Research Statement

Zvi Rosen September 30, 2022

My research applies **algebraic geometry** to questions in **mathematical biology** and other scientific fields. The goal is to isolate algebraic and combinatorial structure in applied problems and leverage that structure to solve those problems and also discover new mathematical insights. My research has primarily focused on three realms, which I will describe in the remainder of this statement:

1. Combinatorial Neural Codes,
2. Sample Frequency Spectra in Population Genetics,
3. Algebraic Matroids in Applications.

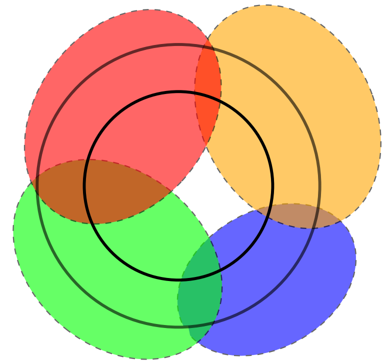
In all three areas, focusing on combinatorial or algebraic aspects of the problem gives insight into the problem in general. In neural codes, we discard most of the data from a spike train and focus only on which subsets of neurons fire together. In population genetics, we study the SFS, which detects fluctuations in the population's size over the course of its history. Instead of looking at all possible demographic histories, we focus on the values that the SFS can take when the population size history is a step function with n steps. Finally, my explorations in algebraic matroids introduce a combinatorial perspective to problems that are often treated non-combinatorially.

1 Combinatorial Neural Codes

My research on neural codes studies how firing patterns of neurons in the brain relate to spatial geometry. In the 1970's, John O'Keefe and colleagues connected electrodes to the brain of a freely moving rat to measure the neural activity in different areas of its hippocampus. They discovered something remarkable: there were neurons (later called *place cells*) that lit up precisely when the rat entered a certain region (called a *place field*) of its cage. As it paced around the cage, different sensors were ignited corresponding to each region; it was like a map inside of the rat's brain. This research won O'Keefe part of the 2016 Nobel Prize in Medicine.

If the brain maintains its internal map using signals from the place cells, these firing patterns must contain some geometric information. This leads naturally to the question: what aspects of the geometry of the space can we reconstruct from the firing patterns alone? For example, consider a rat inhabiting a circular track as in the figure below. Suppose that the four ovals represent place fields, regions where a specific neuron fires. The first reading may report the signal 1000; i.e. only the first neuron is firing. Then as the rat explores, we read 1100 indicating that *both* neurons 1 and 2 are firing. As the rat continues around the track, the signals $\{1000, 1100, 0100, 0110, 0010, 0011, 0001, 1001\}$ are recorded. This list of codewords can be interpreted as faces of a simplicial complex on four vertices, whose homotopy type is a loop, much like the track.

Realizability of Codes



The neural firing patterns, often recorded as a set of binary strings, are referred to as *neural codes*. An arrangement of open sets $\{U_1, \ldots, U_n\}, U_i \subseteq \mathbb{R}^d$ is called a *realization* of a code $\mathcal{C}$ if a d-dimensional rat would produce precisely the set of signals in $\mathcal{C}$ as it explored every region. If the collection of U_i is required to have property P, then the arrangement is called a P-realization, e.g. a convex realization, a connected realization, etc. A code is said to be *realizable* or *P-realizable* if it has a P-realization. My research has tackled the following question:

Problem 1. What *combinatorial* properties of a code $\mathcal{C}$ correspond to P-realizability of $\mathcal{C}$?

Results from combinatorics, convex geometry, and algebraic topology can then be leveraged to obtain properties of these neural codes. Here is a sample of the results we have proved in three geometric regimes: First, in the one-dimensional regime, Yan Zhang and I demonstrated that convex-realizable codes are a modified form of "consecutive-ones matrices" from the combinatorics literature.

Theorem 1.1. *([21, Proposition 2.1.2]) $\mathcal{C}$ is realizable in $\mathbb{R}^1$ as an arrangement of open intervals if and only if $\mathcal{C}$ is the column set of an "harmonious consecutive-ones matrix."*

In that paper, we also prove that this property can be checked in $O(n+k)$ time, by extending the PQ-tree algorithm from [3] for standard consecutive-ones matrices.

In a collaboration initiated by Dr. Carina Curto at an AMS Math Research Community, my co-authors and I provided a necessary condition for convex realizability. Our results focused on the simplicial complex of the code $\Delta(\mathcal{C})$, whose facets are the maximal codewords. We defined *local obstructions* which rest on applications of the Nerve Lemma to links of the complex.

Theorem 1.2. *([5, Theorem 1.3]) For each simplicial complex Δ, there is a unique minimal code $\mathcal{C}_{\min}(\Delta)$ with the following properties:*

1. *the simplicial complex of $\mathcal{C}_{\min}(\Delta)$ is Δ, and*
2. *for any code $\mathcal{C}$ with simplicial complex Δ, $\mathcal{C}$ has no local obstructions if and only if $\mathcal{C} \supseteq \mathcal{C}_{\min}(\Delta)$.*

In other words, given a set of maximal codewords, there is a corresponding set of "make-or-break" codewords that need to be present; if they are present, only non-local obstructions may exist.

Finally, with Vladimir Itskov and Alex Kunin in [10], I examined the consequences of constraining our place fields to be half-spaces whose boundaries are generic hyperplanes. We defined $\Gamma(\mathcal{C})$, another simplicial complex associated to a neural code; we proved the following necessary condition on $\Gamma(\mathcal{C})$ for the code to be hyperplane.

Theorem 1.3. *([10, Theorem 4]) A neural code $\mathcal{C}$ is a generic hyperplane code only if $\Gamma(\mathcal{C})$ is shellable.*

These results all contributed to a young but fast-growing literature using topology, combinatorics, and convex geometry to prove realizability or non-realizability of various classes of neural codes.

Codes, Rings, and Oriented Matroids

Another approach to neural codes has been through the lens of commutative algebra. The idea is to interpret the binary codewords as points in $\mathbb{F}_2^n$ and then analyze the coordinate ring associated to that reducible variety; this approach was first taken in the article [7].

The neural ring appeared as a useful tool in our paper [5]: we applied Hochster's formula relating link topology and the Betti numbers of a free resolution to detect candidates for local obstructions to convexity. Then, in a second collaboration with the MRC group [6], we proved several results establishing properties of the neural ring and its defining ideal as signatures of convexity and non-convexity.

Given the assignment of a ring to a neural code, a natural goal was to define a category of codes so that this map to rings would be functorial. Amzi Jeffs first defined morphisms of neural codes in [11], establishing the category **Code** and proved that the neural ring construction is a functor from the category **Code** $\to$ **Ring**.

In my recent preprint [15], joint with Kunin and Lienkaemper, we introduced a functor from the category **OM** of oriented matroids with strong maps to **Code**. The oriented matroid ideal $O_{\mathcal{M}}$, introduced by [17], then arises naturally as the polarization of the neural ideal of the image of $\mathcal{M}$ in **Code**. We also used the connection to oriented matroids to prove that the question of convex realizability is NP-hard. In particular, we proved the following:

Theorem 1.4. *([15, Theorem 4]) Let $\mathcal{M} = (E, \mathcal{L})$ be a uniform oriented matroid. Then $\mathcal{M}$ is representable if and only if the code $\mathcal{L}^{\pm}(\mathcal{M}) \subseteq 2^{\pm E}$ is convex.*

This result implied as a corollary that the problem of determining convexity of a neural code is NP-hard. Indeed, Mnëv-Sturmfels universality implies that the problem is $\exists\mathbb{R}$-complete.

This project raised many questions that we are pursuing in subsequent research:

1. Given the connection between topes of oriented matroids and codewords in the neural code, are there any useful generalization of other matroid features– e.g. chirotope, circuits, and convex closure function?
2. Oriented matroids live in proximity to a menagerie of combinatorial objects: affine oriented matroids, COMs, lopsided sets, and convex geometries to name a few. Can these objects yield similarly fruitful associations in the neural code realm?
3. Another result we proved in [15] (Theorem 2) is that every neural code realizable by convex polytopes is the image of an oriented matroid under a surjective morphism. Is this sufficient to prove the result for convex realizations more generally?

2 Population Genetics

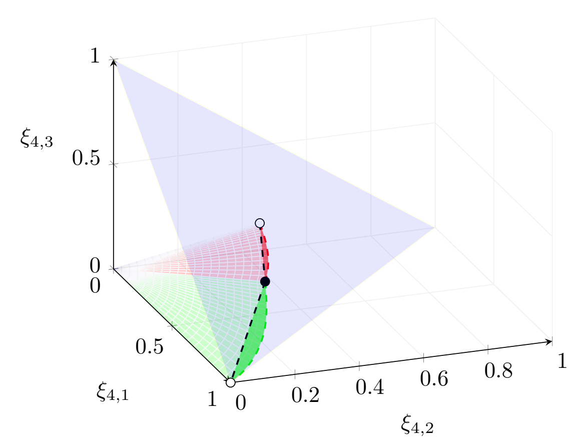
Suppose five individuals are selected at random from a population of amoebas, and their distinct genomes sequenced. We posit that at some point in the distant past, there is a common ancestor shared by all five individuals. Given that amoebas reproduce asexually, one might ask: why do these five individuals have different DNA sequences? Shouldn't they just be copies of the ancestor, and thus identical to one another? The answer is that DNA replication, like copying text by hand, is an error-prone process. Most of the DNA gets copied from one generation to the next, but mutations lead to variations in each offspring. The pairwise differences indicate how much time has elapsed since the most recent common ancestor. In particular, if amoeba 1 and amoeba 2 have very different sets of mutations, then their most recent common ancestor (MRCA) was probably a long time ago. If they share most mutations, their MRCA may have been more recent. With this perspective, mutation statistics can describe aspects of the genealogy underlying our sample.

We now introduce population size into the equation. If mutation statistics describe the genealogy, that genealogy is (in part) determined by the size of the population over time. Looking backwards in time, we can view two branches of the family tree joining up (*coalescence*) as a random process. This process occurs at a rate inversely proportional to the size of the population: if children choose parents at random, they are more likely to choose the same parent when the population is small. A central question studied in population genetics is: can we use mutation statistics to learn the population size history producing those mutations? Or as posed in the title of a 2014 article [13], "can one hear the shape of a population history?"

In our joint project with Bhaskar and Song [18], we did not focus on the preimage of each SFS to infer population history; instead, we considered the full set of SFS generated by all possible histories, addressing the following problem:

Problem 2. Describe the image of the map from demographies, or population size histories, to sample frequency spectra.

Our approach was based on extrapolating from the **piecewise-constant** function with at most k pieces. For example, the plot at right shows the expected SFS vectors of length 4 when limiting the population size to two values. Under this restriction, the mapping from the finite-dimensional parameter space of population size functions to the set of SFS vectors is polynomial and nicely structured. We were able to relate the set of SFS to both: (a) the secant variety of a rational curve, and (b) the cone over a Hadamard product of rational normal curves. By applying a dimension count to the secant variety, we obtained a lower bound, and by a modified Carathéodory argument we obtained an upper bound, leading to the following theorem.



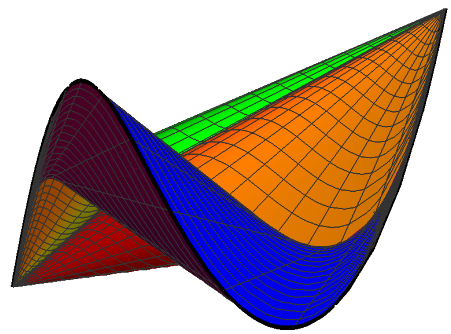
Theorem 2.1. *([18, Theorem 4.3]) Given a fixed sample size n, the selection-neutral sample frequency spectrum for any Lebesgue-measurable demography can be obtained from a piecewise-constant demography with at most $\kappa(n)$ epochs, where*

$$\left\lceil \frac{n}{2} \right\rceil \leq \kappa(n) \leq 2n-1.$$

The SFS as a Projection of a Moment Vector

In my recent publication [19], joint with Georgy Scholten and Cynthia Vinzant, we demonstrated that the entire problem could be reframed as a variation of the classical moment problem in statistics. We used tools from convex algebraic geometry to prove that $\kappa(n)$ referenced above is in fact $n-2$. We also characterized the image as a projection of a spectrahedron. This property enabled us to implement a semidefinite program to find the closest point in the image to some fixed vector outside. If real-life SFS data does not lie in the set of expected SFS, we can thus compute a semidefinite program to find the nearest expected SFS to our data.

For step functions with fewer than $\kappa(n)$ pieces, the set of projected moment vectors is nonconvex and complicated. The image at right depicts the set of moments $\{0, 2, 5, 9\}$ for functions with three constant pieces. Most singular curves on the boundary correspond to lines in parameter space. However, there are additional "accidental" curves where two of the boundary surfaces intersect. In future work, we aim to characterize the relationship between the facial structure of parameter space and the geometry of the image.



In addition to following up on this recent work, I am pursuing additional projects studying population genetics and the SFS. In particular, Bhaskar and Song [2] established injectivity of the SFS under bounded sign complexity. With collaborators, we plan to generalize these results to more complicated models: in particular, 1) isolated populations with common ancestors 2) populations that were isolated but experienced migration events, and 3) populations with selective sweeps. Population genetics, as a branch of science, has had limited interaction with algebraic geometry, so there are many avenues to take.

3 Algebraic Matroids in Applications

Consider a set of elements $E = \{x_1, \ldots, x_n\}$ in an extension of fields $k \subset K$. A subset $S = \{x_{a_1}, \ldots, x_{a_s}\} \subseteq E$ is called *algebraically dependent* if the homomorphism from $k[T_1, \ldots, T_s] \to K$ sending $T_i \mapsto x_{a_i}$ has nontrivial kernel (i.e. S has a nontrivial polynomial relation with coefficients in k.) The algebraically *in*-dependent sets defined as sets that are not dependent, and are denoted by $\mathcal{I}$, satisfy the matroid axioms. Matroids are combinatorial structures designed to serve as the common generalization of linear independence, algebraic independence, and acyclicity in graphs. When a matroid originates in the setting of algebraic dependence in a field extension, we call it an **algebraic matroid**.

Algebraic matroids arise naturally in any application involving polynomial relations. Here's a typical example: suppose I am given only a subset of the entries of an $m \times n$ matrix, and then asked to determine if the matrix might have rank-one.

$$\begin{pmatrix} 1 & \cdot & -3/2 & \cdot & \cdot \\ 7 & 4 & \cdot & \cdot & 10 \\ \cdot & \cdot & -2 & 1 & \cdot \end{pmatrix}$$

This is a question about an algebraic matroid. In particular, the set of rank-one matrices with entries in $\mathbb{C}$ is an irreducible algebraic variety, defined by the prime ideal $\mathcal{I}_2$ of 2×2 minor determinants of a generic matrix. The entries of the matrix $E = \{x_{11}, \ldots, x_{35}\}$ live in the function field of that variety; therefore, they define an algebraic matroid. We want to know: Is this subset of entries an independent set of the matroid? It turns out that it is, since this matroid is isomorphic to the graphic matroid on $K_{3,5}$. Entries of the matrix corresponding to acyclic sets in $K_{3,5}$ are independent in the algebraic matroid as well.

This theme inspired the expository article [20] with coauthors Sidman and Theran, which won the MAA's Merten M. Hasse Prize. I have also explored various algebraic matroids arising in specific applications; I describe some of these below.

Chemical Reaction Networks

The algebraic theory of chemical reaction networks (CRNs) is an active and dynamic field of study. In high school chemistry, we wrote down chemical reactions in the following manner:

$$2\mathrm{H} + \mathrm{O} \to \mathrm{H_2O}.$$

Inside a cell, according to mass-action kinetics, the probability of this reaction taking place is proportional to the product of reactants' concentrations. After all, if there are lots of H's and O's bouncing around, they are more likely to react. This leads to a differential equation as follows:

$$\frac{\mathrm{d}[\mathrm{H_2O}]}{\mathrm{d}t} = k[\mathrm{H}]^2[\mathrm{O}],$$

where the brackets indicate the concentration of each chemical. On the other hand, when the reaction happens, the concentrations of reactants decrease, leading to the equations

$$\frac{\mathrm{d}[\mathrm{H}]}{\mathrm{d}t} = -2k[\mathrm{H}]^2[\mathrm{O}], \quad \text{and } \frac{\mathrm{d}[\mathrm{O}]}{\mathrm{d}t} = -k[\mathrm{H}]^2[\mathrm{O}],$$

since two H's are lost for each O.

When there are many more chemicals floating around with numerous options for reactions to occur, we obtain a system of polynomial ODE's similar to those above governing the concentration of each chemical. As an algebraist, I prefer to wait until the reactions settle down to equilibrium so that time-derivatives are zero. What remains is a system of polynomial equations which can be studied using algebraic geometry. The recent proof of the global attractor conjecture by Craciun [4] demonstrates the power of sophisticated algebraic techniques to prove results about CRN dynamics. The majority of my work in this area focuses on the following problem:

Problem 3. What properties of the steady-state solutions can be inferred from the combinatorial and algebraic properties of the ODE's?

In the paper [16], we studied the Wnt signalling pathway which is related to the development of cancer in the cell. We used algebraic matroids to distinguish between competing models for the Wnt pathway, explaining what experiments might be performed to help assess the fit of each model. In a follow-up paper [9], we delved deeper into the underlying geometry of the set of steady-states for the pathway. Assuming generic parameters and allowing the conserved quantities to vary freely leads to a steady-state variety of dimension 5. Structured polynomial elimination and degree computation in `Macaulay2` [8] and `Bertini` [1] led to the identification of 416 chemical subsets of size 5 such that the conservation relations translate to a polynomial system with mixed volume 9 – the true number of expected solutions for the full system. This result allows us to select chemicals for measurement to make computations as robust as possible to parameter error.

Matrix and Tensor Completion

Consider the joint probability density function of k independent discrete random variables; this corresponds to a tensor $T_{i_1,\ldots,i_k} = P(X_1 = i_1, \ldots, X_k = i_k)$ with entries in $[0,1]$ adding up to one. By independence, the tensor will factor as the outer product of the vectors $P(X_j = i_j)_{j\in[n_j]}$; therefore, it will have rank one. Our work was motivated by the completion problem: given a subset of the tensor entries, is it possible to fill in the remaining probabilities so that the tensor has rank 1 and has entries within the desired bounds? This is a special instance of the following problem:

Problem 4. Given a rank constraint and specified linear constraints, what is the algebraic matroid defined on the entries of a tensor? How do semi-algebraic constraints affect completability of bases of the matroid?

In my first paper on the topic with Kaie Kubjas, we obtained a nice description for the semi-algebraic constraints in terms of the combinatorics of the set of entries. As mentioned above, we can consider matrix entries $\{M_{ij}\}$ as edges in a bipartite graph with row vertices and column vertices; then sets of entries are like subgraphs of $K_{m,n}$.

Theorem 3.1. *([14, Theorem 4]) The projection of all independence matrices onto a forest with n trees has boundary given by coordinate hyperplanes and a hypersurface $\sum_{i=1}^{n} \sqrt{b_i} = 1$, where each b_i is a rational function in the entries of a single connected component.*

The situation for tensors is predictably more complicated. Still, the case of diagonal partial tensors (where all coordinates of all given entries are distinct) is relatively simple, and we proved the following result in the same paper:

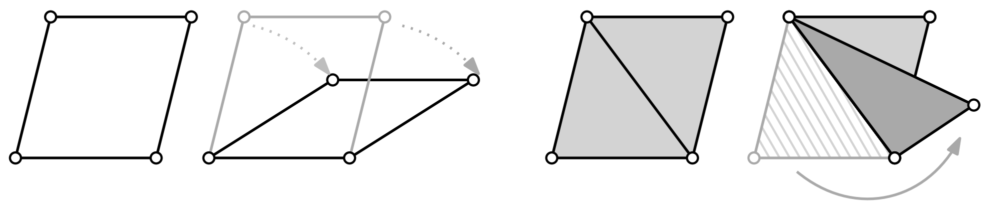
Theorem 3.2. *([14, Theorem 7]) Let $a_1, \ldots, a_k$ be the entries in the diagonal of an order-d tensor. Then the diagonal partial tensor is completable to a rank-1 probability tensor if and only if*

$$\sum_{i=1}^{k} a_i^{1/d} \leq 1.$$

In follow-up work with Kahle and Kummer [12], we answered the question for general partial tensors and gave polynomial inequalities equivalent to the algebraic inequalities above. In future work, we will analyze more general variants of this problem: matrices and tensors of higher rank, and with more general linear constraints.

Rigidity Theory

A bar-and-joint framework is like a physical manifestation of a graph: a collection of vertex-like joints connected by linear bars. Rigidity theory studies which graphs lead to frameworks in d-dimensional space that do not change shape when pressure is applied. For example, a 4-cycle is flexible in the plane: you can squash a parallelogram without bending any of the bars. On the other hand, K_4 minus an edge is rigid in the plane; however, even this graph is flexible in 3-space, where you can fold it like a hinge.



A reframing of this question makes the algebraic matroid apparent. Let $p_1, \ldots, p_n \in \mathbb{R}^d$ be a collection of n points; these define a vector of pairwise distances $d_{ij} = ||p_i - p_j||^2$. This can be considered as a map from $\mathbb{R}^{dn} \to \mathbb{R}^{\binom{n}{2}}$ whose image is an irreducible algebraic variety cut out by the Cayley-Menger ideal. The spanning sets of the algebraic matroid are precisely the rigid graphs on n vertices.

In forthcoming work with Sidman, Theran and Vinzant, we study the pure condition, a polynomial in the edge lengths of a rigid graph which describes conditions under which the framework degenerates. We combine symbolic and numerical computation to obtain explicit polynomials in edge lengths, where previously only bracket polynomials were available. In another upcoming joint project with Urizar, we translate the algebraic matroid perspective on rigidity from the set of edges to the set of *angles* in a framework.

References

[1] Daniel J Bates, Jonathan D Hauenstein, Andrew J Sommese, and Charles W Wampler. Bertini: Software for numerical algebraic geometry (2006). *Software available at http://bertini. nd. edu*.

[2] Anand Bhaskar and Yun S Song. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Annals of Statistics*, 42(6):2469, 2014.

[3] Kellogg S. Booth and George S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms. *Journal of Computer and System Sciences*, 13(3):335–379, 1976.

[4] Gheorghe Craciun. Toric differential inclusions and a proof of the global attractor conjecture. *arXiv preprint arXiv:1501.02860*, 2015.

[5] Carina Curto, Elizabeth Gross, Jack Jeffries, Katherine Morrison, Mohamed Omar, **Zvi Rosen**, Anne Shiu, and Nora Youngs. What makes a neural code convex? *SIAM Journal on Applied Algebra and Geometry*, 1(1):222–238, 2017.

[6] Carina Curto, Elizabeth Gross, Jack Jeffries, Katherine Morrison, **Zvi Rosen**, Anne Shiu, and Nora

Youngs. Algebraic signatures of convex and non-convex codes. *Journal of pure and applied algebra*, 223(9):3919–3940, 2019.

[7] Carina Curto, Vladimir Itskov, Alan Veliz-Cuba, and Nora Youngs. The neural ring: an algebraic tool for analyzing the intrinsic structure of neural codes. *Bulletin of mathematical biology*, 75(9):1571–1611, 2013.

[8] Daniel R Grayson and Michael E Stillman. Macaulay 2, a software system for research in algebraic geometry, 2002.

[9] Elizabeth Gross, Heather A Harrington, **Zvi Rosen**, and Bernd Sturmfels. Algebraic systems biology: a case study for the Wnt pathway. *Bulletin of Mathematical Biology*, 78(1):21–51, 2016.

[10] Vladimir Itskov, Alexander Kunin, and **Zvi Rosen**. Hyperplane neural codes and the polar complex. In *Topological Data Analysis*, pages 343–369. Springer, 2020.

[11] R Amzi Jeffs. Morphisms of neural codes. *SIAM Journal on Applied Algebra and Geometry*, 4(1):99–122, 2020.

[12] Thomas Kahle, Kaie Kubjas, Mario Kummer, and **Zvi Rosen**. The geometry of rank-one tensor completion. *SIAM Journal on Applied Algebra and Geometry*, 1(1):200–221, 2017.

[13] Junhyong Kim, Elchanan Mossel, Miklós Z Rácz, and Nathan Ross. Can one hear the shape of a population history? *Theoretical population biology*, 100:26–38, 2015.

[14] Kaie Kubjas and **Zvi Rosen**. Matrix completion for the independence model. *Journal of Algebraic Statistics*, 8(1), 2017.

[15] Alexander Kunin, Caitlin Lienkaemper, and **Zvi Rosen**. Oriented matroids and combinatorial neural codes. *arXiv preprint arXiv:2002.03542*, 2020.

[16] Adam L MacLean, **Zvi Rosen**, Helen M Byrne, and Heather A Harrington. Parameter-free methods distinguish wnt pathway models and guide design of experiments. *Proceedings of the National Academy of Sciences*, 112(9):2652–2657, 2015.

[17] Isabella Novik, Alexander Postnikov, and Bernd Sturmfels. Syzygies of oriented matroids. *Duke Mathematical Journal*, 111(2):287–317, 2002.

[18] **Zvi Rosen**, Anand Bhaskar, Sebastien Roch, and Yun S Song. Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*, 210(2):665–682, 2018.

[19] **Zvi Rosen**, Georgy Scholten, and Cynthia Vinzant. Sparse moments of univariate step functions and allele frequency spectra. *Vietnam Journal of Mathematics*, 2021. To appear in Special Issue celebrating Bernd Sturmfels' 60th birthday.

[20] **Zvi Rosen**, Jessica Sidman, and Louis Theran. Algebraic matroids in action. *The American Mathematical Monthly*, 127(3):199–216, 2020.

[21] **Zvi Rosen** and Yan X Zhang. Convex neural codes in dimension 1. *arXiv preprint arXiv:1702.06907*, 2017.