manuscript (Methods). The "Used" column indicates where this specific feature was first used in the context of computational SL prediction.

Held-out	Cancer Type	Pan-cancer Unbalanced	Pan-cancer Balanced	Cancer-specific
BRCA	BRCA	$.64 \pm .02$	$.75 \pm .01$	$\mathbf{.83 \pm .01}$
	COAD	$.52 \pm .02$	$.51 \pm .02$	$\mathbf{.60 \pm .02}$
	LUAD	$.73 \pm .03$	$.79 \pm .02$	$\mathbf{.83 \pm .02}$
	OV	$.40 \pm .04$	$.53 \pm .04$	$\mathbf{.58 \pm .03}$
COAD	BRCA	$.65 \pm .02$	$.77 \pm .02$	$\mathbf{.84 \pm .01}$
	COAD	$.52 \pm .02$	$.53 \pm .02$	$\mathbf{.60 \pm .02}$
	LUAD	$.74 \pm .02$	$.80 \pm .02$	$\mathbf{.85 \pm .02}$
	OV	$.40 \pm .04$	$.50 \pm .04$	$\mathbf{.59 \pm .03}$
LUAD	BRCA	$.76 \pm .01$	$.82 \pm .02$	$\mathbf{.86 \pm .01}$
	COAD	$.62 \pm .02$	$.60 \pm .01$	$\mathbf{.64 \pm .01}$
	LUAD	$.81 \pm .02$	$.83 \pm .02$	$\mathbf{.86 \pm .01}$
	OV	$\mathbf{.55 \pm .06}$	$.52 \pm .04$	$.54 \pm .07$
OV	BRCA	$.75 \pm .02$	$.80 \pm .02$	$\mathbf{.86 \pm .01}$
	COAD	$.62 \pm .02$	$.61 \pm .02$	$\mathbf{.63 \pm .02}$
	LUAD	$.80 \pm .02$	$.83 \pm .02$	$\mathbf{.87 \pm .02}$
	OV	$.55 \pm .04$	$.53 \pm .05$	$\mathbf{.57 \pm .07}$

Tab. 5.S2: Performance of one-cancer and pan-cancer SL prediction models (with unbalanced and balanced cancer representation) tested on heldout examples of each cancer type. Mean and standard deviation of AUROC for 10 repetitions.

Held-out	Cancer Type	Pan-cancer Unbalanced	Pan-cancer Balanced	Cancer-specific
BRCA	BRCA	$.68 \pm .02$	$.78 \pm .01$	$\mathbf{.88 \pm .01}$
	COAD	$.53 \pm .02$	$.52 \pm .02$	$\mathbf{.59 \pm .02}$
	LUAD	$.78 \pm .02$	$.84 \pm .01$	$\mathbf{.87 \pm .02}$
	OV	$.46 \pm .02$	$.51 \pm .03$	$\mathbf{.58 \pm .05}$
LUAD	BRCA	$.68 \pm .02$	$.81 \pm .03$	$\mathbf{.87 \pm .01}$
	COAD	$.53 \pm .02$	$.52 \pm .02$	$\mathbf{.59 \pm .01}$
	LUAD	$.79 \pm .02$	$.83 \pm .02$	$\mathbf{.87 \pm .02}$
	OV	$.45 \pm .03$	$.50 \pm .02$	$\mathbf{.58 \pm .04}$
COAD	BRCA	$.82 \pm .01$	$.85 \pm .01$	$\mathbf{.89 \pm .01}$
	COAD	$.61 \pm .03$	$.60 \pm .02$	$\mathbf{.62 \pm .01}$
	LUAD	$.84 \pm .03$	$.84 \pm .02$	$\mathbf{.87 \pm .01}$
	OV	$\mathbf{.54 \pm .05}$	$.52 \pm .05$	$.51 \pm .05$
OV	BRCA	$.81 \pm .01$	$.84 \pm .02$	$\mathbf{.89 \pm .01}$
	COAD	$.60 \pm .02$	$.60 \pm .03$	$\mathbf{.63 \pm .02}$
	LUAD	$.81 \pm .03$	$.83 \pm .02$	$\mathbf{.87 \pm .02}$
	OV	$\mathbf{.55 \pm .05}$	$.52 \pm .05$	$.54 \pm .05$

Tab. 5.S3: Performance of one-cancer and pan-cancer SL prediction models (with unbalanced and balanced cancer representation) tested on heldout examples of each cancer type. Mean and standard deviation of AUPRC for 10 repetitions.

Feature	VIF
discoversl_mutex_amp	1.388619555
discover_mutex	1.039665442
mutex_alt	1.381683071
RNAi_avg	2.224074149
RNAi_cor_pvalue	1.011676996
RNAi_cor_stat	1.013070426
RNAi_dep_pvalue	1.109236607
RNAi_dep_stat	4.917394579
CRISPR_avg	2.247942347
CRISPR_cor_pvalue	1.007564922
CRISPR_cor_stat	1.006988081
CRISPR_dep_pvalue	1.141351776
CRISPR_dep_stat	4.953596929
gtex_corr	1.334004715
gtex_corr.pvalue	1.071894916
tumour_corr	1.170333149
tumour_corr.pvalue	1.048525217
normal_corr	1.289831761
normal_corr.pvalue	1.430066751
diff_exp_logFC	1.023954926
diff_exp_pvalue	1.072218472
pathway_coparticipation	1.018279983
logrank_pvals	1.347337851

Tab. 5.S4: Variance inflation factors (VIF) [25] per feature for the combined dataset. The variance inflation factor indicates how many times higher the variance of the feature coefficient is than one would expect if there was no collinearity. A VIF of 1 indicates that a feature does not correlate with any other features. A VIF of 2 indicates that the variance of a particular feature coefficient is two times higher than one would expect if there was no collinearity, indicating moderate correlation with other features. A VIF value larger than 5 is considered to indicate high correlation with other features.

Cancer Type	Gene Pair	Functional Association
BRCA	TP53/BOP1	BOP1 is a regulator of p53 pathway function. [48]
	TP53/CRYGS	CRYGS is a target gene of the TP53 transcription factor. [31]
	TP53/EEF1D	Not found.
	TP53/AKAP8L	Physical interaction between AKAP8L and TP53 proteins. [36]
	TP53/CPSF1	Not found.
LUAD	KRAS/ANAPC4	ANAPC4 is synthetic lethal with KRAS. [43]
	KRAS/FBL	Not found.
	KRAS/PSMA1	PSMA1 is synthetic lethal with KRAS. [11]
	KRAS/PSMA5	Not found.
	KRAS/RRM2	KRAS is a regulator of RRM2. [65]

Tab. 5.S5: Functional associations found for the top 5 ranked SBSL-L0L2 predictions for BRCA and LUAD. As evidence for functional associations specific to the cancer type in question was rarely available, we instead report functional associations in general. We report the top 5 ranked SL pairs as predicted by the L0L2 model, as the SBSL linear models generalised better across SL gold standard datasets and cancer types. There were no significant differences between the predictions of the L0L2 and Elastic Net models.

References

[1] Michael Ashburner, Catherine A Ball, Judith A Blake, et al. "Gene Ontology: tool for the unification of biology". In: *Nature genetics* 25.1 (2000), pp. 25–29 (cit. on p. 140).

[2] Özgün Babur, Mithat Gönen, Bülent Arman Aksoy, et al. "Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations". In: *Genome Biology* 16.1 (Dec. 2015), p. 45 (cit. on pp. 138, 140, 176).

[3] Shrikant I Bangdiwala. "The Wald statistic in proportional hazards hypothesis testing". In: *Biometrical Journal* 31.2 (1989), pp. 203–211 (cit. on p. 141).

[4] Fiona M. Behan, Francesco Iorio, Gabriele Picco, et al. "Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens". In: *Nature* 568.7753 (Apr. 2019), pp. 511–516 (cit. on pp. 138, 139).

[5] Graeme Benstead-Hume, Xiangrong Chen, Suzanna R Hopkins, et al. "Predicting synthetic lethal interactions using conserved patterns in protein interaction networks". In: *PLoS Computational Biology* 15.4 (2019), e1006888 (cit. on p. 137).

[6] Viv Bewick, Liz Cheek, and Jonathan Ball. "Statistics review 12: survival analysis". eng. In: *Critical Care* 8.5 (Oct. 2004), pp. 389–394 (cit. on pp. 141, 176).

[7] Ruichu Cai, Xuexin Chen, Yuan Fang, Min Wu, and Yuexing Hao. "Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers". In: *Bioinformatics* 36.16 (2020), pp. 4458–4465 (cit. on pp. 136, 137).

[8] Sander Canisius, John W.M. Martens, and Lodewyk F.A. Wessels. "A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence". In: *Genome Biology* 17.1 (2016), pp. 1–17 (cit. on pp. 138, 140, 153, 175).

[9] Nuria Conde-Pueyo, Andreea Munteanu, Ricard V Solé, and Carlos Rodríguez-Caso. "Human synthetic lethal inference as potential anti-cancer target gene detection". In: *BMC Systems Biology* 3.1 (2009), pp. 1–15 (cit. on p. 136).

[10] The Gene Ontology Consortium. “The Gene Ontology resource: enriching a GOld mine”. In: *Nucleic acids research* 49.D1 (2021), pp. D325–D334 (cit. on p. 140).

[11] Kyle R Cron, Kaya Zhu, Deepa S Kushwaha, et al. “Proteasome inhibitors block DNA repair and radiosensitize non-small cell lung cancer”. In: *PloS one* 8.9 (2013), e73710 (cit. on p. 182).

[12] Shaoli Das, Xiang Deng, Kevin Camphausen, and Uma Shankavaram. “DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers”. In: *Bioinformatics* 35.4 (Feb. 2019). Ed. by Russell Schwartz, pp. 701–702 (cit. on pp. 137, 138, 140, 142, 165, 166, 175, 177, 178).

[13] Barbara De Kegel, Niall Quinn, Nicola A Thompson, David J Adams, and Colm J Ryan. “Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines”. In: *Cell Systems* 12.12 (2021), pp. 1144–1159 (cit. on p. 136).

[14] Joshua M. Dempster, Jordan Rossen, Mariya Kazachkova, et al. “Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines”. In: *bioRxiv* (2019). doi: 10.1101/720243. eprint: `https://www.biorxiv.org/content/early/2019/07/31/720243.full.pdf` (cit. on pp. 138, 139, 173, 174).

[15] Houtao Deng and George Runger. “Feature selection via regularized trees”. English (US). In: *2012 International Joint Conference on Neural Networks, IJCNN 2012*. Proceedings of the International Joint Conference on Neural Networks. doi: 10.1109/IJCNN.2012.6252640. 2012 (cit. on pp. 142, 158, 167, 168).

[16] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874 (cit. on p. 159).

[17] Xiaodong Feng, Nadia Arang, Damiano Cosimo Rigiracciolo, et al. “A platform of synthetic lethal gene interaction networks reveals that the GNAQ uveal melanoma oncogene controls the hippo pathway through FAK”. In: *Cancer Cell* 35.3 (2019), pp. 457–472 (cit. on p. 136).

[18] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. “All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously”. In: *Journal of Machine Learning Research* 20.177 (2019), pp. 1–81 (cit. on pp. 143, 169–172).

[19] Ori Folger, Livnat Jerby, Christian Frezza, et al. “Predicting selective drug targets in cancer through metabolic networks”. In: *Molecular Systems Biology* 7.1 (2011), p. 501 (cit. on p. 136).

[20] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22 (cit. on pp. 142, 157, 167, 168).

[21] Mahmoud Ghandi, Franklin W. Huang, Judit Jané-Valbuena, et al. “Next-generation characterization of the Cancer Cell Line Encyclopedia”. In: *Nature* 569.7757 (May 2019), pp. 503–508 (cit. on pp. 139, 173, 174).

[22] Hussein Hazimeh and Rahul Mazumder. “Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms”. In: *Operations Research* 68.5 (2020), pp. 1517–1537 (cit. on pp. 142, 156, 167, 168).

[23] Jiang Huang, Min Wu, Fan Lu, Le Ou-Yang, and Zexuan Zhu. “Predicting synthetic lethal interactions in human cancers using graph regularized self-representative matrix factorization”. In: *BMC Bioinformatics* 20.S19 (Dec. 2019), p. 657 (cit. on pp. 136, 137, 142).

[24] Alexandra Jacunski, Scott J Dixon, and Nicholas P Tatonetti. “Connectivity homology enables inter-species network models of synthetic lethality”. In: *PLoS Computational Biology* 11.10 (2015), e1004506 (cit. on p. 136).

[25] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013 (cit. on pp. 144, 181).

[26] Bijay Jassal, Lisa Matthews, Guilherme Viteri, et al. “The reactome pathway knowledgebase”. In: *Nucleic Acids Research* 48.D1 (2020), pp. D498–D503 (cit. on p. 139).

[27] Livnat Jerby-Arnon, Nadja Pfetzer, Yedael Y. Waldman, et al. “Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality”. In: *Cell* 158.5 (Aug. 2014), pp. 1199–1209 (cit. on pp. 136, 137, 142, 174).

[28] Minoru Kanehisa and Susumu Goto. “KEGG: Kyoto Encyclopedia of Genes and Genomes”. In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. eprint: `https://academic.oup.com/nar/article-pdf/28/1/27/9895154/280027.pdf` (cit. on p. 139).

[29] Hesham M Korashy, AFM Motiur Rahman, and Mohammed Gabr Kassem. “Dasatinib”. In: *Profiles of Drug Substances, Excipients and Related Methodology* 39 (2014), pp. 205–237 (cit. on p. 145).

[30] T Kranthi, SB Rao, and P Manimaran. “Identification of synthetic lethal pairs in biological systems through network information centrality”. In: *Molecular BioSystems* 9.8 (2013), pp. 2163–2167 (cit. on p. 136).

[31] Alexander Lachmann, Huilei Xu, Jayanth Krishnan, et al. “ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments”. In: *Bioinformatics* 26.19 (2010), pp. 2438–2444 (cit. on p. 182).

[32] Michael S Lawrence, Petar Stojanov, Paz Polak, et al. “Mutational heterogeneity in cancer and the search for new cancer-associated genes”. In: *Nature* 499.7457 (2013), pp. 214–218 (cit. on pp. 175–177).

[33] Joo Sang Lee, Avinash Das, Livnat Jerby-Arnon, et al. “Harnessing synthetic lethality to predict the response to cancer treatment”. In: *Nature Communications* 9.1 (Dec. 2018), p. 2546 (cit. on pp. 136, 138, 165, 166).

[34] Herty Liany, Anand Jeyasekharan, and Vaibhav Rajan. “Predicting synthetic lethal interactions using heterogeneous data sources”. In: *Bioinformatics* 36.7 (2020), pp. 2209–2216 (cit. on pp. 137, 142).

[35] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, et al. “Molecular Signatures database (MSigDB) 3.0”. In: *Bioinformatics* 27.12 (May 2011), pp. 1739–1740. eprint: `https://academic.oup.com/bioinformatics/article-pdf/27/12/1739/716159/btr260.pdf` (cit. on pp. 139, 178).

[36] Jiang Liu, Di Guan, Maogong Dong, et al. “UFMylation maintains tumour suppressor p53 stability by antagonizing its ubiquitination”. In: *Nature cell biology* 22.9 (2020), pp. 1056–1063 (cit. on p. 182).

[37] Yong Liu, Min Wu, Chenghao Liu, Xiao-Li Li, and Jie Zheng. “SL 2 MF: Predicting synthetic lethality in human cancers via logistic matrix factorization”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17.3 (2019), pp. 748–757 (cit. on pp. 136, 137).

[38] Yahui Long, Min Wu, Yong Liu, et al. “Graph contextualized attention network for predicting synthetic lethality in human cancers”. In: *Bioinformatics* 37.16 (Feb. 2021), pp. 2432–2440. eprint: `https:`

//academic.oup.com/bioinformatics/article-pdf/37/16/2432/39947241/btab110.pdf (cit. on pp. 137, 142).

[39] John Lonsdale, Jeffrey Thomas, Mike Salvatore, et al. "The genotype-tissue expression (GTEx) project". In: *Nature Genetics* 45.6 (2013), p. 580 (cit. on pp. 139, 177).

[40] Christopher J Lord, Andrew NJ Tutt, and Alan Ashworth. "Synthetic lethality and cancer therapy: lessons learned from the development of PARP inhibitors". In: *Annual Review of Medicine* 66 (2015), pp. 455–470 (cit. on p. 136).

[41] Xiaowen Lu, Wout Megchelenbrink, Richard A Notebaart, and Martijn A Huynen. "Predicting human genetic interactions from cancer genome evolution". In: *PloS ONE* 10.5 (2015), e0125795 (cit. on pp. 136, 137).

[42] H. B. Mann and D. R. Whitney. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". In: *Annals of Mathematical Statistics* 18.1 (Mar. 1947), pp. 50–60 (cit. on p. 140).

[43] Timothy D Martin, Danielle R Cook, Mei Yuk Choi, et al. "A role for mitochondrial translation in promotion of viability in K-Ras mutant cells". In: *Cell reports* 20.2 (2017), pp. 427–438 (cit. on p. 182).

[44] James M McFarland, Zandra V Ho, Guillaume Kugener, et al. "Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration". In: *Nature Communications* 9.1 (2018), pp. 1–13 (cit. on pp. 138, 139, 174).

[45] Craig H Mermel, Steven E Schumacher, Barbara Hill, et al. "GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers". In: *Genome Biology* 12.4 (2011), R41 (cit. on pp. 139, 175, 176).

[46] Robin M Meyers, Jordan G Bryan, James M McFarland, et al. "Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells". In: *Nature Genetics* 49.12 (Dec. 2017), pp. 1779–1784 (cit. on pp. 138, 139, 173, 174).

[47] Sebastian MB Nijman. "Synthetic lethality: general principles, utility and detection using genetic screens in human cells". In: *FEBS Letters* 585.1 (2011), pp. 1–6 (cit. on p. 136).

[48] Dimitri G Pestov, Žaklina Strezoska, and Lester F Lau. "Evidence of p53-dependent cross-talk between ribosome biogenesis and the cell cycle: effects of nucleolar protein Bop1 on G1/S transition". In: *Molecular and cellular biology* 21.13 (2001), pp. 4246–4255 (cit. on p. 182).

[49] Mumtahena Rahman, Laurie K. Jackson, W. Evan Johnson, et al. “Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results”. In: *Bioinformatics* 31.22 (July 2015), pp. 3666–3672. eprint: `https://academic.oup.com/bioinformatics/article-pdf/31/22/3666/17122567/btv377.pdf` (cit. on pp. 139, 176–178).

[50] Karthik Raman, Aditya Pratapa, Omkar Mohite, and Shankar Balachandran. “Computational prediction of synthetic lethals in genome-scale metabolic models using Fast-SL”. In: *Metabolic Network Reconstruction and Modeling*. Springer, 2018, pp. 315–336 (cit. on p. 136).

[51] Florian Richoux, Charlène Servantie, Cynthia Borès, and Stéphane Téletchéa. “Comparing two deep learning sequence-based models for protein-protein interaction prediction”. In: *arXiv preprint* (2019). arXiv: 1901.06268 (cit. on p. 154).

[52] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010), pp. 139–140 (cit. on pp. 141, 177).

[53] Carl F Schaefer, Kira Anthony, Shiva Krupa, et al. “PID: the pathway interaction database”. In: *Nucleic Acids Research* 37.Database Issue (2009), pp. D674–D679 (cit. on p. 139).

[54] Lin Shi, Johan A Westerhuis, Johan Rosén, Rikard Landberg, and Carl Brunius. “Variable selection and validation in multivariate modelling”. In: *Bioinformatics* 35.6 (2019), pp. 972–980 (cit. on pp. 142, 167, 168).

[55] Sriganesh Srihari, Jitin Singla, Limsoon Wong, and Mark A Ragan. “Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer”. eng. In: *Biology Direct* 10 (Oct. 2015), p. 57 (cit. on p. 138).

[56] Thomas Stoeger, Martin Gerlach, Richard I Morimoto, and Luís A Nunes Amaral. “Large-scale investigation of the reasons why potentially important genes are ignored”. In: *PLoS Biology* 16.9 (2018) (cit. on p. 137).

[57] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, et al. “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles”. In: *Proceedings of the National*

Academy of Sciences 102.43 (2005), pp. 15545–15550. eprint: https://www.pnas.org/content/102/43/15545.full.pdf (cit. on p. 139).

[58] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, et al. "The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets". In: *Nucleic Acids Research* 49.D1 (2021), pp. D605–D612 (cit. on p. 139).

[59] TCGA GDAC. "Firehose stddata__2016_01_28 run". In: *Broad Institute of MIT and Harvard* (2016). doi: 10.7908/C11G0KM9 (cit. on p. 139).

[60] Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, et al. "Defining a Cancer Dependency Map". In: *Cell* 170 (3 2017), 564–576.E16 (cit. on p. 174).

[61] Saman Maleki Vareki. "High and low mutational burden tumors versus immunologically hot and cold tumors and response to immune checkpoint inhibitors". In: *Journal for immunotherapy of cancer* 6.1 (2018), pp. 1–5 (cit. on p. 144).

[62] Fangping Wan, Shuya Li, Tingzhong Tian, et al. "EXP2SL: A Machine Learning Framework for Cell-Line-Specific Synthetic Lethality Prediction". In: *Frontiers in Pharmacology* 11 (2020), p. 112 (cit. on pp. 136, 137).

[63] Mark Wappett, Austin Dulak, Zheng Rong Yang, et al. "Multi-omic measurement of mutually exclusive loss-of-function enriches for candidate synthetic lethal gene pairs". In: *BMC Genomics* 17.1 (2016), p. 65 (cit. on pp. 136, 137).

[64] Min Wu, Xuejuan Li, Fan Zhang, et al. "In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer". In: *Cancer Informatics* 13 (2014), CIN–S14026 (cit. on p. 136).

[65] Yasuhiro Yoshida, Toshiyuki Tsunoda, Keiko Doi, et al. "KRAS-mediated up-regulation of RRM2 expression is essential for the proliferation of colorectal cancer cell lines". In: *Anticancer research* 31.7 (2011), pp. 2535–2539 (cit. on p. 182).

[66] Fan Zhang, Min Wu, Xue-Juan Li, et al. "Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates". In: *Journal of Bioinformatics and Computational Biology* 13.03 (2015), p. 1541002 (cit. on p. 136).

6

Discussion

DNA repair is one of the fundamental functions of the genome, essential for maintaining stability and integrity. Without it, mutations would accumulate, genes would be disrupted, and chromosomes would fragment, ultimately leading to premature cellular aging, failure to replicate, and death. Such aberrations would be unavoidable, as DNA repair protects the genome from pervasive mutagenic threats, ranging from intrinsic processes at subcellular level to galactic cosmic rays originating from the distinct reaches of our universe [5].

Perhaps less catastrophic, but more immediately problematic, defective DNA repair can lead to an increased risk of contracting diseases like cancer due to unchecked spontaneous mutagenesis. However, current cancer treatments like chemotherapy and radiation therapy can also exploit this link between cancer cells and compromised DNA repair to selectively kill cancer cells. Defective DNA repair thus represents a "double-edged sword": increasing the risk factor for developing cancer, yet offering a mechanism to treat the disease. However, some of these therapies are still toxic to healthy cells, with significant side effects for patients. Therefore, advancing our understanding of DNA repair mechanisms is crucial for refining cancer therapeutics, reducing treatment toxicity, and improving patient quality of life. Understanding DNA repair and the associated pathways is an ambitious task; the pathways are highly complex, constituting an intricate network of redundancies and interconnected machinery that address different types of DNA damage. This dissertation has focused on the critical DNA double-strand break (DSB) repair pathways, each responsible for producing distinct mutational patterns during repair —from the high-fidelity repairs of homologous recombination to the rapid and direct ligations of non-homologous end joining, or the sequence-selective characteristics of microhomology-mediated end joining.

Due to the highly interconnected nature of the DNA repair pathways and the overlap in the mutations they produce, it is difficult to isolate mutational patterns and reverse engineer the causal mechanisms and behaviours, requiring vast quantities of mutational data covering different sequence contexts and genotypes. This dissertation takes advantage of the fact that each repair pathway leaves unique mutational signatures, which can now be examined across a large data landscape through CRISPR technology. By integrating large-scale CRISPR datasets with novel computational approaches, we aimed to maximise insights while addressing specific challenges inherent to this new data type. The dissertation explored four main aspects: (1) predicting CRISPR repair outcomes in data-rich cell lines and adapting these predictions to data-scarce contexts; (2) identifying novel genes involved in DSB repair through outlier analysis of genome-wide CRISPR knockout mutational screens without using traditional controls; (3) elucidating functions of candidate DSB repair genes using knockout mutational spectra by identifying signatures shared with known DSB repair genes and attributing functions based on guilt-by-association approach.; and (4) identifying synthetic lethality interactions through genome-wide functional CRISPR screens, and further presenting experimental methods to assess synthetic lethality-prediction model robustness to selection bias in current gold standard label sets.

6.1 Challenges in detailed CRISPR repair outcome prediction modeling

Predicting detailed CRISPR repair outcome distributions, as we do in Chapter 2, poses unique challenges to computational modelling. The first challenge is that the set of outcomes comprising a valid outcome distribution differs per input target sequence, in contrast to classic multi-label problems where the set of output labels would be the same across all possible input sequences. Complex rules define the possible outcomes for each sequence (discussed in Chapter 2), and most current models – including our proposed X-CRISP model – circumvent explicitly incorporating these rules into the model. Instead, they assume independence between each indel and score them separately, aggregating across all possible indels to produce the final distribution. Yet, this independence assumption is inherently flawed, as outcome frequencies are interdependent based on the input sequence.

Ignoring this interdependence and interactions between outcomes may be a limiting factor on predictive performances. As noted in Chapter 2, while the average predictive performances are quite good across current models, there is yet a large amount of variability in performance across different target sites, and perhaps modeling approaches would benefit from predicting the entire distribution and considering these interactions. Currently, only one model predicts the entire distribution, Lindel, yielding good yet sub-state-of-the-art performances. The reason Lindel does not achieve top performances even while considering the entire distribution may be due to the second challenge of predictive modeling of CRISPR repair outcomes – engineering input features.

When designing repair outcome prediction models, it is difficult to avoid coupling input features to the output labels, as both their cardinalities vary according to the target sequence. If one tries to model the entire distribution (hundreds of repair outcomes), adding a new feature to describe an outcome results in an explosion in dimensionality as a feature is replicated for each possible outcome. To avoid the curse of dimensionality (having far more features than samples to train on), popular choices have been to limit the number of features used to describe each outcome as Lindel does, or take the X-CRISP approach and model each outcome independently. One consequence is that current models may not encode sequence characteristics or interactions important for predicting repair outcome likelihood, again potentially limiting performance.

Another consequence is that, by missing certain features or interactions between potential outcomes, these models are not capable of fully describing what sequence characteristics are indicative of repair outcome likelihood. Furthermore, as the features in use are typically engineered from domain expertise and not derived from data, their ability to capture novel sequence characteristics that influence outcome prediction is limited. None of the current detailed repair outcome prediction models encode the sequence directly as input, which itself presents an opportunity for the next advancement in repair outcome prediction. Transformer-based architectures could enhance CRISPR modeling by extracting sequence features in a data-driven manner while leveraging the attention mechanism to provide interpretable explanations for model decisions at the sequence level. Hybridising transformer-based neural network models with external domain knowledge to manage sequence-specific outcome possibilities may allow us to build interpretable,

data-driven models that are not based on faulty assumptions and address the key challenges of repair outcome interdependence and feature engineering described above.

6.2 Clinical potential of DSB-induced mutational spectra

Studies have shown that precision therapies targeting specific tumor vulnerabilities can be more effective than broader DNA-damage inducing treatments that affect healthy cells as well. One such vulnerability is deficiency in DSB repair, for which multiple drugs are already in clinical use. For instance, patients with BRCA-deficient or HDR-compromised tumors can reportedly benefit from PARP-inhibitor therapy exploiting synthetic lethality between the PARP and BRCA genes [1].

To identify further vulnerabilities leading to opportunities for targeted treatments, computational models have been proposed to infer repair deficiencies from mutational signatures based on somatic mutations throughout the genome, such as HRDetect [4] and CHORD [6]. For example, HRDetect has demonstrated predictive capabilities on patient tumours in laboratory settings, identifying breast cancer sensitivity to anthracyclines [4] and being prognostic of platinum therapy duration in thoracic and gastrointestinal cancers [7]. However, these models are not yet validated for routine clinical use, and we anticipate that they face a key limitation should they advance to the clinical stage: analysing mutational signatures based on WGS/WES mutational catalogue requires the accumulation of a substantial mutational burden to generate a reliable signal, which may limit the detection of recently acquired but clinically relevant repair deficiencies. Given the time-sensitive nature of cancer treatment, early detection of repair deficiencies could be crucial.

CRISPR-induced mutational spectra may offer a more targeted, immediate alternative. Our transfer learning approaches in Chapter 2 demonstrated how repair outcome prediction models could be tuned to predict mutational spectra in DSB repair-deficient cells. Thus, developing models to reverse the prediction task, (i.e. predict cellular DSB repair deficiencies from observed mutational spectra) should be feasible. Furthermore, our NMF signature analysis in Chapter 4 revealed how available repair mechanisms shape mutational spectra, thus it may even be possible to model the DSB repair capabilities of

the cell in some continuous multidimensional latent space. By inducing DSBs at specific loci and sequencing tumour cells before and after induction, we could directly model and assess DSB repair capabilities by analyzing the mutational outcomes. This strategy promises a cleaner signal, possibly allowing easier capture of rare or subtle repair patterns often missed in whole-genome analyses. Further, mutational spectra analysis of tumours in the clinic could enable on-demand assessment of repair status, potentially making it more effective for timely clinical decision-making when compared to whole-genome somatic mutation analysis.

6.3 Controls for mutational spectra

These studies highlight key considerations for researchers using large-scale CRISPR mutational data in DNA repair research. A primary challenge is selecting appropriate controls, as the datasets in Chapters 3 and 4 either lacked controls or relied on non-targeting sgRNAs. While non-targeting sgRNAs approximate baseline mutational distributions, they typically exhibit lower variance than targeting sgRNAs (knockouts). This discrepancy may arise from DDR stress responses induced by gene knockouts, which perturb cells before DSB induction. Known variations in sgRNA efficacy [3] within multi-guide knockout strategies may further contribute to increased variability in knockout mutational spectra. Other factors may also play a role, but it is clear that non-targeting controls do not fully account for the variability introduced by gene knockouts and subsequent DSB induction.

Recognizing these inherent differences in variance is essential for guiding data-driven research questions and methodological decisions. For instance, statistical analyses that assume homoscedasticity (equal variance) may yield misleading results when comparing knockout and non-targeting control groups.

Until all hidden sources of variability between groups can be identified, the extent to which this variability can be controlled or accounted for experimentally may be limited. Future experiments generating mutational spectra could incorporate a well-designed panel of sgRNAs targeting non-essential intronic regions [2] to better capture mutational variability by inducing similar DDR stress as knockouts while avoiding genotypic alterations. This approach may improve statistical robustness, though variability due to other factors – such

as the differences in sgRNA efficacy across knockout guides – may remain. Therefore, while experimental strategies may help reduce these discrepancies, accounting for residual variability will remain necessary in future studies and analyses.

6.4 Sequencing depth for mutational spectra

Another consideration that warrants discussion is the sequencing depth for mutational spectra analysis. Our outlier detection and non-negative matrix factorisation signatures analyses draw on CRISPR knockout screens with targeted loci for DSB induction. These screens vary in scope: one covers knockouts across the entire genome (Chapter 3), while the other examines about 750 genes with substantially greater read depth (Chapter 4). This depth provides finer resolution for detecting rare mutations and local relationships, capturing details that broader but shallower genome-wide screens miss. Therefore, sequencing depth emerges as a critical factor in capturing robust mutational signatures over broad coverage.

Another challenge to consider when approaching large-scale CRISPR screening as the tool of choice for examining the minutiae of DNA repair processes. Broad, well-documented patterns tend to be representative of the more understood aspects of these mechanisms. To identify novel genes or functional nuances, the focus shifts from global patterns to smaller, localised deviations in mutational behaviour. Our algorithmic approaches to analyse this data from Chapters 3 and 4 reflect this, with capabilities to identify highly correlated and highly localized patterns in the dataset taking precedence, over identifying singular and globally occurring mutations. This shift underscores, again, the importance of high sequencing depth in studies aimed at characterizing mutational spectra, as deeper data enable better differentiation of rare but informative mutations.

In general, when considering resource allocation in experimental designs focused on generating CRISPR mutational spectra, we suggest that researchers face two main trade-offs when allocating their sequencing read budget: depth versus breadth concerning *(i)* gene knockouts and *(ii)* targeted loci for mutational spectra generation. This introduces the problem of selection bias, which was tackled in Chapter 5, albeit within the domain of synthetic lethality prediction. First, reducing the number of targeted loci can bias results toward

repair processes that are most active or yield the most distinct signatures in those specific sequence contexts. Second, limiting the number of gene knockouts may restrict the detection of certain latent signatures or could lead to the misattribution of mutations to particular repair processes. In either case, analyses should carefully account for the selection bias introduced by narrowing the scope to sub-genome levels or restricting the number of targeted loci.

6.5 Final remarks

In summary, this dissertation contributes to the field of DNA repair by developing and applying computational approaches tailored to the challenges of CRISPR-generated mutational data. By refining predictive models and exploring methods to identify gene associations and repair signatures, this work offers practical tools for analyzing DSB repair at a greater level of specificity. These methods underscore the value of deep sequencing and thoughtful control selection in experimental designs, offering guidance for future research aiming to characterize DNA repair pathways more accurately. Altogether, this work advances our ability to interpret complex CRISPR data, adding insights that can support broader efforts in DNA repair research and therapeutic development.

References

[1] Alice Chen. “PARP inhibitors: its role in treatment of cancer”. In: *Chinese journal of cancer* 30.7 (2011), p. 463 (cit. on p. 194).

[2] Chen-Hao Chen, Tengfei Xiao, Han Xu, et al. “Improved design and analysis of CRISPR knockout screens”. In: *Bioinformatics* 34.23 (2018), pp. 4095–4101 (cit. on p. 195).

[3] Giulia I Corsi, Kunli Qu, Ferhat Alkan, et al. “CRISPR/Cas9 gRNA activity depends on free energy changes and on the target PAM context”. In: *Nature Communications* 13.1 (2022), p. 3006 (cit. on p. 195).

[4] Helen Davies, Dominik Glodzik, Sandro Morganella, et al. “HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures”. In: *Nature medicine* 23.4 (2017), pp. 517–525 (cit. on p. 194).

[5] Z Li, KK Jella, Lahcen Jaafar, et al. “Exposure to galactic cosmic radiation compromises DNA repair and increases the potential for oncogenic chromosomal rearrangement in bronchial epithelial cells”. In: *Scientific reports* 8.1 (2018), p. 11038 (cit. on p. 191).

[6] Luan Nguyen, John WM Martens, Arne Van Hoeck, and Edwin Cuppen. “Pan-cancer landscape of homologous recombination deficiency”. In: *Nature communications* 11.1 (2020), p. 5584 (cit. on p. 194).

[7] Erica S Tsang, Veronika Csizmok, Laura M Williamson, et al. “Homologous recombination deficiency signatures in gastrointestinal and thoracic cancers correlate with platinum therapy duration”. In: *NPJ Precision Oncology* 7.1 (2023), p. 31 (cit. on p. 194).

Acknowledgements

Honestly, writing this section of my PhD thesis is something I never believed I would achieve: to say the period of the PhD has been the toughest of my life would be no exaggeration, and would not do the difficulties justice. First, undertaking this endeavour against the backdrop of the COVID-19 pandemic magnified the anxiety and impostor syndrome that most who undertake a PhD understand all too well. Following this, dealing with personal and mental health troubles that only my family and the closest of friends will ever comprehend. And finally, and most tragically, experiencing the loss of both of my parents only a year apart from each other, leaving a hole in my heart that I can never again hope to fill.

And yet, while I have lost, I have gained much also. It is certainly not all gloom. I have made new friends. I have grown as a person, a friend, and yes, even as a researcher, and so here we are – marking the end of my PhD journey. Where this thesis ends, a whole new unwritten chapter of my life begins. I would like to believe that this work lays testament to what can be achieved with the love and compassion of friends and family, with self-reflection and healing with oneself, and with some old-fashioned hard work and persistence. This adventure has been one of emotional fortitude, even more so than intellectual aptitude, and here I wish to recognise the people who provided the support I so desperately needed to carry me this far.

First, to my best friends; **Yildiz**, you have been there to help me through it all, to hold me up when I was down, to inspire me when I did not want to go further, and to understand me when I thought no one could. I hope that I can repay you in full the kindness that you have shown me, and I look forward with great excitement to see what the future has in store for you; **Shane**, for me, you have always been an ever-steady rock. You are dependable, always there when needed, and never ask for anything in return. Your family and love make me feel like I have a second home whenever I return to Ireland. I

am sure you will continue to be an amazing father to Marc and Meadbh, and husband to Eimear, as you have been a friend to me; **Sharlene**, you were one of the first people who got me to open up to view and think differently about the world. With you, I can talk high and low and always feel safe, from the warmth of your living room to some of the highest peaks in the world, and I look forward to each and every conversation; **Miranda**, it is amazing to me to consider how such a short time in your company changed the direction of my life. You were the biggest single inspiration for encouraging me to take this path and to push myself further than I ever thought I could. Never change. I love you all dearly.

To my promoters and lab colleagues throughout the years; **Joana**, my supervisor, your support and patience through this process, especially during the tragedies my family has endured, is one of the main reasons I can write these words today. I hope in the future we may find a chance to work together again; **Marcel**, thank you for your help and patience, especially towards the end of my PhD where you were very generous with your time; **Yasin**, since we started this journey together, I have always admired your friendliness, your openness, and your curiosity. We persisted through some very challenged times together, and I think we are all the better for it, and I am excited to see what becomes of your future; **Sander**, thank you for both listening to me and talking to me throughout the PhD when times got hard, it was generous and kind and won't be forgotten; **Roy**, thanks for all the coffee breaks, and I am sorry I never got to DM a game of DnD for you; **Ivan**, thanks for the debates, your interesting insights, and the silly accents, "come on"; **Sara**, thanks for infecting the whole lab with your joy and energy, you truly brighten any room you enter, but I'm still going to get you for that joke that made me look like a fool.

To my collaborators at the Leiden University Medical Centre: **Marcel**, thank you for your patience and guidance throughout the years, working with you was an absolute pleasure; **Marco**, thank you for being so generous with your time, I really enjoyed the cross-disciplinary experience of working hands-on with a biologist, and it has helped to influence me to make this a part of my career going forward; **Robin**, thanks for your time throughout the PhD, you were always approachable and willing to help.

To my other DBL and DBL-adjacent colleagues, I apologise for the tonal shifts as you read through the next few passages directed towards you all. They

bounce back and forth between sincerity and sarcasm and it's your job to figure out which is which. The clue is that it is all meant with love. To: **Alex**, you just started so we haven't interacted much, but best of luck in your PhD, and try to keep one eye open during BioTalks; **Ahmed**, thanks for the chats and assistance throughout the years; **Amelia** and **Chirag**, thank you for inviting me to share in your wedding day, it was truly something special and I appreciate being invited to be part of it, to Chirag especially for all your questions about martial arts, it's always fun to discuss it with you; **Azza**, thank you for your amazing help and technical support, I hope you will visit Ireland again soon; **Bianca**, forever 65, always a chore, etc, etc; **Bram**, give me a shout if you ever want to take a trip to the bins; **Chengyao**, thanks for all the chats, confidence boosts, and inviting to order lunch with you all those times; **Daan**, it was always a pleasure, you've got a great sense of humour and I hope to see you around at a conference in the future; **Daniyal** thanks for the chats and for not dying that one time in the office; **Ellie**, thank you for showing me cool people can like DnD too; **Gabriel**, thanks for all the SMILES; **Gerard**, thanks for the talks and especially the support you've provided since I started work at the AUMC; **Inez**, thanks for the chats and the salsa dances, let's do more; **Jana**, thanks for the talks through the last few years and best of luck as you embark on your journey of being a new mom; **Jasmijn**, thanks for showing an interest in my mental health through the years, it really was appreciated; **Jasper**, thanks for saying offensive things and then asking if it was OK that you did so, just so you could do it again; **Kirti**, thanks for the chats and for trying DnD with me, and for your company in India; **Madaleon**, you are still my arch-nemesis and my thirst for vengeance (for the thing that I can't remember that you did) will never cease; **Lorenzo**, thanks for the chats about martial arts, it's always fun to discuss; **Marunka**, thanks for making the department a brighter place, your uplifting presence was always welcome; **Ojas**, it was a pleasure getting to know you more at the wedding, best of luck in the rest of your PhD; **Paul**, thanks for never taking yourself too seriously, for your off-beat humour and on-the-beat timing, I had a real pleasure getting to know you and chatting about life, love, martial arts, politics, but mostly just really stupid things, and you're not really that much of a racist after all; **Ruud**, thanks for your assistance, I always found you approachable and kind; **Stavros**, thanks for lending the office your questionably-appropriate sense of humour; **Stephanie**, thanks for listening to me and sharing with me, and for being open with me and making me feel like I always had someone to turn to, and thanks all the drama, may there be many more years to come (or maybe less, we could both do with less drama these days); **Swier**, thanks for

the rugby and fitness chats through the years; **Ramin**, thanks for being the first senior PhD in the office to tell me it was OK to make mistakes; **Tamim**, thanks for your guidance throughout the early years of my PhD; **Thomas**, thanks for the nice discussions at each bioretreat, I always enjoyed them; **Timo**, thanks for making me jealous by showing me you can be a rock star, AND a PhD candidate, you unreasonably talented jerk.

Anyone who knows me knows one of my great passions, which I came to find in Delft, is martial arts. To my coach and friend, **Pilao**, thank you for your selflessness and time. You have never asked for anything in return, and I can not repay the kindness you have shown me to help me achieve some of my dreams. I hope we will remain training partners for years to come. To everyone I have trained with throughout the years, it's been a privilege, and I would like to make some special mentions; **Alex**, your first win in BJJ competition showed me how much I enjoyed not just learning BJJ, but coaching too, thank you; **Riki**, thanks for sharing the experience of my first BJJ competition together in Germany back in the day, and for all the training since then; **Anton** thanks for all the pancakes, coffee, fight analysis, and friendship, I look forward to standing in your corner some day; **Emran**, even though we only interacted a short while, you left a big impression with your impressive jiu-jitsu skills, EQ, and my elbow still hurts; **Gilbert** thanks for your endless curiosity about the sport which helped me to think differently about it too, you deserve your blue belt; **Hew Wei**, thank you for honouring me by asking me to be in your corner; **Javier**, thank you for all the kickboxing tips and MMA sparring all the way back to the pandemic; **Jesus**, thanks for minding my house on occasion and for inviting me to the birthday party that I think you don't remember; **Julien**, thank you thank you thank you thank you thank you; **Lætitia**, from the moment we bonded complaining about you know who, I knew we would be mates; **Pepijn**, thanks for all the BJJ chats, it's really gratifying to see how your skills have developed; **Qichen**, thanks for training and interesting BJJ discussion; **Sara**, thanks for all the coffee, weekend lunches, drama, sparring, training, and being an uplifting presence in my life; **Serdar**, thanks for our weekly waffles and chats, you have given me plenty to think about over the years; **Tom**, thanks for all the training; **Tomás**, thanks for the beers, company, and weekend brunches. I've always felt welcome in your home, and I always get a reminder of what I miss about Ireland when I'm in your presence; **Wojtek**, thanks for all the times you say "did you know...?". because I know two things; I don't know, and the answer will be gold; **Leon** thank you for leaving a permanent mark on my face to

remember you by; **Adnan** and **Ran**, I put this together because I have the same note for both - thanks for the offers to BBQs and drinks, I may not always attend but I appreciate every one of them.

For some of my friends from Sosalsa: **Elena**, thanks for all the gossip, but also for all the listening, I feel like we speed-ran years of friendship in a few months and I really appreciate you; **Nienke**, thank you for introducing me to the joys of dancing, for being the first friend I made at SoSalsa, for all the long talks and chats and being there to listen, I truly appreciate you; **Savanne**, thanks for your wicked and addictive sense of humour, I always enjoy our chats and dances.

To my other good friends from my gaming group back in Ireland; **Brian**, I enjoy the rare moments you make it to play these days, you should neglect your family and play more; **Eoin**, thanks for introducing me to DnD and giving me yet another hobby to sink more money into; **Gary**, I will always appreciate the endless number of ways you have invented to inform an opponent, boss, or entire game to ingest various parts of your anatomy; **Kieran**, your rage-baiting, trash talk knows no equal, but you're also a pretty good buddy to go drinking with at a concert too; **Matt**, honest to God, I think I can no longer disassociate you from Britney Spears now, they occupy the same neuron in my brain, and that's not OK; and **Paddy**, we don't play together as much as in the Rainbow Six Seige days, but I still appreciate your banter and creativity when we get together for DnD. To you all, I would like to say that I really could not ask for a more dependable group. No matter how stressful things get, you guys are always there to listen each evening, distract with random banter, shoot the next bad guy, or guess the next movie, and throughout the years have provided a constant through-line as everything else around me changed. I hope this remains the same for many more years to come.

To other friends I have made throughout the years around the world; **Barbara** thanks for the coffees and drinks and chats through the years, you have been an inspiring presence in my life and I know that you will go on to do great things; **Bianca** and **Matteo**, I am grateful to have met you both and to have had your support through our studies and beyond, even as our lives have taken their own paths; **Ines** thanks for the banter, gossip, and tea, and always having another story to tell; **Senuri**, I delight in every conversation, phone call, and visit, and hope that the fact that I am finishing my PhD journey and

gives you the hope that: if even this guy can do it, surely you can too; **Tom Hardiman**, thanks for your friendship throughout the years, and though we may not keep in touch as often anymore, I look forward to each catch-up when we do.

To my therapist, **Pirkko**, I want to appreciate your hard work and patience, for helping me open up, learn about myself, deal with my grief, and help me to grow.

During the tragic loss of our parents, there were many people who helped me to survive that period. I want to acknowledge all my aunts, uncles, cousins, and friends of the family who shared in our loss. There are some people who I would like to pay a special mention: to **Mary** and **Frank Haslem**, I will never be able to thank you enough for how you stepped up to help us through this time, you both taught me the value of having neighbours you can rely on; to **Margaret** and **Micheal Keegan**, through different stages of my life, you have both share your wisdom and helped me to grow as a human, from bringing me out of my shell as a kid working in the butchers, to learning to being an adult learning to talk and deal with life, love, and loss. Thank you.

To my siblings, **Adrian**, **David**, and **Kathy**, thank you for your love and support. Time helps heal all wounds, but some scars remain forever. I hope we can all continue to learn and grow together and that I can be there to help you achieve your dreams in the same way you helped me to achieve mine. To my uncle, **Tony**, thank you for teaching me that sometimes life is simpler than we perceive it to be, and that all you can do is to try to do your best.

Finally, to my mother, **Kathleen**, you instilled into me my love for music and cooking, and demonstrated every day your endless selflessness in how you cared for your family. To my father, **Gerry**, you taught me the value of hard work, of curiosity, of patience, of humour, and of kindness. I will never be able to write enough words, eloquently enough, or elaborate enough to convey exactly how much I miss you both. I wish you were here to read these words and see me at the end of this journey, but I know you would have been proud no matter the outcome. My one regret is that I did not express to you more how much I love you while you were with me, but at least I can do it here. I love you both, I miss you dearly, and I dedicate this work in its entirety to you.

Curriculum Vitae

Colm Seale

1991	Born in Portlaoise, Ireland

Education

2008-2012	**BEng in Electronic and Computer Engineering** National University of Ireland, Galway Galway, Ireland
2018-2020	**MSc in Computer Science** Delft University of Technology Delft, The Netherlands
2020-2025	**Doctoral Candidate** Delft University of Technology Delft, The Netherlands

Professional Experience

2012-2016	**Software Engineer** Fidelity Investments Galway, Ireland
2016-2018	**Senior Software Engineer** Altocloud Galway, Ireland

List of Publications

- **Colm Seale**, Yasin Tepeli, & Joana P. Gonçalves. (2022). "Overcoming selection bias in synthetic lethality prediction." Bioinformatics, 38(18), 4360-4368, 10.1093/bioinformatics/btac523.

- Yasin Tepeli, **Colm Seale**, & Joana P Gonçalves. (2022). "ELISL: early–late integrated synthetic lethality prediction in cancer." Bioinformatics, 40(1), 10.1093/bioinformatics/btad764.

- Sander Goossens, Yasin Tepeli, **Colm Seale**, & Joana P. Gonçalves. "SNMF: Integrated Learning of Mutational Signatures and Prediction of DNA Repair Deficiencies." bioRxiv, 10.1101/2024.11.27.624656.

- **Colm Seale** and Joana P. Gonçalves. "X-CRISP: Domain-Adaptable and Interpretable CRISPR Repair Outcome Prediction." Bioinformatics Advances, 10.1093/bioadv/vbaf157.

- **Colm Seale**, Marco Barazas, Robin van Schendel, Marcel Tijsterman, and Joana P. Gonçalves. "MUSICiAn: Detecting Gene-DNA Repair Associations via Control-Free Mutational Spectra Analysis." bioRxiv, 10.1101/2025.01.27.635038. Submitted to Nucleic Acids Research: Genomics and Bioinformatics.

- **Colm Seale**, Marco Barazas, Robin van Schendel, Marcel Tijsterman, and Joana P. Gonçalves "Signatures in CRISPR Mutational Spectra Reveal Role and Interplay of Genes in DNA Repair."