Contributions to Science:

Contribution 1: Discovery of a new type of carbon pump that drives CO_2 concentration in bacteria:

1.1: Historical background Rubisco is the enzyme responsible for fixing the vast majority of all the CO_2 that enters the biosphere each year and is essential for all plants, algae, and most autotrophic bacteria. However, Rubisco evolved before the great oxygenation event and is competitively inhibited by the presence of O_2. This is a problem that all autotrophs that use rubisco must overcome to survive in the the modern atmosphere, with its 20% O_2 and only 0.04% CO_2. Many autotrophic bacteria including a wide range of chemolithoautotrophs and most cyanobacteria overcome this issue using an α-carboxysome based CO_2 concentrating mechanism (CCM). These CCMs work by creating a compartment inside the bacteria where the concentration of CO_2 is raised high enough to saturate rubisco with CO_2 and out-compete O_2. While our theoretical understanding and previous experimental work suggested that an inorganic carbon (C_i) transporter was absolutely required for the functioning of the CCM, no such transporter was known in the model chemolithoautotroph *H. neapolitanus*.

1.2: Central finding I performed a genome-wide genetic screen for CCM components in *H. neapolitanus*. I identified the essential components of the CCM in *H. neapolitanus*. I cloned putative C_i transporters, called DABs, then confirmed their activity with reporter strain and C_i uptake assays in *E. coli*. I showed that data was consistent with these transporters acting not as direct transporters but as a new family of energy coupled carbonic anhydrase (CA) enzymes. These CAs concentrate C_i by converting membrane permeable CO_2 into membrane impermeable HCO_3^- causing a net flow of C_i into the cell. I further showed that similar operons in the human pathogens *V. Cholera* and *B. anthracis* had the same function.

1.3: Influence/Application This work was instrumental in ongoing study of the CCM including a successful effort to reconstitute a functional α-carboxysome CCM in *E. coli*, and work on CCM evolution that I will discuss in contribution 2. Since this work was published, it was shown that a similar operon in *S. aureus* has the same function and is essential for growth in atmospheric CO_2 concentrations. Considering the activity of energy coupled CA transporters has been shown to be necessary to understand the carbon isotope fractionation data in very old rock strata. There has been interest in investigating the potential applications of DABs in engineering crop plants and autotrophic bio-fuel production hosts (like *C. necator*) for increased yield.

1.4: My role I was the primary author on this work. I conceived and designed the experiments, performed the genetic screen, analyzed the sequencing data, did cloning, performed the biochemistry experiments, and performed the reporter strain experiments.

John J Desmarais, ... et al. (2018). *The essential gene set for bacterial carbon concentration*. Western Photosynthesis Conference. Biosphere 2, Oracle, Arizona.

John J Desmarais, ... et al. (2019a). *DABs Accumulate Bicarbonate*. Gordon Research Conference - Photosynthesis. Sunday River Resort, Maine.

John J Desmarais, ... et al. (Dec. 2019b). "DABs are inorganic carbon pumps found throughout prokaryotic phyla". en. In: *Nat Microbiol* 4.12, pp. 2204–2215. ISSN: 2058-5276. DOI: 10.1038/s41564-019-0520-8.

Contribution 2: Characterization of potential evolutionary paths for developing carbon dioxide concentrating mechanisms:

2.1: Historical background The α-carboxysome based CO_2 concentrating mechanism (CCM) required several major evolutionary steps to evolve. These were acquiring uncoupled CA activity, gaining C_i transport, and encapsulating CA and rubisco in an α-carboxysome. However, all of these components are need for the effect of the CCM and removing even one of these components is lethal in modern autotrophs. Since there is no apparent fitness benefit for a partial system, it is not clear how the system could have evolved. Geochemical evidence suggests that the atmosphere was very different when the CCM first evolved, with much higher levels of CO_2 and much lower levels of O_2 and a mixture of biological and geochemical processes has slowly changed the atmospheric composition to the current mixture. This information lead us to the hypothesis that evolutionary intermediates of the CCM may

have provided fitness benefits at intermediate atmospheric compositions.
2.2: Central finding We compared the effect of CCM gene knockouts in *H. neapolitanus* across different intermediate CO_2 concentrations. We also compared the growth phenotypes of partial CCM constructs in CO_2 concentration dependent strains of *E. coli* and *C. necator* that we constructed. We found that introduction of an uncoupled CA (non-transporter) or a coupled CA (C_i transporter) provided a fitness benefit at intermediate CO_2 concentrations. We also found that while combining either of these activities with the α-carboxysome on their own provided no benefit, combining them with each other provided benefits in some situations. This was unexpected, because including a coupled and an uncoupled CA in the same cell without encapsulation is expected to produce a futile cycle. Our mathematical modeling suggests that at intermediate CO_2 concentrations and high growth rates the cell can become limited by both CO_2 and HCO_3 and C_i consumption is high enough that cycling is actually beneficial for providing both C_i species for growth. This revealed a possible path to evolving a CCM through only fitness positive steps by acquiring first either a coupled or uncoupled CA, then acquiring the other type of CA as CO_2 fall further, and finally evolving an α-carboxysome as CO_2 concentrations approach modern levels.
2.3: Influence/Application This work provides insight into the evolution of the CCM and into possible life strategies of modern organisms living in high CO_2 environments. These strategies might also be useful for improving the growth of industrial autotrophs at intermediate CO_2 concentrations. Further, showing the functional expression of the DAB in *C. necator* offers the potential for using the DAB to improve bioplastics production.
2.4: My role My role in this project was to measure the effect of all CCM gene knockouts on the growth of *H. neapolitanus* across a panel of intermediate CO_2 concentrations to establish which components of it's CCM are needed for growth at intermediate CO_2 concentrations.

Avi I Flamholz, ..., John J Desmarais, ... et al. (2022). "Trajectories for the evolution of bacterial CO_2-concentrating mechanisms". In: *Proceedings of the National Academy of Sciences* 119.49, e2210539119. DOI: 10.1073/pnas.2210539119.

Contribution 3: Random sampling general epistasis protein fitness landscape mapping and design:
3.1: Historical background Deep mutational scanning allows close mapping of a protein's fitness landscape by using massively parallel assays to measure the fitness of all single mutants of a protein of interest. This provides information useful for predicting the effects of mutants and in designing new mutants. However, deep mutational scanning experiments are difficult to perform because they require you to first make a library of all single mutants of a protein. They also only provide information on fitness in the protein's local context limiting their utility for prediction and design of diverse proteins. Fitness landscape mapping efforts would benefit greatly from being able to learn from data sets that are easier to generate and provide information over a wider area. Efforts to learn from random mutants for prediction and design have tended to rely on large neural networks of linear methods. However, linear methods miss nonlinearities in the data's true structure and neural networks operate as a black box that cannot be directly inspected to understand the results of the experiment. General epistatic models capture simple nonlinear genotype-phenotype relationships and may be a perfect fit for these applications, but have not been applied to protein design applications. Further, approaches trained on phylogentic data have been used for the same purposes but models that are able to fully leverage both phylogenetic and experimental data have been rare.
3.2: Central finding In this work we generate error prone PCR mutagenesis libraries of dihydrofolate reductase (DHFR) and measured enzyme activity in a massively parallel growth assay. I trained models including simple linear models, general epistatic models, and large neural nets with and without access to phylogenetic data. I showed that the general epistatic models had simmilar performance to the neural net on held out training data and out performed the phylogenetic data alone or the linear models. I also used each of these models to design new DHFR variants while systematically varying both number of mutations and optimization strategy to evaluate how well each model is able to extrapolate beyond it's training distribution. Evaluation of designed proteins was performed in a massively parallel growth

assay. This work is currently in the final data analysis and writing phase, but we hope to submit soon.
3.3: Influence/Application This work will be useful in providing new methods for rapidly mapping the genotype-phenotype landscape of proteins and designing new mutants. It will also provide simple directly inspectable models that produce performance on par with neural net approaches for some prediction and design tasks.
3.4: My role I am the primary author on this work. I designed and conceived of the experiments, produced mutant libraries, performed massively parallel growth assays, wrote analysis code, wrote model code, trained models, tested model performance, and evaluated optimized sequence behavior.

Contribution 4: Development of new CRISPR tools:
4.1: Historical background CRISPR associated proteins have fuel a biotechnological revolution. They have found uses ranging from knocking out genes in human cells, to controlling expression in bacteria, to detecting nucleic acids in clinical samples. Crafting new CRISPR tools relies on either discovery of new natural proteins that have evolved the activity we require or combining already known proteins in new combinations. During my graduate career I helped with two main CRISPR tool projects, the identifying CasX as a new RNA guided DNA nuclease and the development of Cas13 and Csm6 based RNA diagnostics.
4.2: CasX genome editing RNA guided DNA cleaving nucleases were the foundational discovery of the CRISPR field. At the time of this work, cas9 and cas12a were the only two families proteins known to perform this function. However, these proteins are large and can be difficult to deliver as therapeutics, so there was interest in identifying new smaller nucleases. CasX was discovered as a ¡1,000 Amino acids RuvC containing protein in CRISPR loci and shown to be capable of protecting from plasmid transformation. We demonstrated guided cutting activity *in vitro* as well as gene knockout and CRISPRi *in vivo*. We then solved a Cryo-EM structure of the complex with target DNA and identified two new domains the non-target strand binding (NTSB) domain and the target strand loading (TSL) domain. We showed that the NTSB is needed for unwinding dsDNA. We also detected a putative zinc binding motif in the TSL and showed that casX purifies bound to zinc. My role in this project was to use x-ray fluorescence spectroscopy to identify the zinc bound to the purified protein. CasX provides a new modality for genome editing that is proving useful in a variety of applications and provides hope for aiding the treatment of a wide variety of diseases. Scribe Therapeutics is pursuing casX based therapies.
4.3: Csm6 boosted Cas13 RNA detection CRISPR based diagnostics offer a quick and simple method for detecting nucleic acids that can be coupled to a fluorescence or lateral flow assay without multiple liquid handling or incubation steps. This makes them attractive for use in at-home or point-of-care diagnostics. However existing CRISPR diagnostics were not sensitive enough to detect clinically relevant levels of SARS-COV2 in patient samples without pre-amplification of target sequences which negated all of these advantages. Class III CRISPR systems include a cyclic-oligo-A activated nuclease csm6 that is used to amplify the CRISPR immune response. We were interested in determining if we could link cas13 detection of viral RNAs to csm6 activation to improve limit of detection and allow SARS-COV2 detection in SARS-COV2 clinical samples. We designed collateral substrates for cas13 that would activate Csm6. We used modeling and spike in assays to determine that activator self cleavage was limiting sensitivity and chemically modified the activator inhibit this process, achieving detection of SARS-COV2 clinical samples. My role in this work was to perform kinetic modeling of different potential reaction setups to predict potential improvement in time to detection or limit of detection, this included the analysis that suggested that using an uncleavable activator would remove self-limitation. I also wrote analysis pipelines to process data and perform statistical analysis of data, including developing methods for detecting positive samples from microfluidic device data. This work has demonstrated a new method of improving time to detection and sensitivity in CRISPR diagnostics. This work will contribute to new and improved diagnostics technologies for the detection of a variety of clinically and scientifically relevant nucleic acids.

Jun-Jie Liu, ..., John Desmarais, ... et al. (Feb. 2019). "CasX enzymes comprise a distinct family of RNA-guided genome editors". en. In: *Nature*, p. 1. ISSN: 0028-0836. DOI: 10.1038/s41586-019-0908-x.

Tina Y Liu, ..., John J Desmarais, ... et al. (Aug. 2021). "Accelerated RNA detection using tandem CRISPR nucleases". en. In: *Nat. Chem. Biol.*, pp. 1–7. ISSN: 1552-4450. DOI: 10.1101/2021.03.19.21253328.

Contribution 5: Driving carbon flux toward chemical production by engineering glucose uptake during nitrogen starvation:
5.1: Historical background Using microbial hosts to produce chemicals offers the potential to produce a variety of compounds from renewable feed-stock. However, production hosts frequently loose production efficiency as natural selection drives evolution towards redirecting carbon and energy flux towards growth not production. This can be overcome by coupling production of the desired chemical to cell fitness, but in many cases this is not possible. An alternative strategy is growth decoupling, in which production hosts are grown up, then growth is stopped and all metabolic flux is directed towards production. However, when growth is stopped, many chassis organisms including *E. coli* slow and eventually halt their metabolism stopping production. A general strategy for enhancing metabolic rate during growth decoupling would dramatically improve prospects for engineered chemical production in biological hosts.
5.2: Central finding We found that by over-expressing PstI we were able to increase glucose uptake after growth was stopped by nitrogen limitation. However we did not find this increased glucose consumption increased yield significantly, and it is likely additional work will be needed to direct this increased flux towards chemical production.
5.3: Influence/Application This work provides another tool that metabolic engineers can use to optimize the production of their compound of interest.
5.4: My role My role in this project was to perform growth and chemical production assays with modified strains and to prepare samples for mass spectrometry.

Victor Chubukov, John James Desmarais, ... et al. (Jan. 2017). "Engineering glucose metabolism of Escherichia coli under nitrogen starvation". In: *NPJ Syst Biol Appl* 3, p. 16035. ISSN: 2056-7189. DOI: 10.1038/npjsba.2016.35.

Contribution 6: Development of nuclease amplification for cas13 viral diagnostics:
6.1: Historical background CRISPR based diagnostics offer an attractive option for rapid point of care or at home detection of nucleic acids, such as virus genomes. Their advantages include fast detection times, ease of conversion into either lateral flow or fluorescence assays, and simple operation (they do not require multiple liquid handling or incubation steps). However existing CRISPR diagnostics were not sensitive enough to detect clinically relevant levels of SARS-COV2 in patient samples without pre-amplification of target sequences which negated all of these advantages. Class III CRISPR systems include a cyclic-oligo-A activated nuclease csm6 that is used to amplify the CRISPR immune response. We were interested in determining if we could link cas13 detection of viral RNAs to csm6 activation to improve limit of detection and allow SARS-COV2 detection in clinical samples.
6.2: Central finding We found that cleavage of an A4-U6 substrate by cas13 produced a good activator for csm6. However, the reaction appeared to be self limiting. Kinetic modeling and reagent spike in experiments suggested that csm6's intrinsic activator cleavage activity was causing the self-limitation. By using a single-fluoro modified activator, we were able to remove this self limitation and detect SARS-COV2 genomes in clinical samples. We also developed a microfluidic device for automating sample processing and assay performance.
6.3: Influence/Application This work has demonstrated a new method of improving time to detection and sensitivity in CRISPR diagnostics. This work will contribute to new and improved diagnostics technologies for the detection of a variety of clinically and scientifically relevant nucleic acids.
6.4: My role My role in this work was to perform kinetic modeling of different potential reaction setups to predict potential improvement in time to detection or limit of detection, this included the analysis that suggested that using an uncleavable activator would remove self-limitation. I also wrote analysis

pipelines to process data and perform statistical analysis of data, including developing methods for detecting positive samples from microfluidic device data.

Tina Y Liu, ..., John J Desmarais, ... et al. (Aug. 2021). "Accelerated RNA detection using tandem CRISPR nucleases". en. In: *Nat. Chem. Biol.*, pp. 1–7. ISSN: 1552-4450. DOI: 10.1101/2021.03.19.21253328.

Contribution 7: Discovery of CasX:
7.1: Historical background The field of genome editing was launched by the discovery of RNA guided DNA cleaving nucleases. At the time there were only two such families, cas9 and cas12a, and there was interest in finding new programmable DNA nucleases with smaller sizes that would make them easier to deliver in therapeutics and potentially new mechanisms. CasX was discovered as a ¡1,000 Amino acids RuvC containing protein in CRISPR loci and shown to be capable of protecting from plasmid transformation.
7.2: Central finding We demonstrated that CasX was capable of cleaving dsDNA in cis and ssDNA in trans when provided with a complementary sgRNA *in vitro*. We showed induction of NHEJ at the targeted locus in human HEK293 T cells with wild type CasX and CRISPRi knockdown of genes in *E. coli* with dead casX. We solved cryo-EM to solve the structure of dcasX in complex with DNA and showed that there were two new domains the non-target strand binding (NTSB) domain and the target strand loading (TSL) domain. We showed that the NTSB is needed for unwinding dsDNA. We also detected a putative zinc binding motif in the TSL and showed that casX purifies bound to zinc.
7.3: Influence/Application CasX provides a new modality for genome editing that is proving useful in a variety of applications and provides hope for aiding the treatment of a wide variety of diseases. Scribe Therapeutics is pursuing casX based therapies.
7.4: My role My role in this project was to measure zinc content in purified protein using x-ray fluorescence spectroscopy.

Jun-Jie Liu, ..., John Desmarais, ... et al. (Feb. 2019). "CasX enzymes comprise a distinct family of RNA-guided genome editors". en. In: *Nature*, p. 1. ISSN: 0028-0836. DOI: 10.1038/s41586-019-0908-x.