Contributions to Science:

Contribution 1: A new type of inorganic carbon pump that drives CO_2 concentration:

1.1: *Historical background* The enzyme rubisco fixes >99.5% of the CO_2 entering the biosphere each year and is essential in plants, algae, and most autotrophic bacteria. However, Rubisco is inhibited by O_2, a problem in the the modern atmosphere with its 20% O_2 and only 0.04% CO_2. Many bacteria overcome this using an α-carboxysome based CO_2 concentrating mechanism (α-CCM). These systems rely on HCO_3^- pumping, however, the mechanism of HCO_3^- pumping was unknown in chemotrophs.

1.2: *Central finding* To identify HCO_3^- pumps, I screened for α-CCM genes in the model chemotroph *H. neapolitanus.* I identified two putative transporter operons, then showed sufficiency for pumping in *E. coli.* Unexpectedly, the data were consistent with energy coupled carbonic anhydrase (CA) activity not direct pumping. This causes HCO_3^- flux by converting membrane permeable CO_2 into membrane impermeable HCO_3^- trapping it in the cell. I showed homologs in the pathogens *V. Cholera* and *B. anthracis* had the same activity.

1.3: *My role* I conceived and designed the experiments, performed the genetic screens, analyzed sequencing data, performed the mechanistic experiments, and purified protein.

1.4: *Influence/Application* This work identified a new family of energy coupled CAs, only the second such family known. This work enabled reconstitution of a functional α-CCM in *E. coli.* A homolog found in *S. aureus* has the same function and is essential for growth in air. Factoring in energy coupled CAs aided interpretation of carbon isotope fractionation in rock strata. There are proposed applications of these pumps in engineering crop plants and autotrophic bio-fuel production hosts.

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: Potential evolutionary paths of carbon dioxide concentrating mechanisms:

2.1: *Historical background* The α-carboxysome based CO_2 concentrating mechanism (α-CCM) required several major evolutionary steps to evolve. However, none of the potential intermediates are expected to provide a fitness benefit in modern conditions so it is not clear how it evolved. The atmosphere was very different when the α-CCM evolved, with much higher levels of CO_2 and much lower levels of O_2. This lead us to hypothesize that evolutionary intermediates of the CCM may have provided fitness benefits at intermediate atmospheric compositions.

2.2: *Central finding* Evolving an α-CCM required acquiring a CA, gaining a HCO_3^- pump, and co-encapsulating CA and rubisco. Removing any of these stops α-CCM function in normal atmosphere. We measured the effect of α-CCM gene knockouts in *H. neapolitanus* across different CO_2 concentrations. We also measured the CO_2 dependent phenotypes of potential evolutionary intermediates in reporter strains of *E. coli* and *C. necator* that we constructed. We modeled carbon fluxes as a function of growth rate and CO_2 concentration. This data suggested that as CO_2 concentrations fall, HCO_3^- becomes limiting before CO_2. This suggested that as CO_2 started to fall either a pump or CA can help. As levels fall further CO_2 and HCO_3^- become co-limiting and having both a CA and a pump provides a benefit despite the potential for producing a futile cycle Eventually, only a full α-CCM will work. This provides a potential path for the evolution of a α-CCM.

2.3: *My role* I performed a massively parallel growth assay of gene knockouts in *H. neapolitanus* across intermediate CO_2 concentrations to identify CCM genes needed at intermediate CO_2 concentrations.

2.4: *Influence/Application* This work provides insight into the evolution of the α-CCM and into possible life strategies of modern organisms living in high CO_2 environments. These advances might also be useful for improving the growth of industrial autotrophs. Further, showing expression of the pumps in *C. necator* offers the potential to improve bio-plastics production.

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: General epistasis protein fitness landscape mapping and design:
3.1: Historical background Mutational scanning maps a protein's fitness landscape by measuring the fitness of all single mutants. This information is used for variant effect prediction and design. However, mutational scanning experiments require production of all single mutants and only provide information on fitness in the protein's local context, limiting their utility for divergent proteins. Being able to learn from data sets that are easier to generate and provide information over a wider area would be greatly beneficial. Current efforts to learn from random mutagenesis have relied on neural networks or linear methods. However, linear methods miss nonlinearities in the data's true structure and neural networks cannot be inspected to gain insight into function. General epistatic models capture nonlinear genotype-phenotype relationships without sacrificing interpretability but have not been applied to protein design. Phylogentic data also provide insight into fitness landscapes but inspectable models that are able to leverage both phylogenetic and experimental data have been rare.
3.2: Central finding We performed random mutagenesis on dihydrofolate reductase and measured enzyme activity in a massively parallel growth assay. I trained models including 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 test data. Both these models showed improvement when provided with phylogenetic data but out performed the phylogenetic data alone or with the linear model. I also used each model to design new variants while systematically varying both number of mutations and optimization strategy to evaluate how well each model is able to extrapolate. I used a massively parallel growth assay to validate 12,000 designed proteins. This work is currently awaiting sequencing results from the final experiment before submission.
3.3: Influence/Application This work provides new methods for mapping the genotype-phenotype landscape of proteins and designing new variants. It also provides 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 The ever expanding suite of CRISPR tools has helped drive a biotech revolution. I have been part of two tool development projects, identifying CasX as a new RNA guided DNA nuclease and the development of Cas13/Csm6 RNA diagnostics.
4.2: CasX genome editing RNA guided DNA nucleases launched the CRISPR field. However, nuclease size has hindered therapeutic applications. CasX is a <1,000 amino acid RuvC containing protein from CRISPR loci. We demonstrated guide directed cutting activity *in vitro* as well as cutting and CRISPRi *in vivo*. We solved a structure of the complex with target DNA and identified two new domains. We also detected a zinc binding motif and showed that CasX binds zinc. My role was to use x-ray fluorescence to detect zinc bound to the purified protein. CasX provides a new modality for genome editing that is proving useful in a variety of applications. Scribe Therapeutics is pursuing casX based therapies.
4.3: Csm6 boosted Cas13 RNA detection CRISPR diagnostics can detect nucleic acids in one-pot isothermal reactions. This makes them attractive for at-home or point-of-care diagnostics. However, poor sensitivity meant pre-amplification was required to detect SARS-COV2 in patient samples. Class III CRISPR systems include a cyclic-oligo-A activated nuclease, Csm6. We hypothesized that linking cas13 and csm6 using cas13 targets that release Csm6 activator upon cleavage would improve sensitivity. We used kinetic modeling and spike in assays to show that secondary activator cleavage was limiting sensitivity. Using an activator resistant to secondary cleavage allowed detection of SARS-COV2 in clinical samples. My role in this work was kinetic modeling and writing the data analysis and statistical pipelines. This work has improved time to detection and sensitivity in CRISPR diagnostics.

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.