COMPUTATIONAL ANALYSIS OF OLFACTION AND ARTIFICIAL NOSE TECHNOLOGIES

By

JULIA TSITRON

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

written under the direction of

Alexandre V. Morozov

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2013

ABSTRACT OF THE DISSERTATION

Computational Analysis of Olfaction and Artificial Nose Technologies

By JULIA TSITRON

Dissertation Director:

Alexandre V. Morozov

Cross-selective receptor arrays coupled with higher-order neural processing can be seen in naturally occurring olfactory systems, where a large number and variety of analytes can be detected, distinguished, even quantified using relatively few receptors and clever combinatorial odor decoding. This strategy has been imitated in artificial sensor array systems that are paired with computational signal-processing tools in diverse applications that range from vintage wine year discrimination to disease diagnosis. However, the complexity of receptor response patterns to even a single analyte, coupled with non-linearity of response to mixtures of analytes, makes quantitative inference of individual compound concentrations within mixtures a challenging task. In this work, I show how output from two distinct types of sensor arrays, each combined with Bayesian analysis, can be used to predict component concentrations in complex mixtures. In the first case, the array consists of four engineered G-protein-coupled receptors used for deciphering mixtures of highly related sugar nucleotides. We employ a biophysical model that explicitly takes receptor-ligand interactions into account in order to quantify mixture constituents. Furthermore, we develop a universal metric of receptor array performance, and use it to study the fundamental limits imposed on the accuracy of ligand recognition by the physics and biology of

receptor-ligand interactions. This provides design guidelines for sensor arrays optimized for mixture analysis. Antagonistic receptor response, well-known to play an important role in biological systems, proves to be essential for precise recognition of mixture components. The second array consists of a mixed-potential electrochemical sensor operating under different applied bias currents to monitor gas mixtures in diesel engine exhaust. Here, the sensitivity and selectivity of a device is tuned by current application, thus a single sensor serves as an entire array when operated under multiple conditions. Both a linear and non-linear model are used to quantify ammonia gas in the presence of propylene interference. While more data-intensive, the nonlinear model captures cross-interference between analytes and yields more accurate predictions. Our Bayesian methodology is easily generalized to other 'artificial nose' applications by the inclusion of additional models.

Acknowledgments

First and foremost, I would like to thank my advisor, Alexandre V. Morozov, for giving me the opportunity to work on some very cool science under his expert leadership. Alex is a highly effective teacher - knowledgable, passionate, always accessible, and with a keen ability to tap into each of his student's strengths. This combination is every graduate student's dream for a mentor, and I am extremely grateful to have been a part of his group.

I would like to acknowledge our collaborators, each of whom made an invaluable contribution to the work presented here. Addison (Tad) Ault designed and performed the experiments described in Chapter 2 in the laboratory of Professor James R. Broach. I am grateful for the enthusiastic conversations we had surrounding this project and for their willingness to discuss biology with physicists.

The work presented in Chapter 4 was done in collaboration with Eric L. Brosha, Cortney R. Kreller, Praveen K. Sekhar, Rangachary Mukundan, and Fernando H. Garzon. Particularly, I would like to acknowledge Eric and Cortney for their patience and persistence as we went back and forth from their data to our models and even asked them to repeat some experiments under different conditions. I would also like to thank Professor Broach and Tad Ault for introducing the work of the Electrochemical Sensors and Devices Group at Los Alamos National Laboratory (LANL) to Alex and me and for sparking this collaboration.

I would like to thank Eduardo Sontag for his generous support during my last year at Rutgers.

A great, big thank you to my mom for allowing me to go back to work with a clear mind and a clear heart so soon after M was born. I had no doubt that he was in good hands and without that, completion of this thesis would have been impossible.

Finally, I need to mention the friends and family who have supported me throughout. Particularly, my brother Lev for being such a great uncle and babysitter. Eliane and Alice

for inspiring me, each in their own way, and for keeping me sane. Allan Haldane, who is not only a reliable source for fruitful scientific discussions and technical prowess (and advice) but who also became a good friend during our shared time at Rutgers. And Dim, whose love, support, and friendship helps me in everything I do.

Dedication

To my family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Mammalian and insect olfactory systems are capable of recognizing tens of thousands of odors – mostly organic compounds with diverse chemical structures and properties [1–4]. The olfactory tasks commonly faced by such systems include detecting odors, estimating their strength, identifying their source, and recognizing one specific odor in the background of another [5]. The sense of smell exhibits amazing sensitivity and discriminatory power, distinguishing between closely related compounds and detecting vanishingly small odorant concentrations [6]. Olfactory signaling is mediated by a superfamily of several hundred G-protein-coupled receptors (GPCRs) – a significant fraction of the total number of genes in many higher eukaryotes [7–11]. In mammals, GPCRs are located on the surfaces of the cilia projected from olfactory receptor neurons (ORNs); typically receptors of only one type are expressed in a given neuron [12] (Figure 1.1a). Odor recognition is combinatorial, with one odorant activating multiple receptors and one receptor responding to multiple odorants (Figure 1.1b) [12–16]. The resulting complex patterns of receptor activation enable robust identification of many more odors than would have been possible with "lock and key" receptors reacting to only one analyte. Moreover, several studies provide evidence for widespread inhibitory responses in which receptors are antagonized by odorants [14, 15, 17–20].

The idea of combinatorial recognition has been adapted to artificial arrays in which multiple sensors with partially overlapping selectivities respond to a given analyte [22–25]. While the output of these cross-specific arrays in response to single compounds can generally be interpreted through pattern recognition algorithms [25–28], computational analysis becomes more difficult when the array is presented with a mixture of compounds. Indeed,

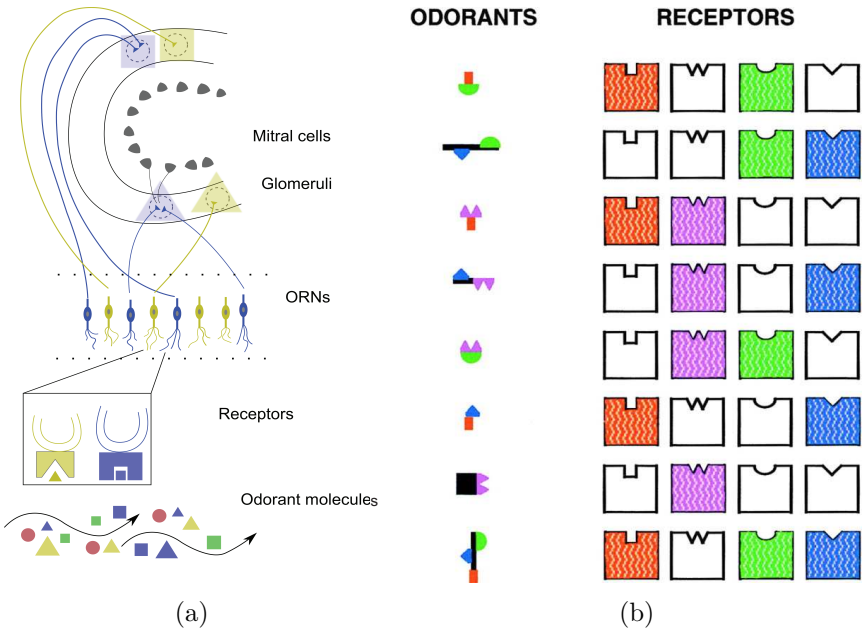


Figure 1.1: **Overview of olfactory system and the combinatorial scheme for coding odors.** (a) Schematic of the olfactory epithelium and olfactory bulb (adapted from [21]). Typically, one receptor is expressed in each ORN. These receptors bind odorant molecules as the first step of signal transduction to higher brain areas. (b) Adapted from [12]. An odorant can activate multiple receptors and each receptor is capable of responding to multiple odorants. In this scheme, each receptor contributes a component of a combinatorial code for discriminating a large number and variety of odorant molecules.

the non-linear nature of sensor responses to multiple ligands makes it hard to train discriminatory algorithms on a "typical" subset of patterns. The non-linear dependence of sensor output on ligand concentrations is generic in reporter systems and may be compounded by potential binding interference of the two ligands, saturation of the sensor output [29] and, of particular concern, potential antagonistic action of one ligand on another's activity [19]. As a result, responses to complex mixtures have primarily been used to "fingerprint" specific mixtures rather than identify their constituents quantitatively [30–33]. There are relatively few studies which focus on the quantitative analysis of mixtures: for example, Heilig et al. used a single sensor and Fourier transformation techniques to analyze a binary mixture of CO and NO_2 [34], White et al. trained artificial neural networks to identify relative concentrations in binary mixtures [35], and Woodka et al. used a non-negative least squares method to quantify the composition of analyte mixtures with up to five components [36].

In Chapter 2, we present a physical model of receptor-ligand recognition that explicitly relates observed response patterns to component concentrations and receptor properties,

making it easier to quantify mixture constituents. We use Bayesian inference to predict absolute concentrations of each ligand in arbitrary mixtures of uridine diphosphate (UDP) sugar nucleotides applied to a combinatorial array of four GPCRs. We introduce RANSA (Receptor Array Nested Sampling Algorithm), the software developed for calibration, inference, and optimization of sensor arrays that was used throughout the work described in this thesis. RANSA is freely available at http://olfaction.rutgers.edu.

In Chapter 3 we derive design principles that enable accurate mixture discrimination with cross-specific GPCR-based sensor arrays. We find that optimal sensor parameters exhibit relatively weak dependence on component concentrations, making a single designed array useful for analyzing a sizable range of mixtures and that the maximum number of mixture components that can be successfully discriminated is twice the number of receptors in the array. Our *in silico* studies of sensor array design also reveal that antagonistic responses, in which the receptor is bound by the ligand but there is no downstream reporter activity, are necessary for precise recognition of mixture components. This conclusion provides a rationale for the widespread inhibitory responses observed in natural olfactory systems. The work described in Chapters 2 and 3 was done in collaboration with James R. Broach and Addison (Tad) D. Ault and was published in [37].

In Chapter 4 we turn our attention to solid-state electrochemical devices. Au/YSZ/Pt (Au and Pt electrodes and Yttria-stabilized Zirconia electrolyte) sensors are a promising technology for monitoring levels of nitrogen oxides and ammonia in diesel engine exhaust. However, in addition to the target gases these sensors react to unburned hydrocarbons present in the gas mixture. The observed crossinterference between target and non-target gases cannot be fully mitigated by applying different bias currents to the sensor. On the other hand, sensor sensitivity and selectivity towards various components of the mixture depend on the bias current setting, allowing us to effectively create an array of sensors by applying dfferent bias currents to the same device. We show how such an array can be used to predict absolute concentrations of ammonia in the presence of propylene. The procedure consists of two steps: the calibration step, in which the parameters of the model are determined *a priori* in the laboratory setting, and the prediction step, which mimics the deployment of the device in real-world conditions. We investigate a linear model, in

which response of the sensor to each gas is assumed to be additive, and a nonlinear model, which takes interference between gases into account. We find that the nonlinear model, although more complex, yields more accurate predictions. We also find that relatively few sensor readings and bias current settings are required to make reliable predictions of gas concentrations in the mixture, making our approach feasible in a variety of automotive and other technological settings. In summary, we show that our methodology can be employed with both biological systems and synthetic receptor arrays ("artificial noses") designed for various industrial needs. The work described in Chapter 4 was performed in collaboration with the Electrochemical Sensors and Devices Group at Los Alamos National Laboratory and is presented in the following manuscript: J. Tsitron, C.R. Kreller, P.K. Sekhar, R. Mukundan, F.H. Garzon, E.L. Brosha, and A.V. Morozov, "Bayesian decoding of the ammonia response of a zirconia-based mixed-potential sensor in the presence of hydrocarbon interference," which is currently *in review*.

Chapter 2

Decoding output of GPCR-based sensor arrays

2.1 Biological implementation of the sensor array

Our sensor array is comprised of four engineered receptors (L-3, H-20, K-3 and 2211) with distinct but overlapping specificities for four types of nucleotide sugars: UDP-glucose (UDP-Glc), UDP-galactose (UDP-Gal), UDP-glucosamine (UDP-GlcNAc) and UDP. The receptors were evolved *in vitro* from the human UDP-glucose receptor using directed mutagenesis of the residues involved in ligand binding (see § 2.1.3) [38]. Nucleotide sugars and their derivatives are key constituents in polysaccharide synthesis and other cellular processes. Their structural similarity makes them a challenging target for array-based discriminatory analysis. To assess receptor-ligand interactions quantitatively in our sensor array, we functionally expressed the receptors in *S.cerevisiae*. To do so, we replaced the yeast pheromone receptor with one of the sensor GPCRs in strains in which the pheromone response pathway was modified to respond to the heterologous receptor by inducing transcription of the *E. coli* lacZ gene [39]. In this fashion, the extent of GPCR activation following ligand addition could be directly monitored as the level of β-galactosidase produced in the cell, which we measured using a fluorescence-based assay (§ 2.1.4). Applying a mixture of nucleotide sugars to the receptor array yields a complex pattern of responses of the four receptor-bearing strains. The response of each receptor depends on the concentration of all components in the mixture, on the receptor-ligand binding affinities, and on the efficacy with which each ligand activates the receptor. Nonetheless, the contents of arbitrary nucleotide sugar mixtures can be deciphered using array readout as input to a physical model of receptor-ligand interactions.

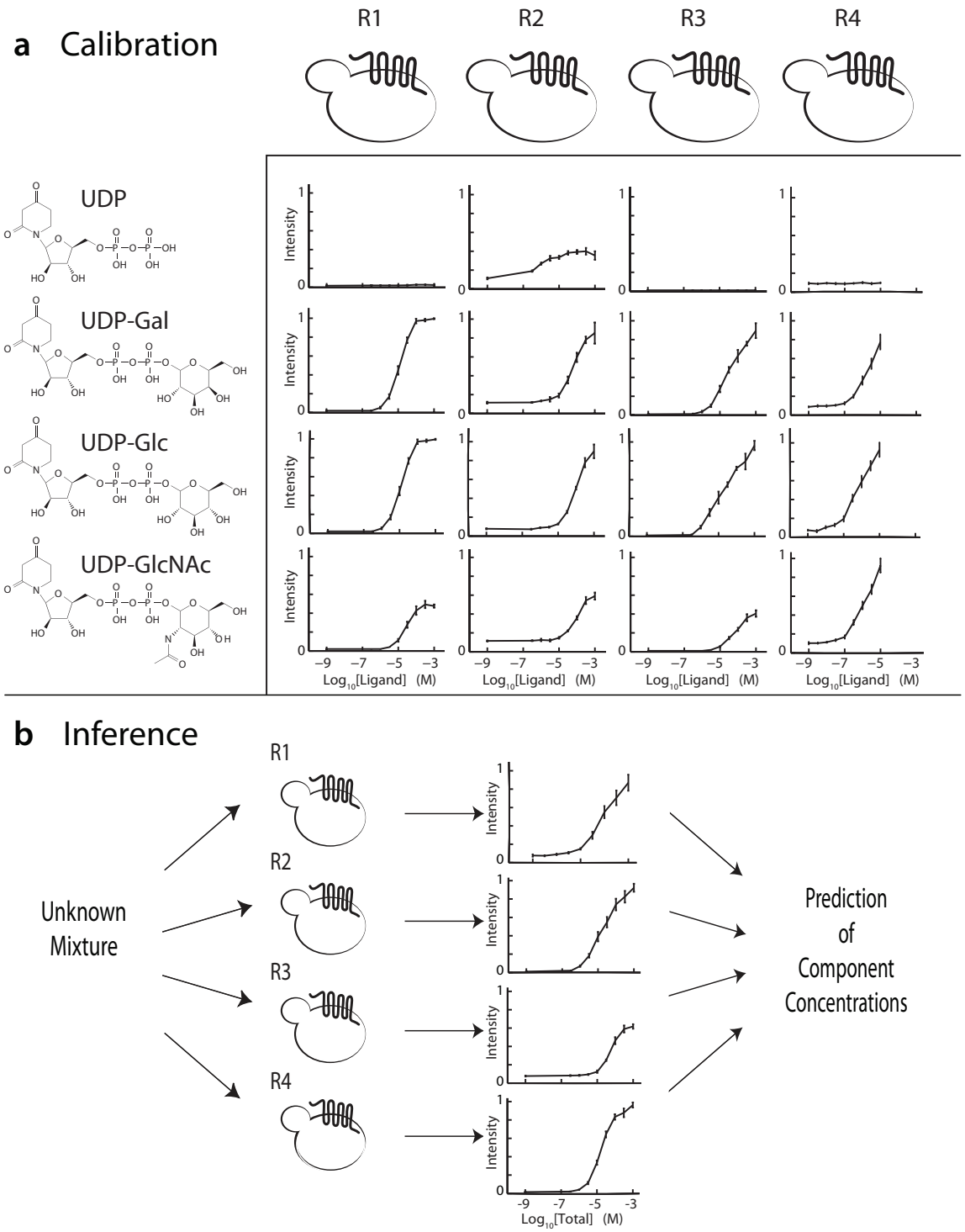


Figure 2.1: **Overview of the GPCR-based biosensor array.** (**a**) Each receptor-ligand combination is tested for functional activation, yielding 16 binding curves. For each curve, intensity normalized by the maximum intensity on the plate is plotted against $\log_{10} n$ (n is the ligand concentration in M). The error bars on each curve are from four biological replicates. The single-receptor, single-ligand binding curves are used to calibrate the physical model by inferring ΔG, A and b separately for each receptor-ligand combination. (**b**) An unknown mixture of four ligands is applied to each of the four receptors. The resulting fluorescent response curves together with the $\{\Delta G, A, b\}$ predictions are used as input to the Bayesian algorithm designed to predict absolute concentrations of each ligand in the mixture. R1: H-20, R2: K-3, R3: L-3, R4: 2211.

2.1.1 Reagents.

UDP-glucose (UDP-Glc), UDP-galactose (UDP-Gal), UDP-N-acetylglucosamine (UDP-GlcNAc), uridine diphosphate (UDP) and fluorescin (FDG) were purchased from Sigma-Aldrich (St. Louis, MO).

2.1.2 Strains and plasmids.

Mutagenesis and selection were performed in yeast strain CY10560 (P_{FUS1}-$HIS3$ $ade2\Delta3447$ $ade8\Delta3457$ $can1$-100 $far1\Delta1442$ $his3\Delta200$ $leu2$-$3,112$ $lys2$ $sst2\Delta1056$ $ste14$ :: $trp1$ :: $LYS2$ $ste18\gamma6$-3841 $ste3\Delta1156$ $trp1$-1 $ura3$-52). β-galactosidase assays were performed using yeast strain CY10981 ($PFUS1$-$HIS3$ $can1$-100 $far1\Delta1442$ $his3\Delta200$ $leu2$-$3,112$ $lys2$ $sst2\Delta2$ $ste14$ :: $trp1$ :: $LYS2$ $ste3\Delta1156$ $trp1$-1 $ura3$-52) carrying plasmid Cp1021 (P_{FUS1}-$LacZ$ $2\mu m$ $URA3$). The UDP-glucose receptor and 2211, H-20 and K-3 mutants were cloned as previously described [38].

2.1.3 Targeted mutagenesis and selection of functional receptor mutants

The L-3 mutant was isolated using a procedure similar to that previously employed with the H-20 and K-3 mutants [38]: oligonucleotides with randomized sequences corresponding to the codons to be mutagenized were utilized to generate overlapping PCR products. The L-3 motif corresponds to amino acid residues LLxSA on TM7. Mutant libraries were generated by gap repair using overlapping PCR products and transformed to media selective for recombined plasmids. To select for functional mutants, libraries were replica-plated to selective SC-His media [40] containing one of six ligands: UDP-Gal, UDP-Glc, UDP-galNAc, UDP-GlcNAc, UDP or dTDP-glucose (50 μL of 1 mM solution spread on 30 mL of SC–Leu-His agar medium in 8.5 cm Petri plates). Yeast growth media was supplemented by 1 mM 3AT, a competitive inhibitor of the *HIS3* reporter gene product, which sets the threshold for reporter gene activation. Functional receptor mutants that showed qualitatively disparate responses to the panel of ligands were selected for further analysis. Among these, the H-20 and K-3 mutants, described earlier [38], and the L-3 mutant, described here, were selected to be utilized alongside the 2211 “parent” in a four-receptor array for analysis of mixtures

of UDP-Glc, UDP-Gal, UDP-GlcNAc and UDP.

2.1.4 β-galactosidase assays.

Our β-galactosidase assays were based on microtiter assays described previously [41]. Yeast strains expressing each of the four mutant receptors were diluted to OD_{600} of ~ 0.05 in flasks. Cultures were then grown overnight in 100 mL selective media to an OD_{600} of $0.1 - 0.2$. Serial dilutions of each ligand or mixture of ligands were prepared in yeast culture medium in 96-well culture blocks. Ligands or mixtures of ligands were transferred in 20 μL aliquots in quadruplicate to deep-well polypropylene 384-well plates using a BioMek robotic liquid handler. 180 μL of suspended yeast cells in medium (undiluted from the overnight cultures) were then aliquotted into each well and mixed. The cultures were sealed with foil tape and incubated at 30°C on a plate shaker at 400-500 rpm for 4 hours (H-20, K-3 and L-3 receptors) or overnight (2211 receptor). After incubation, β-galactosidase substrate [41] (FDG solution; 0.5 mM fluorescein di-beta-D-galactopyranoside, 2.3% Triton X-100, and 0.127 M Pipes, pH 7.2) was mixed with an equal volume of Pierce Y-PER solution (Thermo Scientific) and distributed in 25 μL aliquots to black 384-well plates. 50 μL aliquots of the yeast/ligand cultures were then transferred into the black 384-well plates and mixed gently but thoroughly by pipetting, taking care to avoid generating bubbles. A single layer of paper towel was placed on top of each plate and the plates were then individually wrapped in aluminum foil and incubated without shaking at 37°C for approximately one hour before reading on an automated fluorescent plate reader (Perkin Elmer EnVision).

Microtiter plate-based assays are often subject to edge- or plate-bias due to uneven heating or discrepancies in timing across a single plate or among plates. While no obvious plate effects were seen, it is very difficult to control for all possible variations in a single experiment. Due to the number of samples and the need to make efficient use of materials, each of the mixture experiments was split across two plates per receptor. In mixtures of equal proportions, samples containing UDP, UDP-Gal and UDP-GlcNAc but lacking UDP-Glc were on Plate 1, while all mixtures containing UDP-Glc were on Plate 2. In the UDP-Gal/UDP-Glc binary mixtures of unequal proportions, samples containing 90%, 80% or 60% UDP-Glc were on Plate 1, while samples containing 40%, 20% or 10% UDP-Glc

were on Plate 2.

For each single ligand or combination of ligands, a series of measurements was performed at several values of the total concentration $n^l = \sum_{i=1}^{N_{lig}} n_i^l$ (M): $\log_{10} n^l = \{-3.0, -3.5, \ldots, -6.5, -9.0\}$ for H-20, K-3, L-3 and $\{-5.0, -5.5, \ldots, -9.0\}$ for 2211. The total chemical potential $\mu^l = k_B T \log n^l$ is then given by $\mu^l = \mu - \sum_{i=1}^{l-1} \Delta\mu^i \quad (l = 1 \ldots N)$, where N is the number of measurements in the series, $\Delta\mu^l$ are known chemical potential differences between two consecutive measurements, and $\mu = -3k_B T \log(10)$ is the chemical potential at the 1 mM reference point. Note that in order to reconstruct the total chemical potential for all points in the series, only μ needs to be predicted. Each series of measurements was replicated four times; fluorescence counts were normalized to 1.0 separately for each plate (Dataset.xls).

2.2 Physical model of the sensor array

2.2.1 Single-receptor, single-ligand model

We start with the simplest case in which a receptor interacts with a single ligand. We assume that the observed signal in our receptor-bearing reporter strain is proportional to the probability that the receptor is bound by the ligand. This proportionality value, A, which we refer to as the receptor efficacy, can range from 1, for a full agonist, to 0, for a full antagonist. Thus, for a single receptor interacting with a single ligand, the amount of activation of the reporter in the receptor-bearing strain is given by eq. (2.1). Reporter activation measurements as a function of single ligand concentration are shown in Figure 2.1a We use these data to estimate the parameters of eq. (2.1) (Figure 2.2a) and the amount of experimental noise $\tilde{\sigma}$ for each single-receptor, single-ligand combination using Bayesian inference with nested sampling [42] – a Bayesian Monte-Carlo (MC) technique that yields an ensemble of models from which the average value of each parameter and its standard deviation can be computed. Unlike other methods such as MC sampling of the product of likelihood and priors, nested sampling allows us to keep track of the evidence, yielding absolute values of the posterior probability. See Appendix A for details. The most likely values of the parameters (Table B.1) are then used for subsequent evaluation of mixtures of

compounds. The accuracy of parameter predictions depends on the range of concentrations available for these calibration experiments (Figure B.1) and on the amount of experimental noise (Figure B.2)

2.2.2 Multiple-receptor, multiple-ligand model

Once all receptor-ligand interaction parameters have been determined through the analysis of single-ligand calibration experiments, we can proceed to interrogating mixtures of ligands with receptor arrays. In considering the response of receptor-bearing strains to ligand mixtures, we note that each ligand contributes to the overall receptor occupancy and that each receptor molecule on the cell surface activates the reporter with an efficacy specified by the ligand to which it is bound, which is often different for different ligands (Table B.1). Assuming that all ligands bind competitively to the same site on the receptor, we model the response of the receptor-bearing strain to mixtures of compounds by calculating the total intensity as a sum of fractional occupancies of the receptor by each ligand weighted by the corresponding efficacies (eq. (2.6)). We treat each of the receptor-bearing strains with an unknown mixture, sequentially diluted to provide a series of samples across a million-fold range of concentrations (Figure 2.1b) We carry out Bayesian inference for the entire receptor array, predicting the total concentration of all ligands and the concentration ratios of ligand pairs (Figure 2.2b). From these values we can deduce the absolute concentration of each ligand in the mixture.

2.3 Mathematical details of receptor-ligand binding models; Bayesian formalism

2.3.1 Single-receptor, single-ligand

For a single receptor interacting with a single ligand, we model the normalized reporter fluorescent intensity as:

$$I^l(\Delta G, A, b) = A\frac{e^{-\beta(\Delta G-\mu^l)}}{1+e^{-\beta(\Delta G-\mu^l)}} + b, \tag{2.1}$$

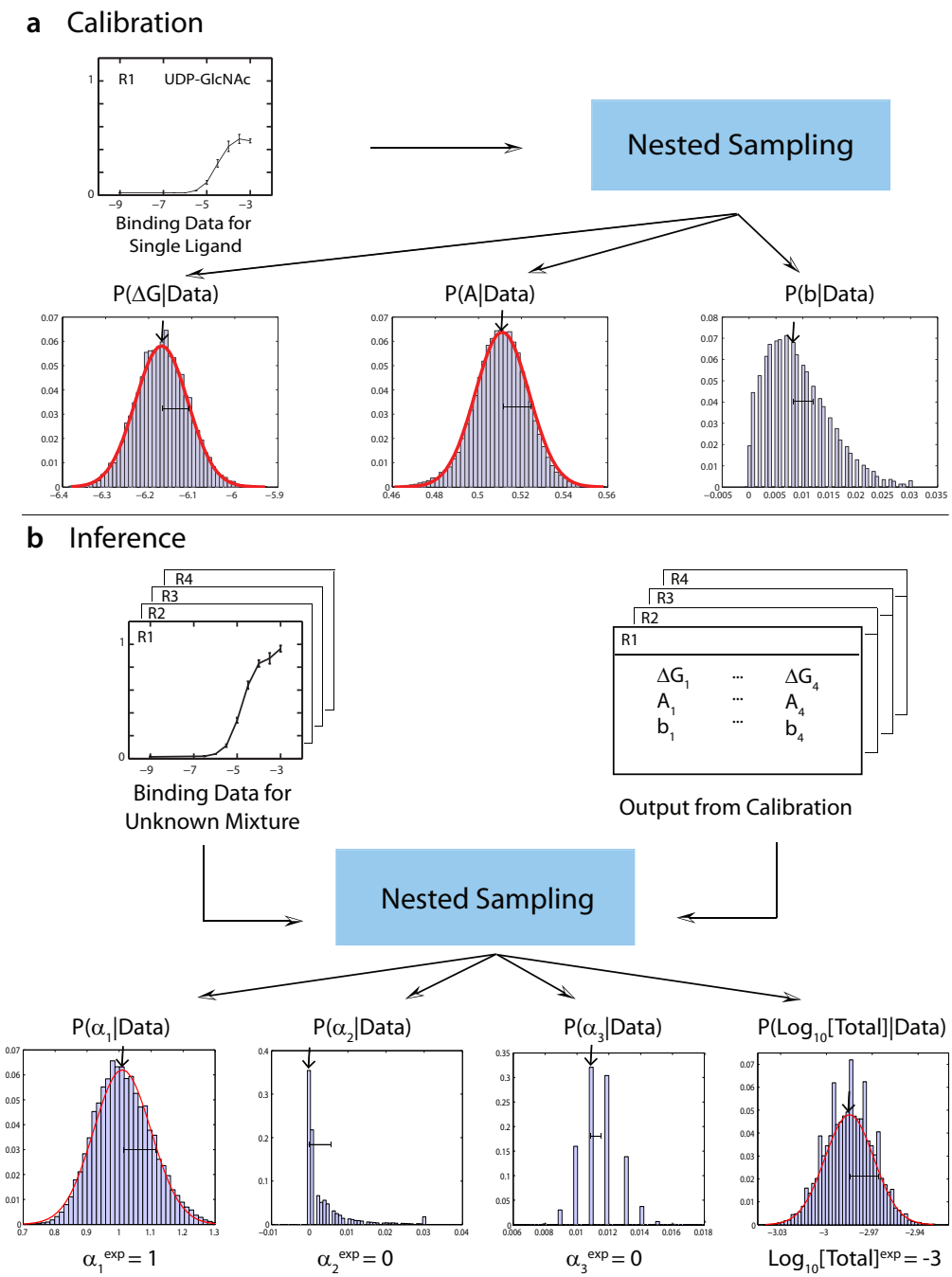


Figure 2.2: **Bayesian algorithm for predicting ligand concentrations in mixtures.** (**a**) Calibration of the algorithm: single-ligand, single-receptor binding curves are used to infer binding free energy ΔG, efficacy A and background intensity b for every receptor-ligand combination. Histograms for each predicted parameter are based on an ensemble of 50000 models sampled by Metropolis Monte Carlo [43] starting from the log-likelighood maximum found by nested sampling [42]. Arrows and error bars indicate the most likely value of each parameter and its standard deviation. (**b**) Inference of ligand concentrations in an unknown mixture. Model parameters from (**a**) together with the response curves for all receptors serve as input to the nested sampling algorithm which predicts relative concentrations α_i for each component (with respect to one arbitrarily chosen component, cf. eq. (2.7)) and the total concentration $\log_{10} n$ of all ligands in the mixture. Together these predictions yield absolute concentrations for each constituent ligand. Histograms, arrows and error bars have the same meaning as in (**a**), and experimental values are shown below each panel ([Total] = 1 mM at the reference point). For each binding curve, intensity normalized by the maximum intensity on the plate is plotted against $\log_{10} n$ (n is the total ligand concentration in M).

where A is the receptor efficacy, b is the background intensity (a small amount of background fluorescence observed in the absence of ligand binding), ΔG is the free energy of receptor-ligand binding, $\beta = 1/k_B T$ (k_B is the Boltzmann constant, and T is the temperature), and μ^l is the chemical potential.

We compute the log-likelihood of the data by assuming that fluorescence measurements are Gaussian-distributed around values from eq. (2.1):

$$\mathcal{L} = \log P(\{\tilde{I}\}|\Delta G, A, b, \tilde{\sigma}) = -\frac{1}{2\tilde{\sigma}^2}\sum_{l=1}^{N}\left[I^l(\Delta G, A, b) - \tilde{I}^l\right]^2 - \frac{N}{2}\log(2\pi\tilde{\sigma}^2), \tag{2.2}$$

where $\tilde{I}^l$ ($l = 1 \ldots N$) are measured intensities and $\tilde{\sigma}$ is the noise parameter. The log-likelihood is used to estimate the posterior probability of all model parameters according to the Bayes' formula [42]:

$$P(\Delta G, A, b, \tilde{\sigma}|\{\tilde{I}\}) = \frac{P(\{\tilde{I}\}|\Delta G, A, b, \tilde{\sigma})P(\Delta G)P(A)P(b)P(\tilde{\sigma})}{P(\{\tilde{I}\})}, \tag{2.3}$$

where on the right-hand side the likelihood from eq. (2.2) is multiplied by the product of priors for each model parameter and divided by evidence. $\{\tilde{I}\}$ combines data from all experimental replicates. We use uniform priors (invariant with respect to translations, $x \to x + a$):

$$P(x) = \begin{cases} 1/(x_{max} - x_{min}) & \text{if } x \in [x_{min}, x_{max}], \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

for ΔG, A and b, and Jeffrey's priors (invariant with respect to rescaling, $x \to ax$) for $\tilde{\sigma}$:

$$P(x) = \begin{cases} 1/(x\log(x_{max}/x_{min})) & \text{if } x \in [x_{min}, x_{max}], \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

We have used $(\Delta G_{min}, \Delta G_{max}) = (-20.0, 5.0)$ kcal/mol, $(A_{min}, A_{max}) = (0.0, 1.0)$, $(b_{min}, b_{max}) = (0.0, 1.0)$, $(\tilde{\sigma}_{min}, \tilde{\sigma}_{max}) = (0.001, 100.0)$ in our calculations.

2.3.2 Single-receptor, multiple-ligands

The reporter response to a mixture of ligands is given by

$$I_k^l = \sum_{m=1}^{N_{lig}} A_m^k p_m^{k,l} + b^k, \tag{2.6}$$

where $p_m^{k,l} = e^{-\beta(\Delta G_m^k - \mu_m^l)}/\mathcal{Z}^{k,l}$ is the probability that receptor k is bound by ligand m and $\mathcal{Z}^{k,l} = 1 + \sum_{i=1}^{N_{lig}} e^{-\beta(\Delta G_i^k - \mu_i^l)}$ is the partition function. ΔG_m^k is the binding free energy between receptor k and ligand m ($k = 1 \dots N_{rec}$, $m = 1 \dots N_{lig}$), A_m^k is the efficacy, and b^k is the background intensity. The background intensity for receptor k is the average from all calibration experiments involving that receptor. $\mu_m^l = k_B T \log n_m^l$ is the chemical potential of ligand m, which can be expressed through the total chemical potential μ^l and the relative concentrations $\alpha_m = n_{m+1}^l / n_1^l$ ($\forall l$, $m = 1 \dots N_{lig} - 1$):

$$\begin{aligned} \mu_1^l &= \mu^l + k_B T \log \frac{1}{S}, \\ \mu_m^l &= \mu^l + k_B T \log \frac{\alpha_{m-1}}{S}, \quad m = 2, 3, \dots \end{aligned} \tag{2.7}$$

where $S = 1 + \sum_{i=1}^{N_{lig}-1} \alpha_i$. Note that an arbitrary choice of the ligand in the denominator leads to several equivalent representations of the relative concentrations.

2.3.3 Multiple-receptors, multiple-ligands

The log-likelihood of the observed pattern of fluorescence intensities from multiple receptors interrogated by a mixture of ligands is given by

$$\mathcal{L} = \log P(\{\tilde{I}\}|\{\alpha\}, \mu, \{\tilde{\sigma}\}) = -\sum_{k=1}^{N_{rec}} \left\{ \frac{1}{2\tilde{\sigma}_k^2} \sum_{l=1}^{N_k} [I_k^l(\{\alpha\}, \mu) - \tilde{I}_k^l]^2 + \frac{N_k}{2} \log(2\pi\tilde{\sigma}_k^2) \right\}. \tag{2.8}$$

Here $I_k^l(\{\alpha\}, \mu)$ is defined in eq. (2.6) (in the interests of brevity, we suppress its dependence on $\{A, b, \Delta G\}$ for each receptor-ligand combination). $\tilde{I}_k^l$ denotes fluorescence measured for receptor k at the total chemical potential μ_l, N_k is the total number of measurements, and $\tilde{\sigma}_k$ is the noise parameter. Similarly to eq. (2.3), the log-likelihood is used to estimate the posterior probability $P(\{\alpha\}, \mu, \{\tilde{\sigma}\}|\{\tilde{I}\})$. We employ a uniform prior for μ with $(\mu_{min}, \mu_{max}) = (-10.0, -2.0)$ and a Jeffrey's prior for α's with $(\alpha_{min}, \alpha_{max}) = (0.0001, 100.0)$.

2.4 Tests (and results) of the physical model

We have tested our approach using a series of assays in which a known combination of ligands was applied to the receptor-bearing strains. As an initial test, we mixed equal proportions of two, three and four ligands in all possible combinations and predicted absolute ligand

concentrations. We used a model in which four ligands interacted with four receptors, even if only one, two or three ligands were actually present in the mixture. As can be seen in Figure 2.3 and Table B.2, our approach is generally quite successful in identifying both zero and non-zero ligand concentrations in the mixtures. For example, with single ligands and binary mixtures the correct chemical or pair of chemicals is predicted to have the highest concentrations in all 10 cases. However, the inference is consistently less accurate with UDP-containing mixtures, due in part to larger errors in the predicted total concentration. Thus UDP-related efficacies and binding free energies are less optimal than those of other ligands, as will be demonstrated in detail below.

Our second test involved combining UDP-Glc and UDP-Gal in several unequal proportions and applying the resulting mixture to the four-receptor array (Figure 2.4 and Table B.3). As before, we use a four-ligand model, which should predict zero concentrations for UDP-GlcNAc and UDP. The predicted values of α_1= [UDP-Glc]/[UDP-Gal] show that the ratio of [UDP-Glc] to [UDP-Gal] is successfully ranked in all cases except for the 60/40 and 40/60 mixtures. Apart from the excessive values of α_2= [UDP-GlcNAc]/[UDP-Gal] in the 90/10 and 80/20 cases, which are nonetheless not as large as α_1, concentrations of all ligands absent from the mixture are correctly inferred to be close to zero. We obtain similar results with the alternative definition of α's (α_1= [UDP-Gal]/[UDP-Glc], etc.) (Table B.4), showing that our approach is not overly sensitive to the arbitrary definition of relative concentrations.

Increasing the number of receptors should improve prediction accuracy by providing additional information about the mixture. To see the extent of these improvements, we have used a variable number of receptors to infer component concentrations in six equal-proportion mixtures of two nucleotide sugars from Figure 2.3 (Figures 2.5 and B.3). As expected, the errors rapidly get smaller as the number of receptors is increased, making larger arrays unnecessary. Surprisingly, in several cases adding extra receptors makes the errors somewhat worse before they become better again (see e.g. the R3 and R3/R4 error bars in the UDP + UDP-Gal α_1 panel of Figure B.3, indicating that the noise in the new data outweighs the benefit of additional measurements.

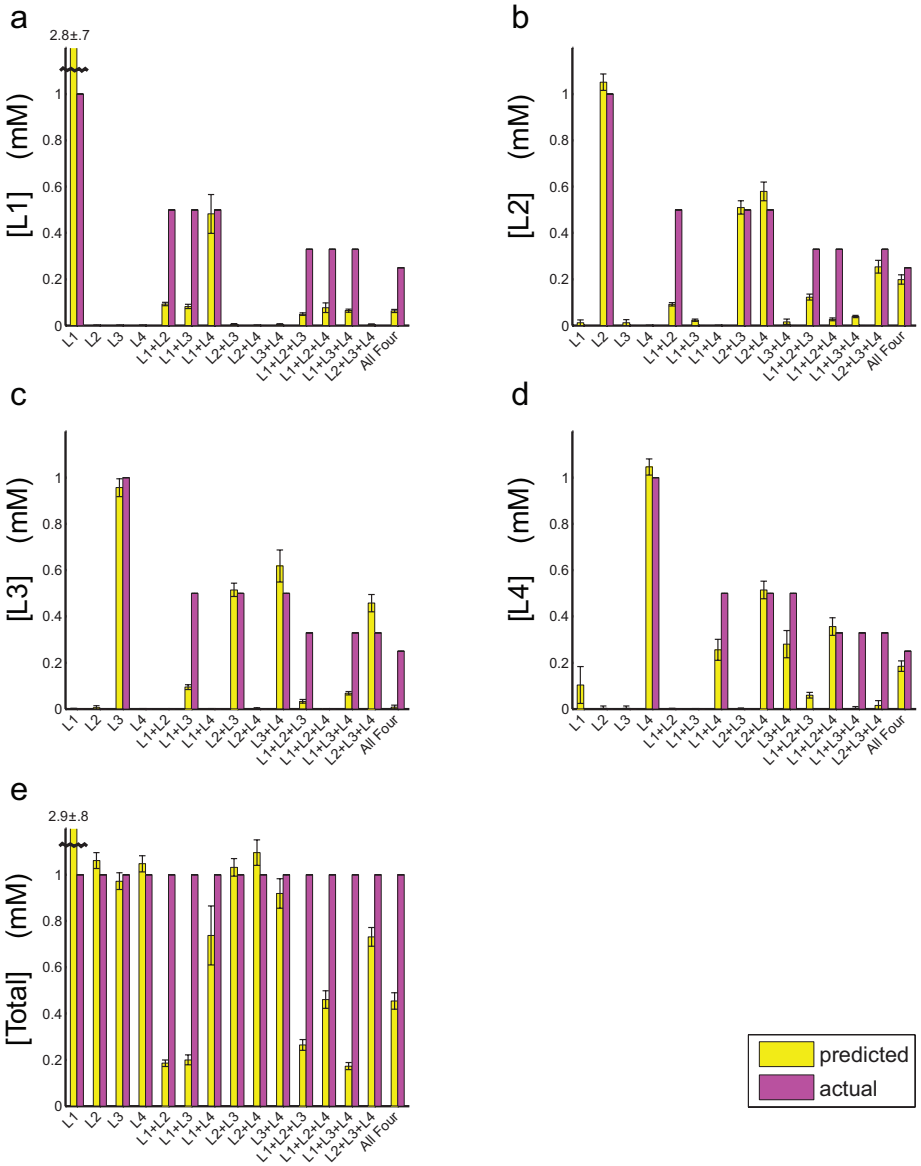


Figure 2.3: **Prediction of ligand concentrations in equal-proportion mixtures.** We used nested sampling of a four-receptor, four-ligand model to estimate means and standard deviations for the relative concentrations of all ligands in the mixture and the total ligand concentration at the 1 mM reference point (see § 2.1.4)These predictions were converted into absolute concentrations (mM) for each ligand at the 1 mM reference point.
L1: UDP, L2: UDP-Gal, L3: UDP-Glc, L4: UDP-GlcNAc.

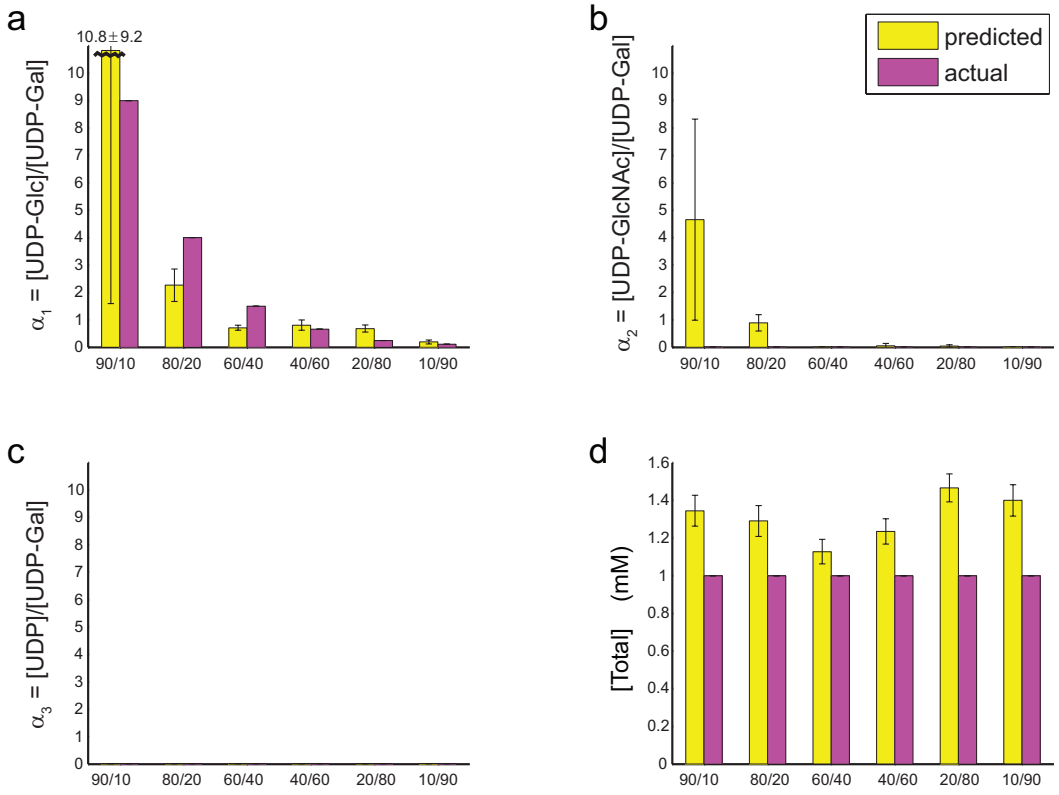


Figure 2.4: **Prediction of ligand concentrations in unequal-proportion binary mixtures of [UDP-Gal] and [UDP-Glc].** We used nested sampling of a four-receptor, four-ligand model to estimate means and standard deviations for the relative concentrations α_1= [UDP-Glc]/[UDP-Gal], α_2= [UDP-GlcNAc]/[UDP-Gal], α_3= [UDP]/[UDP-Gal] and the total concentration (M) [Total] = [UDP-Gal] + [UDP-Glc] + [UDP-GlcNAc] + [UDP] at the 1 mM reference point (see § 2.1.4).We found that our predictions were improved if A's and b's were refit to account for "plate bias" (cf. header of Table B.4): small deviations in the values of A and b (from the standard values shown in Table B.1 and used everywhere else) between measurements 1-3 (Plate 1) and 4-6 (Plate 2).

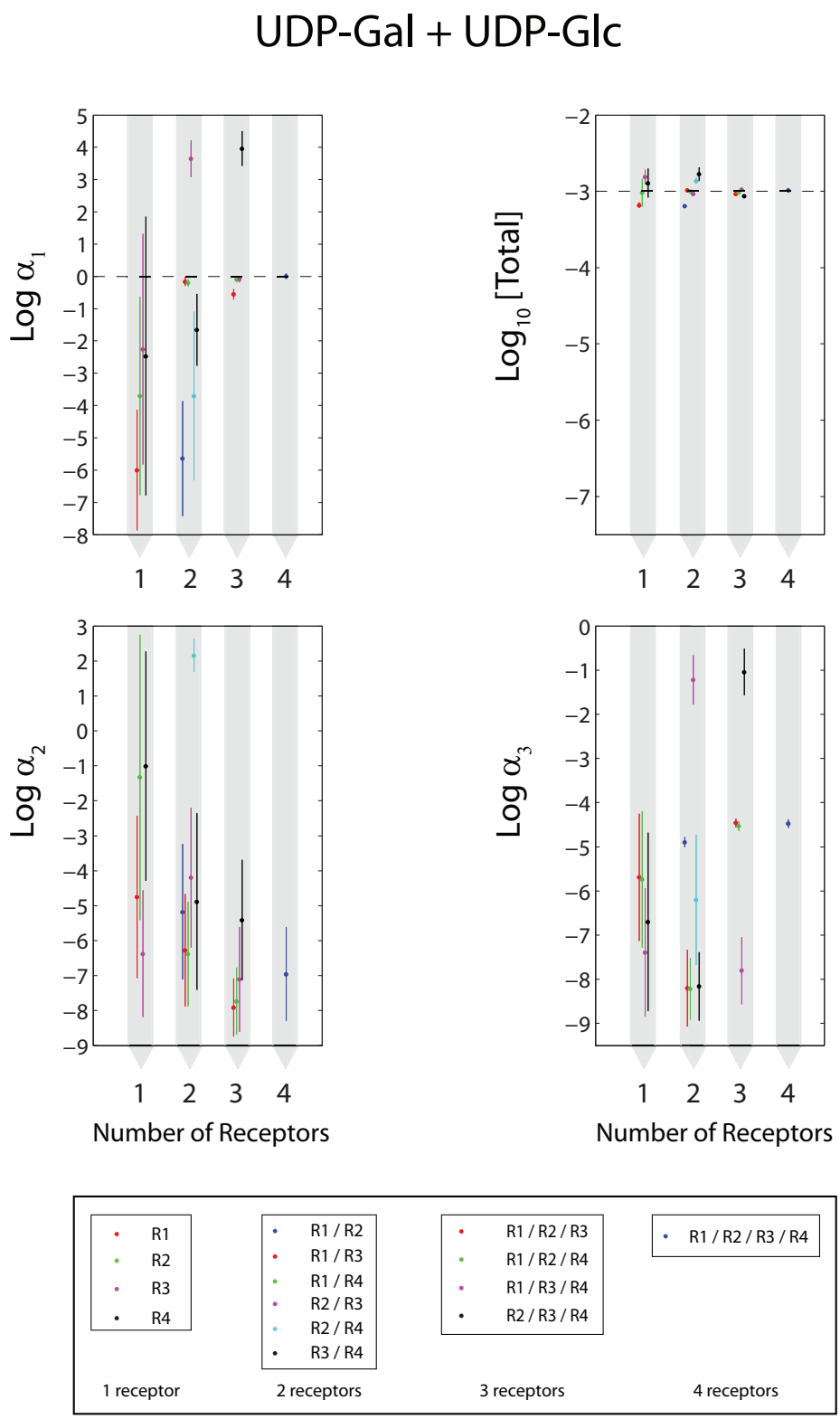


Figure 2.5: **Inference of ligand concentrations is improved with the number of receptors interrogating the mixture.** Shown on the log-scale are means and standard deviations for α_1= [UDP-Glc]/[UDP-Gal], [Total] = [UDP-Glc] + [UDP-GlcNAc] + [UDP-Gal] + [UDP], α_2= [UDP-GlcNAc]/[UDP-Gal], and α_3= [UDP]/[UDP-Gal]. The data are for the 50-50 [UDP-Glc]-[UDP-Gal] binary mixture, leading to $\alpha_1 = 1$, $\alpha_2 = \alpha_3 = 0$, and $\log_{10}$[Total] $= -3$ at the reference point. The means and standard deviations were predicted by RANSA using the four-ligand model and up to four receptors: H-20 (R1), K-3 (R2), L-3 (R3), 2211 (R4).

As evident from the activation profile of each receptor in response to each ligand (Figure 2.1a), the receptors differ from each other in fairly subtle ways. In particular, different ligands do not invoke markedly orthogonal profiles of receptor responses. Nonetheless, even with this suboptimal array design, our algorithm provides accurate identification of ligands present in a mixture and a reasonable assessment of the relative amounts of each.

Chapter 3

Optimization of sensor array performance

3.1 Hessian analysis

Our Bayesian approach estimates posterior probabilities for the concentration of each component in an arbitrary mixture. With sufficient data, variation of the posterior probability with model parameters is determined by the corresponding log-likelihood (eq. (2.8)), which can be visualized as a multidimensional landscape. The global maximum on this landscape corresponds to the model that best describes the data, while the curvature at the maximum shows how sensitive the likelihood is to the change in each parameter. Narrow peaks result in precisely defined parameter values, whereas wide plateaus yield many nearly equivalent predictions and therefore sizable uncertainties in parameter estimates. Expanding the log-likelihood in the vicinity of its maximum yields a Hessian matrix (eq. (C.2)), which contains information about standard deviation σ_i of each model parameter γ_i (eq. (C.4)) [44]. For example, if the observed receptor response does not depend on γ_i, zero entries appear in the Hessian, leading to the infinite uncertainty σ_i. Making all Hessian matrix elements uniformly larger leads to the smaller σ_i for each predicted parameter γ_i.

Hessian analysis relies on the quadratic expansion in the vicinity of the log-likelihood maximum and hence it is important to check how well it captures the behavior of the more general but computationally intensive nested sampling approach. To create a test case for which the answer is known, we have used eq. (2.6) to generate synthetic data for 15 equal-proportion mixtures from Figure 2.3 in the low-noise limit ($\tilde{\sigma} = 0.01$ for all receptors, several times smaller than experimental values from Table B.1). We observe close correspondence between parameter uncertainties inferred from nested sampling vs. Hessian analysis (Figure D.1). Moreover, since larger uncertainties make it easier for the average

values of predicted parameters to be incorrect, there is also correlation between Hessian errors and the absolute differences between mean predicted and true values (Figure D.2). The Hessian-based approach remains useful when experimental data, for which the precise model is unknown and the noise is substantially higher (Table B.1), is analyzed in the same way: the average over 4 correlation coefficients between Hessian errors and standard deviations from nested sampling (computed for α_1, α_2, α_3 and $\log_{10} n$) is 0.85, and the average over 4 correlation coefficients between Hessian errors and absolute differences between predicted and true values is 0.69. In both real and synthetic cases, the Hessian matrix was computed with correct relative and total concentrations and $\{\Delta G, A, b\}$ values from Table B.1. We conclude that Hessian errors are a reasonable measure of sensor array performance.

Not all receptors are equally good candidates for inclusion into biosensor arrays – for example, receptors with similar sets of efficacies and binding affinities should be less useful than receptors with more orthogonal binding and activation patterns. Here we make such qualitative insights precise by developing a Hessian approach to biosensor array design. That is, given a certain number of measurements with an array of fixed size (typically, a series in which the total concentration is changed step-by-step within a certain range), we wish to derive the most optimal choice of receptor properties for deciphering the mixture. From the Hessian point of view, the best array will have the smallest errors in predicting component concentrations (eq. (C.4)). Because each error is inversely proportional to the determinant of the Hessian, we maximize the determinant instead of minimizing the errors directly. Similarly to the prediction of constituent concentrations, the maximization is carried out by nested sampling [42]. In general, the most optimal receptor parameters and their robustness will depend on the relative concentration of each component in a mixture and on the number of measurements made with the array. For example, an array fine-tuned to detect small admixtures of compound B in the background of compound A may function less well if the concentrations of A and B become approximately equal.

3.2 Optimal parameters for single-receptor, two-ligand system

To demonstrate our approach, we first optimize parameters of a single receptor discriminating a mixture of two ligands. By maximizing the determinant of the Hessian, in this

case a 2×2 matrix, as a function of two efficacies and two binding energies, we find that the best discrimination is achieved if one ligand acts as an agonist and the other as an antagonist: $A_1 = 1$ and $A_2 = 0$ *or* $A_1 = 0$ and $A_2 = 1$ (for simplicity, background intensities were set to 0 in all sensor array designs). Although in both cases each ligand binds strongly to the receptor, there is a unique set of optimal binding energies ΔG_1 and ΔG_2 for each agonist-antagonist scenario (Figures 3.1 and 3.5a). The actual values of the binding energies depend on the relative concentration α; for unequal ligand concentrations the two $\{\Delta G_1, \Delta G_2\}$ sets will in general be distinct. This is not surprising since exchanging ligand labels amounts to exchanging relative concentrations of the agonist and the antagonist in the mixture. The height of the peak in both determinant landscapes is the same, indicating that the two alternative solutions lead to equally acceptable array designs as long as the ΔG's are tuned appropriately.

The fine-tuning of binding energies is not necessary if either the total concentration $\log_{10} n$ is known and the task is to minimize the error in predicting the relative concentration α, or vice versa (Figures 3.1 and 3.2). The single-peak landscape structure appears only if the absolute concentrations of both components need to be predicted together. Strikingly, simultaneous prediction of the total and relative concentrations is impossible with the agonist-agonist receptor response (Figure 3.3). See Appendix C.1 for details.

The dependence of the optimal binding energies on the value of α is fairly weak (Figure 3.4). Thus one set of ΔG's optimized for a specific value of α provides a near-optimal solution for a sizable range of ligand concentrations.

3.3 Design of multiple-receptor, multiple-ligand arrays

The agonist-antagonist pattern observed in the one-receptor, two-ligand case plays the role of a basic building block when two or more receptors interact with multiple ligands: nested sampling maximization of the Hessian determinant with respect to binding energies ΔG and efficacies A reveals that the array as a whole performs best if each receptor binds one agonist and one antagonist. For example, in the two-receptor, four-ligand case ($N_{rec} = 2$, $N_{lig} = 4$) receptor 1 strongly binds ligands 1 and 3 with $A_1 \simeq 1$ and $A_3 \simeq 0$, whereas

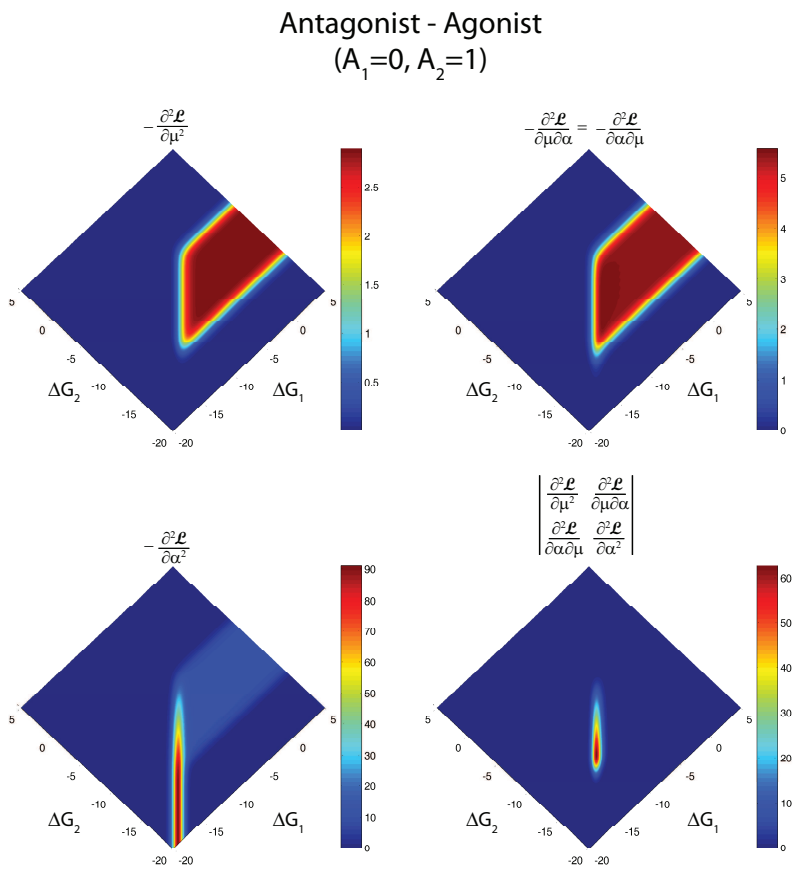


Figure 3.1: **Matrix elements and the determinant of the Hessian in the antagonist-agonist case,** plotted as a function of binding energies ΔG_1 and ΔG_2 in the one-receptor, two-ligand system. The efficacies are fixed at $A_1 = 0$, $A_2 = 1$; $\alpha = 0.25$. We used 7 replicates with $\log_{10} n^l = \{-3.0, -3.5, \ldots, -8.5, -9.0\}$. The values of the binding energies at the peak of the determinant landscape are $\Delta G_1 = -11.05$ kcal/mol and $\Delta G_2 = -12.30$ kcal/mol.

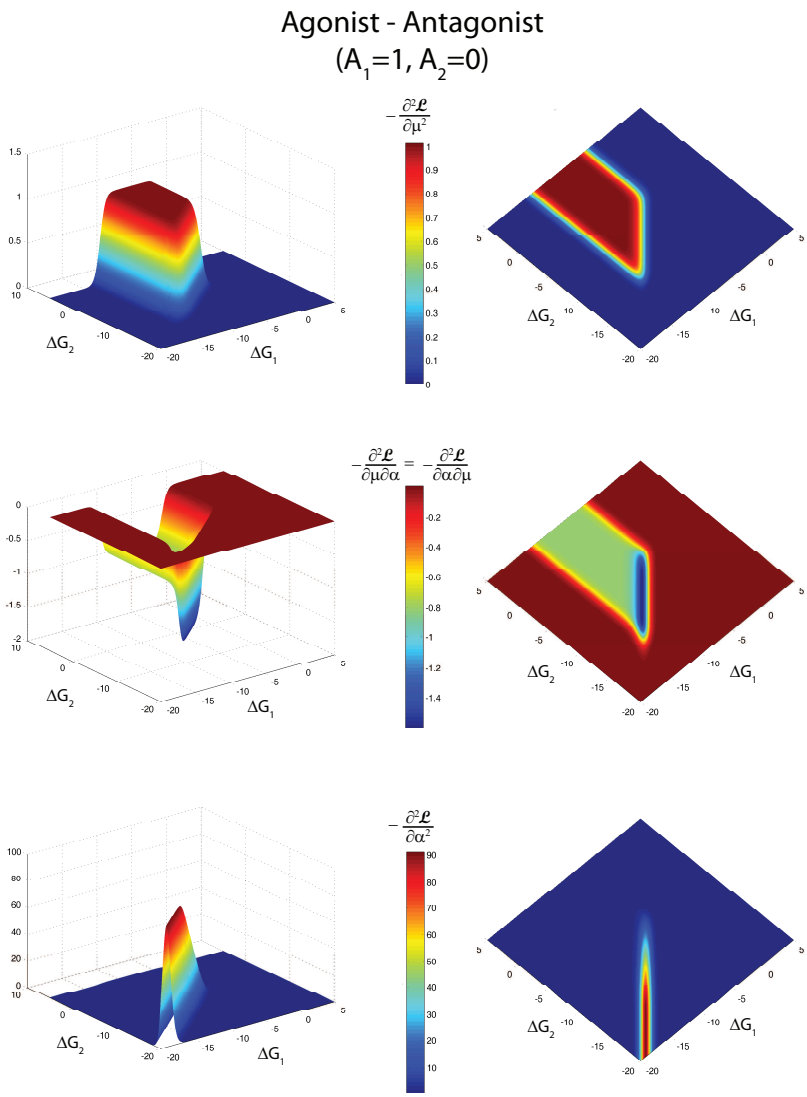


Figure 3.2: **Matrix elements of the Hessian in the agonist-antagonist case,** plotted as a function of binding energies ΔG_1 and ΔG_2 in the one-receptor, two-ligand system. The efficacies are fixed at $A_1 = 1$, $A_2 = 0$; $\alpha = 0.25$. We used 7 replicates with $\log_{10} n^l = \{-3.0, -3.5, \ldots, -8.5, -9.0\}$.

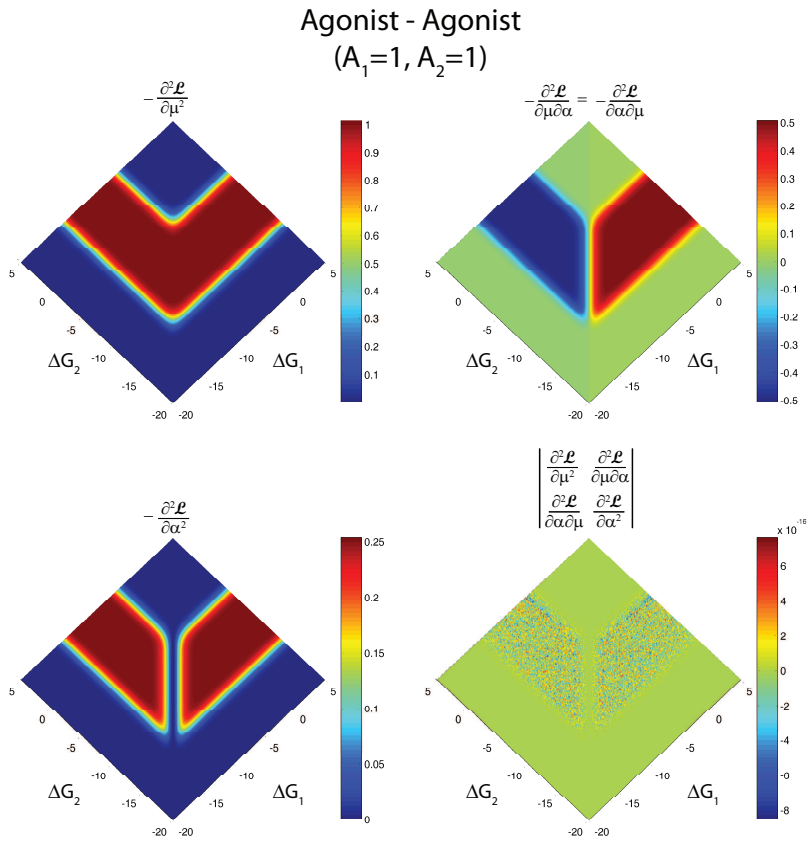


Figure 3.3: **Matrix elements and the determinant of the Hessian in the agonist-agonist case,** plotted as a function of binding energies ΔG_1 and ΔG_2 in the one-receptor, two-ligand system. The efficacies are fixed at $A_1 = 1$, $A_2 = 1$; $\alpha = 1$. We used 7 replicates with $\log_{10} n^l = \{-3.0, -3.5, \ldots, -8.5, -9.0\}$.

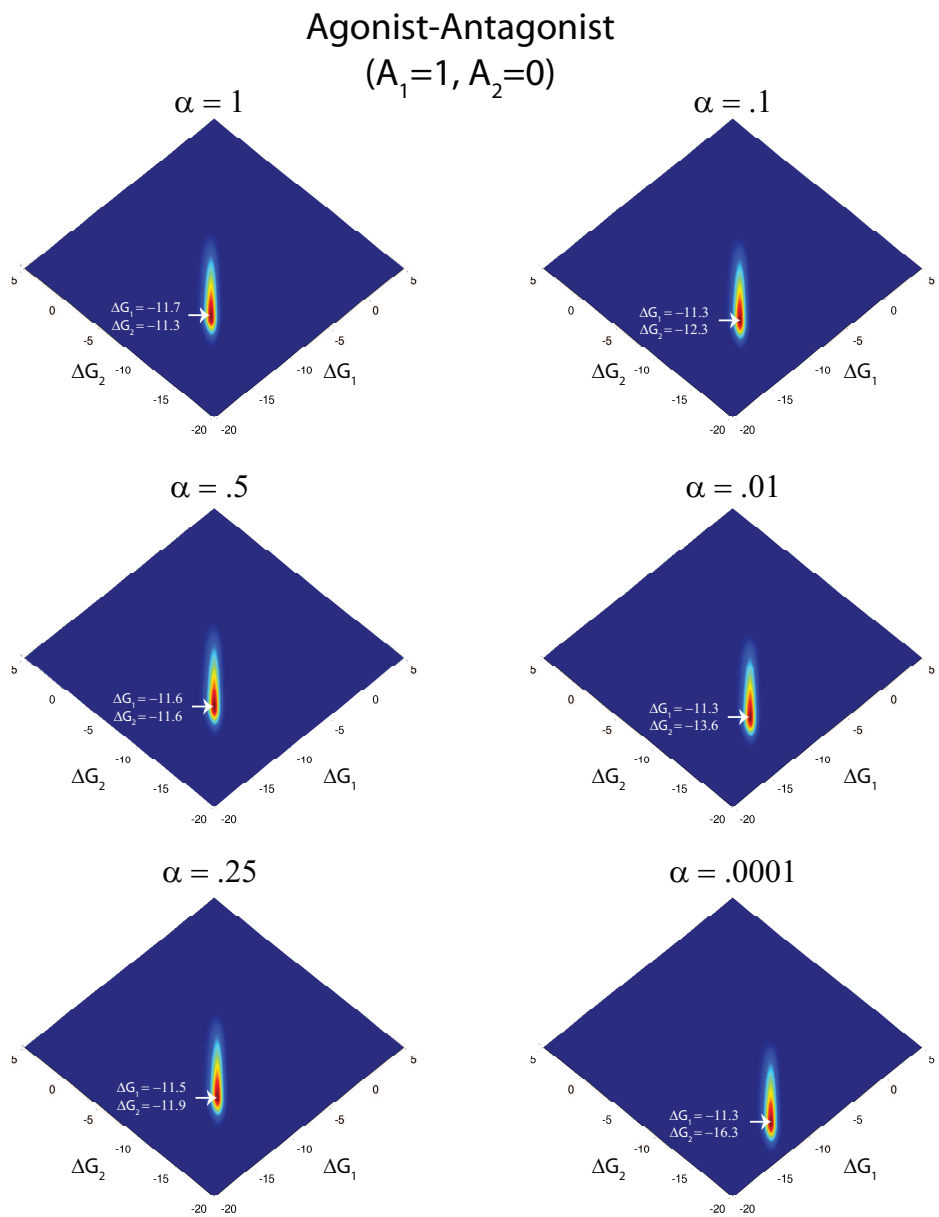


Figure 3.4: **Changes in the Hessian determinant with the concentration of the second ligand.** Each determinant is plotted as a function of binding energies ΔG_1 and ΔG_2 for a given value of $\alpha = n_2/n_1$. The efficacies are fixed at $A_1 = 1$, $A_2 = 0$. We used 7 replicates with $\log_{10} n^l = \{-3.0, -3.5, \ldots, -8.5, -9.0\}$. Shown in each panel are the optimal ΔG_1 and ΔG_2 corresponding to the maximum value of the determinant.

receptor 2 strongly binds ligands 2 and 4 with $A_2 \simeq 0$ and $A_4 \simeq 1$ (Figure 3.5b (*i*)). Each ligand preferentially binds to only one receptor. When another receptor is added to the array, the optimal binding and activation pattern becomes strikingly different: each receptor once again binds both an agonist and an antagonist but ligand 1 now acts as an antagonist to all three receptors (Figure 3.5b (*ii*)). Each of the other three ligands is an agonist to one of the receptors. In the $N_{rec} = 4$, $N_{lig} = 4$ case each ligand is an agonist for one receptor and an antagonist for another (Figure 3.5b (*iii*)). Once again, each receptor binds both an agonist and an antagonist. The determinant of the Hessian is dominated by these agonist-antagonist patterns, and is less sensitive to the changes in efficacies and binding energies that do not affect them.

In light of the observed agonist-antagonist behavior, it is not surprising to see that each receptor can identify concentrations of at most two ligands (Figure 3.5c, blue dots). The uncertainty in predicting components of the mixture is minimized if for every receptor one ligand binds strongly as a full agonist and another as a full antagonist. As we have seen, when receptor parameters are less than optimal, the discrimination is still possible but additional receptors may be required: three or four rather than two in the four-ligand case (Figures 2.5 and B.3). If we eliminate the agonist-antagonist degree of freedom by setting all efficacies to 1, discriminating N_{lig} requires twice as many receptors (Figure 3.5c, red dots). In this case each receptor is strongly bound by only one ligand, measuring its concentration independently of the other members of the array. Having access to the full range of receptor responses makes it possible to double the number of ligands in the mixture, but the relationship between N_{rec} and N_{lig} remains linear.

3.4 Symmetry properties of optimal sensor arrays

The patterns shown in Figure 3.5b are not unique – indeed, alternative agonist-antagonist patterns can be generated simply by exchanging receptor labels. Less trivially, a given ligand can be an agonist or an antagonist for different combinations of receptors. In the simplest case of one receptor interacting with two ligands, this symmetry generates two equivalent global maxima discussed above: $A_1 = 0$, $A_2 = 1$ and $A_1 = 1$, $A_2 = 0$ (Figure 3.5d (*i*)). In the two-receptor, three-ligand case symmetry arguments combined with extensive

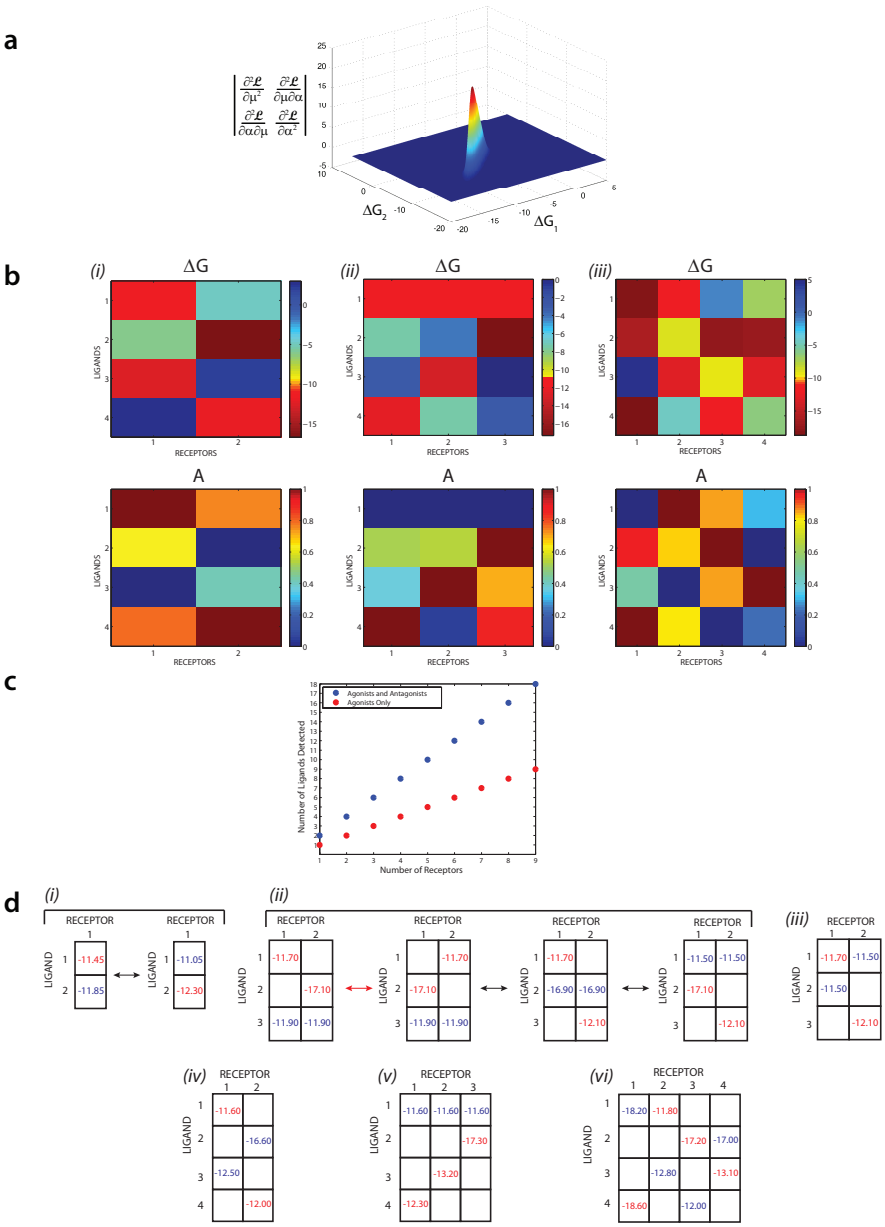


Figure 3.5: **Optimal design of receptor arrays.** (**a**) Determinant of the Hessian in the one-receptor, two-ligand case, plotted as a function of binding energies ΔG_1 and ΔG_2. The binding energies at the peak are $\Delta G_1 = -11.45$ kcal/mol and $\Delta G_2 = -11.85$ kcal/mol. The efficacies are fixed at $A_1 = 1$, $A_2 = 0$; $\alpha = 0.25$. (**b**) Optimal free energies ΔG and efficacies A obtained by maximizing the determinant of the Hessian in the (*i*) two-receptor, four-ligand, (*ii*) three-receptor, four-ligand and (*iii*) four-receptor, four-ligand cases. $\alpha_1 = 10^{-4}, \alpha_2 = 0.1, \alpha_3 = 0.5$. (**c**) The number of successfully discriminated ligands increases linearly with the number of receptors: $N_{lig} = 2N_{rec}$ if the determinant of the Hessian is maximized with respect to all binding energies ΔG and efficacies A (separately for each N_{rec}); $N_{lig} = N_{rec}$ if all ligands are forced to be full agonists with unit efficacies. We call all ligands successfully discriminated if $\sigma_\mu^2 < 2$, $\sigma_{\alpha_i}^2 < 2$ ($\forall i$) in a given nested sampling run. Two alternative choices of relative concentrations: $\alpha_i = 1$ ($\forall i$) and $\alpha_i = 0.25$ ($\forall i$) yielded the same linear dependence on the number of receptors. (**d**) Optimal ΔG values are shown for several cases: (*i*) one-receptor, two-ligand, (*ii*) two-receptor, three-ligand, (*iii*) two-receptor, three-ligand (at a local maximum), (*iv*) two-receptor, four-ligand, (*v*) three-receptor, four-ligand and (*vi*) four-receptor, four-ligand. Values in blue correspond to an efficacy of $A = 0$ (full antagonist), while values in red correspond to $A = 1$ (full agonist). $N_{lig} = 2$: $\alpha = 0.25$; $N_{lig} = 3$: $\alpha_1 = 10^{-4}, \alpha_2 = 0.5$; $N_{lig} = 4$: $\alpha_1 = 10^{-4}, \alpha_2 = 0.1, \alpha_3 = 0.5$. In all cases shown in (**a**)-(**d**), we used 91 datapoints for each receptor (7 replicates with $\log_{10} n^l = \{-3.0, -3.5, \ldots, -8.5, -9.0\}$).

sampling yield three global maxima of the Hessian determinant. Each global maximum corresponds to the situation where one of the three ligands acts as an antagonist to both receptors (Figure 3.5d (*ii*)). The red arrow in Figure 3.5d (*ii*) indicates a trivial exchange of receptor labels, whereas the black arrows connect three different globally optimal solutions. In addition, there are 9 local maxima with one of the ligands acting either as an agonist to both receptors, or as an agonist to one receptor and an antagonist to the other (e.g. Figure 3.5d (*iii*); see Appendix C.3 for a complete enumeration).

In general, $N_{rec} \times N_{lig}$ ΔG's are necessary to characterize all the global and local peaks on the Hessian determinant landscape, with $2N_{rec}$ binding energies describing any given agonist-antagonist pattern. The values of the binding energies depend on the component concentrations in the interrogated mixture. In the $N_{lig} = 2N_{rec}$ case all maxima are global and each receptor interacts with two unique ligands. To estimate the benefit of additional receptors, we increased the number of receptors from two to three to four in the four-ligand case (Figure 3.5d (*iv*), (*v*), (*vi*)). After adding the third receptor the average uncertainty of one total and three relative concentrations, $\langle \sigma_i^2 \rangle$, decreased from 0.439 to 0.125. However, only a slight gain was seen when the fourth receptor was added, with $\langle \sigma_i^2 \rangle$ becoming 0.101. Thus adding more and more receptors to the array yields increasingly marginal improvements after a certain threshold.

The agonist-antagonist rules described above create readout patterns that are not a simple sum of array responses to single-ligand binding. For one receptor optimized to discriminate two ligands (Figure 3.5d (*i*)), fluorescent response to the mixture is intermediate between full activation by the agonist and full repression by the antagonist (Figure 3.6a). This intensity modulation provides enough information for decoding the contents of the mixture. Similarly, in the two-receptor, three-ligand case (Figure 3.5d (*ii*)) a mixture of all three ligands induces a response with intermediate fluorescense levels (Figure 3.6b). This pattern is distinct from those induced by single ligands and by binary mixtures with the same relative concentrations as in the three-ligand case.

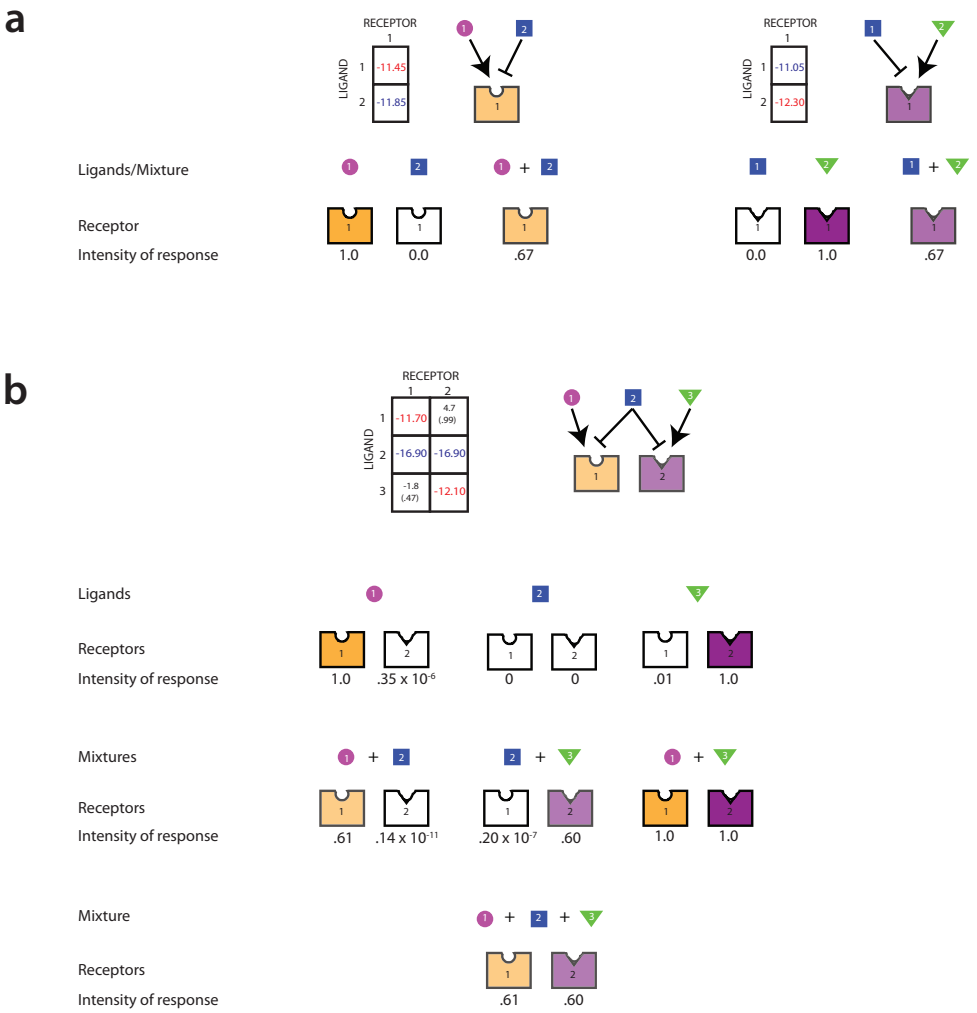


Figure 3.6: **Schematic diagram of receptor activation by single ligands and ligand mixtures with optimized binding affinities and efficacies.** (**a**) Two cases of agonist-antagonist single-receptor arrays designed to discriminate a mixture of two ligands (Figure 3.5d (*i*). (**b**) Array of two receptors designed to discriminate a mixture of three ligands Figure 3.5d (*ii*). In both cases we show relative intensities (normalized to 1.0) corresponding to $\log_{10} n = -3.0$. $N_{lig} = 2$: $\alpha = 0.25$; $N_{lig} = 3$: $\alpha_1 = 10^{-4}, \alpha_2 = 0.5$ (leading to $\alpha = 10^{-4}$, 0.5 and 5×10^3 for the binary combinations of ligands 1-2, 2-3 and 1-3, respectively). 4.7(0.99) and $-1.8(0.47)$ are ΔG's and A's (in parentheses) for receptor-ligand interactions outside of the dominant agonist-antagonist pattern.

3.5 Performance analysis and improvement of the experimental biosensor array

The design guidelines described above can be used to predict which parameter changes lead to most significant improvements in performance compared to our currently implemented array. Although we do not have direct experimental control over the values of A and ΔG, such insights are useful e.g. for choosing the best combination of several receptors from a larger library. Familiar agonist-antagonist patterns emerge when ΔG's and A's are optimized either separately or together to discriminate an equal-proportion, four-ligand mixture (Figure 3.7). In particular, if A's are kept fixed, ΔG's for the most distant pair of A's become more favorable for each receptor, creating an agonist-antagonist pair (Figure 3.7c). Conversely, if ΔG's are fixed, the values of A corresponding to the two lowest ΔG's become more distant from each other (Figure 3.7d). Not surprisingly, the agonist-antagonist patterns are even more pronounced if both ΔG's and A's are allowed to relax (Figure 3.7e and Figure 3.7f). Because two and certainly three optimized receptors are sufficient for discriminating four-ligand mixtures (Figure 3.5c), the fourth receptor, which does not follow the usual pattern as strongly as the other three, appears to be superfluous. Similarly to the cases shown in Figure 3.5d, Figure 3.7e and Figure 3.7f represent only one solution from a large family of local and global maxima of the Hessian determinant, which are related by permutations of receptor and ligand indices. Optimizing receptor-ligand parameters leads to a sizable improvement in array performance: with $\tilde{\sigma} = 1$ for all receptors, $\sigma_\mu = 2.64,\ \sigma_{\alpha_1} = 5.46,\ \sigma_{\alpha_2} = 16.47,\ \sigma_{\alpha_3} = 3.19$ for the original array, whereas $\sigma_\mu = 0.69,\ \sigma_{\alpha_1} = 0.68,\ \sigma_{\alpha_2} = 0.95,\ \sigma_{\alpha_3} = 0.68$ for the array in which both A's and ΔG's have been optimized.

For the experimentally implemented four-receptor GPCR array, nested sampling errors are consistently larger when UDP is present in the mixture (Figure 2.3, Table B.2). This observation is consistent with Hessian analysis: for example, the average Hessian uncertainties for three UDP-free binary mixtures are $\langle\sigma_\mu\rangle = 0.77,\ \langle\sigma_{\alpha_1}\rangle = 6.12,\ \langle\sigma_{\alpha_2}\rangle = 11.09,\ \langle\sigma_{\alpha_3}\rangle = 0.0002$. For three UDP-containing binary mixtures, the average Hessian errors are $\langle\sigma_\mu\rangle = 19.09,\ \langle\sigma_{\alpha_1}\rangle = 5.57,\ \langle\sigma_{\alpha_2}\rangle = 113.44,\ \langle\sigma_{\alpha_3}\rangle = 57.72$ (as before, all Hessian

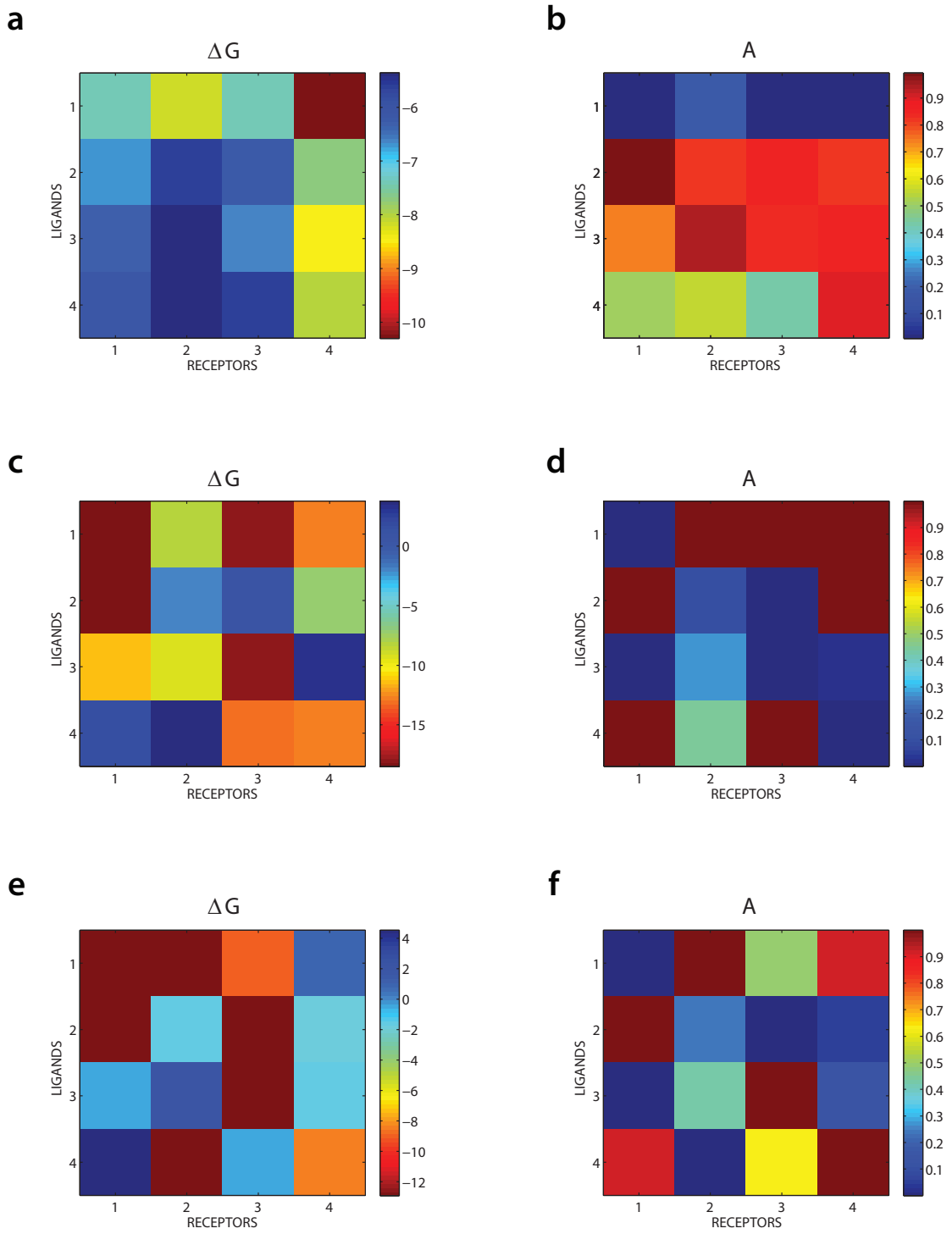


Figure 3.7: **Improving performance of the experimentally implemented sensor array.** Free energies ΔG (**a**) and efficacies A (**b**) in the experimentally implemented sensor array with parameters from Table B.1. Free energies ΔG (**c**) and efficacies A (**d**) from two sensor arrays in which the determinant of the Hessian was maximized only with respect to ΔG's and A's, respectively. Free energies ΔG (**e**) and efficacies A (**f**) in the optimal sensor array in which the determinant of the Hessian was maximized with respect to both ΔG's and A's. The determinant was computed using four replicates of an equal-proportion mixture of four ligands and $\tilde{\sigma} = 1$ for all receptors. For each receptor, concentrations were taken from the corresponding experiment (§ 2.1.4). In panels **a-d**, the order of ligands is L1: UDP, L2: UDP-Gal, L3: UDP-Glc, L4: UDP-GlcNAc. The order of receptors is R1: H-20, R2: K-3, R3: L-3, R4: 2211. Note that ligand and receptor identities are lost in panels **e**,**f** since all parameters have been optimized.

errors are computed with correct concentrations and $\tilde{\sigma} = 1$ for all receptors). The Hessian determinants are also consistently smaller for UDP-containing binary mixtures. These observations indicate that UDP parameters are further away from the optimal four-receptor array designed to analyze an equal-proportion binary mixture: either A's or ΔG's need to be changed in order to create stronger agonist-antagonist patterns.

3.6 Discussion

We have developed a Bayesian algorithm that allows determination of all the constituents in an unknown mixture from the output of a cross-specific sensor array. Our algorithm employs a physical picture of sensor-analyte interactions to model the non-linear relationship between ligand concentrations and the reporter response. After appropriate calibration of each sensor's response to each analyte of interest, the algorithm interprets the integrated output of the entire array and, with a sufficient number of variably tuned sensors, reliably returns the amount of each chemical in a complex mixture.

We also provide quantitative guidelines for designing optimal sets of sensors. Three general principles emerged from our computational and theoretical studies of array design. First, the optimal parameters of the sensors exhibit weak dependence on the relative amounts of compounds in a mixture. Thus a given set of optimal sensors will remain near-optimal through a sizable range of ligand concentrations. Nonetheless, analyzing a mixture where both compounds are present in roughly similar amounts is better accomplished with a set of sensors different from those fine-tuned to measure a small amount of one compound in the presence of a large excess of the other.

Second, the maximum number of ligands in a mixture whose levels can all be determined simultaneously is simply twice the number of sensors in the array. This linear relationship is different from the exponential relationship between ligands and receptors in olfactory systems [4, 12]. The problem addressed by the olfactory system, to recognize a very large number of individual odors with a limited repertoire of receptors, is not the same as that solved by our algorithm, to determine all the constituents in a complex mixture. In fact, even the most skilled human nose can simultaneously detect and distinguish no more than

a handful of odorants.

Third, the optimum design of receptors for the array demands that one of the ligands function as a strong agonist of a receptor and a second ligand as a strong antagonist of that receptor. Antagonists sharpen the discriminatory powers of the array by heightening the differences in the receptor response to individual compounds. As a result, a mixture of chemicals produces an array readout which is not a superposition of responses to individual ligands, and whose intensity pattern may be fine-tuned for maximum recognition through receptor-ligand binding energies. Accordingly, odors that function as antagonists to a subset of olfactory receptors could potentially increase the discriminatory power of the olfactory system, and in particular enable it to resolve mixtures that contain those odors. Recent analysis of olfactory receptors suggests that some odorants do possess antagonist activity [14, 17–20]. Our theoretical framework provides a rationale for the existence of such antagonists and underscores their role in both olfactory systems and artificial receptor arrays.

Chapter 4

Another application: RANSA used to mitigate hydrocarbon interference when targeting ammonia gas with a mixed-potential sensor array

4.1 Motivation

The potential health consequences resulting from exposure to diesel engine exhaust have been well documented; these transportation-related pollutants are responsible for a number of negative health and environmental impacts [45,46]. However, new studies have shown that recently strengthened federal requirements governing diesel engines of new manufactured heavy-duty diesel vehicles have resulted in major cuts in emissions of particulate matter (PM) and nitrogen oxides (NO_x) [47]. After implementation of new standards in the US in 2010, the NO_x and PM emissions from the diesel vehicles studied had been reduced upwards of 98% and 95% respectively per gallon of diesel fuel [47]. These reductions stem from the use of selective reduction catalyst (SRC) [48,49] and exhaust gas recirculation (EGR) [50,51] systems paired with regenerative particulate traps [52] located in the exhaust systems of heavy-duty diesel trucks. With SRC type systems, an ammonia source is provided by the on-board storage of diesel exhaust fluid, typically a mixture of urea and de-ionized water. SRC systems also incorporate limited EGR.

In contrast, the enhanced-EGR (EEGR) approach uses extensive exhaust gas recirculation in order to deprive the combustion event of oxygen by introducing cooled exhaust gas into the intake system and therefore does not use, or require, urea for on-board ammonia generation [53]. Both systems rely on particulate traps to reduce PM emissions; the traps are regenerated periodically on-board by injection of diesel fuel at elevated temperatures. The inclusion of PM filters and traps made necessary the transition to ultra-low sulfur levels in diesel fuels to prevent excessive PM formation and premature clogging. The EGR

approach to reducing NO_x emissions has a greater negative impact on overall diesel fuel efficiency since more particulates are produced by design in order to reduce NO_x formation within the combustion chamber and thus the PM traps must be regenerated more frequently. SRC systems, on the other hand, may offset added weight and extra system costs and complexity by offering greater fuel savings. The exact approach used by vehicle manufacturers is left to their discretion.

Despite these recent advances, there remains a demand for suitable exhaust gas sensor technologies to monitor tailpipe emissions and to control and maintain efficient operation of SRC and EGR systems [54]. For example, uncontrolled ammonia injection into the SRC can cause more serious air pollution than excess NO_x resulting from the combustion event [55]. Ideally, implementation of closed-loop systems that keep ammonia use close to stoichiometry in SRC systems in order to monitor tailpipe-out NO_x emissions and insure regulatory compliance is desired [54–57]. However, the development of successful sensors for diesel applications has lagged in comparison to the advances made in the implementation of oxygen lambda sensors for on-board diagnostics (OBD-II) of emission systems [58] that have become mandatory for gasoline, spark-ignition vehicles. Because of the successful implementation of robust, zirconia-based mixed-potential electrochemical sensors for gasoline engine control and OBD in conjunction with three-way catalytic converters, derivatives of this technology for diesel applications naturally attract considerable interest. However, the implementation of the diesel analogue to OBD has proved to be more complicated.

Mixed-potential sensors are electrochemical devices that measure the non-Nernstian potential of a mixture of gases, where the mixed potential is fixed by the rates of different electrochemical reactions occurring simultaneously at an electrode/electrolyte interface [59, 60]. The employment of electrode materials that possess varying electro-catalytic activities towards the redox half reactions has been shown to further increase the mixed-potential response. Leveraging the success of commercial zirconia sensors, NO_x [61, 62] and NH_3 [63, 64] sensors based on non-equilibrium electrochemical principles have been investigated; however, commercialization of this technology has been hindered by its poor reproducibility and stability [65, 66]. Because the mixed-potential response is highly dependent on the kinetics of electrochemical reactions occurring at each electrode, stable

electro-active areas are required to maintain a constant sensor response over the lifetime of the device. Additionally, device sensitivity is improved by minimizing the amount of heterogeneous reactions occurring prior to the analyte gas reaching the electrochemical interface [67, 68]. To address these issues, the Electrochemical Sensors and Devices Group at Los Alamos National Laboratory (LANL) has developed a patented sensor design that incorporates dense electrodes and porous electrolytes in bulk [68–70], thin film [71, 72], and tape cast forms [73] (Figure 4.1). By using dense electrodes, heterogeneous catalysis is reduced, increasing sensitivity, and the increased morphological stability of dense electrodes yields a robust electrochemical interface, increasing lifetime durability.

In general, the scientific community has addressed only some of the various issues with electrochemical sensors such as accuracy, temperature-dependent sensitivity, response time, sensor drift (baseline and signal), flow-rate dependence, poisoning from fuel constituents, cross-sensitivity, and thermal cycling durability and shock. Cross-sensitivity to non-target chemical species within a diesel emission system is an area of concern that this work specifically focuses on. Existing commercial solid-state electrochemical technology (Au and Pt electrodes and Yttria-stabilized Zirconia (YSZ) electrolyte) exhibits significant cross-sensitivity to hydrocarbons (HCs) when exposed to a complex gas mixture of diesel engine exhaust [74]. In this study, we investigate how the application of small bias currents to Au/YSZ/Pt mixed-potential sensors (Figure 4.1) may be used to address this issue. Prior studies have shown that sensor selectivity and sensitivity to a given analyte may be tuned by using current biasing [72, 75–77]. Although the response of the Au/YSZ/Pt sensor used in this study towards both the target gas, NH_3, and HCs is strongly influenced by current biasing, we find that cross-interference effects cannot be entirely eliminated with this method.

Here we focus on detecting NH_3 in the presence of propylene (C_3H_6) as an HC interference. Since the sensor reacts to both interfering and target gases regardless of the magnitude of the bias current, the presence and concentration of NH_3 in the gas mixture cannot be straightforwardly deduced from the sensor output. Thus we employ Bayesian inference techniques [42] to analyze sensor readouts. In order to achieve robust predictions, we use an

array of sensors, each under a different bias current and thus exhibiting differential sensitivity to NH_3 and C_3H_6. The array-based approach is reminiscent of mammalian and insect olfactory systems in which, as a rule, several kinds of olfactory receptors respond to a given analyte [2, 4]. Similar principles have been employed to create cross-reactive synthetic sensor arrays, whose output is typically analyzed using pattern-recognition algorithms [22, 24]. We have previously developed a Bayesian framework which explicitly takes receptor-ligand interactions into account, and applied it to infer concentrations of four highly related sugar nucleotides from the output of four bioengineered G-protein-coupled receptors [37, 78]. Here we apply a similar methodology to deduce NH_3 and C_3H_6 concentrations simultaneously, using readouts from an array of mixed-potential sensors as input. We also study the accuracy of our predictions as a function of the number of sensors in the array and the number of independent measurements. This allows us to deduce the minimal set of requirements for deployment of our sensor system in real-world conditions.

4.2 Preparation of sensors

Mixed-potential sensors based on 8 mol% yttria stabilized zirconia (Tosoh TZ-8YS, Japan) and dense, metal wire electrodes (Au, Pt) with preferential selectivity to ammonia were fabricated using a tape cast approach. A detailed description of the LANL method for preparation of stable, reproducible mixed-potential sensors has been published elsewhere [73, 79]. Briefly, the YSZ electrolyte powder was dried in air at $100-150°C$ for approximately one hour. The dried powder was mixed with solvents xylene, ethyl alcohol and fish oil (Blown Menhaden) and ball-milled for 24 hours. Plasticizers and binders (S-160 butyl benzyl phthalate, polyalkylene glycol and polyvinyl butyral) were then added to the mixture, which was further ball-milled for 24 hours. The mill was then discharged and de-aired for approximately 10 minutes at the reduced pressure of 20-25” of mercury. Prior to casting of the electrolyte slip, pieces of pure Au and Pt wire (0.01” in diameter) were cut into 1” lengths and straightened. Alternating pairs of Au and Pt were stuck to tape such that the spacing between wire pairs was fixed; each set of Au/Pt electrode pairs would constitute a single device. This time-saving approach facilitated handling the electrodes during the upcoming time-critical portions of sensor forming process, while maintaining the

same electrode separation and eventually the same electrode penetration depth into the solid electrolyte. After de-gassing, the slurry was cast onto a Si-coated Mylar (G10JRM) carrier film using a standard doctor blade apparatus with a gap of approximately 0.05” to 0.2”.

Once the tape was cast, it was allowed to partially dry for several minutes. This resulted in a tape that was dry on the outside (facing air) but wet on the inside (in contact with the carrier film). This tape was then turned upside down, exposing the wet side to air while the dry side was in contact with the carrier film. The pre-fabricated electrode pairs were placed on top of this wet tape such that a length of 7.5 mm of each wire was embedded in the electrolyte. The tape was folded onto itself in order to enclose the electrodes within the electrolyte. The remaining portions of the wires were left uncovered and exposed to air, and would eventually serve as both electrode and signal-out leads attached to the data acquisition system.

Next, the electrolyte tape with the partially enclosed electrodes was allowed to air-dry fully, forming a mechanically stable, green sensor body (the green sensor body refers to the unsintered and flexible form of the tape before it is heated to a high temperature to densify). Individual sensors were cut from the green tape using a razor. The tape used to fix the electrode gap and to facilitate handling during the sensor-forming procedure was then removed, and the sensor green bodies were placed into an aluminum boat and transferred to a tube furnace for removal of the organics and sintering using a multistep temperature profile. While previous sensor work of this nature conducted at LANL showed sensor response characteristics such as level of response and response time to be dependent on sintering temperature, [67] the use of gold electrodes precludes sintering temperatures above 1064°C. Thus the maximum sintering temperature used for sensors in this work was set at 1000°C.

4.3 Sensor response characterization

The sensor response or sensitivity (defined as the difference in voltage generated upon exposure to the test and base gases, respectively) was recorded using a Keithley 2400 source

measurement unit with the Pt electrode connected to its positive terminal. The base gas mixture was prepared by mixing pure streams of N_2 and O_2 using MKS 1179A mass-flow controllers and the oxygen partial pressure was maintained at 10% PO_2 during single-gas sensor testing (e.g. NH_3 testing in the concentration range of $0 - 100$ ppm). The C_3H_6 interference (introduced to the test gas stream using another mass-flow controller and 2500 ppm C_3H_6/balance N_2 stock concentration) was mixed with the $NH_3/O_2/N_2$ mixture at the inlet to the furnace tube upstream of the sensor, imparting a small dilution effect to O_2 and NH_3.

4.4 Models of sensor voltage

4.4.1 Linear Model

The response voltage of electrochemical sensors has previously been modeled for a single compound, under some simplifying assumptions, as a linear function of analyte gas concentration [60]. Starting with this simple model, we assume that mixtures of ammonia and propylene produce a response voltage in the device which is a linear sum of the component gas concentrations, thus disregarding any interaction between analyte gases. We express the sensor response voltage at each bias current setting k as:

$$V_k = V_0^k + A_k[NH_3] + B_k[C_3H_6], \tag{4.1}$$

or, equivalently,

$$V_k = V_0^k + (A_k + B_k\alpha)\left(\frac{[TOTAL]}{1+\alpha}\right), \tag{4.2}$$

where $\alpha = [C_3H_6]/[NH_3]$, $[TOTAL] = [NH_3] + [C_3H_6]$, and A_k, B_k are the linear coefficients to be determined. Note that the advantage of expressing voltage in terms of α and $[TOTAL]_i$ ($i = 1...N_\alpha$, where N_α is the number of different total concentrations for a given α) rather than individual gas concentrations $[NH_3]_i$ and $[C_3H_6]_i$ is that the total concentration can be controlled experimentally, e.g. through a series of dilutions.[1]

Sensor response voltage readings were taken at 4 bias current settings ($0, -1.5, -3.5$, and -6 μA), with the sensor exposed to different mixtures of NH_3 and C_3H_6. At each setting,

[1]Equivalently, we could have chosen to define the mixing ratio as $\alpha' = [NH_3]/[C_3H_6]$. This arbitrary choice of representation of the mixing ratio does not affect final predictions.

a series of measurements corresponding to different total concentrations was extracted for several values of α:

$$[TOTAL]_i \text{ (ppm)} = \begin{cases} \{0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\} & \text{for } \alpha = 0, \\ \{0, 40, 60, 120, 160, 180, 200\} & \text{for } \alpha = 1, \\ \{0, 50, 100, 150\} & \text{for } \alpha = 1.5, \\ \{0, 30, 90, 120, 150\} & \text{for } \alpha = 2, \\ \{0, 40, 80, 120\} & \text{for } \alpha = 3, \\ \{0, 20, 30, 60, 80, 90, 100\} & \text{for } \alpha' = 0 \end{cases}$$

Note that in order to reconstruct $[TOTAL]_i$ for each series, only $[TOTAL]_{ref}$ needs to be predicted since

$$[TOTAL]_i = [TOTAL]_{ref} - \sum_{j=1}^{i-1} \Delta c_j, \tag{4.3}$$

and Δc_j, the differences in total concentration between consecutive measurements, are assumed to be known. We choose $[TOTAL]_{ref} = 200$ ppm to be the total concentration at an arbitrary reference point (chosen to be a mixture of 100 ppm $[NH_3]$ and 100 ppm $[C_3H_6]$). Once predictions for $[TOTAL]_{ref}$ and α are made, they can be used to estimate $[NH_3]_i$ and $[C_3H_6]_i$ for each individual mixture.

We use voltage measurements collected during sensor exposure to each test gas separately ($\alpha = 0$ and $\alpha' = 0$) to estimate linear coefficients A_k and B_k for each bias current k, as well as the offset parameter $V_0^k = V_{0_1}^k + V_{0_2}^k$. When $\alpha = 0$, eq. (4.2) becomes

$$V_k = V_{0_1}^k + A_k[TOTAL]. \tag{4.4}$$

When $\alpha' = 0$, eq. (4.2), written in terms of α':

$$V_k = V_0^k + (A_k\alpha' + B_k)\left(\frac{[TOTAL]}{1+\alpha'}\right),$$

becomes

$$V_k = V_{0_2}^k + B_k[TOTAL]. \tag{4.5}$$

Equations (4.4) and (4.5) are used as input to RANSA, which predicts the average values of A_k, B_k, and V_0^k to be used in eq. (4.2) (Table E.1; least-squares fits yield very similar

results). Note that although the baseline values have been subtracted from all voltages, $V_{0_1}^k$ and $V_{0_2}^k$ are allowed to adopt non-zero values as dictated by the global fits of non-linear curves to straight lines. In practice, these offsets are small (Table E.1).

4.4.2 Nonlinear Model

We can potentially improve our predictions by incorporating interference between NH_3 and C_3H_6 via a more general nonlinear model :

$$V_k = C_k(\alpha)\left(\frac{[TOTAL]}{1+\alpha}\right)^{p_k(\alpha)}, \tag{4.6}$$

where $C_k(\alpha)$ and $p_k(\alpha)$ are unknown functions of α, with parameters to be estimated from the data. Since full Bayesian treatment of these functions is too laborious, we employ least-squares fits with cross-validation. Specifically, we randomize each set of 140 measurements with the same $[NH_3]_i$ and $[C_3H_6]_i$ and partition it into 5 training sets with 120 measurements each and 5 non-overlapping test sets with 20 measurements each. For each training set and each value of α, we find average $C_k(\alpha)$ and $p_k(\alpha)$ using RANSA. Manual inspection of these results shows that, for each bias current k, C_k is an approximately quadratic function of α (Figure E.1), and p_k is an approximately linear function of α between $\alpha = 1$ and $\alpha = 3$ (Figure E.2):

$$\begin{aligned} C_k(\alpha) &= a'_k + b'_k\alpha + c'_k\alpha^2, \\ p_k(\alpha) &= a_k + b_k\alpha. \end{aligned} \tag{4.7}$$

We use a least-squares fit to estimate parameters a_k, b_k and a'_k, b'_k, c'_k on a given training set and test our model on the corresponding test set by predicting α and $[TOTAL]_{ref}$ for all available mixtures, again using RANSA. Figures E.1 and E.2 show variation in the predicted $C_k(\alpha)$ and $p_k(\alpha)$, as well as the respective linear and quadratic fits based on 5 training sets. For a complete list of model parameters, see Tables E.2 to E.4.

4.4.3 Bayesian estimation of model parameters

We assume that the sensor voltage measurements are Gaussian-distributed around values predicted by the model, and compute the log-likelihood of the data:

$$\mathcal{L}_k = \log P(\{\tilde{V}_k\}|\{\gamma_k\}, \alpha, [TOTAL]_{ref}, \tilde{\sigma}_k) = -\frac{1}{2\tilde{\sigma}_k^2}\sum_{l=1}^{N}(V_k^l - \tilde{V}_k^l)^2 - \frac{N}{2}\log(2\pi\tilde{\sigma}_k^2), \quad (4.8)$$

where k is the bias current index, $V_k^l(\{\gamma_k\}, \alpha, [TOTAL]_{ref}, \tilde{\sigma})$ is predicted voltage which depends on a set of parameters $\{\gamma_k\}$, $\{\tilde{V}_k\}$ is a set of voltage measurements $\tilde{V}_k^l$ ($l = 1 \ldots N$, where N is the total number of measurements with a given α and bias current), and $\tilde{\sigma}_k$ is the noise parameter. The parameter set $\{\gamma_k\}$ consists of A_k, B_k, and V_0^k for the linear model (eq. (4.2)), and $a_k, b_k, a'_k, b'_k, c'_k$ for the nonlinear model (eqs. (4.6) and (4.7)). Note that predicted voltage V_k^l is a function of the measurement counter l because the total concentration, which depends on $[TOTAL]_{ref}$ through eq. (4.3), may be different for each measurement.

The total log-likelihood is given by $\mathcal{L} = \sum_{k=1}^{N_s}\mathcal{L}_k$, where $N_s = 4$ is the total number of bias current settings. In principle, the log-likelihood can be used to estimate the posterior probability for all unknowns according to the Bayes rule [42]:

$$P(\{\gamma\}, \alpha, [TOTAL]_{ref}, \{\tilde{\sigma}\}|\{\tilde{V}\}) = \frac{P(\{\tilde{V}\}|\{\gamma\}, \alpha, [TOTAL]_{ref}, \{\tilde{\sigma}\})P(\{\gamma\})P(\alpha)P([TOTAL]_{ref})P(\{\tilde{\sigma}\})}{P(\{\tilde{V}\})}, \quad (4.9)$$

where on the right-hand side the likelihood from eq. (4.8) is multiplied by the product of priors for each model parameter and divided by evidence. $\{\tilde{V}\}$ combines data from all bias current settings, $\{\tilde{\sigma}\}$ is a set of N_s noise parameters, and $\{\gamma\}$ is the union of all parameter sets $\{\gamma_k\}$. $P(\{\gamma\})$ and $P(\{\tilde{\sigma}\})$ are products of prior probabilities for each parameter in the set. However, in practice we find it more reliable to estimate model parameters first in a separate calibration step:

$$P(\{\gamma_k\}, \tilde{\sigma}_k|\{\tilde{V}_k\}, \alpha, [TOTAL]_{ref}) = \frac{P(\{\tilde{V}_k\}|\{\gamma_k\}, \alpha, [TOTAL]_{ref}, \tilde{\sigma}_k)P(\{\gamma_k\})P(\tilde{\sigma}_k)}{P(\{\tilde{V}_k\})}. \quad (4.10)$$

Specifically, with the linear model eq. (4.4) is used to predict A_k and $V_{0_1}^k$ and, similarly, eq. (4.5) is used to predict B_k and $V_{0_2}^k$. With the nonlinear model, eq. (4.6) is employed

to estimate $C_k(\alpha)$ and $p_k(\alpha)$. Note that α and $[TOTAL]_{ref}$ are assumed to be known for the training datasets. Finally, we compute ensemble averages $\bar{\gamma}$ for all model parameters γ, and use them to predict α and $[TOTAL]_{ref}$:

$$P(\alpha, [TOTAL]_{ref}, \{\tilde{\sigma}\}|\{\tilde{V}\}, \{\bar{\gamma}\}) = \frac{P(\{\tilde{V}\}|\{\bar{\gamma}\}, \alpha, [TOTAL]_{ref}, \{\tilde{\sigma}\})P(\alpha)P([TOTAL]_{ref})P(\{\tilde{\sigma}\})}{P(\{\tilde{V}\})}. \tag{4.11}$$

We use uniform priors (invariant with respect to translations, $x \to x + a$):

$$P(x) = \begin{cases} 1/(x_{max} - x_{min}) & \text{if } x \in [x_{min}, x_{max}], \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

for $\{\gamma\}$ and $[TOTAL]_{ref}$, and Jeffrey's priors (invariant with respect to rescaling, $x \to ax$) for α and $\{\tilde{\sigma}\}$:

$$P(x) = \begin{cases} 1/(x \log(x_{max}/x_{min})) & \text{if } x \in [x_{min}, x_{max}], \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

In each case, x_{min} and x_{max} are chosen to encompass all plausible solutions and thus have no effect on final results.

We estimate all posterior probabilities by nested sampling [42].

4.5 Effects of bias current application

Figure 4.1 shows a schematic of the Au/YSZ/Pt tape-cast, mixed-potential sensor used in this work, along with a photograph of the actual device. The sensor was placed into a tube furnace and tested at a series of temperatures in order to determine optimum performance specified by the tradeoff of NH_3 sensitivity and response time. Figure 4.2a shows the open-circuit response of the sensor for individual five-minute exposures to 100 ppm of NO, NO_2, NH_3, and C_3H_6. While the sensor has relatively little sensitivity to NO and NO_2, there is substantial cross-sensitivity to C_3H_6. Figure 4.2b shows the effect of the application of a -3 μA bias current on the relative selectivity for this device. We observe that while the response to 100 ppm NO_2 has been removed, the response to 100 ppm NO has been augmented and that to C_3H_6 reduced by roughly 30% relative to NH_3.

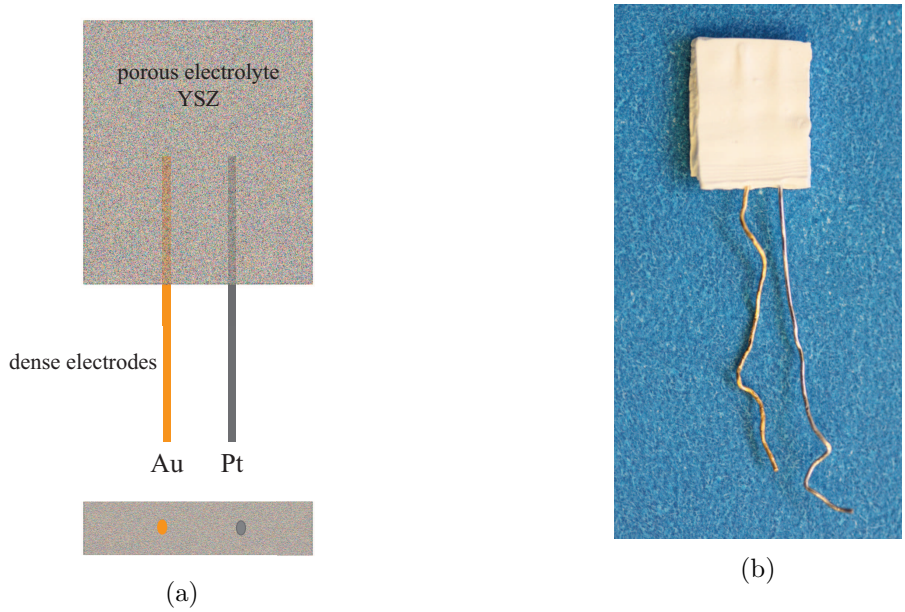


Figure 4.1: **Au/YSZ/Pt mixed-potential sensor.** (a) Schematic illustration of the sensor. The sensor uses wire electrodes embedded into tape-cast, porous YSZ solid electrolyte. Top: side view, bottom: view from below. (b) Photograph of the actual device. Electrolyte body is 1 cm in length.

Figure 4.2c shows that in the absense of hydrocarbon gases, this sensor could serve as a highly selective ammonia sensor for SRC applications, as it exhibits limited sensitivity to NO and especially NO_2. Figure 4.2d shows voltage response to 100 ppm of the four test gases as the bias current is varied from -4.5 μA to $+3.5$ μA. With the application of increasing negative current bias, the NO, NO_2, and C_3H_6 levels decrease with respect to the ammonia response; however, the voltage for 100 ppm of C_3H_6 does not decrease below -4.5 μV and therefore this particular hydrocarbon remains a substantial source of interference with the ammonia signal. Thus the ability to alter response selectivity with the Au/YSZ/Pt mixed-potential sensor is limited by HC interference. This fact, taken together with preliminary sensor design studies of the effect of electrode geometry and the ratio of bulk to interfacial resistance on Au/YSZ/Pt sensor response, [77] shows that some level of undesirable cross-interference with non-target gas species such as C_3H_6 will be inevitable with this sensor construct.

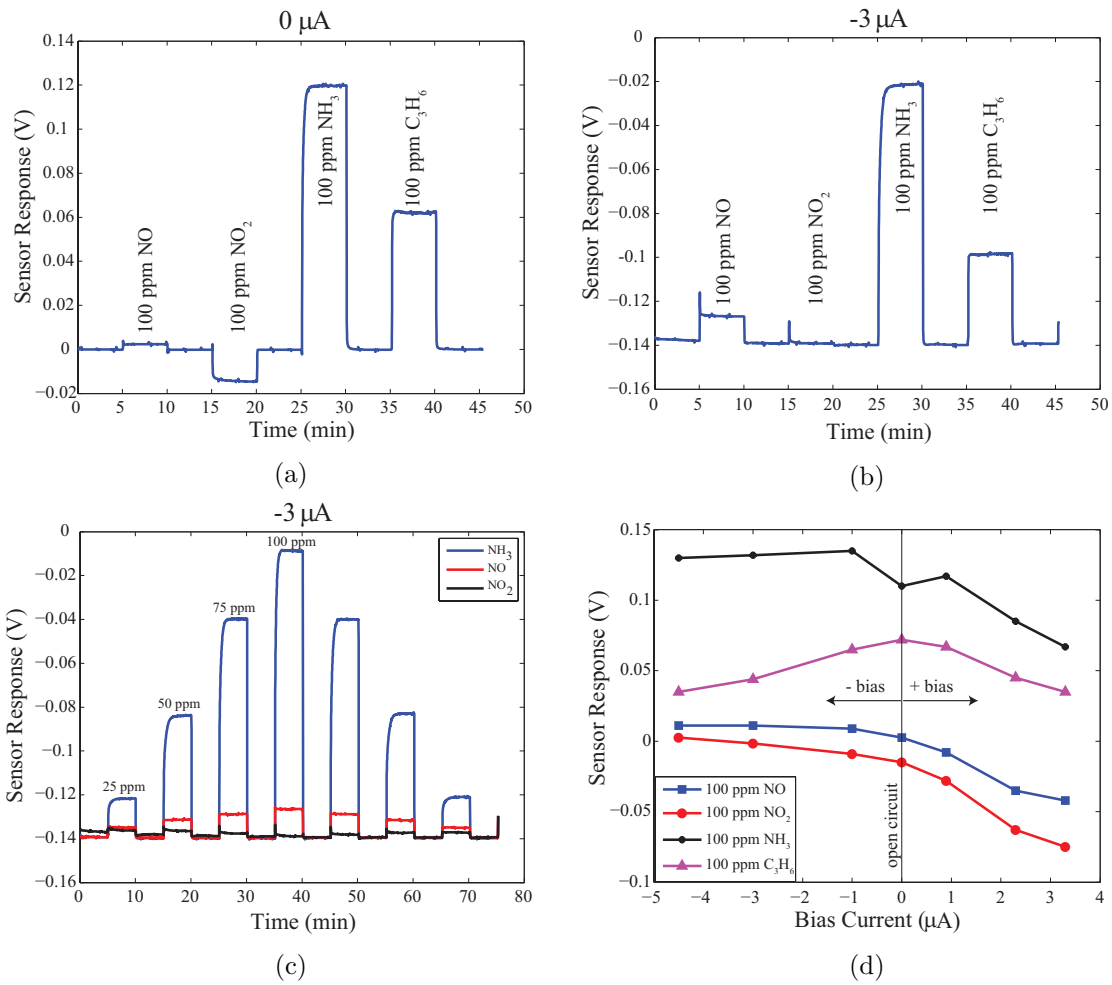


Figure 4.2: **Effects of bias current on sensor response.** (a) Open-circuit (0 μA) response of the sensor to 100 ppm of NO_2, NO, NH_3, and C_3H_6. (b) Same as (a), at -3 μA current bias. (c) Sensor response to 0, 25, 50, 75, and 100 ppm of NH_3 and equivalent amounts of NO_2 and NO at -3 μA applied current bias. (d) Sensor response to 100 ppm of NO_2, NO, NH_3, and C_3H_6 at various levels of positive and negative current bias. Note that the baseline voltage (measured with no test gas present) is subtracted from the voltage measured with 100 ppm of each test gas. All measurements were taken with sensors at 575°C and in 10% O_2/balance N_2, flowing at 500 ml/min.

4.6 Deconvoluting ammonia response in the presence of hydrocarbon interference

In light of the unavoidable HC interference discussed above, we carried out a series of experiments aimed at collecting NH_3/C_3H_6 interference data for subsequent mathematical modeling. Each series of measurements was performed at four different bias current settings. Since in this regime sensor selectivity and sensitivity depend on the magnitude of the bias current (Figure 4.2d), using four settings is equivalent to creating an array of four sensors, each of which produces a different response to a given mixture of NH_3 and C_3H_6. Figure 4.3a shows the baseline-corrected sensor response to various levels of NH_3 in the presence of 0, 20, 30, 60, 80, 90 and 100 ppm of the interfering gas, C_3H_6, with an applied bias current of -6 μA (data for bias current settings of 0, -1.5, and -3.5 μA is shown in Figures E.3a, E.4a and E.5a, respectively). These measurements are used as input to RANSA, which predicts NH_3 and C_3H_6 concentrations in each mixture.

4.6.1 Using linear Model

Since the magnitude of hydrocarbon interference depends on the relative concentration of C_3H_6 with respect to NH_3, we find it convenient to replot the data in terms of the relative concentration (or mixing ratio) $\alpha \equiv [C_3H_6]/[NH_3]$, and the total concentration $[TOTAL] \equiv [NH_3] + [C_3H_6]$ (Figures 4.3b, E.3b, E.4b and E.5b). Each series of points represents baseline-corrected sensor voltage measurements at a given bias current, as a function of the total mixture concentration $[TOTAL]$, for a given value of α (we only consider gas mixtures for which four or more total gas concentrations are available, (see § 4.4.1). We see that, to a first approximation, the response of each sensor to a given mixture is linear.

The linear model assumes that the total response to a mixture of gases is simply a sum of responses to each individual gas in the mixture (eq. (4.1)). Thus, to fit mixture data to a linear model, we first determine its parameters (V_0^k, A_k, and B_k, where $k = 1 \ldots 4$ labels bias current settings; cf. eqs. (4.4) and (4.5)) by using sensor responses to each separate gas (black and red squares in Figures 4.3b, E.3b, E.4b and E.5b. Although we obtain full

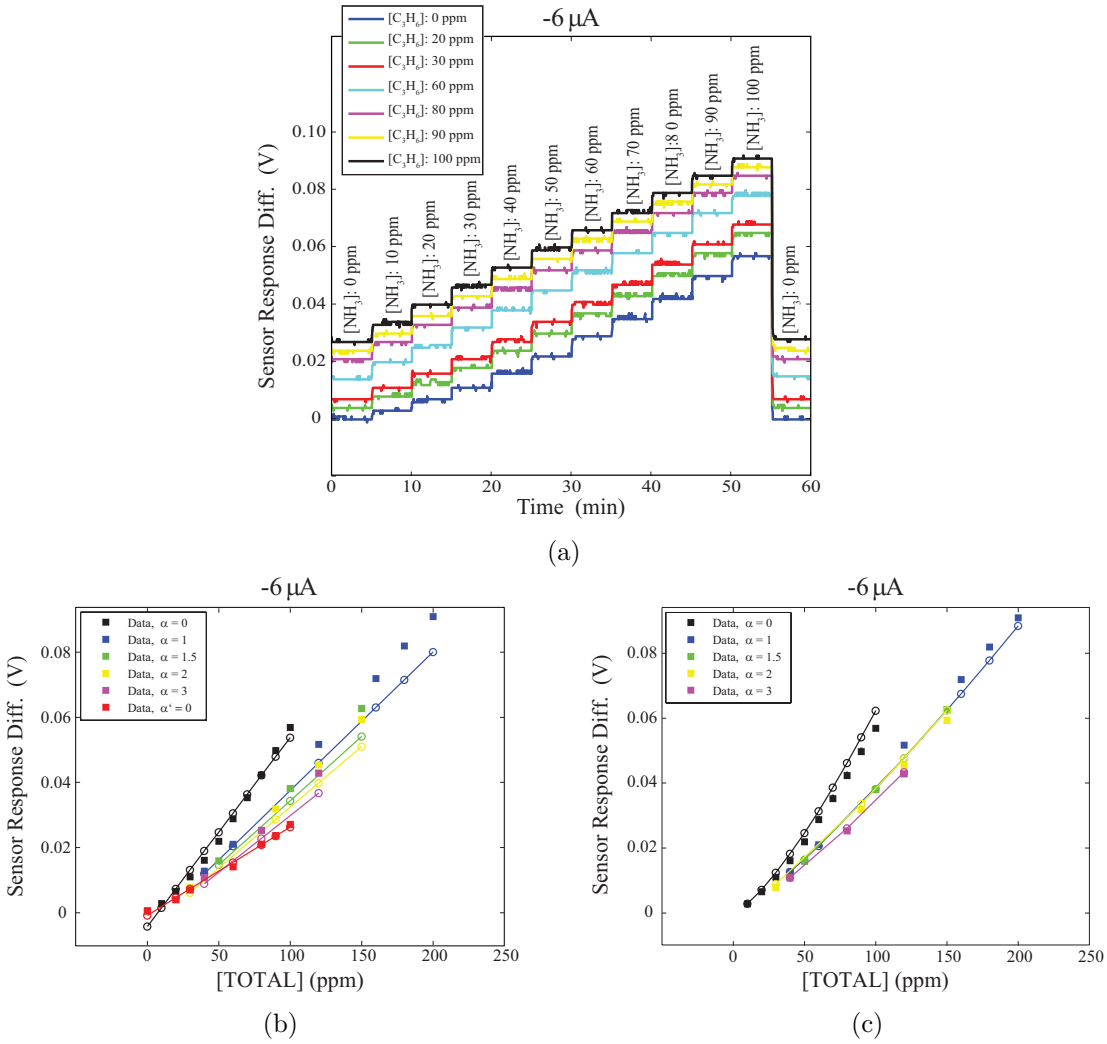


Figure 4.3: **Sensor response to mixtures of NH_3 and C_3H_6.** (a) Sensor response voltage plotted as a function of time and $[C_3H_6]$. $[NH_3]$ levels increase from 0 to 100 ppm in 10 ppm increments in the same time course, and each curve corresponds to a different level of $[C_3H_6]$, as indicated in the legend. (b) Sensor response vs. total concentration $[TOTAL]$, sorted by the mixing ratio α (squares). Lines with open circles represent linear model predictions (eq. (4.2)). Raw interference data in (a) corresponds to $\alpha = 0$, 1, 1.5, 2, 3, and $\alpha' \equiv [NH_3]/[C_3H_6] = 0$. Each data point is an average of the approximately 140 voltage readings for each pair of concentrations in (a) (standard errors are omitted for clarity). For example, the 200 ppm data point for $\alpha = 1$ (blue square), is an average of all measurements taken when the sensor was exposed to 100 ppm of NH_3 and 100 ppm of C_3H_6, i.e., the black line in (a) from $t = 51$ min to $t = 54$ min. Note that $\alpha = 0$ and $\alpha' = 0$ series of measurements correspond to individual gases rather than mixtures. (c) Same data as in (b) (squares), with the $\alpha' = 0$ case omitted. Lines with open circles represent non-linear model predictions (eq. (4.6)). With the exception of the $\alpha = 0$ (black) curve, C_k and p_k (cf. eq. (4.6)) are functions of $\bar{\alpha}$, the mean value of α as predicted across five training/test data subsets (see § 4.4.2). The curve coefficients in eq. (4.7) $(a_k, b_k; a'_k, b'_k, c'_k)$ are also averaged over five training/test subsets. In the case of $\alpha = 0$, C_k is estimated as before but $p_k(\alpha = 0)$ is the mean value inferred by RANSA in the calibration step rather than predicted by a least-squares fit (see Figure E.2 and Table E.2). In (b) and (c), each predicted curve is color-coded to correspond to different mixing ratios, as defined in the legends. Note that $[TOTAL] = 0$ measurements and predictions were omitted for clarity, except for the $\alpha = 0$ and $\alpha' = 0$ cases in (b) which were used to calibrate the linear model. For all panels, the bias current was set at -6μA, the baseline voltage (voltage measured with no test gases present) was subtracted from all voltage measurements, and all measurements were taken at 575°C and in 10% O_2/balance N_2, flowing at 500 ml/min.

Bayesian estimates for all model parameters by applying RANSA to eq. (4.10), we only employ their average values in subsequent predictions, as integrating over the ensemble of models would be prohibitively expensive. We then apply RANSA to eq. (4.11) to predict the mixing ratio α and the total concentration at the reference point $[TOTAL]_{ref}$ (equivalent to predicting absolute concentrations of NH_3 and C_3H_6, see § 4.4.1) for each of the four mixtures. Despite neglecting all nonlinearity in the data, we nonetheless do reasonably well at predicting individual analyte concentrations (Figure 4.4).

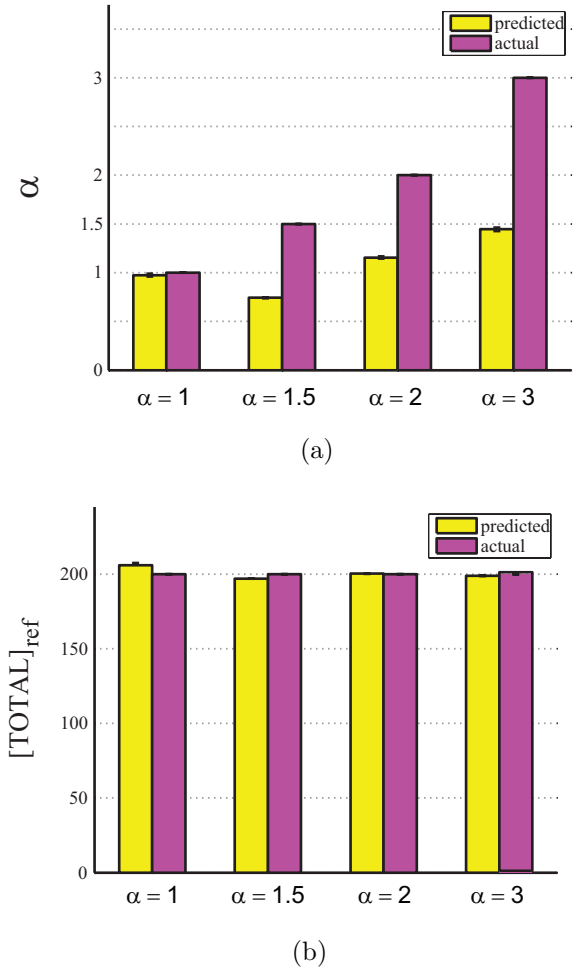


Figure 4.4: **Prediction of gas concentrations in mixtures of NH_3 and C_3H_6.** Using a linear model (eq. (4.2)) and a Bayesian nested sampling algorithm RANSA, we predict means and standard deviations (shown as error bars) of the mixing ratio, α (a), and the total mixture concentration at the 200 ppm reference point, $[TOTAL]_{ref}$ (b) for four different values of α from Figure 4.3b ($\alpha = 0$ and $\alpha' = 0$ are omitted since they were used to calibrate the model).

4.6.2 Using nonlinear Model

Although the sensor response is approximately linear, interference between NH_3 and C_3H_6 becomes more prominent at higher gas concentrations, making the additivity approximation less accurate. Interestingly, cross-interference between the two gases depends on both the mixing ratio α and the bias current. For example, at 0 μA the voltage response is sub-linear, indicating a preferential response to one of the two gas components (Figure E.3c). This trend is gradually reversed with the increase in the negative current bias, resulting in voltage responses at -6 μA that are inverted compared with their counterparts at 0 μA (Figures 4.3c, E.4c and E.5c).

To model nonlinear sensor responses, we employ a straightforward generalization of eq. (4.2) shown in eq. (4.6). The nonlinear model depends on two independent parameters, $C_k(\alpha)$ and $p_k(\alpha)$, which are functions of both the mixing ratio α and the bias current, as observed in the data. To predict these functions and thus model the sensor response over a range of α, we first divide all measurements into 5 training/test subsets as described in § 4.4.1. For each training subset, we predict $C_k(\alpha)$ and $p_k(\alpha)$ using RANSA, for all available values of α: 1.0,1.5,2.0, and 3.0. In this calibration step, we assume α and $[TOTAL]_{ref}$ to be known, and employ eq. (4.10) to estimate means and standard deviations of the model parameters.

Next, we employ polynomial fits to infer C_k and p_k as explicit functions of α (eq. (4.7); cf. Figures E.1 and E.2) for each training subset, using RANSA-predicted mean values of $C_k(\alpha)$ and $p_k(\alpha)$ as input. These steps yield a set of 5 models for $C_k(\alpha)$ and $p_k(\alpha)$, allowing us to cross-validate our predictions. Optionally, a single model can be produced by averaging the polynomial parameters of $C_k(\alpha)$ (a'_k, b'_k, c'_k) and $p_k(\alpha)$ (a_k, b_k) over 5 subsets of training data, as in Figure 4.3c. Since the polynomial fits are done by least squares, they represent a non-Bayesian step which was implemented because we do not have sufficient data for full Bayesian inference (eq. (4.9)). This necessitates the use of cross-validation. We note that, unlike the linear model, the nonlinear model is limited to a certain range of α's because $C_k(\alpha)$ and $p_k(\alpha)$ predictions cannot be reliably extrapolated too far beyond the set of α's used in the calibration step. Thus the nonlinear model should be trained using a set of

mixing ratios that cover the entire range of interest in subsequent applications.

Models trained on 5 training data subsets are then tested on their test set counterparts. RANSA is utilized again to predict individual analyte concentrations (or, more specifically, α and $[TOTAL]_{ref}$ which encode the same information, see § 4.4) using $C_k(\alpha)$ and $p_k(\alpha)$ functions as input to eq. (4.11) (Figure 4.5). Note that, as in the linear model, these predictions combine data from all bias current settings. We see a sizable improvement in our predictions of α and $[TOTAL]_{ref}$ when the nonlinear model is employed: the mean absolute error in predicting α for four mixtures is 0.80 with the linear model, but is reduced to 0.29 with the nonlinear model. The error becomes 0.23 when the $\alpha = 0$ case is included in the average for nonlinear predictions. Note that this case cannot be used to test the linear model because $\alpha = 0$ (single gas) data was used to calibrate it. Similarly, the mean absolute error in predicting $[TOTAL]_{ref}$ is 2.6 ppm with the linear model, and 2.1 ppm with the nonlinear model when the $\alpha = 0$ case is included. The essential differences between linear and nonlinear models also manifest themselves in the quality of predictions shown in panels (b) and (c) of Figures 4.3 and E.3 to E.5. In the nonlinear case, the model predictions employ *predicted* mixing ratios and *predicted* C_k and p_k functions, even in the $\alpha = 0$ case. In contrast, $\alpha = 0$ and $\alpha' = 0$ data were used to calibrate the linear model as described in § 4.4.1, and thus linear model predictions closely reproduce the data in those two cases.

4.7 Minimum data requirements for robust concentration predictions

Our computational framework for predicting gas analyte concentrations consists of two steps: the calibration step, in which the parameters of the model are inferred in a laboratory setting using an array of sensors, and the prediction step, in which the same sensor setup is employed to predict α and $[TOTAL]_{ref}$ (and thus absolute concentrations of each constituent analyte) in real-world conditions. (In this study, all measurements were done in the lab, with some of the data set aside to mimic real-world measurements.) While we can assume that sufficient training data is available for gas mixtures of interest for subsequent practical applications, the number of measurements taken in the prediction step must be minimized in order to improve reaction times and efficiency of the device.

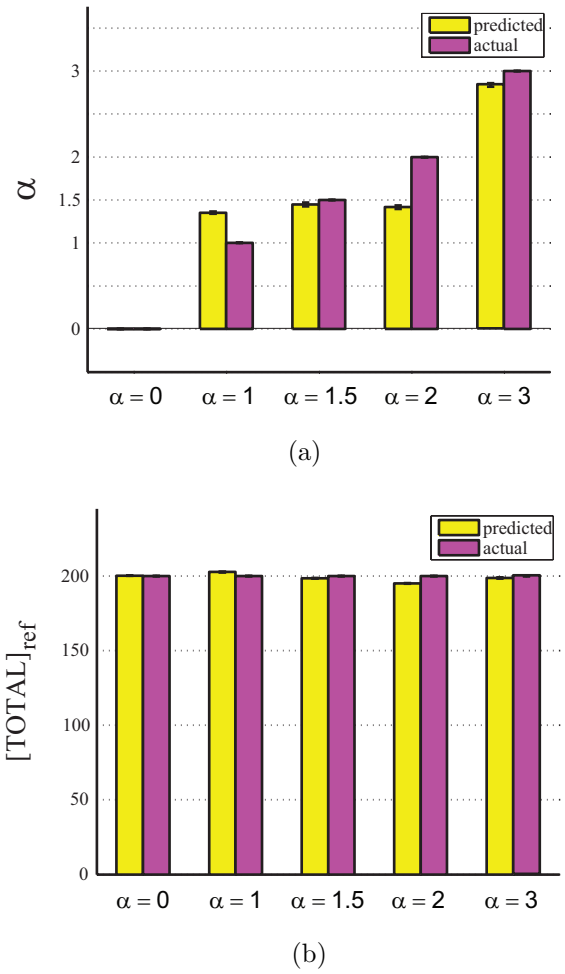


Figure 4.5: **Prediction of gas concentrations in mixtures of NH_3 and C_3H_6.** Using a nonlinear model (eq. (4.6)) and a Bayesian nested sampling algorithm RANSA, we predict means and standard deviations of the mixing ratio, α (a), and the total mixture concentration at the 200 ppm reference point, $[TOTAL]_{ref}$ (b) for five different values of α from Figure 4.3c. All means and standard deviations are averaged over five training/test data subsets (see § 4.4.2).

Our predictions in Figure 4.5 were based on 20 independent voltage readings for each mixture and each bias current setting (for the nonlinear model which was previously calibrated on 5 training datasets as described above). Since in practice a single device under different bias current settings rather than a physical array of sensors is likely to be employed, it is important to minimize the number of measurements in each series of readings. Surprisingly, we find that the quality of our predictions does not decline significantly when we gradually reduce the number of measurements from 20 to 1 for each k, α, and $[TOTAL]$. This may be due to the fact that each measurement, taken at the steady-state flow of the gas mixture, conveys roughly the same information to the algorithm. Indeed, the mean absolute error in predicting all α's is 0.38 for 2 measurements, 0.22 for 5, and 0.28 for 20. The mean absolute error is computed by first averaging mean values of α predicted by RANSA for each mixture using 5 test sets with the number of points indicated above as input. Next, the absolute error is computed as the absolute magnitude of the difference between predicted and exact values, and averaged over all 5 mixtures (with $\alpha = 0.0, 1.0, 1.5, 2.0, 3.0$). Similarly, the mean absolute error in predicting $[TOTAL]_{ref}$ is 2.54 for 2 measurements, 2.16 for 5, and 2.36 for 20. Note that, as expected, standard deviation in each individual RANSA prediction of α and $[TOTAL]_{ref}$ increases as the number of measurements is reduced, although the uncertainties remain small.

Our algorithm depends on voltage measurements taken at more than one value of the total concentration for a given mixture (cf. eq. (4.3)). In real-world applications, controlled dilution of the mixture of ammonia and hydrocarbons in a neutral buffer such as fresh air will make the device more complex to build and operate. Thus, we investigate how the quality of predictions depends on the number of different total concentrations or, equivalently, the number of dilutions for the same mixture. Since among 4 mixtures with $\alpha \neq 0$ we have the most data for the $\alpha = 1$ case (seven total concentrations; blue squares in Figures 4.3c, E.3c, E.4c and E.5c, S3(c),S4(c),S5(c)), we focus on this dataset for our analysis. We start with two points, corresponding to $[TOTAL] = 0$ ppm and $[TOTAL] = 40$ ppm ('Range 1'), and then gradually extend the range by including points with higher concentrations until the full dataset is recovered. We predict $[TOTAL]_{ref}$ and α using the nonlinear model, and plot absolute errors of our predictions as blue crosses in Figures 4.6a and 4.6b. In

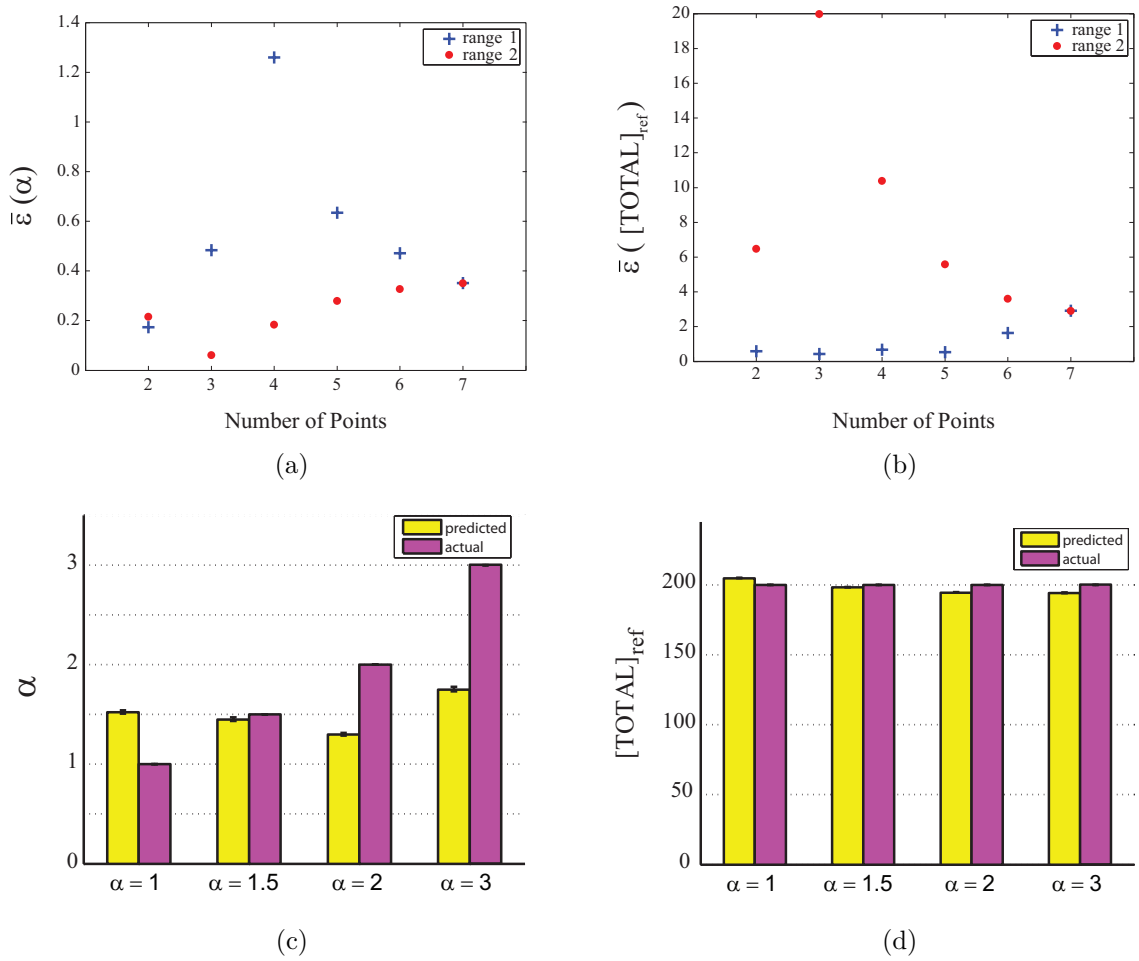


Figure 4.6: **Robustness of gas concentration predictions with respect to the number of measurements in training and test datasets.** (a) Average absolute error $\bar{\epsilon}(\alpha)$ for predictions of α, plotted as a function of the number of points with different total concentrations. 'Range 1' and 'Range 2' are defined in the text. Mean values of α estimated by RANSA were averaged across 5 training datasets. The average absolute error is given by the absolute magnitude of the difference between this average and the exact value $\alpha = 1$. (b) Same as (a) but the average absolute errors are shown for predictions of $[TOTAL]_{ref}$. The exact value is $[TOTAL]_{ref} = 200$ ppm. (c) Means and standard deviations of α predicted by RANSA after omitting the mixture for which the predictions were made from the training data, and recalibrating the model. All means and standard deviations are averaged over 5 training subsets. (d) Same as (c) but for means and standard deviations of $[TOTAL]_{ref}$. Predictions in (a)-(d) use data from all 4 bias current settings.

each individual calculation, we use the nonlinear model trained on 5 complete calibration datasets.

We repeat the calculation starting at the high concentration end ('Range 2'), first using two points corresponding to $[TOTAL] = 200$ ppm and $[TOTAL] = 180$ ppm and then adding measurements with lower concentrations until all seven points are included. Errors in $[TOTAL]_{ref}$ and α for this set of calculations are shown as red circles in Figures 4.6a and 4.6b. We observe that the quality of predictions strongly depends on which total concentrations are included in the analysis. It appears that low-concentration measurements (Range 1), where the contributions from both gases are approximately additive, can be used to infer $[TOTAL]_{ref}$ more accurately than α. In contrast, high-concentration measurements, where cross-interference effects are more prominent, appear to be more beneficial for the inference of α. In both cases, average errors exhibit non-uniform trends but ultimately start to converge as the range is extended. Thus the number of times each mixture is diluted should be chosen carefully depending on the level of accuracy desired for each predicted variable.

Next, we check the ability of our algorithm to predict α and $[TOTAL]_{ref}$ for mixtures that were not part of the training dataset. This ability is crucial because our framework must yield reasonably accurate predictions for a range of α's rather than just the discrete set of mixtures on which it was trained. To this end, we leave out all data corresponding to one of the $\alpha \neq 0$ ($\alpha = 1$, 1.5, 2, or 3), and refit the nonlinear model using the remaining 3 mixtures, plus the data for $\alpha = 0$. The model is then used to predict the total gas concentrations in the mixture that was left out of the calibration step (Figures 4.6c and 4.6d). We see that, encouragingly, the errors for $\alpha = 1.5$ are reasonably small, likely because the data for both $\alpha = 1$ and $\alpha = 2$ were available to train the model. On the other hand, our ability to extrapolate the range of α is limited, as manifested by the $\alpha = 3$ example. Thus the model calibration step should be based on training data that covers the entire range of α's to be expected in subsequent applications, with reasonably fine spacing between adjacent values of α.

Finally, we evaluate the benefit of combining data collected at multiple bias current settings. We monitor errors in our predictions as additional sensors are incorporated into

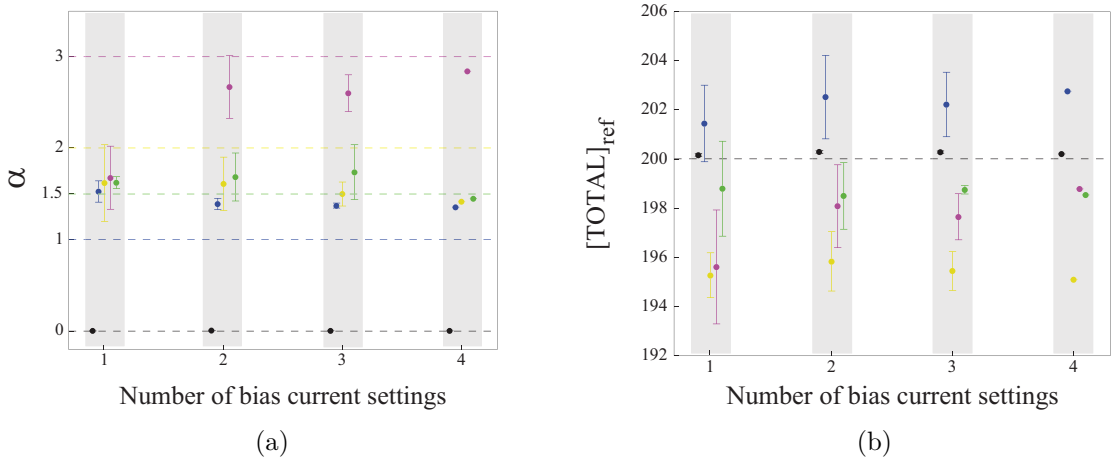


Figure 4.7: **Prediction accuracy as a function of the number of bias current settings.** (a) Each point represents the average predicted value of α. First, mean α's are obtained for each of the 5 training subsets and for each combination of sensors. These mean α's are averaged over the training subsets and, next, over the sensor combinations. The error bars represent standard deviations with respect to different sensor combinations. Exact values are shown as horizontal dashed lines of the same color. Light grey bars indicate the number of bias current settings (predictions are slightly offset for clarity). (b) Same as (a) but for $[TOTAL]_{ref}$.

the array. Using fewer current settings would simplify construction and operation of the device. Here, we use the nonlinear model calibrated earlier on all 5 training datasets and all 4 bias current settings. However, instead of combining the data from all 4 bias currents, we estimate the errors in α and $[TOTAL]_{ref}$ as the number of bias current settings is increased from 1 to 4. With 4 settings, there are 4 combinations with a single sensor in the array, 6 combinations with two sensors, 4 with three, and a single one with the full dataset. For each combination, we predict α and $[TOTAL]_{ref}$ (Figure 4.7). We see that, as expected, the accuracy of our inference improves as additional bias current settings are used to collect the measurements. However, if necessary, reasonable accuracy can be achieved even with $2-3$ sensors, especially if the total concentration is of primary interest. Overall, the differential sensitivity and selectivity of the device which results from the use of multiple bias currents can be exploited for improved discrimination of gas mixtures.

4.8 Discussion

Modern diesel engines require continuous monitoring of NO_x and NH_3 levels in tailpipe emissions to ensure regulatory compliance and minimize air pollution. Au/YSZ/Pt mixed-potential sensors represent a promising technology which could address this need. The sensors are compact and rugged (Figure 4.1) and, moreover, their sensitivity and selectivity toward target gases can be tuned by applying a bias current (Figure 4.2). Thus a single sensor under different bias current settings can be used to create an entire array of sensors, yielding more robust and accurate predictions of gas concentrations in mixtures. Unfortunately, currently available zirconia-based electrochemical sensors exhibit cross-interference with hydrocarbons such as propylene (C_3H_6) (Figure 4.2d). Thus sensor voltage readings are not simply proportional to the amount of target gas in diesel exhaust.

Here we show that target gas concentrations can nonetheless be inferred from sensor readouts if data is collected under several bias current settings. We focus on mixtures of NH_3 (target gas) and C_3H_6 (interfering gas), and develop a Bayesian framework for predicting α, the ratio of concentrations between the two gases, and the sum of the two concentrations. This information is equivalent to predicting absolute concentration of each gas. Our Bayesian approach allows us to build an ensemble of models that explain the data (rather than a single maximum-likelihood model), and thus evaluate both the mean value of each parameter of interest and its uncertainty. The method is not limited to Au/YSZ/Pt mixed-potential sensors and can be employed to infer concentrations for more than two gases simultaneously. We expect our approach to be widely applicable in automotive as well as other industrial applications which rely on the accurate knowledge of gas concentrations in complex chemical mixtures.

The simplest model we have investigated assumes that gases in the mixture do not interefere with one another, and hence the total voltage is a sum of contributions from each individual gas. We train the parameters of this linear model using sensor readings from individual gases, and then apply it to predict gas concentrations in mixtures. The linear model is easy to set up and calibrate, and has few fitting parameters. Its accuracy does not depend on the relative proportions of gases in the mixture. However, it neglects

cross-interference between gases at higher gas concentrations, which causes nonlinearities in sensor response (Figures 4.3c, E.3c, E.4c and E.5c).

In order to check whether modeling this effect improves the accuracy of our predictions, we have built a nonlinear model as a straightforward generalization of the linear model (cf. eqs. (4.2) and (4.6)). The nonlinear model has fitting parameters that are explicit functions of α (see § 4.4.2), and therefore must be trained on data that uniformly covers the range of mixtures to be encountered in subsequent applications. This makes the training procedure more data-intensive, and limits the applicability of the model. In addition, the model is not fully Bayesian and as a result needs to be cross-validated. The degree of polynomials used in curve fitting likely depends on the application (sensor types, range of α, etc.) and needs to be re-examined in each case. Despite these difficulties, prediction accuracy improves when nonlinearities are taken into account, which justifies using the nonlinear model in situations where it can be first calibrated in a laboratory setting using extensive training datasets, and then deployed in the field. Interestingly, the nonlinear model captures cross-interference between gases without explicit knowledge of reaction activity at molecular level.

Using our sensors and models in real-world conditions will be greatly facilitated by making as few measurements as possible. To check what the minimal requirements are for the deployment of our system, we have carried out four tests in which we reduce the number of test datapoints in various ways and monitor prediction accuracy of the nonlinear model. First, we kept all 4 bias current settings and the entire available range of total concentrations of NH_3 and C_3H_6, but reduced the number of datapoints for each given total concentration and bias current setting. We found that, encouragingly, just $1-2$ datapoints are necessary, which should improve response times. More crucially, our algorithm depends on the availability of data for several total concentrations, which in practice entails diluting the mixture of interest in a neutral buffer. Since this step adds complexity to the system, we checked the minimum number of total concentrations that are necessary for acceptably accurate predictions. It appears that the number of dilutions can be reduced by judicious choice of neutral buffer volumes, especially if only α or the total concentration are of interest.

Next, we tested the robustness of our framework with respect to extrapolating to the values of α which do not appear in training data. We found that prediction accuracy is

sensitive to such extrapolations, which argues for carefully choosing the α range in the calibration step. Finally, we checked whether the number of sensors (i.e., bias current settings) can be reduced below four. We saw that although adding sensors to the array is clearly beneficial, it may be acceptable to use $2-3$ or even a single sensor without too much accuracy loss. Overall, we conclude that it is possible to achieve acceptable accuracy levels, system complexity, and reaction times with our sensors and algorithms.

Appendices

Appendix A

Nested Sampling

The information contained in this Appendix is meant to serve as a brief summary of the nested sampling approach. For a complete treatment please refer to [42].

A.1 Motivation

Bayes' theorem tells us the relationship between the posterior probability density, $P(\boldsymbol{\theta}|\boldsymbol{D}, H)$, the likelihood function, $P(\boldsymbol{D}|\boldsymbol{\theta}, H)$, and the prior probability density, $P(\boldsymbol{\theta}|H)$ of a set of parameters, $\boldsymbol{\theta}$:

$$P(\boldsymbol{\theta}|\boldsymbol{D}, H) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta}, H)P(\boldsymbol{\theta}|H)}{P(\boldsymbol{D}|H)} \tag{A.1}$$

where $\boldsymbol{D}$ is the observed data and H is the underlying hypothesis, or model. The denominator on the right side of eq. (A.1) is often referred to as the *evidence*($\equiv \mathcal{Z}$), or as the *marginal likelihood* (which describes how it is obtained, see below). This quantity is overlooked in many parameter estimation problems as simply the normalization constant since it does not explicitly depend on $\boldsymbol{\theta}$. However, when we are concerned with model selection (which, it can be argued, should always be the case), this term takes center stage:

$$R = \frac{P(H_2|\boldsymbol{D})}{P(H_1|\boldsymbol{D})} = \frac{P(\boldsymbol{D}|H_2)P(H_2)}{P(\boldsymbol{D}|H_1)P(H_1)} = \frac{\mathcal{Z}_2 P(H_2)}{\mathcal{Z}_1 P(H_1)} \tag{A.2}$$

Here, R is the ratio of posterior probabilities of two competing models, H_1 and H_2 and $P(H_2)/P(H_1)$ is the ratio of the two priors. The latter can usually be set to 1 since there is no obvious reason to prefer one model over another *a priori*. In general, the evidence will be higher for a model in which much of its parameter space has associated with it high likelihood values, and smaller for a model if large areas of its parameter space are unlikely.

In light of the significance of Bayesian evidence in model selection (eq. (A.2)), $\mathcal{Z}$ joins the posterior ($P(\boldsymbol{\theta})$, in compressed notation) as a desired output of eq. (A.1), the inputs

being the likelihood and prior, $\mathcal{L}(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$, respectively. Traditionally, techniques such as Markov Chain Monte Carlo (MCMC) sampling have been designed for the posterior and used to estimate parameters of a model given data. However, convergence to stationarity becomes incredibly slow when the posterior distribution is highly multi-modal. Moreover, calculation of the evidence poses a computational challenge that is typically an order of magnitude higher than that of parameter estimation since it involves a multi-dimensional integration over the prior density:

$$\mathcal{Z} = \int d^k\boldsymbol{\theta}\mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{A.3}$$

Nested sampling mitigates this challenge by statistically "sorting" the points $\{\theta\}$ in parameter space by their likelihood values[1] and then summing these likelihoods to find the evidence. This avoids the multi-dimensional integral of eq. (A.3). The posterior is then obtained as a by-product of this procedure, as described in the following sections.

A.2 Implementation

A.2.1 Sorting

Introduce the function

$$x(\lambda) = \int_{\mathcal{L}(\boldsymbol{\theta})>\lambda} \pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{A.4}$$

defined as the proportion of cumulant prior with likelihood greater than λ. As λ increases, the restriction on likelihood becomes tighter. Therefore, x is a decreasing function of λ. Note that $x_{min} = 0$ at $\lambda = \mathcal{L}_{max}$ (if a global maximum exists). At $\lambda = \mathcal{L}_{min} \geq 0$, $x_{max} = 1$ since in this case we are describing the cumulant mass of the entire prior and must satisfy the normalization requirement for probability distributions:

$$\int \pi(\boldsymbol{\theta})d^k\boldsymbol{\theta} = 1. \tag{A.5}$$

We can look at the inverse of eq. (A.4), namely, $\mathcal{L}(x(\lambda)) \equiv \lambda$, thus making a transformation from $\mathcal{L}(\boldsymbol{\theta})$, with vector argument $\boldsymbol{\theta}$, to $\mathcal{L}(x)$ which has a scalar argument. Since

[1] In order to resolve ties between points of equal likelihood, a small amount of random noise can be added to $\mathcal{L}$. This imposes a ranking even if none existed before.

$dx = \pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, eq. (A.3) tells us that the evidence is simply a sum over these elements and can be computed using $\mathcal{L}(x)$:

$$\mathcal{Z} = \int_0^1 \mathcal{L}(x)dx. \tag{A.6}$$

The sorted likelihood function $\mathcal{L}(x)$ is also the key to finding the posterior:

$$P(x) = \frac{\mathcal{L}(x)}{\mathcal{Z}}. \tag{A.7}$$

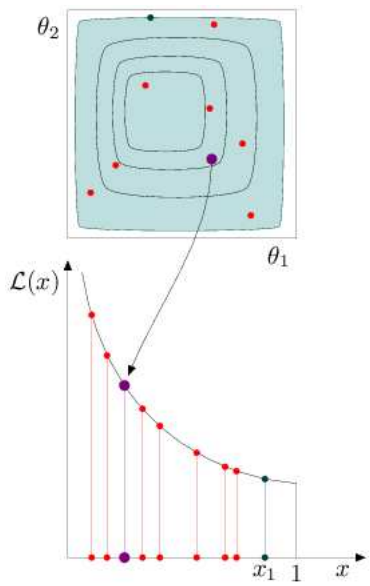


Figure A.1: **Nested sampling.** Adapted from [80]. A mapping of a contour $\mathcal{L}(\boldsymbol{\theta})$ to $\mathcal{L}(x)$.

A.2.2 The nested sampling technique

In practice, nested sampling tabulates the sorted likelihood function by sampling n objects $\boldsymbol{\theta}$ (usually by MC methods) from the prior π subject to an evolving constraint $\mathcal{L}(\boldsymbol{\theta}) > \mathcal{L}^*$. Initially, $\mathcal{L}^* = 0$ to ensure complete coverage. In terms of x this is equiavalent to sampling uniformly and subject to the constraint $x < x^*$, where x^* corresponds to $\mathcal{L}^*$ (Figure A.1). At each iteration, the move is inwards in x and upwards in $\mathcal{L}$ with the goal of finding and quantifying the small region of high likelihood where most of the posterior lies. This is done by replacing $\{x^*, \mathcal{L}^*\}$ with the worst object's $\{x, \mathcal{L}\}$ before that worst object is discarded. There are now $n - 1$ surviving objects, still uniformly distributed in x but confined to a shrunken domain (characterized by the new x^*) which is nested within the old. The

shrinkage ratio is distributed as $P(t) = nt^{n-1}$, with statistics $log(t) = (-1 \pm 1)/n$. The final step is to generate a replacement object within the nested domain. At the kth iterate, $\mathcal{L}_k = \mathcal{L}^*$ and $x_k = x^* = \prod_{j=1}^{k} t_j$. So we have a crude approximation of $x_k = e^{-k/n}$ (ignoring uncertainty) and the sequence $\mathcal{L}(x)$ is now available - no sorting is needed at all!

The evidence in eq. (A.6) can now be estimated by any convenient numerical recipe as a weighted sum of the likelihoods:

$$\mathcal{Z} \approx \sum_k w_k \mathcal{L}_k, \tag{A.8}$$

where $w_k \approx \Delta x$ is the width associated with each object in the sequence. The simplest assignment of the width is $w_k = x_{k-1} - x_k$. The estimation of $\mathcal{Z}$ can be slightly improved with the trapezoidal rule, or similar.

In addition to the evidence and posterior, the *information*, or negative entropy can be found:

$$\mathcal{H} = \int P(x) log[P(x)] dx \tag{A.9}$$

The information measure provides a guideline for termination of iterations: the nested sampling calculation should be continued until most of $\mathcal{Z}$ has been found. The bulk of this area occurs in the region of $x \approx e^{-\mathcal{H}}$ which suggests that iteration should continue until the count k significantly exceeds $n\mathcal{H}$.

Appendix B

Additional Figures and Tables, Chapter 2

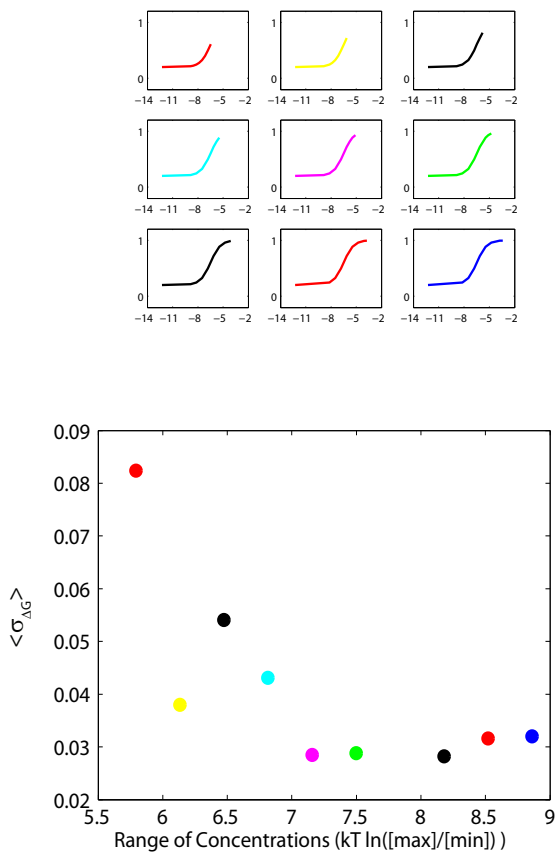


Figure B.1: **Prediction accuracy increases with the number of measurements at different total concentrations.** Synthetic data was generated using eq. (2.1), with $A = 0.8$, $b = 0.2$, and $\Delta G = -6.5$ kcal/mol. To account for experimental error, Gaussian noise with $\tilde{\sigma} = 0.02$ was added to the intensity from eq. (2.1). The maximum total concentration of the ligand was gradually increased as shown in the nine panels on top, yielding more and more complete binding curves: $\log_{10}[\text{max}] = \{-4.75, -4.5, -4.25, -4.0, -3.75, -3.5, -3.0, -2.75, -2.5\}$. $\log_{10}[\text{min}]$ was -9.0 in all cases, and 4 replicates with 9 datapoints per curve were created for each concentration range. In each panel I^l is plotted as a function of μ^l in the absence of noise. For each concentration range, 1000 nested sampling runs were carried out to predict ΔG, A and b. The standard deviation $\sigma_{\Delta G}$ from each run was averaged and plotted in the bottom panel as a function of the total range of ligand concentrations $k_B T \ln([\text{max}]/[\text{min}])$. Each dot in the bottom panel is color-coded to correspond to a particular binding curve on top.

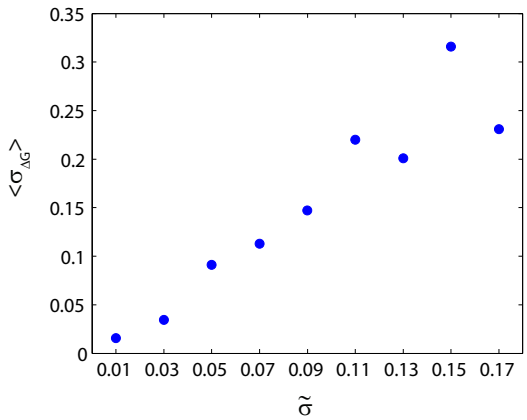
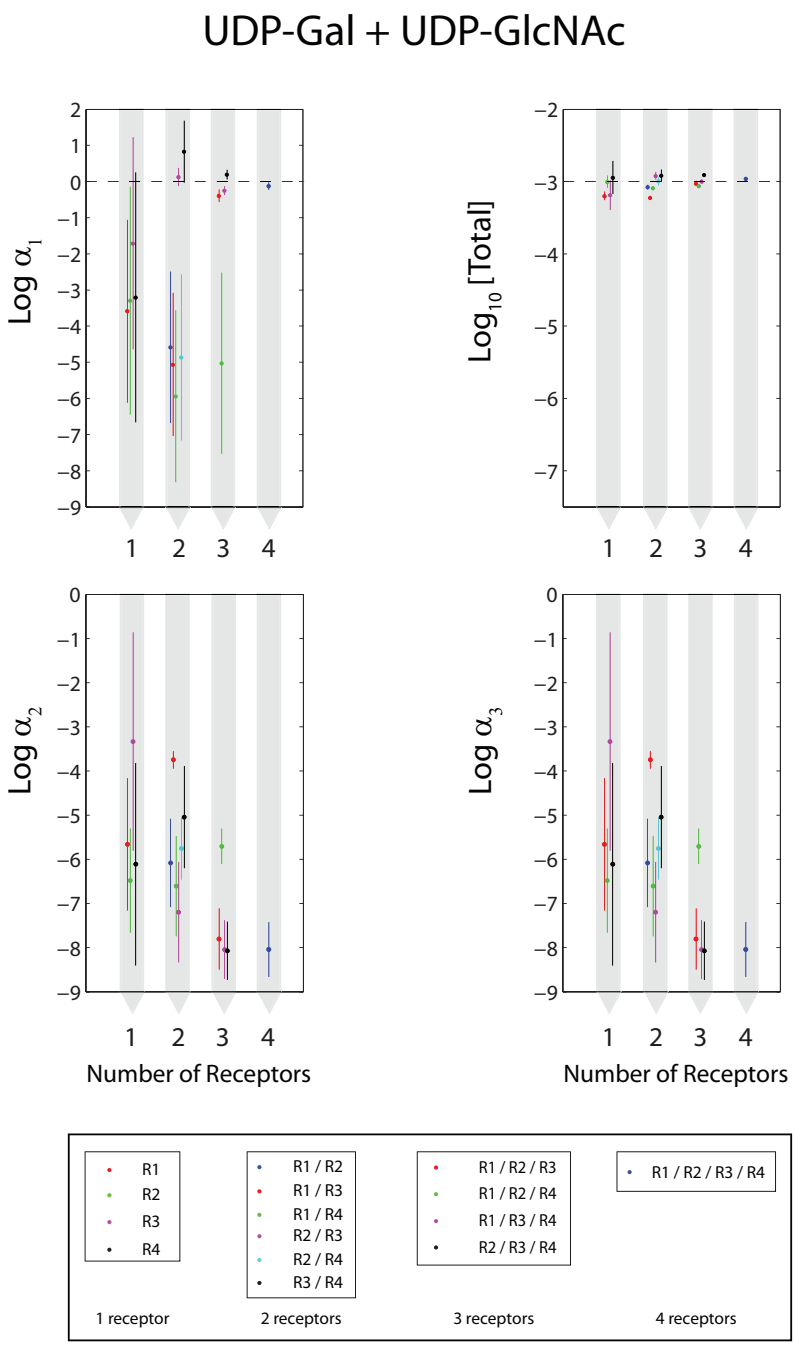
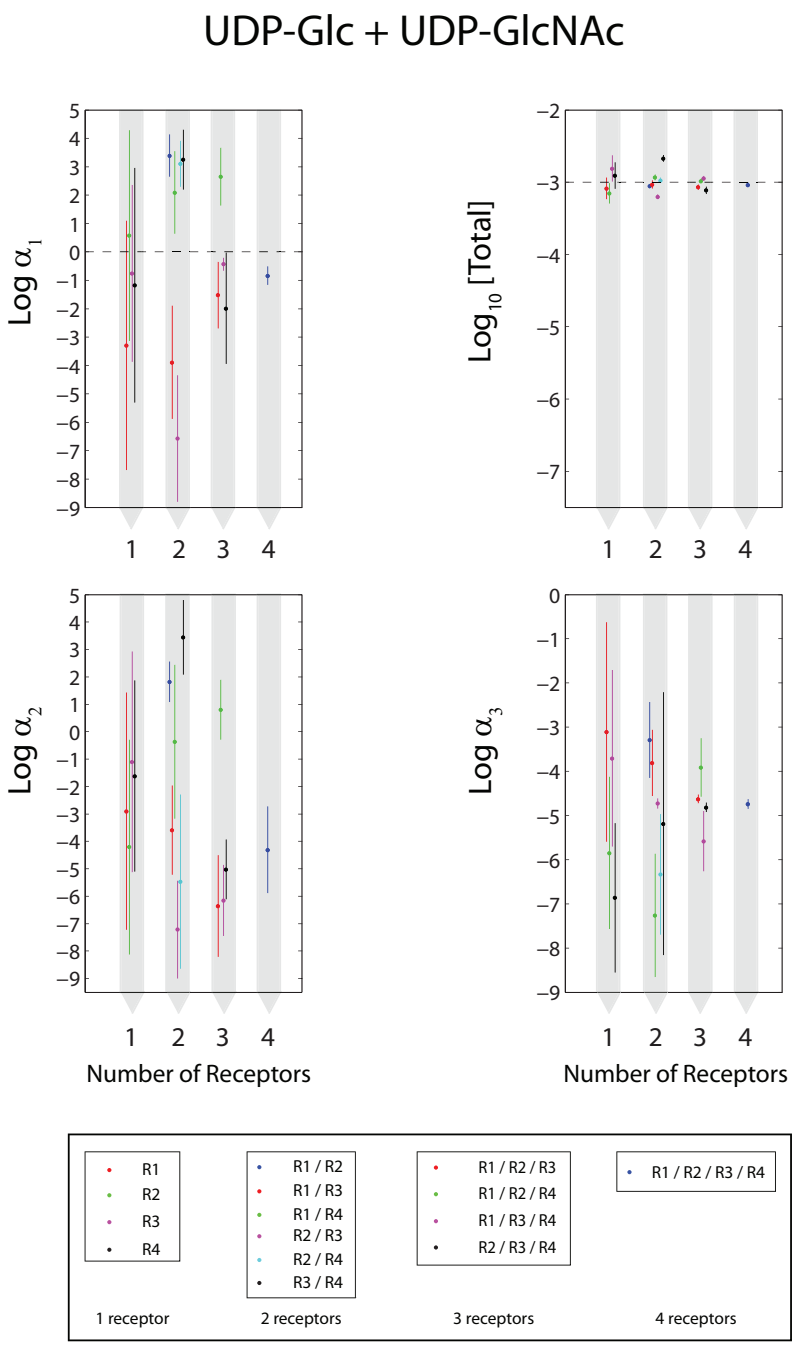
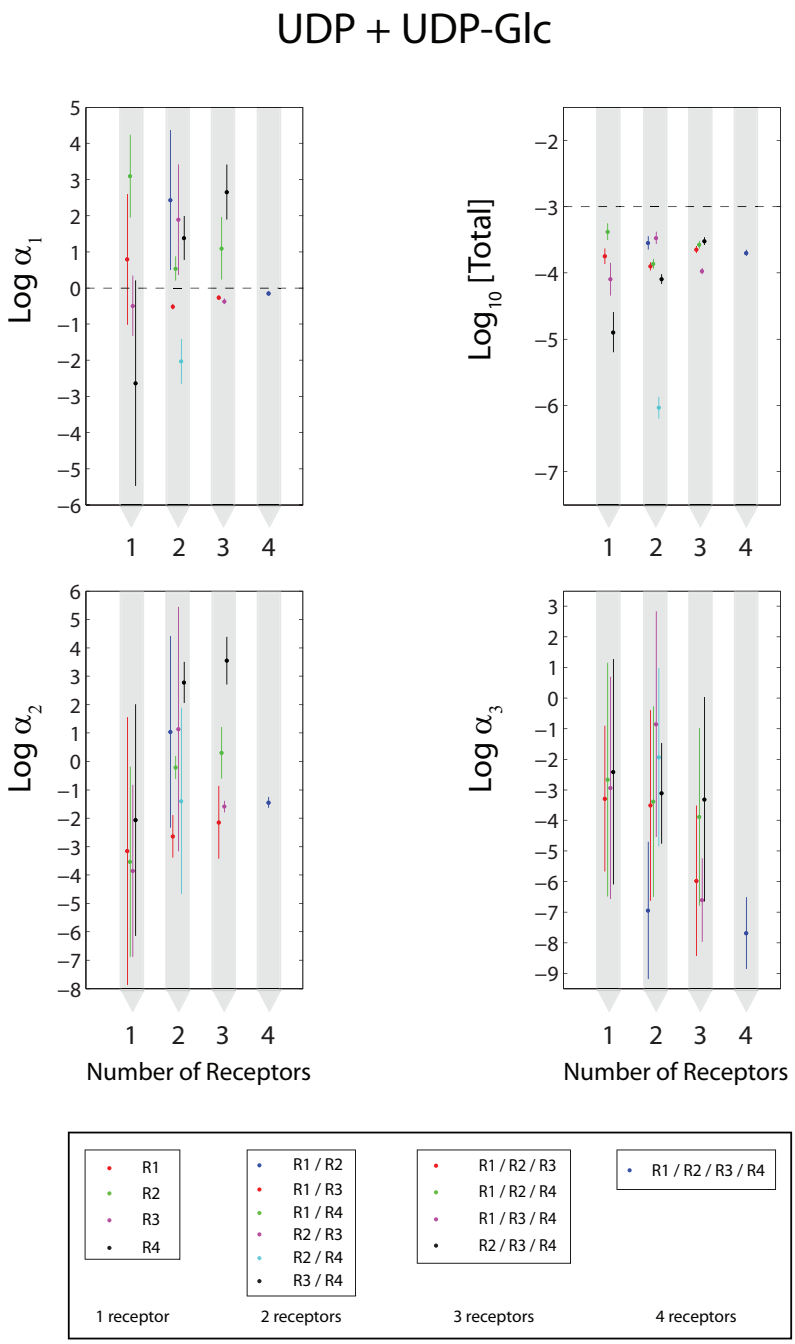
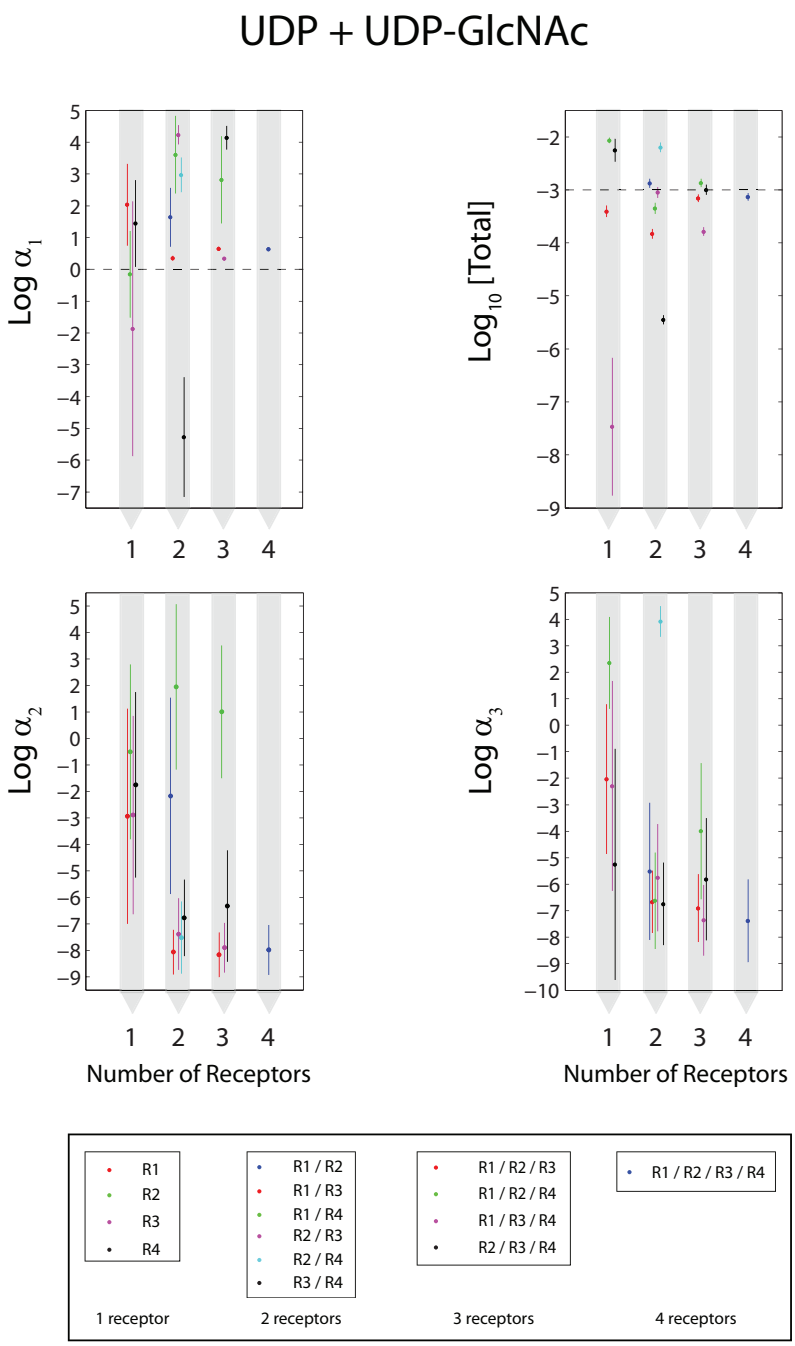


Figure B.2: **Prediction accuracy decreases with the amount of noise in the data.** Synthetic data was generated using eq. (2.1), with $A = 0.8$, $b = 0.2$, and $\Delta G = -6.5$ kcal/mol. In analogy with the experiments, we used the concentration range $\log_{10} n^l = \{-3.0, -3.5, \ldots, -6.5, -9.0\}$ and created 4 replicates, yielding 36 datapoints. To model the increase in experimental error, Gaussian noise with $\tilde{\sigma}$ ranging from 0.01 to 0.17 was added to the intensity from eq. (2.1). For each value of $\tilde{\sigma}$, 1000 nested sampling runs were carried out to predict ΔG, A and b. The standard deviation $\sigma_{\Delta G}$ from each run was averaged and plotted as a function of $\tilde{\sigma}$.


UDP-Gal + UDP-GlcNAc
Log α1
Log10 [Total]
Log α2
Log α3
Number of Receptors
Number of Receptors
R1
R2
R3
R4
R1 / R2
R1 / R3
R1 / R4
R2 / R3
R2 / R4
R3 / R4
R1 / R2 / R3
R1 / R2 / R4
R1 / R3 / R4
R2 / R3 / R4
R1 / R2 / R3 / R4
1 receptor
2 receptors
3 receptors
4 receptors


UDP-Glc + UDP-GlcNAc
Log α1
Log10 [Total]
Log α2
Log α3
Number of Receptors
Number of Receptors
R1
R2
R3
R4
R1 / R2
R1 / R3
R1 / R4
R2 / R3
R2 / R4
R3 / R4
R1 / R2 / R3
R1 / R2 / R4
R1 / R3 / R4
R2 / R3 / R4
R1 / R2 / R3 / R4
1 receptor
2 receptors
3 receptors
4 receptors


UDP + UDP-Glc
Log α1
Log10 [Total]
Log α2
Log α3
Number of Receptors
Number of Receptors
R1
R2
R3
R4
R1 / R2
R1 / R3
R1 / R4
R2 / R3
R2 / R4
R3 / R4
R1 / R2 / R3
R1 / R2 / R4
R1 / R3 / R4
R2 / R3 / R4
R1 / R2 / R3 / R4
1 receptor
2 receptors
3 receptors
4 receptors


UDP + UDP-GlcNAc
Log α1
Log10 [Total]
Log α2
Log α3
Number of Receptors
Number of Receptors
R1
R2
R3
R4
R1 / R2
R1 / R3
R1 / R4
R2 / R3
R2 / R4
R3 / R4
R1 / R2 / R3
R1 / R2 / R4
R1 / R3 / R4
R2 / R3 / R4
R1 / R2 / R3 / R4
1 receptor
2 receptors
3 receptors
4 receptors

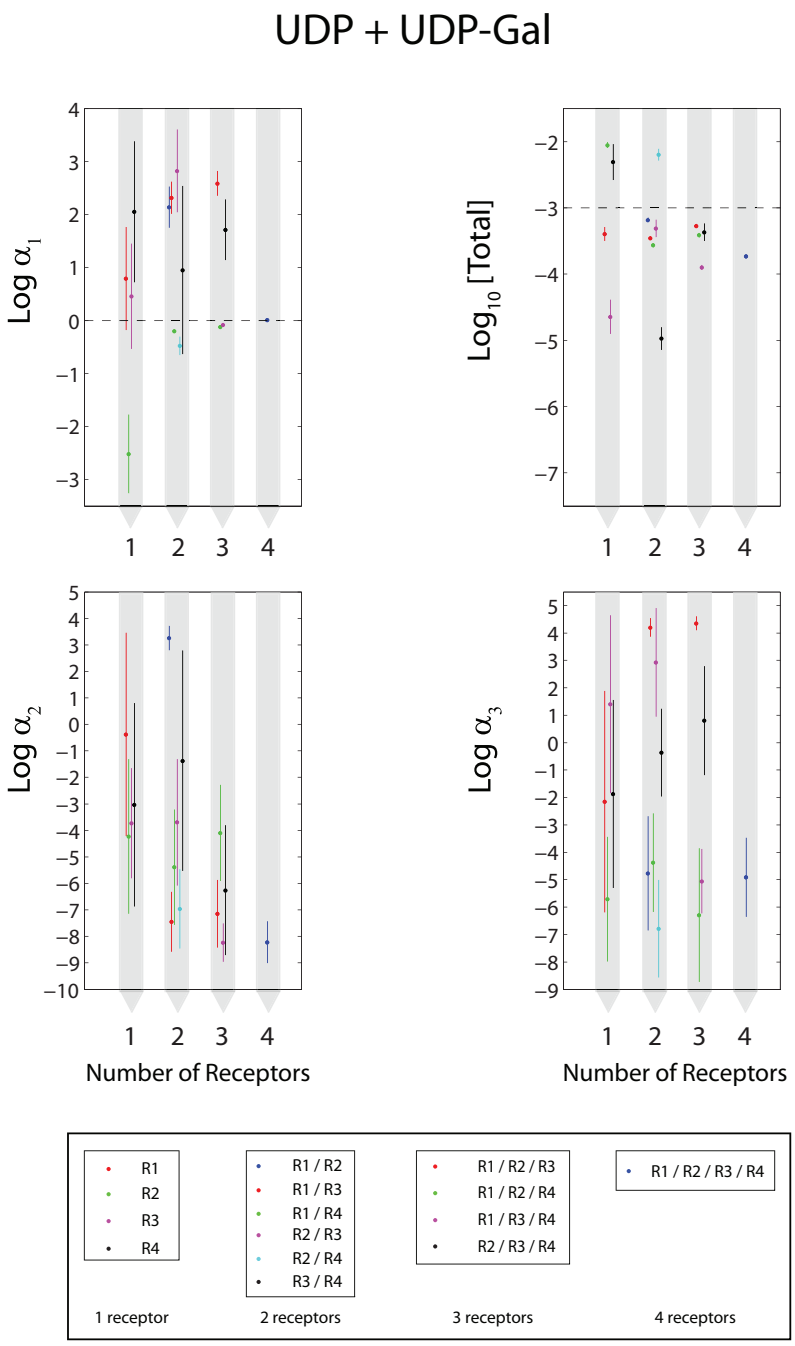


Figure B.3: **Inference of ligand concentrations is improved with the number of receptors interrogating the mixture.** Shown on the log-scale are means and standard deviations for three relative concentrations (α_1, α_2, α_3) and the total concentration, predicted by RANSA using the four-ligand model and up to four receptors: H-20 (R1), K-3 (R2), L-3 (R3), 2211 (R4). Each experiment has a 50-50 binary mixture of two ligands indicated on top of each panel, leading to $\alpha_1 = 1$, $\alpha_2 = \alpha_3 = 0$, and $\log_{10}[\text{Total}] = -3$ at the reference point ([Total] = [UDP-Glc] + [UDP-GlcNAc] + [UDP-Gal] + [UDP]). **UDP-Gal+UDP-GlcNAc mixture:** α_1= [UDP-GlcNAc]/[UDP-Gal], α_2= [UDP-Glc]/[UDP-Gal], α_3= [UDP]/[UDP-Gal]. **UDP-Glc+UDP-GlcNAc mixture:** α_1= [UDP-GlcNAc]/[UDP-Glc], α_2= [UDP-Gal]/[UDP-Glc], α_3= [UDP]/[UDP-Glc]. **UDP+UDP-Glc mixture:** α_1= [UDP]/[UDP-Glc], α_2= [UDP-Gal]/[UDP-Glc], α_3= [UDP-GlcNAc]/[UDP-Glc]. **UDP+UDP-GlcNAc mixture:** α_1= [UDP]/[UDP-GlcNAc], α_2= [UDP-Glc]/[UDP-GlcNAc], α_3= [UDP-Gal]/[UDP-GlcNAc]. **UDP+UDP-Gal mixture:** α_1= [UDP]/[UDP-Gal], α_2= [UDP-Glc]/[UDP-Gal], α_3= [UDP-GlcNAc]/[UDP-Gal].

Table B.1: **Parameters of receptor-ligand interactions predicted from one-receptor, one-ligand binding curves (UDP-Gal, UDP-Glc, UDP-GlcNAc) and one-receptor, two-ligand binding curves (UDP).** ΔG is the receptor-ligand binding free energy (kcal/mol), A is the receptor efficacy, b is the background intensity, and $\tilde{\sigma}$ is the noise parameter which quantifies the discrepancy between the model and the observed binding curves. Due to antagonistic activity of UDP, 50/50 UDP+UDP-Glc binary mixture was used with K-3, L-3, 2211 and 50/50 UDP+UDP-Gal binary mixture was used with H-20 to predict UDP parameters (compound concentrations were set to their exact values for these calibration predictions). In each case, the mixture was chosen on the basis of the smallest standard deviation of ΔG.

		UDP (L1)	**UDP-Gal (L2)**	**UDP-Glc (L3)**	**UDP-GlcNAc (L4)**
H-20 (R1)	ΔG	$-7.42 \pm .05$	$-6.69 \pm .04$	$-6.29 \pm .06$	$-6.16 \pm .06$
	A	$.008 \pm .009$	$.994 \pm .005$	$.74 \pm .01$	$.51 \pm .01$
	b	$.006 \pm .005$	$.015 \pm .009$	$.014 \pm .009$	$.008 \pm .006$
	$\tilde{\sigma}$	$.036 \pm .004$	$.041 \pm .005$	$.037 \pm .005$	$.029 \pm .004$
K-3 (R2)	ΔG	$-8.2 \pm .2$	$-5.64 \pm .07$	$-5.36 \pm .04$	$-5.36 \pm .06$
	A	$.18 \pm .01$	$.81 \pm .02$	$.94 \pm .02$	$.55 \pm .01$
	b	$.10 \pm .02$	$.11 \pm .01$	$.069 \pm .008$	$.111 \pm .006$
	$\tilde{\sigma}$	$.022 \pm .003$	$.043 \pm .005$	$.030 \pm .004$	$.022 \pm .003$
L-3 (R3)	ΔG	$-7.4 \pm .1$	$-6.21 \pm .07$	$-6.6 \pm .1$	$-5.60 \pm .06$
	A	$.02 \pm .02$	$.85 \pm .03$	$.83 \pm .03$	$.43 \pm .01$
	b	$.01 \pm .01$	$.02 \pm .01$	$.04 \pm .02$	$.007 \pm .004$
	$\tilde{\sigma}$	$.064 \pm .009$	$.049 \pm .007$	$.070 \pm .009$	$.018 \pm .002$
2211 (R4)	ΔG	$-10.3 \pm .3$	$-7.69 \pm .09$	$-8.48 \pm .08$	$-8.00 \pm .09$
	A	$.010 \pm .009$	$.82 \pm .04$	$.85 \pm .02$	$.89 \pm .04$
	b	$0.05 \pm .01$	$.094 \pm .008$	$.08 \pm .01$	$.11 \pm .01$
	$\tilde{\sigma}$	$.019 \pm .002$	$.035 \pm .005$	$.046 \pm .006$	$.044 \pm .006$

Table B.2: **Prediction of ligand concentrations in equal-proportion mixtures (data for Figure 2.3).**

Ligand(s) in mixture		**[L1]**	**[L2]**	**[L3]**	**[L4]**	**[Total]**
L1	predicted	$2.8 \pm .7$	$.01 \pm .01$	$.001 \pm .001$	$.10 \pm .08$	2.9 ± 0.8
	actual	1.0	0	0	0	1.0
L2	predicted	$.0002 \pm .0001$	$1.05 \pm .04$	$.006 \pm .008$	$.004 \pm .009$	$1.06 \pm .03$
	actual	0	1.0	0	0	1.0
L3	predicted	$.0003 \pm .0002$	$.01 \pm .01$	$.96 \pm .04$	$.003 \pm .009$	$.97 \pm .04$
	actual	0	0	1.0	0	1.0
L4	predicted	$2 \times 10^{-4} \pm 6 \times 10^{-6}$	$.0004 \pm .0005$	$.0008 \pm .0007$	$1.05 \pm .04$	$1.05 \pm .04$
	actual	0	0	0	1.0	1.0
L1 + L2	predicted	$.093 \pm .007$	$.092 \pm .007$	$4 \times 10^{-5} \pm 4 \times 10^{-5}$	$.001 \pm .002$	$.19 \pm .01$
	actual	.5	.5	0	0	1.0
L1 + L3	predicted	$.082 \pm .009$	$.023 \pm .004$	$.09 \pm .01$	$.000 \pm .001$	$.20 \pm .02$
	actual	.5	0	.5	0	1.0
L1 + L4	predicted	$.48 \pm .08$	$.0001 \pm .0002$	$.001 \pm .001$	$.26 \pm .05$	0.7 ± 0.1
	actual	.5	0	0	.5	1.0
L2 + L3	predicted	$.006 \pm .001$	$.51 \pm .03$	$.52 \pm .03$	$.001 \pm .003$	$1.03 \pm .04$
	actual	0	.5	.5	0	1.0
L2 + L4	predicted	$.0002 \pm .0002$	$.58 \pm .04$	$.002 \pm .004$	$.51 \pm .04$	$1.10 \pm .06$
	actual	0	.5	0	.5	1.0
L3 + L4	predicted	$.0054 \pm .0008$	$.02 \pm .01$	$.62 \pm .07$	$.28 \pm .06$	$.92 \pm .06$
	actual	0	0	.5	.5	1.0
L1 + L2 + L3	predicted	$.049 \pm .006$	$.12 \pm .01$	$.033 \pm .008$	$.06 \pm .01$	$.26 \pm .02$
	actual	.33	.33	.33	0	1.0
L1 + L2 + L4	predicted	$.08 \pm .02$	$.027 \pm .006$	$1 \times 10^{-5} \pm 3 \times 10^{-5}$	$.36 \pm .04$	$.46 \pm .04$
	actual	.33	.33	0	.33	1.0
L1 + L3 + L4	predicted	$.064 \pm .007$	$.038 \pm .005$	$.068 \pm .007$	$.003 \pm .007$	$.17 \pm .02$
	actual	.33	0	.33	.33	1.0
L2 + L3 + L4	predicted	$.0049 \pm .0009$	$.25 \pm .03$	$.46 \pm .04$	$.02 \pm .02$	$.73 \pm .04$
	actual	0	.33	.33	.33	1.0
All Four	predicted	$.063 \pm .007$	$.20 \pm .02$	$.008 \pm .008$	$.20 \pm .02$	$.45 \pm .04$
	actual	.25	.25	.25	.25	1.0

Table B.3: **Prediction of [UDP-Gal] and [UDP-Glc] in unequal proportion mixtures (data for Figure 2.4).**

[UDP-Glc]/[UDP-Gal]		α_1	α_2	α_3	$\log_{10}$[Total]
90/10	predicted	11 ± 9	5 ± 4	$.001 \pm .002$	$-2.87 \pm .03$
	actual	9	0	0	-3
80/20	predicted	$2.3 \pm .6$	0.9 ± 0.3	$.0004 \pm .0002$	$-2.89 \pm .03$
	actual	4	0	0	-3
60/40	predicted	$.71 \pm .09$	$.01 \pm .02$	$.0010 \pm .0007$	$-2.95 \pm .03$
	actual	1.5	0	0	-3
40/60	predicted	0.8 ± 0.2	$.05 \pm .09$	$.0005 \pm .0004$	$-2.91 \pm .02$
	actual	0.66	0	0	-3
20/80	predicted	0.7 ± 0.1	$.05 \pm .05$	$.0026 \pm .0007$	$-2.83 \pm .02$
	actual	0.25	0	0	-3
10/90	predicted	$.20 \pm .07$	$.01 \pm .02$	$.0031 \pm .0007$	$-2.85 \pm .03$
	actual	0.11	0	0	-3

Table B.4: **Prediction of [UDP-Gal] and [UDP-Glc] in unequal proportion mixtures using an alternative definition of relative concentrations.** We used RANSA of a four-receptor, four-ligand model to infer relative concentrations α_1 = [UDP-Gal]/[UDP-Glc], α_2 = [UDP-GlcNAc]/[UDP-Glc], α_3 = [UDP]/[UDP-Glc], as well as the total concentration [Total] = [UDP-Gal] + [UDP-Glc] + [UDP-GlcNAc] + [UDP] at the 1 mM reference point. A's and b's were refit to account for "plate bias": small systematic deviations in the values of A and b (from the standard values shown in Table B.1 and used everywhere else) between different plates. For plate 1 (measurements 1-3), $A_{\mathrm{UDP-Gal}}$, $A_{\mathrm{UDP-Glc}}$ and b were set to $\{0.94, 0.79, 0.64, 0.50\}$, $\{0.91, 0.81, 0.90, 0.81\}$ and $\{0.02, 0.15, 0.02, 0.12\}$ for H-20, K-3, L-3 and 2211, respectively. For plate 2 (measurements 4-6), the corresponding values were $\{0.96, 0.93, 0.94, 0.94\}$, $\{0.85, 0.79, 0.86, 0.95\}$ and $\{0.02, 0.13, 0.03, 0.11\}$. $A_{\mathrm{UDP-GlcNAc}}$ and A_{UDP} were taken from Table B.1.

[UDP-Gal]/[UDP-Glc]		$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\alpha_3}$	$\mathbf{log_{10}[Total]}$
10/90	predicted	$.48 \pm .09$	0.5 ± 0.1	$.0002 \pm .0001$	$-2.94 \pm .03$
	actual	.11	0	0	-3
20/80	predicted	$1.0 \pm .2$	0.4 ± 0.3	$.0003 \pm .0002$	$-2.94 \pm .04$
	actual	.25	0	0	-3
40/60	predicted	$3.2 \pm .4$	$.02 \pm .04$	$.003 \pm .002$	$-2.95 \pm .03$
	actual	.66	0	0	-3
60/40	predicted	3.3 ± 0.8	$.01 \pm .03$	$.0003 \pm .0002$	$-2.89 \pm .02$
	actual	1.5	0	0	-3
80/20	predicted	7 ± 2	$.02 \pm .03$	$.003 \pm .002$	$-2.80 \pm .03$
	actual	4	0	0	-3
90/10	predicted	54 ± 25	0.0 ± 0.1	$.03 \pm .03$	$-2.78 \pm .03$
	actual	9	0	0	-3

Appendix C

Hessian analysis, details

Hessian matrix as a measure of sensitivity to model parameters. Differentiating log-likelihood (eq. (2.8)) twice with respect to model parameters yields:

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma_i \partial \gamma_j} = -\sum_{k=1}^{N_{rec}} \frac{1}{\tilde{\sigma}_k^2} \sum_{l=1}^{N_k} \left\{ \frac{\partial I_k^l}{\partial \gamma_i} \frac{\partial I_k^l}{\partial \gamma_j} + \left[I_k^l(\{\gamma\}) - \tilde{I}_k^l \right] \frac{\partial^2 I_k^l}{\partial \gamma_i \partial \gamma_j} \right\}, \tag{C.1}$$

where $\{\gamma\} = (\{\alpha\}, \mu)$. The second term can be omitted in the low-noise limit ($I_k^l(\{\alpha\}, \mu) - \tilde{I}_k^l \sim \tilde{\sigma}_k \to 0$), yielding a standard expression for the Hessian matrix [44]:

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma_i \partial \gamma_j} = -\sum_{k=1}^{N_{rec}} \frac{1}{\tilde{\sigma}_k^2} \sum_{l=1}^{N_k} \frac{\partial I_k^l}{\partial \gamma_i} \frac{\partial I_k^l}{\partial \gamma_j}, \tag{C.2}$$

Explicitly,

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \alpha_i \partial \alpha_j} &= -\sum_{k=1}^{N_{rec}} \frac{1}{\tilde{\sigma}_k^2} \sum_{l=1}^{N_k} \left(\frac{p_{i+1}^{k,l}}{\alpha_i} (A_{i+1}^k - I_k^l) - \frac{I_k^l}{\mathcal{Z}^{k,l} S} \right) \left(\frac{p_{j+1}^{k,l}}{\alpha_j} (A_{j+1}^k - I_k^l) - \frac{I_k^l}{\mathcal{Z}^{k,l} S} \right), \\
\frac{\partial^2 \mathcal{L}}{\partial \mu^2} &= -\beta^2 \sum_{k=1}^{N_{rec}} \frac{1}{\tilde{\sigma}_k^2} \sum_{l=1}^{N_k} \left(\frac{I_k^l}{\mathcal{Z}^{k,l}} \right)^2, \\
\frac{\partial^2 \mathcal{L}}{\partial \mu \partial \alpha_i} &= -\beta \sum_{k=1}^{N_{rec}} \frac{1}{\tilde{\sigma}_k^2} \sum_{l=1}^{N_k} \left(\frac{I_k^l}{\mathcal{Z}^{k,l}} \right) \left(\frac{p_{i+1}^{k,l}}{\alpha_i} (A_{i+1}^k - I_k^l) - \frac{I_k^l}{\mathcal{Z}^{k,l} S} \right).
\end{aligned} \tag{C.3}$$

Uncertainties σ_i for each predicted parameter γ_i are given by the diagonal elements of the inverse Hessian matrix:

$$\sigma_i^2 = -\| \frac{\partial^2 \mathcal{L}}{\partial \gamma_i \partial \gamma_j} \|_{ii}^{-1}. \tag{C.4}$$

C.1 Hessian analysis of a mixture of two ligands interacting with a single receptor

For a single receptor and two ligands we obtain from eq. (C.3) (omitting the receptor index and setting $\alpha_1 = \alpha,\ \tilde{\sigma}^2 = 1$ for convenience):

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} &= -\sum_{l=1}^{N} \left(\frac{p_2^l}{\alpha}(A_2 - I^l) - \frac{I^l}{\mathcal{Z}^l(1+\alpha)} \right)^2, \\
\frac{\partial^2 \mathcal{L}}{\partial \mu^2} &= -\beta^2 \sum_{l=1}^{N} \left(\frac{I^l}{\mathcal{Z}^l} \right)^2, \\
\frac{\partial^2 \mathcal{L}}{\partial \mu \partial \alpha} &= -\beta \sum_{l=1}^{N} \left(\frac{I^l}{\mathcal{Z}^l} \right) \left(\frac{p_2^l}{\alpha}(A_2 - I^l) - \frac{I^l}{\mathcal{Z}^l(1+\alpha)} \right).
\end{aligned} \tag{C.5}$$

C.1.1 Agonist-agonist scenario.

If both ligands have unit efficacies ($A_1 = A_2 = 1$), eq. (C.5) gives

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} = -\sum_{l=1}^{N} \left(\frac{p_2^l - \alpha p_1^l}{\mathcal{Z}^l \alpha(1+\alpha)} \right)^2. \tag{C.6}$$

Let us assume for simplicity that μ is known, so that $\sigma_\alpha^2 = -\left(\partial^2 \mathcal{L}/\partial \alpha^2\right)^{-1}$. Furthermore, let us suppose that ΔG_1 is fixed at a finite value, while ΔG_2 varies from $-\infty$ to $+\infty$. It is then easy to see that $\Delta G_2 = \Delta G_1$ is a special case, yielding $p_2^l = \alpha p_1^l$ $(\forall l)$ and thus $\sigma_\alpha^2 = \infty$. So, as expected, discrimination between the two ligands is impossible if they have equal efficacies and binding affinities. If $\Delta G_2 \to -\infty$, $\mathcal{Z}^l \to +\infty$ (and $p_2^l \to 1, \forall l$), making $\partial^2 \mathcal{L}/\partial \alpha^2 = 0$ for finite α. Thus discrimination is impossible if one of the ligands completely saturates the receptor. However, if $\Delta G_2 \to +\infty$, $\mathcal{Z}^l$ remains finite (while $p_2^l \to 0, \forall l$), yielding

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} \simeq -\frac{1}{(1+\alpha)^2} \sum_{l=1}^{N} \left(\frac{p_1^l}{\mathcal{Z}^l} \right)^2. \tag{C.7}$$

Surprisingly, discrimination is still possible in this limit, even if ligand 2 does not bind the receptor (Figure 3.3). This is because the total concentration is known in this example and so the information provided by ligand 1 is sufficient to infer α. Of course, if ligand 1 is either unbound ($p_1^l \to 0, \forall l$) or strongly bound ($\mathcal{Z}^l \to +\infty$), predicting α becomes impossible again.

If α is known, the error in the predicted total chemical potential μ is determined by (eq. (C.5)):

$$\frac{\partial^2 \mathcal{L}}{\partial \mu^2} = -\beta^2 \sum_{l=1}^{N} \left(\frac{\mathcal{Z}^l - 1}{\mathcal{Z}^{l\,2}} \right)^2 . \tag{C.8}$$

Learning the value of the total concentration becomes impossible if both ligands are unbound ($\mathcal{Z}^l \to 1$), or if one ligand is bound so strongly that measuring the concentration of the other ligand becomes problematic ($\mathcal{Z}^l \to \infty$). Note that in both of these limits $\partial^2\mathcal{L}/\partial\alpha^2 = 0$ as well.

Simultaneous discrimination of μ and α is not possible if both ligands have equal efficacies (Figure 3.3). Indeed, the determinant of the 2×2 matrix of second derivatives is always close to 0:

$$\begin{aligned} \det \left(\frac{\partial^2 \mathcal{L}}{\partial \gamma_i \partial \gamma_j} \right) = \left(\frac{\beta}{\alpha(1+\alpha)} \right)^2 \sum_{l,l'=1}^{N} \frac{\mathcal{Z}^{l'} - 1}{\mathcal{Z}^{l\,2} \mathcal{Z}^{l'\,3}} (p_2^l - \alpha p_1^l) \times \\ \left[(p_2^l - \alpha p_1^l)(p_1^{l'} + p_2^{l'}) - (p_2^{l'} - \alpha p_1^{l'})(p_1^l + p_2^l) \right] . \end{aligned} \tag{C.9}$$

This is because appreciable values of p_1^l or p_2^l lead to $\mathcal{Z}^l > 1$, which in turn suppresses the determinant. The small value of the determinant means that at least one of the errors is large. For example, if $\Delta G_2 = \Delta G_1$,

$$\sigma_\alpha^2 = -\frac{1}{\det \left(\frac{\partial^2 \mathcal{L}}{\partial \gamma_i \partial \gamma_j} \right)} \frac{\partial^2 \mathcal{L}}{\partial \mu^2} \tag{C.10}$$

is infinite because $p_2^l = \alpha p_1^l (\forall l)$, whereas $\mathcal{Z}^l$ and thus $\partial^2\mathcal{L}/\partial\mu^2$ are finite. In general, the multidimensional analysis of this type is difficult because zeros in the numerator and denominator of eq. (C.10) and a similar equation for σ_μ^2 have to be handled correctly.

C.1.2 Agonist-antagonist and antagonist-agonist scenarios.

If one of the ligands (*e.g.* ligand 2) acts as a perfect antagonist ($A_1 = 1, A_2 = 0$), eq. (C.5) gives (Figure 3.2):

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial \alpha^2} &= -\sum_{l=1}^{N} \left(p_1^l \right)^2 \left(\frac{p_2^l}{\alpha} + \frac{1}{\mathcal{Z}^l (1+\alpha)} \right)^2 , \\ \frac{\partial^2 \mathcal{L}}{\partial \mu^2} &= -\beta^2 \sum_{l=1}^{N} \left(\frac{p_1^l}{\mathcal{Z}^l} \right)^2 , \\ \frac{\partial^2 \mathcal{L}}{\partial \mu \partial \alpha} &= -\beta \sum_{l=1}^{N} \left(\frac{p_1^{l\,2}}{\mathcal{Z}^l} \right) \left(\frac{p_2^l}{\alpha} + \frac{1}{\mathcal{Z}^l (1+\alpha)} \right) . \end{aligned} \tag{C.11}$$

With ligand 1 acting as an antagonist ($A_1 = 0, A_2 = 1$), we obtain (Figure 3.1):

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} &= -\sum_{l=1}^{N} \left(p_2^l\right)^2 \left(\frac{p_1^l}{\alpha} + \frac{1}{\mathcal{Z}^l \alpha(1+\alpha)}\right)^2, \\
\frac{\partial^2 \mathcal{L}}{\partial \mu^2} &= -\beta^2 \sum_{l=1}^{N} \left(\frac{p_2^l}{\mathcal{Z}^l}\right)^2, \\
\frac{\partial^2 \mathcal{L}}{\partial \mu \partial \alpha} &= -\beta \sum_{l=1}^{N} \left(\frac{p_2^l{}^2}{\mathcal{Z}^l}\right) \left(\frac{p_1^l}{\alpha} + \frac{1}{\mathcal{Z}^l \alpha(1+\alpha)}\right).
\end{aligned} \tag{C.12}$$

If $\alpha = 1$ eq. (C.12) is the same as eq. (C.11) with ΔG_1 and ΔG_2 interchanged. However, for arbitrary α there is no symmetry, so that maximizing the determinant of the Hessian with $A_1 = 0, A_2 = 1$ and $A_1 = 1, A_2 = 0$ yields two distinct solutions in the $\{\Delta G_1, \Delta G_2\}$ space.

C.2 Alternative definitions of the relative concentrations.

Recall that the relative concentrations are defined as $\alpha_m = n_{m+1}/n_1$ ($m = 1 \ldots N_{lig} - 1$), where n_j is the concentration of ligand $j = 1 \ldots N_{lig}$. Different choices of the ligand in the denominator may lead to very different numerical values of α if, for example, $n_1 \ll n_2$ in a two-ligand, one-receptor system. Nevertheless, the uncertainty of both predictions is related. Indeed, if $\alpha = n_2/n_1 \to 0$ one can show that

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} \to -\sum_{l=1}^{N} \left(\frac{e_1^l e_2^l (A_2 - A_1) - A_1 e_1^l + A_2 e_2^l}{(1+e_1^l)^2}\right)^2, \tag{C.13}$$

where $e_i^l = \exp\{-\beta(\Delta G_i - \mu^l)\}$, $i = 1, 2$. Thus the absence of ligand 2 can generally be predicted with finite uncertainty, at least if the total chemical potential μ is known. In the $\alpha \to +\infty$ limit,

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} \to -\frac{1}{\alpha^4} \sum_{l=1}^{N} \left(\frac{e_1^l e_2^l (A_2 - A_1) - A_1 e_1^l + A_2 e_2^l}{(1+e_2^l)^2}\right)^2, \tag{C.14}$$

and $\sigma_\alpha^2 = -\left(\partial^2 \mathcal{L}/\partial \alpha^2\right)^{-1}$ diverges as α^4. This is expected because $\alpha \to +\infty$ is equivalent to $\alpha' = n_1/n_2 = 1/\alpha \to 0$, yielding $\sigma_\alpha^2 = \alpha^4 \sigma_{\alpha'}^2$. Thus $\sigma_{\alpha'}^2$ remains finite as ligand 1 disappears from the mixture. Moreover, the expression for $\partial^2 \mathcal{L}/\partial \alpha'^2$ in the $\alpha' \to 0$ limit should be the same as the expression for $\partial^2 \mathcal{L}/\partial \alpha^2$ in the $\alpha \to 0$ limit, but with ligand labels 1 and 2 interchanged. Indeed, if $1 \leftrightarrow 2$ eq. (C.14) becomes the same as eq. (C.13), apart from the $1/\alpha^4$ factor mentioned above.

C.3 Number of agonist-antagonist patterns in an arbitrary receptor array.

For an arbitrary receptor-ligand combination, there are

$$\sum_{m=1}^{N_{rec}-1} \frac{(m+1)!}{N_{rec}!} n^{N_{rec}-m} + \frac{n(n^{N_{rec}} - 2n^{N_{rec}-1} + 1)}{(n-1)N_{rec}!} \tag{C.15}$$

unique patterns in which each receptor interacts with one agonist and one antagonist (the patterns are unique in a sense that all patterns connected by trivial receptor label permutations are counted only once). Here N_{rec} is the number of receptors, N_{lig} is the number of ligands, and $n = 2\binom{N_{lig}}{2}$ is the number of ways in which one agonist and one antagonist can be bound by a single receptor, so that the total number of patterns (some connected by receptor label permutations) is $n^{N_{rec}}$. With $N_{rec} = 2$ and $N_{lig} = 3$ the total number of patterns is 36, and 21 unique patterns include 3 combinations in which one of the ligands acts as an antagonist for both receptors, 3 more where one of the ligands is the global agonist, 6 mixed patterns where one ligand invokes the agonist-antagonist receptor response and the other two are either an agonist or an antagonist for one of the remaining receptors, and finally 9 patterns in which one of the receptors does not strongly interact with any of the ligands. The remaining 15 patterns are related to the ones listed above through receptor label permutations. The first subclass yields equivalent globally optimal solutions, subclasses 2 and 3 correspond to local maxima, and the fourth subclass in which one of the receptors remains unused does not yield any stable solutions, relaxing into one of the other categories.

Appendix D

Additional Figures, Chapter 3

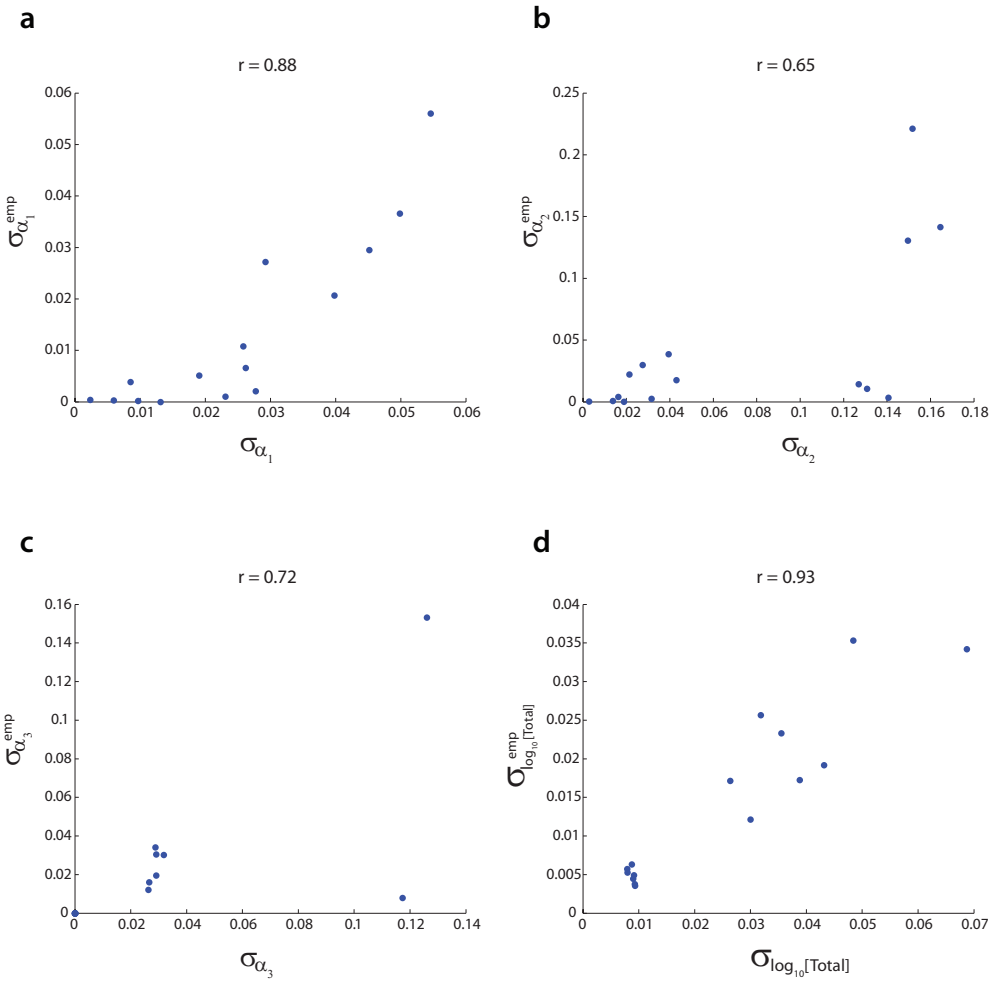


Figure D.1: **Hessian uncertainties vs. standard deviations from nested sampling.** Synthetic data was generated as four replicates for each of the 15 equal-proportion mixtures from Figure 2.3, using parameters from Table B.1 and $\tilde{\sigma} = 0.01$ for all receptors (eq. (4.6)). For each receptor, concentration ranges were taken from the corresponding experiment (see § 2.1.4). For each parameter, a Hessian error computed using eq. (C.4) (x-axis) is compared with the standard deviation from a nested sampling run (y-axis). Nested sampling simultaneously infers relative concentrations α_1, α_2, α_3, the total concentration and $\tilde{\sigma}$'s given receptor-ligand parameters from Table B.1.

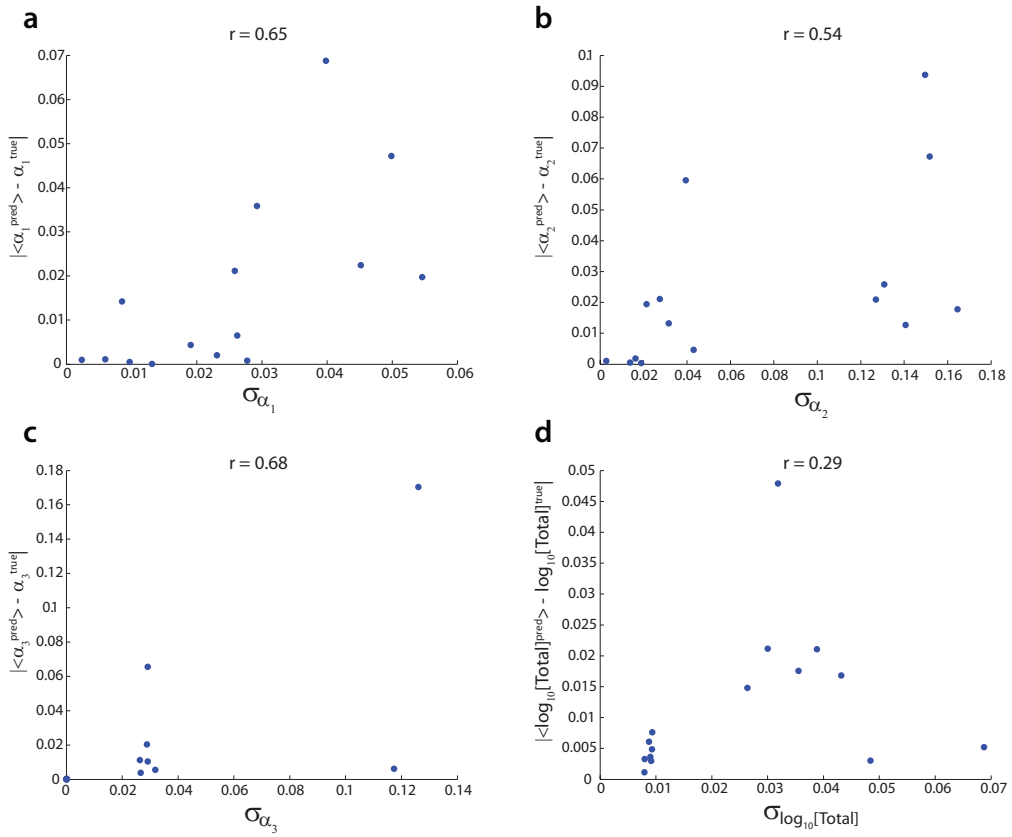


Figure D.2: **Hessian uncertainties vs. errors in parameter predictions.** Synthetic data was generated as four replicates for each of the 15 equal-proportion mixtures from Figure 2.3, using parameters from Table E.1 and $\tilde{\sigma} = 0.01$ for all receptors (eq. (4.6)). For each receptor, concentration ranges were taken from the corresponding experiment (§ 2.1.4).For each parameter, a Hessian error computed using eq. (C.4) (x-axis) is compared with the absolute magnitude of the difference between the mean value predicted by nested sampling and the correct value (y-axis). Nested sampling simultaneously infers relative concentrations α_1, α_2, α_3, the total concentration and $\tilde{\sigma}$'s given receptor-ligand parameters from Table B.1.

Appendix E

Additional Figures and Tables, Chapter 4

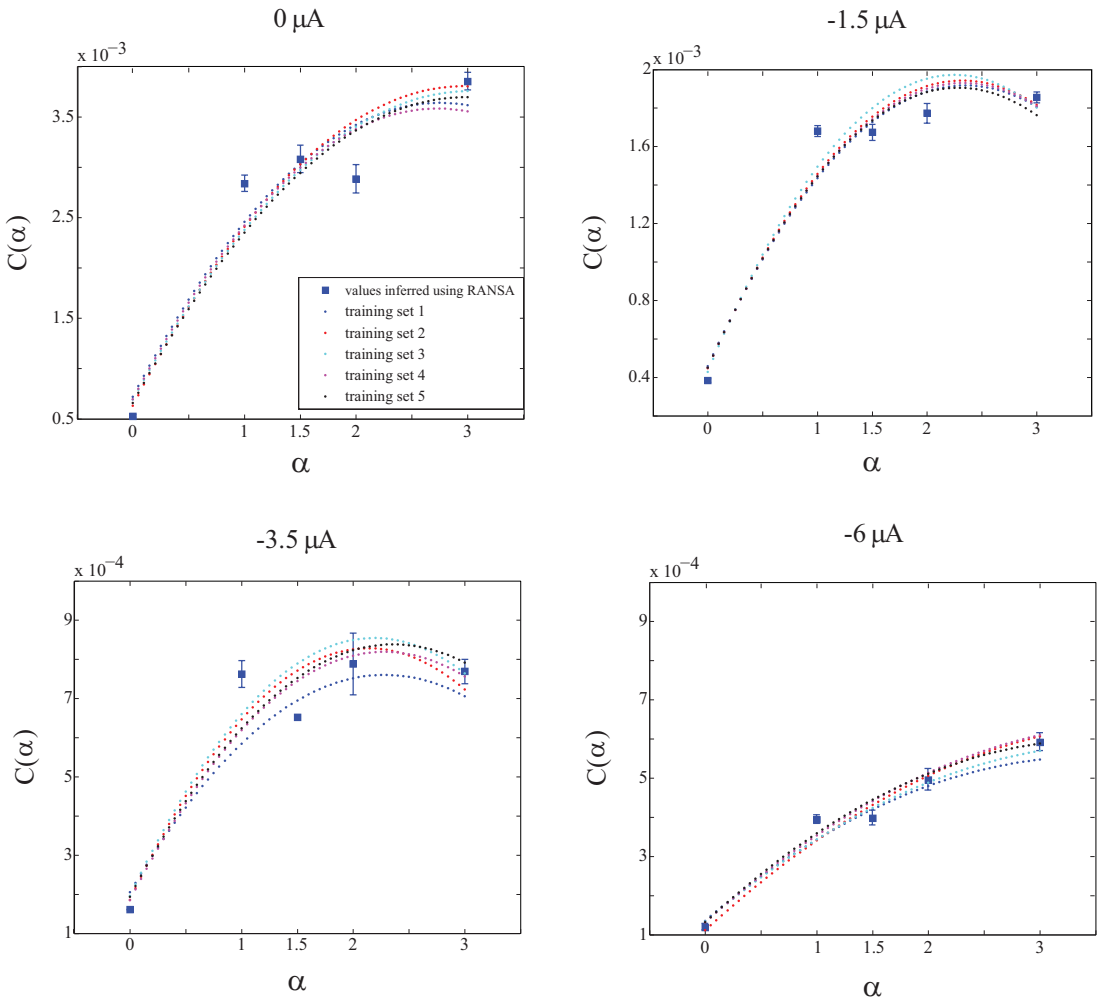


Figure E.1: **Prediction of parameters a'_k, b'_k, c'_k (eq. (4.7)) by a least-squares quadratic fit to** $C_k(\alpha)$. Each point and its error bar represent the mean and the standard deviation of $C_k(\alpha)$ predicted by RANSA for each training data subset (see § 4.4.2), and averaged over all 5 subsets. Each curve is found by a least-squares fit to the mean $C_k(\alpha)$ ($\alpha = 0, 1, 1.5, 2, 3$) inferred from a given training subset.

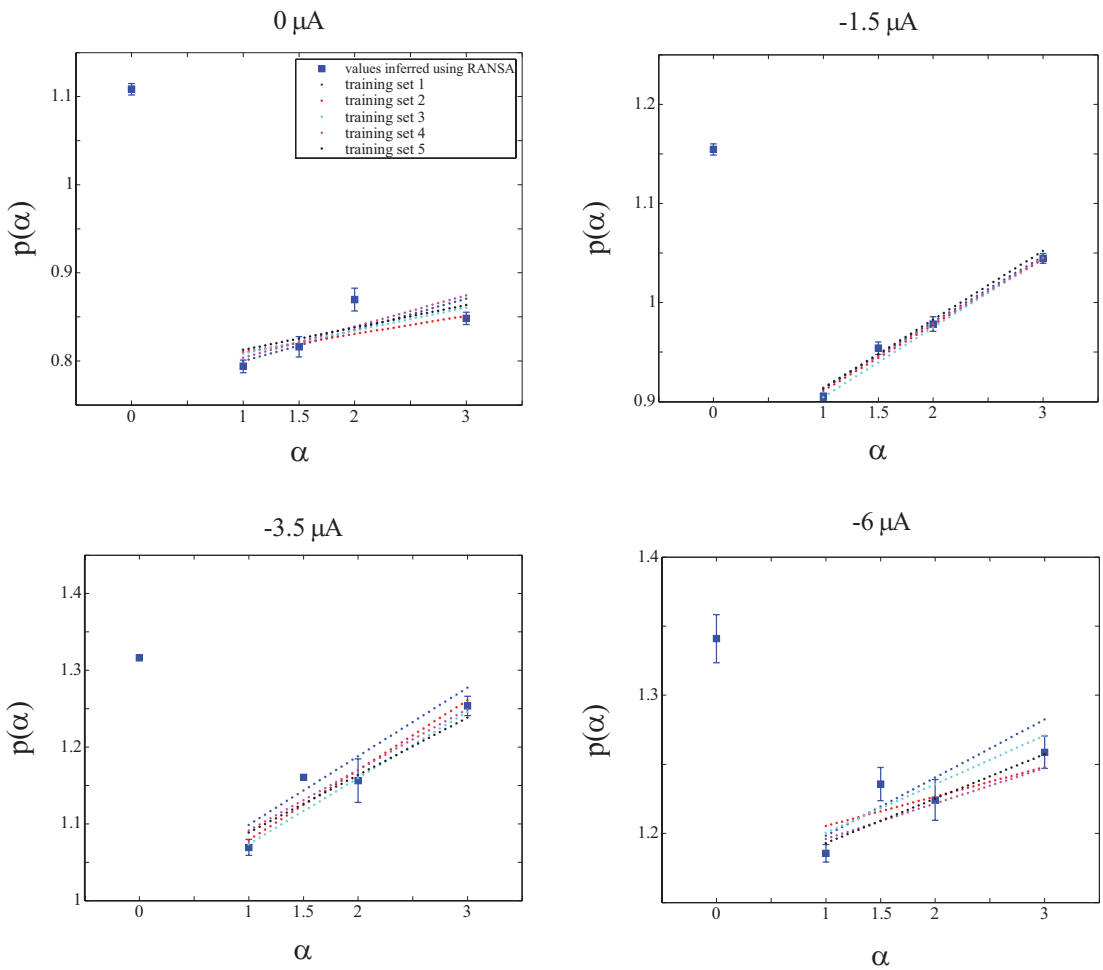


Figure E.2: **Prediction of parameters a_k and b_k (eq. (4.7)) by a least-squares linear fit to $p_k(\alpha)$.** As in Figure E.1, each point and its error bar represent the mean and the standard deviation of $p_k(\alpha)$ predicted by RANSA for each training data subset (see § 4.4.2), and averaged over all 5 subsets. Each curve is found by a least-squares fit to the mean $p_k(\alpha)$ ($\alpha = 1, 1.5, 2, 3$) inferred from a given training subset. Note that the $\alpha = 0$ case, which does not follow the linear trend, is treated separately, as described in the text.

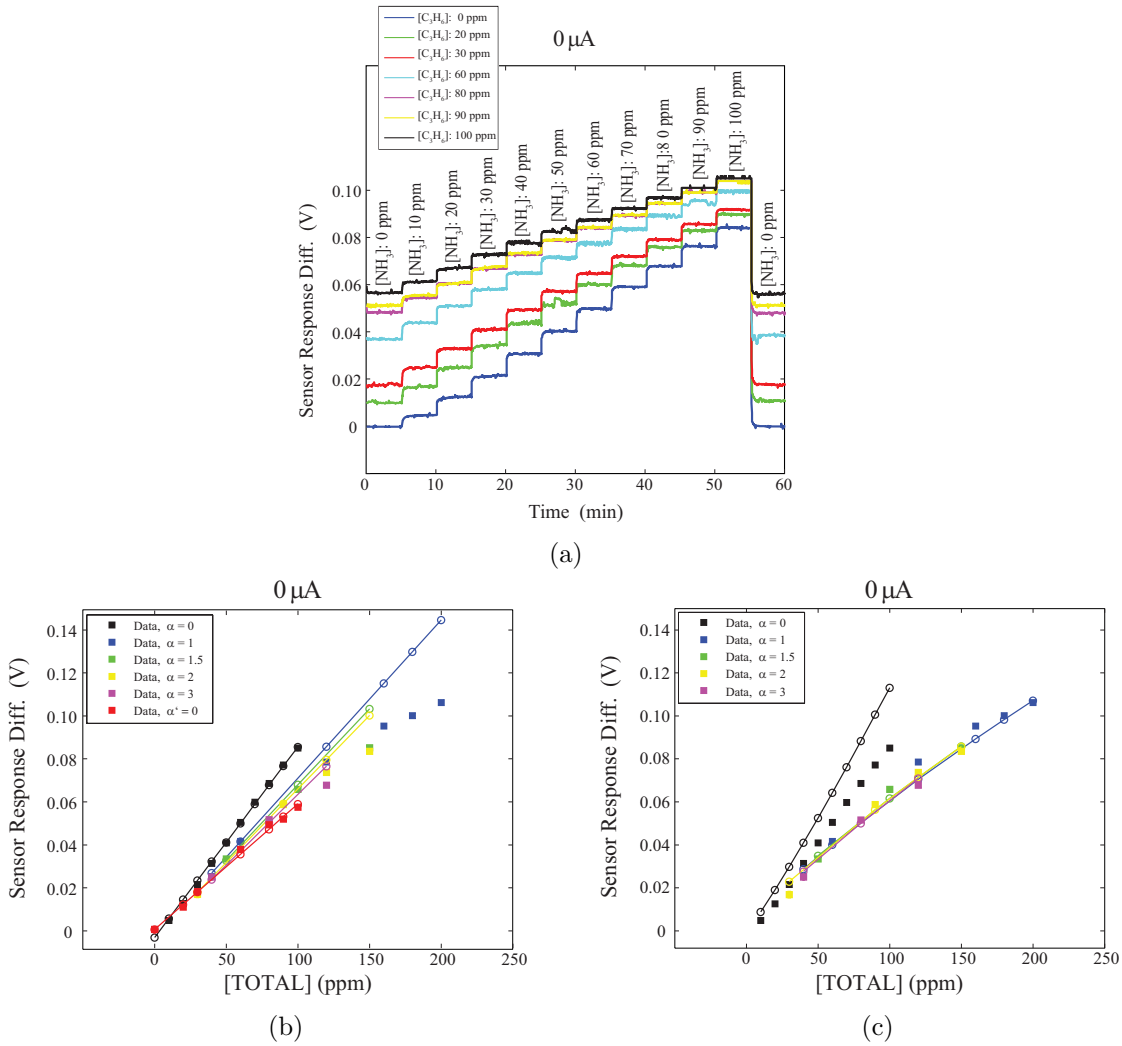


Figure E.3: Same as Figure 4.3, but with 0 μA bias current.

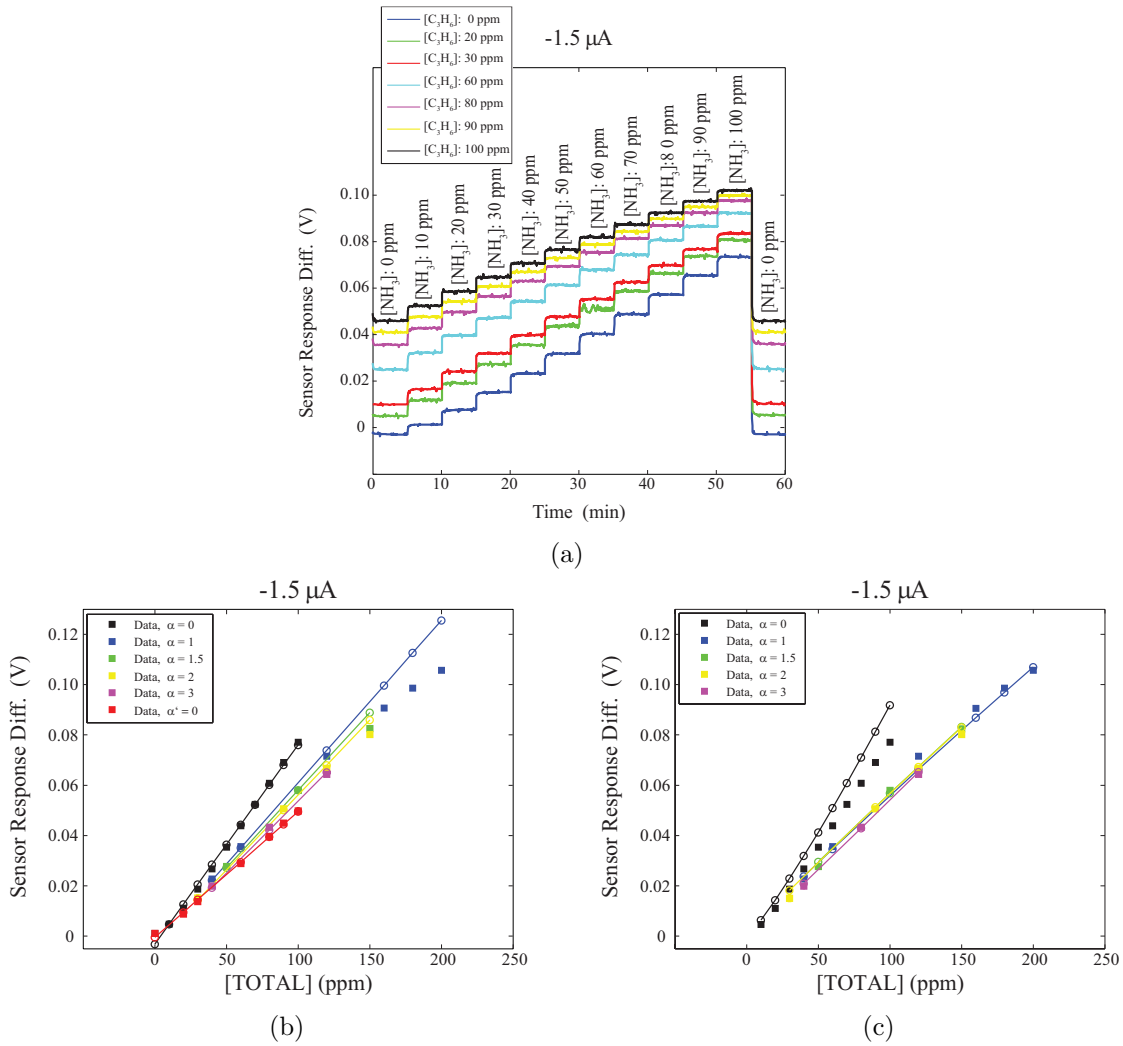
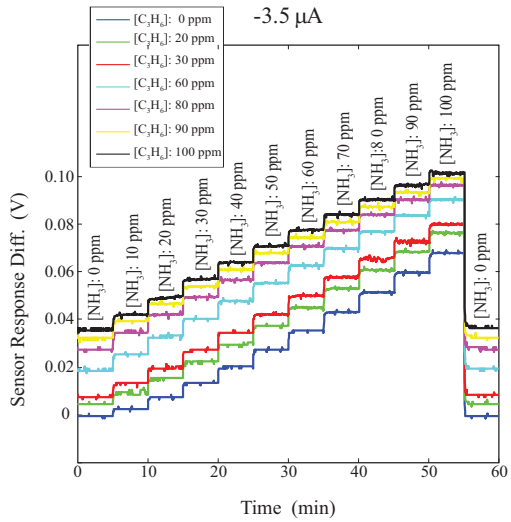
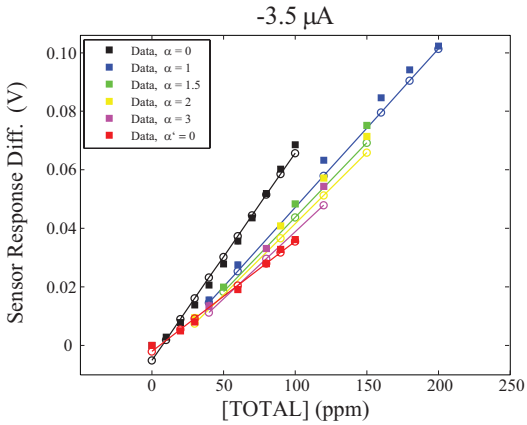


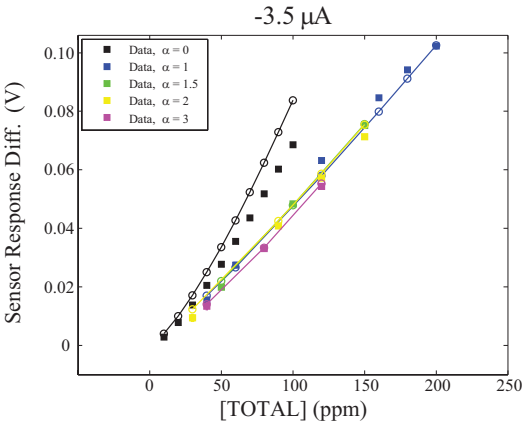
Figure E.4: Same as Figure 4.3, but with -1.5 μA bias current.



(a)



(b)



(c)

Figure E.5: Same as Figure 4.3, but with -3.5 μA bias current.

Table E.1: **Means and standard deviations of linear model parameters $V_{0_1}^k$, $V_{0_2}^k$, A_k, B_k (Eqs. (4) and (5)) inferred by RANSA in the calibration step.**

Bias Current (μA)	$V_{0_1}^k$	$V_{0_2}^k$	A_k	B_k
0.0 ($k = 1$)	$-.0032 \pm 0$	$5.9 \times 10^{-4} \pm 0$	$8.9 \times 10^{-4} \pm 0$	$5.8 \times 10^{-4} \pm 0$
-1.5 ($k = 2$)	$-.0031 \pm 0$	$-4.9 \times 10^{-4} \pm 0$	$7.9 \times 10^{-4} \pm 0$	$5.0 \times 10^{-4} \pm 0$
-3.5 ($k = 3$)	$-.0051 \pm 1.7 \times 10^{-4}$	$-.0020 \pm 1.2 \times 10^{-4}$	$7.1 \times 10^{-4} \pm 0$	$3.8 \times 10^{-4} \pm 0$
-6.0 ($k = 4$)	$-.0042 \pm 0$	$-8.1 \times 10^{-4} \pm 0$	$5.8 \times 10^{-4} \pm 0$	$2.7 \times 10^{-4} \pm 0$

Table E.2: **Means and standard deviatins of nonlinear model parameters $C_k(\alpha)$ and $p_k(\alpha)$ (eq. (4.6)) inferred by RANSA in the calibration step. All means and standard deviations are averaged over 5 training subsets (see § 4.4.2).**

Bias Current (μA)	α	$C_k(\alpha)$	$p_k(\alpha)$
0.0 ($k = 1$)	0	$.0005 \pm 1.0 \times 10^{-5}$	$1.108 \pm .006$
	1	$.0028 \pm 8.0 \times 10^{-5}$	$.794 \pm .007$
	1.5	$.0031 \pm 1.4 \times 10^{-4}$	$.816 \pm .012$
	2	$.0029 \pm 1.4 \times 10^{-4}$	$.870 \pm .013$
	3	$.0039 \pm 9.0 \times 10^{-5}$	$.848 \pm .007$
-1.5 ($k = 2$)	0	$.0004 \pm 9.0 \times 10^{-6}$	$1.154 \pm .006$
	1	$.0017 \pm 2.9 \times 10^{-5}$	$.905 \pm .004$
	1.5	$.0017 \pm 4.1 \times 10^{-5}$	$.954 \pm .006$
	2	$.0018 \pm 5.1 \times 10^{-5}$	$.978 \pm .007$
	3	$.0019 \pm 2.8 \times 10^{-5}$	$1.045 \pm .005$
-3.5 ($k = 3$)	0	$1.6 \times 10^{-4} \pm 9.4 \times 10^{-7}$	$1.316 \pm .001$
	1	$7.6 \times 10^{-4} \pm 3.4 \times 10^{-5}$	$1.07 \pm .01$
	1.5	$6.5 \times 10^{-4} \pm 7.0 \times 10^{-6}$	$1.161 \pm .003$
	2	$7.9 \times 10^{-4} \pm 7.9 \times 10^{-5}$	$1.16 \pm .03$
	3	$7.7 \times 10^{-4} \pm 3.1 \times 10^{-5}$	$1.25 \pm .01$
-6.0 ($k = 4$)	0	$1.2 \times 10^{-4} \pm 8.0 \times 10^{-6}$	$1.34 \pm .02$
	1	$3.9 \times 10^{-4} \pm 1.1 \times 10^{-5}$	$1.186 \pm .006$
	1.5	$4.0 \times 10^{-4} \pm 1.9 \times 10^{-5}$	$1.24 \pm .01$
	2	$5.0 \times 10^{-4} \pm 2.7 \times 10^{-5}$	$1.22 \pm .01$
	3	$5.9 \times 10^{-4} \pm 2.3 \times 10^{-5}$	$1.26 \pm .01$

Table E.3: **Parameters a'_k, b'_k, c'_k (eq. (4.7)) inferred by least-squares fits as shown in Figure E.1. Each parameter is averaged over 5 training subsets (see § 4.4.2).**

Bias Current (μA)	**a'**	**b'**	**c'**
0.0 ($k = 1$)	$-.0004$	.0021	.0007
-1.5 ($k = 2$)	$-.0003$	.0013	.0004
-3.5 ($k = 3$)	$-.00012$	.00056	.00020
-6.0 ($k = 4$)	-3.4×10^{-5}	.00025	.00013

Table E.4: **Parameters a_k, b_k (eq. (4.7)) inferred by least-squares fits as shown in Figure E.2. Each parameter is averaged over 5 training subsets (see § 4.4.2).**

Bias Current (μA)	**a**	**b**
0.0 ($k = 1$)	.029	.778
-1.5 ($k = 2$)	.068	.844
-3.5 ($k = 3$)	.084	1.00
-6.0 ($k = 4$)	.031	1.17

References

[1] L. Buck, "Information coding in the vertebrate olfactory system," *Annu. Rev. Neurosci.*, vol. 19, pp. 517–544, 1996.

[2] S. Firestein, "How the olfactory system makes sense of scents," *Nature*, vol. 413, pp. 211–218, 2001.

[3] H. Breer, "Olfactory receptors: molecular basis for recognition and discrimination of odors," *Anal. Bioanal. Chem.*, vol. 377, pp. 427–433, 2003.

[4] K. Touhara and L. Vosshall, "Sensing odorants and pheromones with chemosensory receptors," *Annu. Rev. Physiol.*, vol. 71, pp. 307–332, 2009.

[5] J. Hopfield, "Odor space and olfactory processing: Collective algorithms and neural implementation," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 12506–12511, 1999.

[6] C.-Y. Su, K. Menuz, and J. Carlson, "Olfactory perception: Receptors, cells, and circuits," *Cell*, vol. 139, pp. 45–59, 2009.

[7] L. Buck and R. Axel, "A novel multigene family may encode odorant receptors: a molecular basis for odor recognition," *Cell*, vol. 65, pp. 175–187, 1991.

[8] X. Zhang and S. Firestein, "The olfactory receptor gene superfamily of the mouse," *Nature Neurosci.*, vol. 5, pp. 124–133, 2002.

[9] B. Malnic, P. Godfrey, and L. Buck, "The human olfactory receptor gene family," *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 2584–2589, 2004.

[10] P. Godfrey, B. Malnic, and L. Buck, "The mouse olfactory receptor gene family," *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 2156–2161, 2004.

[11] Y. Hasin-Brumshtein, D. Lancet, and T. Olender, "Human olfaction: from genomic variation to phenotypic diversity," *Trends Genet.*, vol. 25, pp. 178–184, 2009.

[12] B. Malnic, J. Hirono, T. Sato, and L. Buck, "Combinatorial receptor codes for odors," *Cell*, vol. 96, pp. 713–723, 1999.

[13] H. Saito, M. Kubota, R. Roberts, Q. Chi, and H. Matsunami, "Rtp family members induce functional expression of mammalian odorant receptors," *Cell*, vol. 119, pp. 679–691, 2004.

[14] E. Shirokova, K. Schmiedeberg, P. Bedner, H. Niessen, K. Willecke, J.-D. Raguse, W. Meyerhof, and D. Krautwurst, "Identification of specific ligands for orphan olfactory receptors," *J Biol. Chem.*, vol. 280, pp. 11807–11815, 2005.

[15] E. Hallem and J. Carlson, "Coding of odors by a receptor repertoire," *Cell*, vol. 125, pp. 143–160, 2006.

[16] H. Saito, Q. Chi, H. Zhuang, H. Matsunami, and J. Mainland, "Odor coding by a mammalian receptor repertoire," *Sci. Signal.*, vol. 2, p. ra9, 2009.

[17] M. Spehr, G. Gisselmann, A. Poplawski, J. Riffell, C. Wetzel, R. Zimmer, and H. Hatt, "Identification of a testicular odorant receptor mediating human sperm chemotaxis," *Science*, vol. 299, pp. 2054–2058, 2003.

[18] R. Araneda, Z. Peterlin, X. Zhang, A. Chesler, and S. Firestein, "A pharmacological profile of the aldehyde receptor repertoire in rat olfactory epithelium," *J Physiol.*, vol. 555(Pt 3), pp. 743–756, 2004.

[19] Y. Oka, M. Omura, H. Kataoka, and K. Touhara, "Olfactory receptor antagonism between odorants," *EMBO J*, vol. 23, pp. 120–126, 2004.

[20] G. Sanz, C. Schlegel, J. Pernollet, and L. Briand, "Comparison of odorant specificity of two human olfactory receptors from different phylogenetic classes and evidence for antagonism," *Chem. Senses*, vol. 30, pp. 69–80, 2005.

[21] B. Auffarth, B. Kaplan, and A. Lansner, "Map formation in the olfactory bulb by axon guidance of olfactory neurons," *Front. Syst. Neurosci.*, vol. 5, no. 00084, 2011.

[22] K. Albert, N. Lewis, C. Schauer, G. Sotzing, S. Stitzel, T. Vaid, and D. Walt, "Cross-reactive chemical sensor arrays," *Chem. Rev.*, vol. 100, pp. 2595–2626, 2000.

[23] N. Rakow and K. Suslick, "A colorimetric sensor array for odour visualization," *Nature*, vol. 406, pp. 710–713, 2000.

[24] A. Wright and E. Anslyn, "Differential receptor arrays and assays for solution-based molecular recognition," *Chem. Soc. Rev.*, vol. 35, pp. 14–28, 2006.

[25] S. Shabbir, L. Joyce, G. da Cruz, V. Lynch, S. Sorey, and E. Anslyn, "Pattern-based recognition for the rapid determination of identity, concentration, and enantiomeric excess of subtly different threo diols," *J. Am. Chem. Soc.*, vol. 131, pp. 13125–13131, 2009.

[26] P. Jurs, G. Bakken, and H. McClelland, "Computational methods for the analysis of chemical sensor array data from volatile analytes," *Chem. Rev.*, vol. 100, pp. 2649–2678, 2000.

[27] T. Gao, M. Woodka, B. Brunschwig, and L. N.S., "Chemiresistors for array-based vapor sensing using composites of carbon black with low volatility organicmolecules," *ChemMater*, vol. 18, pp. 5193–5202, 2006.

[28] M. Kitamura, S. Shabbir, and E. Anslyn, "Guidelines for pattern recognition using differential receptors and indicator displacement arrays," *J. Org. Chem.*, vol. 74, pp. 4479–4489, 2009.

[29] T. Feuerstein and N. Limberger, "Mathematical analysis of the control of neurotransmitter release by presynaptic receptors as a supplement to experimental data," *Arch. Pharmacol.*, vol. 359, pp. 345–359, 1999.

[30] C. Zhang, D. Bailey, and K. Suslick, "T colorimetric sensor arrays for the analysis of beers: A feasibility study," *J Agr Food Chem*, vol. 54, p. 4925–4931, 2006.

[31] C. Zhang and K. Suslick, "Colorimetric sensor array for soft drink analysis," *J Agr Food Chem*, vol. 55, p. 237–242, 2007.

[32] R. Dutta, K. Kashwan, M. Bhuyan, E. Hines, and J. Gardner, "Electronic nose based tea quality standardization," *Neural Netw*, vol. 516, p. 847–853, 2003.

[33] M. Pardo and G. Sberveglieri, "Coffee analysis with an electronic nose," *IEEE Trans Inst Measur*, vol. 51, p. 1334–1339, 2002.

[34] A. Heilig, N. Barsan, U. Weimar, M. SchweizerBerberich, and W. e. a. Gopel, "Gas identification by modulating temperatures of sno2-based thick film sensors," *Sens Actuators B*, vol. 43, pp. 45–51, 1997.

[35] J. White, J. Kauer, T. Dickinson, and D. Walt, "Rapid analyte recognition in a device based on optical sensors and the olfactory system," *Anal Chem*, vol. 68, pp. 2191–2202, 1996.

[36] M. Woodka, B. Brunschwig, and N. Lewis, "Use of spatiotemporal response information from sorption-based sensor arrays to identify and quantify the composition of analyte mixtures," *Langmuir*, vol. 23, p. 13232–13241, 2007.

[37] J. Tsitron, A. Ault, J. Broach, and A. Morozov, "Decoding complex chemical mixtures with a physical model of a sensor array," *PLoS Comput. Biol.*, vol. 7, p. e1002224, 2011.

[38] A. Ault and J. Broach, "Creation of GPCR-based chemical sensors by directed evolution in yeast," *Prot. Eng. Des. Sel.*, vol. 19, pp. 1–8, 2006.

[39] C. Klein, J. Paul, K. Sauve, M. Schmidt, L. Arcangeli, J. Ransom, J. Trueheart, J. Manfredi, J. Broach, and A. Murphy, "Identification of surrogate agonists for the human fprl-1 receptor by autocrine selection in yeast," *Nature Biotechnol.*, vol. 16, pp. 1334–1337, 1998.

[40] C. Kaiser, S. Michaelis, and A. Mitchell, *Methods in yeast genetics : a Cold Spring Harbor Laboratory course manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 1994.

[41] J. Chambers, L. Macdonald, H. Sarau, R. Ames, K. Freeman, J. Foley, Y. Zhu, M. McLaughlin, P. Murdock, L. McMillan, J. Trill, A. Swift, N. Aiyar, P. Taylor, L. Vawter, S. Naheed, P. Szekeres, G. Hervieu, C. Scott, J. Watson, A. Murphy, E. Duzic, C. Klein, D. Bergsma, S. Wilson, and G. Livi, "A G protein-coupled receptor for UDP-glucose," *J. Biol. Chem.*, vol. 275, pp. 10767–10771, 2000.

[42] D. Sivia and J. Skilling, *Data analysis: A Bayesian tutorial*. Oxford: Oxford University Press, 2006.

[43] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and T. E., "Equation of state by fast computing machines," *J. Chem. Phys.*, vol. 21, pp. 1087–1092, 1953.

[44] R. Gutenkunst, J. Waterfall, F. Casey, K. Brown, C. Myers, and J. Sethna, "Universally sloppy parameter sensitivities in systems biology models," *PLoS Comp. Biol.*, vol. 3, no. 10, pp. 1871–1878, 2007.

[45] M. Kampa and E. Castanas, "Human health effects of air pollution," *Environ. Pollut.*, vol. 151, pp. 362–367, 2008.

[46] P. Scheepers and R. Bos, "Combustion of diesel fuel from a toxicological perspective," *Int. Arch. Occup. Environ. Health*, vol. 64, pp. 149–161, 1992.

[47] G. Sandhu and H. C. Frey, "Real-world measurement and evaluation of heavy duty truck cycles, fuels, and emissions control technologies," *Transport. Res. Rec.*, 2012.

[48] R. Burch, J. Breen, and F. Meunier, "A review of the reduction of NO_x with hydrocarbons under lean-burn conditions with non-zeolite oxide and Pt group metal catalysts," *Appl. Catal. B*, vol. 39, pp. 283–303, 2002.

[49] S. Noroozi, "Urea enhances safety in SCR applications," *Power Eng.*, vol. 97, pp. 28–29, 1993.

[50] G. H. Abd-Alla, "Using exhaust gas recirculation in internal combustion engines: a review," *Energy Conv. and Management*, vol. 43, pp. 1027–1042, 2002.

[51] A. Maiboom, X. Tauzia, and J.-F. Hetet, "Experimental study of various effects of exhaust gas recirculation (EGR) on combustion and emissions of an automotive direct injection diesel engine," *Energy*, vol. 33, pp. 22–34, 2008.

[52] R. Anand and N. V. Mahalaskmi, "Simultaneous reduction of oxides of nitrogen and smoke emissions by using EGR combined with particulate trap in a direct injection diesel engine," *Proc. 2006 Spring Tech. Conf. ASME Internal Combustion Engine Division*, pp. 209–216, 2006.

[53] F. Schafer and R. van Basshuysen, *Reduced Emissions and Fuel Consumption in Automobile Engines.* New York: Springer-Verlag, 1995.

[54] P. Sekhar, E. Brosha, R. Mukundan, W. Li, M. Nelson, P. Palanisamy, and F. H. Garzon, "Application of commercial automotive sensor manufacturing methods for NO_x/NH_3 mixed potential sensors for on-board emissions control," *Sens. Actuators B: Chem.*, vol. 144, pp. 112–119, 2010.

[55] P. Forzatti, "Present status and perspectives in de-NO_x SCR catalysis," *Appl. Catal.*, vol. 222, pp. 221–236, 2001.

[56] M. Hsieh and J. Wang, "Development and experimental studies of a control-oriented SCR model for a two-catalyst urea-SCR system," *Control Eng. Practice*, vol. 19, pp. 409–422, 2011.

[57] M. Hsieh and J. Wang, "Sliding-mode observer for urea-selective catalytic reduction (SCR) mid-catalyst ammonia concentration estimation," *Int. J. Automotive Tech.*, vol. 12, pp. 321–329, 2011.

[58] J. Visser and R. Soltis, "Automotive exhaust gas sensing systems," *IEEE Trans. Instrum. Measur.*, vol. 50, pp. 1543–1550, 2001.

[59] J. W. Fergus, "Sensing mechanism of non-equilibrium solid-electrolyte-based chemical sensors," *J. Solid State Electrochem.*, vol. 15, pp. 971–984, 2010.

[60] F. H. Garzon, R. Mukundan, and E. Brosha, "Solid-state mixed potential gas sensors: theory, experiments and challenges," *Solid State Ionics*, vol. 136-137, pp. 633–638, 2000.

[61] N. Miura, T. Koga, M. Nakatou, P. Elumalai, and M. Hasei, "Electrochemical NO_x sensors based on stabilized zirconia: comparison of sensing performances of mixed-potential-type and impedancemetric NO_x sensors," *J. Electroceram.*, vol. 17, pp. 979–986, 2006.

[62] S. Zhuiykov and N. Miura, "Development of zirconia based potentiometric NO_x sensors for automotive and energy industries in the early 21st century: what are the prospects for sensors," *Sens. Actuators B: Chem.*, vol. 121, pp. 639–651, 2007.

[63] B. Timmer, W. Olthuis, and A. van den Berg, "Ammonia sensors and their applications – a review," *Sens. Actuators B: Chem.*, vol. 107, pp. 666–677, 2005.

[64] X. Liang, T. Zhong, H. Guan, F. Liu, G. Lu, and B. Quan, "Ammonia sensors based on NASICON and Cr_2O_3 electrodes," *Sens. Actuators B: Chem.*, vol. 136, pp. 479–483, 2009.

[65] W. J. Fleming, "Physical principles governing nonideal behavior of the zirconia oxygen sensor," *J. Electrochem. Soc.*, vol. 124, pp. 21–28, 1977.

[66] P. Pasierb and M. Rekas, "Solid-state potentiometric gas sensors-current status and future trends," *J. Solid State Electrochem.*, vol. 13, pp. 3–25, 2009.

[67] P. K. Sekhar, E. L. Brosha, R. Mukundan, M. A. Nelson, D. Toracco, and F. H. Garzon, "Effect of YSZ sintering temperature on mixed potential sensor performance," *Solid State Ionics*, vol. 181, pp. 947–953, 2010.

[68] R. Mukundan, E. L. Brosha, and F. H. Garzon, "Mixed potential hydrocarbon sensors based on a YSZ electrolyte and oxide electrodes," *J. Electrochem. Soc.*, vol. 150, pp. H279–H284, 2003.

[69] R. Mukundan, E. L. Brosha, and F. H. Garzon, "Electrodes for solid state gas sensor." Patent, 08 2003. 6,605,202.

[70] F. H. Garzon, E. L. Brosha, and R. Mukundan, "Method for forming a potential hydrocarbon sensor with low sensitivity to methane and CO." Patent, 12 2003. 6,656,336.

[71] E. L. Brosha, R. Mukundan, R. Lujan, and F. H. Garzon, "Mixed potential NO_x sensors using thin film electrodes and electrolytes for stationary engine type applications," *Sens. Actuators B: Chem.*, vol. 119, pp. 398–408, 2006.

[72] E. L. Brosha, R. Mukundan, D. R. Brown, Q. X. Jia, R. Lujan, and F. H. Garzon, "Techniques for the thin film growth of $La_{1-x}Sr_xCrO_3$ for solid state ionic devices," *Solid State Ionics*, vol. 166, pp. 425–440, 2004.

[73] R. Mukundan, F. H. Garzon, and E. L. Brosha, "Electrodes for solid state gas sensor." Patent, 05 2007. 7,214,333.

[74] C. Boyer, S. Zemskova, P. Park, L. Balmer-Millar, D. Endicott, and S. Faulkner, "DOE EERE Report," 2003.

[75] R. Mukundan, K. Teranishi, E. L. Brosha, and F. H. Garzon, "Nitrogen oxide sensors based on yttria-stabilized zirconia electrolyte and oxide electrodes," *Electrochemical and Solid-State Letters*, vol. 10, no. 2, pp. J26–J29, 2007.

[76] D. L. West, F. C. Montgomery, and T. R. Armstrong, "Total NO_x sensing elements with compositionally identical oxide electrodes," *J. Electrochem. Soc.*, vol. 153, pp. H23–H28, 2006.

[77] N. Miura, G. Lu, M. Ono, and N. Yamazoe, "Selective detection of NO by using an amperometric sensor based on stabilized zirconia and oxide electrode," *Solid State Ionics*, vol. 117, pp. 283–290, 1999.

[78] J. Broach, A. Ault, A. Morozov, and J. Tsitron, "Chemosensory Arrays." Patent Application, 07 2011. US 2011/0177964 A1.

[79] F. H. Garzon, E. L. Brosha, and R. Mukundan, "Solid state ionic devices for combustion gas sensing," *Solid State Ionics*, vol. 175, pp. 487–490, 2004.

[80] I. Murray, D. J. C. MacKay, Z. Ghahramani, and J. Skilling, "Nested sampling for potts models," *Advances in Neural Information Processing Systems*, vol. 18, 2006.

Vita

Julia Tsitron

2013	Ph.D. in Computational Biology and Molecular Biophysics, Rutgers University
2008	M. A. in Physics from Hunter College, City University of New York (CUNY)
2008	B. A. *(Special Honors Curriculum)* in Physics from Hunter College, CUNY