Biological Strategies

ParetoGP vs. Wide & Ill-Conditioned Data from Nonlinear Systems

Mark Kotanchek, Theresa Kotanchek, Kelvin Kotanchek & etc.

Evolved Analytics LLC

Genetic, proteomic and other biologically derived data sets are often ill-conditioned with many more variables than data records. To compound things, the variables are often highly correlated as well as coupled. These attributes make such data sets very difficult to analyze with conventional statistical and machine learning techniques.

The ParetoGP approach exploring the trade-off between model complexity and accuracy enables attacking such data sets with dual benefits of identifying key variables, associations and metavariables along with providing concise explainable & human-interpretable models. Transparency of key variables, model structures and response behaviors provide a substantial benefit relative to conventional machine learning and the associated black-box models.

In this chapter, we describe the analysis methodology and highlight benefits using real CAR-T, RNA, metabolite, etc. data sets.

Outline

Goals & Objectives

Although ParetoGP has been applied to a wide variety of domains over the past twenty years and there are a number of publications where it has been used for biological data analysis, the methodology and best practices of attack for such data types have not been the explicit focus — which we address herein. We use four real-world data sets as surrogates for typical data seen in cell therapeutics, immune systems and multiomics with the latter representing the integration of genomics, metabolomics, proteomics, etc.

These data sets are typically ill-conditioned in the sense that there are more variables available than data records with a confounding problem that many of the variables are coupled by the biological pathways creating the data. The focus here is the methodology and general benefits; for problem-specific insights, the reader should go to the indicated reference publications.

Illustration Data Sets

To motivate the methodology and workflow we will describe, we will use a variety of data sets against which we have applied ParetoGP . In general, we have a limited number of patients — human or rat surrogates or lab reactors — providing measurements of physical or chemical responses along with system measurements. Inclusion of genomic, proteomic, cytokine or molecular (NMR) data quickly results in a super-wide data set with few records (patients) and thousands of variables. Longitudinal studies over time can further augment the data width.

MSC Metabolites

The immune system can be a blessing or a curse. Mesenchymal Stromal Cells (MSC) have potential for immunomodulation with benefits for regenerative medicine as well as treating immune disorders [*Maughon, et al*]. For this data set [*Facundo*], we have twenty records with an associated functional score and 8,282 metabolites — pretty much a poster child for a super-wide data array. Our goal is to identify those metabolites and cytokines which are most predictive for the functional score (lower is better). Identifying critical quality attributes (CQAs) will help to exploit MSC as a therapy.

CMaT T-cell Quality Control

Cancer escapes detection and response by the body’s immune system. Chimeric antigen receptor (CAR) T-cells are a cancer therapy approach which trains CAR T-cells to recognize cancer cells and attack them — essentially replacing broad-brush chemotherapy with a targeted immune system response. Although this approach has been successfully demonstrated, growing the requisite T-cells is a costly process. The intent of the CMaT (Cell Manufacturing Technology, cellmanufacturingusa.org) data analysis is to identify critical parameters to improve the efficacy and efficiency of production of the targeted CD4 & CD8 T-cells. In this case, an initial data set (18 records and 264 variables) from a designed experiment was analyzed and ParetoGP indicated desirable regions of parameter space outside the original design. This Active Design-of-Experiments exploited the unique capability of model ensembles to extrapolate cleanly.

The expanded data set only features 30 data records with cytokine and NMR measurements sampled at intervals during the batch process with the analytics goal to identify the key factors and develop critical process attributes (CQAs) and critical process parameters (CPPs) for early nondestructive prediction of batch quality as well as possible interventions.

The targeted responses are CD4 and CD8 T cells which correspond to the killing cells and immune response, respectively. There needs to be a balance between these two attributes so we also encounter a need for multi-objective optimization

Bone Regeneration Biomarkers [Guldberg]

After trauma, a risk factor for bone regeneration and successful healing is a dysfunctional immune system. The goal of this data set (from rats) is to identify early biomarkers predictive of the long-term healing success. Early intervention in such situations is important from the patient perspective. In this case T-cells and cytokines are measured at intervals and this data set is a surrogate for longitudinal studies.

Single Cell Multiomics ⟹ Summary Statistics Analysis

The original data set included a mixture of responses and liposome from 23 patients along with single cell RNA data. Since the response and liposome data are patient-specific (23 unique values), we converted the original data set dimension of 7,219 records × 1,333 columns into a 23 record × 8,452 column set by computing summary statistics for the single cell data under the rationale that since the single cell measurements fluctuate so much, a summary statistic would have to be used in practice. The (noisy) RNA data is also modeled directly for comparison.

The Application Space

Biological Data

Biological systems can be viewed as very complex and highly coupled chemical plants. As such, they are intrinsically nonlinear and teasing insights from such systems is difficult and needs to respect the complexity as well as difficulty in collecting the requisite data to identify the key factors driving the system to develop robust predictive models.

Hundreds or thousands of variables and relatively few data records pose an intractable problem for machine learning techniques such as deep learning which implicitly assume every data record matters. Additionally, for biological systems, an assumption of linear or polynomial model structure is equally suspect. Correlated and coupled variables add another degree of difficulty which violate implicit assumptions of many analysis techniques.

Definitions of Success

Success in analyzing these data sets is multifaceted:

- identify the controlling variables/factors and variable set possibilities
- understand the variable associations and interactions
- develop transparent, trustable and interpretable predictive models
- understand response behaviors and trade-offs
- predict optima and possibilities
- explore competing goal trade-offs

Given the difficulty of such data sets, some measurements may be easier than others so while exploring the modeling possibilities we also want to investigate the options for operational use. We also desire ease-of-use for the analyst and avoiding the need/temptation for the vigorous hyperparameter tuning that seems to be endemic to much of the machine learning and statistical model techniques. Chasing over-fitting generally doesn’t end well.

Data Challenges

In addition to the foundational problem of wide/ill-conditioned data sets, analysis is further complicated by correlated and coupled variables, un-captured factors and reproducibility issues across labs/people/time. Data organization and curation is a consistent problem (a data pile of 27 spreadsheets should not be confused with a data set). Surprisingly, simply choosing a target to model can sometimes be difficult since *in vitro* targets may not correspond to *in vivo* behavior. Missing data elements must often be addressed — something that is sometimes concealed by zeros, for example, being substituted for the missing values.

Attempts at standardization can also be problematic with mandates for open-source code or standardized tools (e.g., python or Palantir) precluding innovative analysis solutions.

Data Trends

The path followed by the chemical process industry is being pursued by the regenerative medicine [*SystemsThinking*] and cell and gene manufacturing [*BioInsights*] so, with concepts like systems thinking, multi-scale modeling, Industry 4.0, Quality-by-Design, digital twins, etc being embraced, the amount of data collected and the need to efficiently and effectively convert it into actionable insight will continue to grow.

ParetoGP Foundations

Essential Assumptions

Every modeling approach makes assumptions. We should emphasize that our presumption here is that we are interested in predictive modeling — i.e., we have one or more known target responses that we are interested in modeling.

- Linear modeling assumes that the model structure is known (or reasonably approximated) and the variables are known
- Neural nets and deep learning assume that every variable matters and that a sigmoid is a reasonable basic model structure and that lots of data is available to tune the myriad hidden parameters
- Support Vector Regression and Gradient Boosted Trees assume that a Gaussian located on selected data points is appropriate and that the driving variables are known
- Random forests assume that decision trees are a good model form
- etc.

Obviously, given their adoption, all of these can be very good modeling techniques if their implicit assumptions are aligned to the nature of the application space. In a similar vein, ParetoGP makes fundamental assumptions:

- Simple and accurate models are desirable
- Algebraic models in relatively few (explicit) variables are appropriate
- The proper trade-off between complexity and accuracy is an emergent property and should be determined by the data

Beyond those constraints, the data is free to define the appropriate model form — evolution can be quite innovative as well as not aware, for example, that a proper model should be a second-order polynomial with no cross-terms.

Evolutionary Basics

The model search is stochastic with better models awarded breeding and mutation rights and, hopefully, over the course of multiple generations of evolution, good models are developed. Generally, we start with an initial population of random models; however, the population can be seeded if desired.

Defining Model Quality

If you want to grow a better potato, you first have to be able to characterize a better potato. In a similar fashion, if we want a simple and accurate model, we need to define those terms. For purposes of ParetoGP, an accurate model is generally measured by the correlation of the model, R^2, to the targeted response. Since correlation is a shape-matching function, we need to scale and translate it to fit to the observed data — this is easily and efficiently achieved by a least squares fit, i.e.,

$y = a + b * f\ (vars)$

where f is the evolved expression and a and b are the translation and scaling coefficients, respectively. Since these coefficients are easily computed, we avoid making the search harder than it needs to be [Keijzer]

Model complexity is defined as a structural metric — simply the number of nodes traversed from the root node to all nodes summed together. The attraction of this metric is that it is very simple and fast to compute and a reasonable surrogate for complexity while providing a finer resolution (aka, larger fitness landscape) than simpler metrics such as leaf counts or model depth.

Although not presented to the user, hidden objectives can be used during the model search to reward the use of fewer variables or basis functions as well as to promote novelty (e.g., model age)

Selecting Better Models

The models are characterized by their performance in a multi-dimensional quality space (typically, complexity, accuracy and, optionally, age). For simplicity, we define these so that smaller values are better (e.g., use $1 - R^2$ rather than R^2). The Pareto front of the models identifies those models which are optimal in the sense that, for example, there is no model simpler which is more accurate.

Rather than being greedy and working only with those models on the Pareto front, we choose a stochastic approach where we randomly sample a fraction (default is 10%) of the population and any models which lie upon that subset's Pareto front gets propagation rights. This ParetoTournament random sampling is repeated until we achieve the desired number of models to propagate the next generation. Although identifying the Pareto front does not scale well with large numbers, this divide-and-conquer approach has two benefits: (1) it is computationally efficient for the small number of models sampled in each tourney and (2) the sampling focuses the modeling selection on the knee of the ParetoFront — the knee being the inflection point where increases in model complexity provide diminishing gains in accuracy. Since knee models are most interesting from a practical standpoint, focusing the evolutionary effort on that region is appropriate. (There is no need for explicit bloat control in ParetoGP due to the evolutionary preference for simplicity — although there is a loose upper bound on the allowable complexity for safety reasons.

As an aside, models on the Pareto front of the overall population get a free pass into the next generation to avoid losing quality results across generations — although they may lose their primacy to new models.

Search Time, Population Size and Independent Evolutions

Most data modeling techniques are greedy and try to identify THE model. As we shall see later, the goal of ParetoGP is to identify diverse good-enough models and extract insight as well as trustable model ensembles from that collection. Towards that end, we prefer to spend the computational energy on multiple shorter searches with smaller populations, analyze those results for insight and guidance and repeat the process to iterate towards a final set of variables and models.

The current best practice is to run 15–30 independent evolutions with a population size of 300 models for a duration appropriate for the data size and modeling difficulty — typically 3–30 minutes. Searches can be distributed in parallel to available CPUs on the computer so wall time to achieve results are generally not onerous.

Modeling Nuances

Symbolic regression is an appropriate technology for large but not huge data sets — aka, up to tens of thousands of variables and hundreds of thousands of records. Model search efficacy and efficiency can be boosted by appropriate adjustments as discussed in the following paragraphs.

Direct Model vs Basis Set Search

As mentioned before, we search using correlation since that mitigates the need to discover the easily determined scale and translation coefficients. This is appropriate for multiplicative/monolithic models; however, if we have a system which is intrinsically additive, $y = a + b_1 f_1(\text{vars}_1) + b_2 f_2(\text{vars}_2) + \ldots$, then we are implicitly demanding that the evolutionary process discover both the basis functions, f_i, as well as the corresponding coefficient, b_i. Recognizing the analogy to the direct model search, we can opt to search for the basis sets directly and compute optimal coefficients as needed. This imposes an additional least-squares-fit overhead on every model evaluated but tends, on aggregate, to be beneficial for additive systems.

The results from the basis set search approach can be viewed as linearizing transforms — which can help to make symbolic regression results amenable to statisticians who can, as a result, bring into play their suite of linear analysis tools.

Handling Lots of Data Records & Lumpy Data

Lots of data doesn't necessarily mean lots of information but it does impose a computational burden as we evaluate the model performance against the entire data set. Since our goal during model search is simply to determine which models are better for purposes of genetic propagation, we can use OrdinalGP [*OrdinalGP*] and, for each generation, use a different randomly selected data subset for model assessment. There are two benefits to this approach: (1) we have a computational efficiency gain due to the avoided computational load and (2) we have an efficacy gain since models are rewarded for their generality since the fitness landscape is dynamically changing.

Data sets are often lumpy — especially for systems with closed-loop control (e.g., chemical or biochemical systems) — in the sense that data records may not be uniformly distributed across parameter space. Such data effectively overweights the lumpy regions and, implicitly, underweights other regions — which is contrary to our desire for a global model. Ideally, we would subsample the data to uniformly cover both the targeted response as well as the parameters used by the model. Unfortunately, we generally do not know which of the inputs will be used in a given model so the fallback position is to partition the response into bins (default is to use the Rice rule of $\lceil 2 \sqrt[3]{n} \rceil$ which is used for histograms) and randomly sample from each bin with any deficiency achieved by randomly sampling the overall data set. This variant of OrdinalGP is known as BalancedGP. As implied by the default behavior shown in Figure 1, we can be aggressive in terms of computational savings.

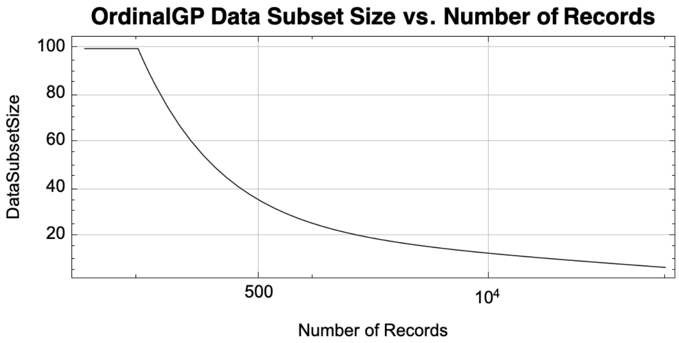


Figure 1: Subsampling the data with each generation can provide a dramatic increase in the efficiency of model search as well as reward model generality

Dealing with Few Data Records

At the opposite end of the data set spectrum are wide data sets where data records are precious. This is a space where ParetoGP provides unique capabilities. As we shall see, data sets with thousands of coupled variables and very few records can be handled. We do have to impose some constraints — foremost is that we want to avoid chasing phantom accuracy. Hence, the default restriction is that any model can contain no more than 20 percent of the number of records for either the number of variables or the number of basis sets. Preferably, even this upper bound will not be used in any deployed model. If we have strictly linear model, $\hat{y} = \sum_{i=1}^{N} \alpha_i x_i$, we can get a perfect model if the number of variables used is one less than the number of observations. Equivalently, N nonlinear transforms of a single variable (aka, a *basis set* of N), we also achieve a perfect model. The default behavior is intended to mitigate both of those risks.

Figure 2 illustrates the risk of spurious correlation with very small data sets so we are making an implicit assumption that the data represents reality rather than noise. Figure 2 also should reinforce the need for the analyst to apply domain knowledge to validate the plausibility of the chosen model variables.

We can also use OrdinalGP and choose different data subsets for each generation — think of it as leave-one-out-cross-validation-on-steroids. Although it is possible to have explicit training/test/validation sets, it is generally better to exploit the diversity of the developed models and have a fully informed ensemble of models when dealing with situations where information is precious.

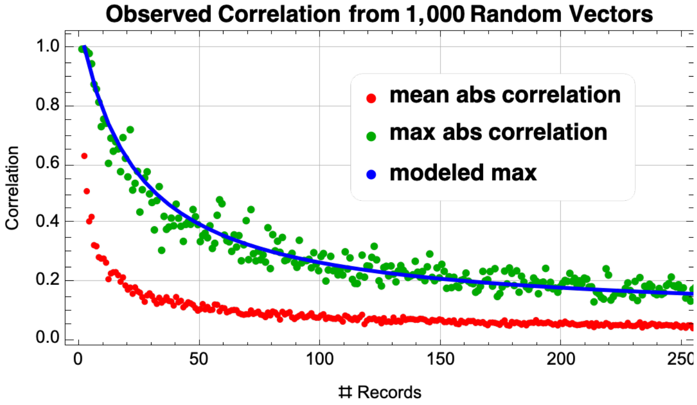


Figure 2: The observed correlation of a random vector relative to 1,000 other random vectors is shown as a function of vector length. For length 2, the correlation is always 100% with the spurious correlation decreasing as additional data records are added. The key is that the risk of spurious correlation decreases with additional data but increases as additional variables are added — assuming the variables are random.

Quantized Data, Interval Arithmetic and ANOVATrim

Models developed against quantized data — and small data sets are implicitly quantized — have the problem that they are unconstrained in the interstitial region. This imposes a risk that we can have singularities in those regions which are not detected by the data assessment. An efficient, albeit conservative, means to address this situation is to use interval arithmetic [Keijzer] (aka, RobustModels) to kill off models which have the potential to have a singularity. This can be applied either during model development or in post-processing. This feature is disabled by default since when working with coupled variables, a singularity can be the correct behavior. However, for very short data sets, the user should probably err on the side of caution.

An ANOVA table identifies those additive top-level terms which are independent but implicitly make assumptions on the distribution (normal) of the response as well as the inputs having a balanced distribution. Although this may not be correct in practice (especially with small data sets), it is also probably conservative. Hence, out of an abundance of caution we can apply ANOVATrim to models — i.e., simply examine the ANOVA table of each model and iteratively remove top-level terms until all satisfy the specified p-value threshold (default is 0.005).

Ensembles — aka,Trustable Models

During the course of an analysis, we will slowly focus from hundreds or thousands of variables to a very few which are interesting, plausible and desirable in terms of measurability and reliability. Using this focused variable set, we can have hundreds or thousands of models in the simple-but-accurate region at the knee of the ParetoFront — all of which would be a viable model from a performance standpoint. Rather than choosing THE model from this candidate set, a better alternative is to choose diverse set of models and use this ensemble as a super-model for operational use. This ensemble will have several properties:

- the models will agree near known data points (otherwise, they would not be good models) and
- the models will diverge when exposed to new regions of parameter space (since they have been chosen for their diversity)

The default ensemble definition process uses the lack of correlation of error residuals as a surrogate for diversity and over-weights the knee region. Because we exploit the model diversity, as a matter of practice we want a loose definition of the knee of the ParetoFront with some candidate models a little too simple and some a little too complex. The median of the ensemble will be a good predictor — and extrapolator — and the spread of the constituent model predictions provides a trust metric on that prediction.

The Analysis Workflow

The analysis flow when dealing with a plethora of candidate variables is fairly simple:

- identify those variables which are useful in predicting the target behavior
- iteratively focus on the most impactful variables considering both modeling performance and mechanistic *a priori* rationales as well as ease of application (e.g., ease of measurement, cost of measurement, etc.)
- once interesting variable sets with few variables have been identified, develop more models using those variable sets
- build ensembles (trustable models) from diverse accurate-but-simple models of the small sets of interesting models

The role of the data owner

Since the results of ParetoGP are white-box algebraic models and, in the presence of coupled and correlated input variables, the data expert can and should be intimately involved in the variable selection and model selection — i.e., we are pursuing a path of augmented intelligence rather than artificial intelligence. The situation of few data records places an additional onus on the analyst to confirm that selected variables are plausible and not a situation of spurious correlation.

At the same time, since the data is being allowed to speak for itself in terms of variables used and model forms developed, the analyst must be willing to listen and learn.

Model Evolution

Each model search typically beings with a random and different collection of models (although, if desired, we could seed with an initial model or population) and follows its own stochastic path as is illustrated in Figure 3. We want to execute a reasonable number of independent searches and let them run for a reasonable number of generations. Our goal is to be pragmatic and identify results upon which we can build to get useful solution.

The evolutionary engine can be viewed as an automated hypothesis generator/refiner whose only constraint is a preference for simplicity and accuracy — presumptions of model form or a need for linearity are not part of the criteria (although, if system can be modeled as a linear system such would emerge). Truly, the data is given an opportunity to speak for itself in terms of possible model structures.

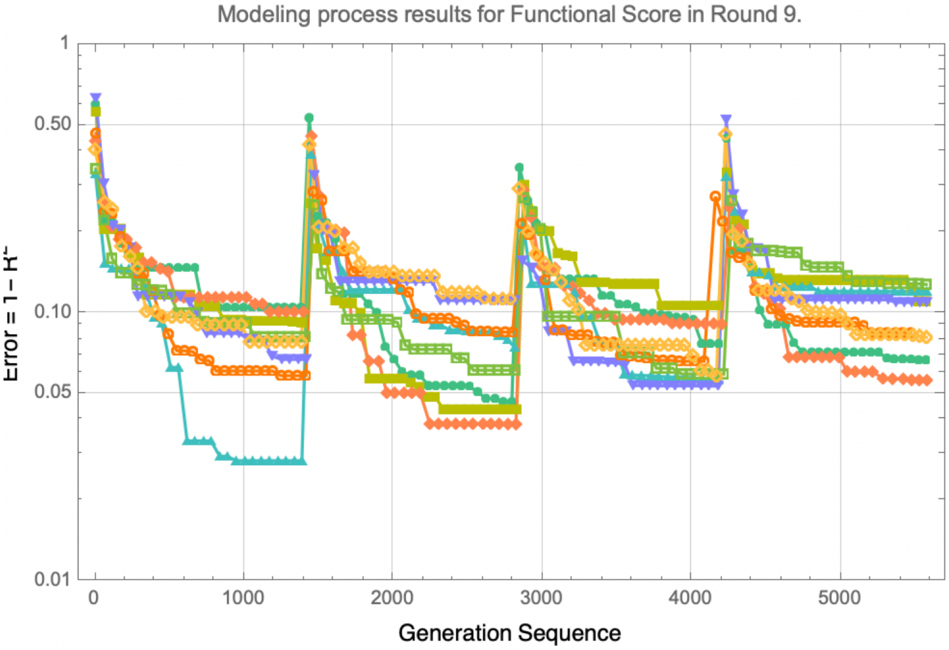


Figure 3: Here we are looking at the trace of the highest accuracy model for 32 evolutions for the MSC metabolite data (8,282 allowed variables) on a MacBook Pro. Each evolution follows its own path from the initial set of randomly generated models with different variables and demonstrates continual improvement as variable combinations and models are refined. Although the initial models are not very accurate, over time better models are developed as models are refined, new variables introduced and new combinations explored. Although the default for a short (20 record) data set like this would impose an upper bound of four variables in any given model, we have imposed a harsher cap of three variables — although, all variables are eligible for model inclusion.

Interesting Models — the Knee of the Pareto Front

The trade-off explored by imposing a preference for simplicity and accuracy focuses the evolutionary effort. As illustrated in Figure 4, the models providing the best bang-for-the-buck are those near the inflection point. These are the set which warrant further study. In defining the interesting region, we need to resist chasing accuracy — instead, we want a mixture of slightly too simple as well as slightly too complex models. Normal practice is to keep some large fraction (50-70%) closest to the Pareto front of the models within the nominal targeted quality box to select a band along the Pareto front.

For small data sets, we probably also want to use interval arithmetic as well as apply ANOVATrim processing to minimize the risk of unwarranted singularities or chasing outliers.

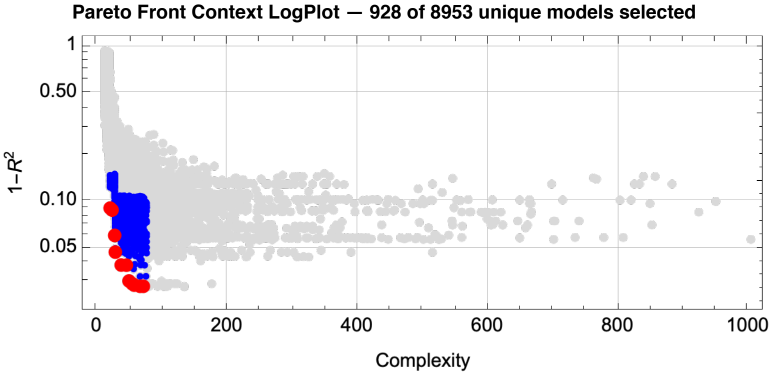


Figure 4: Merging the results from our searches can produce thousands of models. Most are either not acceptably accurate or are inappropriately complex. However, a fraction — generally about ten percent — lie in the region at the knee of the Pareto front and represent a good balance between complexity and accuracy. Such models provide the foundation for variable as well as model selection.

Model Dimensionality

One of the advantages of ParetoGP and its quest for simplicity is that we explicitly explore the number of variables needed for a model and can see the implications of chasing accuracy. Figure 5 shows that although six variables can provide the maximum accuracy models, we may be able to achieve good-enough accuracy with fewer variables and garner more insight in that process.

Model Dimensionality of 1375 ratio•of•CD4•to•CD8•Memory•Cells Day 6 Models

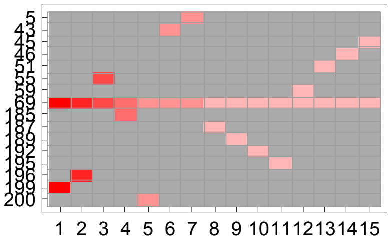
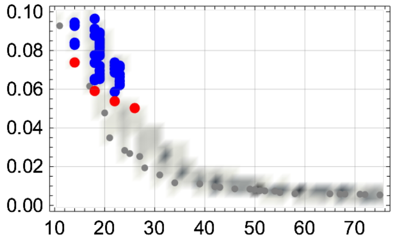
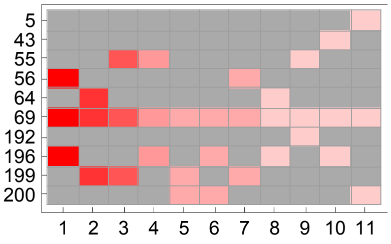
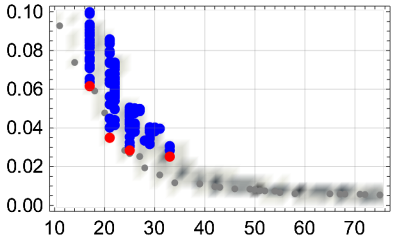
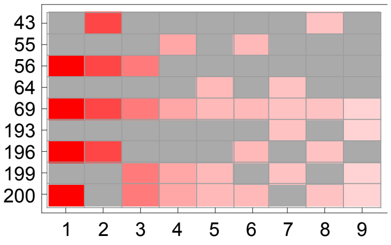
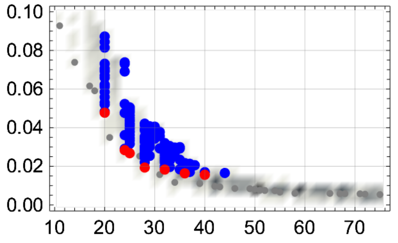
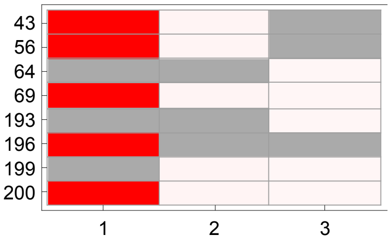
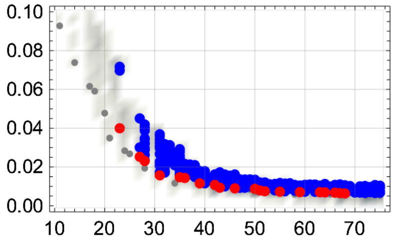
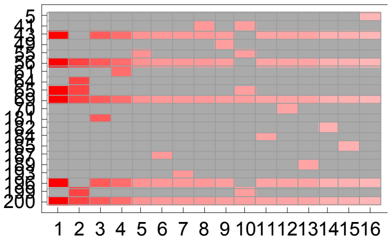
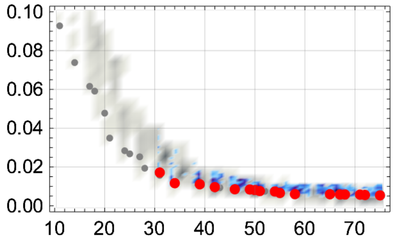
# Vars	# Models ⟹ %	VariableCombinationMap	ParetoFrontPlot
2	73 ⟹ 5.3 %		
3	134 ⟹ 9.7 %		
4	184 ⟹ 13.4 %		
5	456 ⟹ 33.2 %		
6	523 ⟹ 38.0 %		

Figure 5: The search for simplicity means that we can look at the trade-off on number of variables (aka dimensionality) vs. accuracy and complexity. Here we are looking at models developed against the CMaT data set using all the Day 6 variables as well as the process parameters (53 variables total and 30 data records). Since the default requirement of each variable being supported by five data records was used here, the number of variables in any given model was capped at 6; however, we can get quite good models with just two variables and four seem to provide most of the possible accuracy. The variable combination map shows that diverse combinations can provide good predictions.

Variable Selection — Making the Haystack Smaller

Our premise in ParetoGP is that a handful of variables are sufficient for a good predictive model. There are three basic views to aid us in our focusing on the most impactful:

- *Variable Presence* — the preference for simplicity means that the interesting models feature impactful variables. As indicated in Figure 4, even though we have a large search space, quality models can be discovered. In Figure 6 we see that there are different paths to quality due to high levels of variable coupling. Although not exhaustive, we can identify good-enough variables to provide a foundation for further exploration and refinement.
- *Variable Combinations* — We can examine the variables used by individual models or groups of models and select or combine them for additional development. This is illustrated in Figure 7. Proven ability to generate a desirable model is a good thing.
- Variable Associations — We can relax the specificity of the variable combinations and look at which variables associate with another and select those associations for additional investigation. Although pairwise relationships are illustrated in Figure 8, we can also look at triplets as well as higher-order associations.

We are helped in our search for useful variables in the biological space by the natural coupling of metabolites, cytokines and other factors so there is a reasonable chance of stumbling across a useful factor in the stochastic search. However, the evolutionary ability to leverage and build upon those discoveries is remarkable.

Our assumptions in selecting variables is very simple but in situations like the illustrations, we want to proceed slowly and let the computer do the heavy lifting in terms of variable selection rather than jumping directly onto promising early results. Although the evolutionary process is greedy in the sense that useful variables will be rewarded if they are discovered, it is not greedy in terms of only variables with a high linear correlation to the target are allowed to participate in the model development process. The implications of this are illustrated in Figure 9.

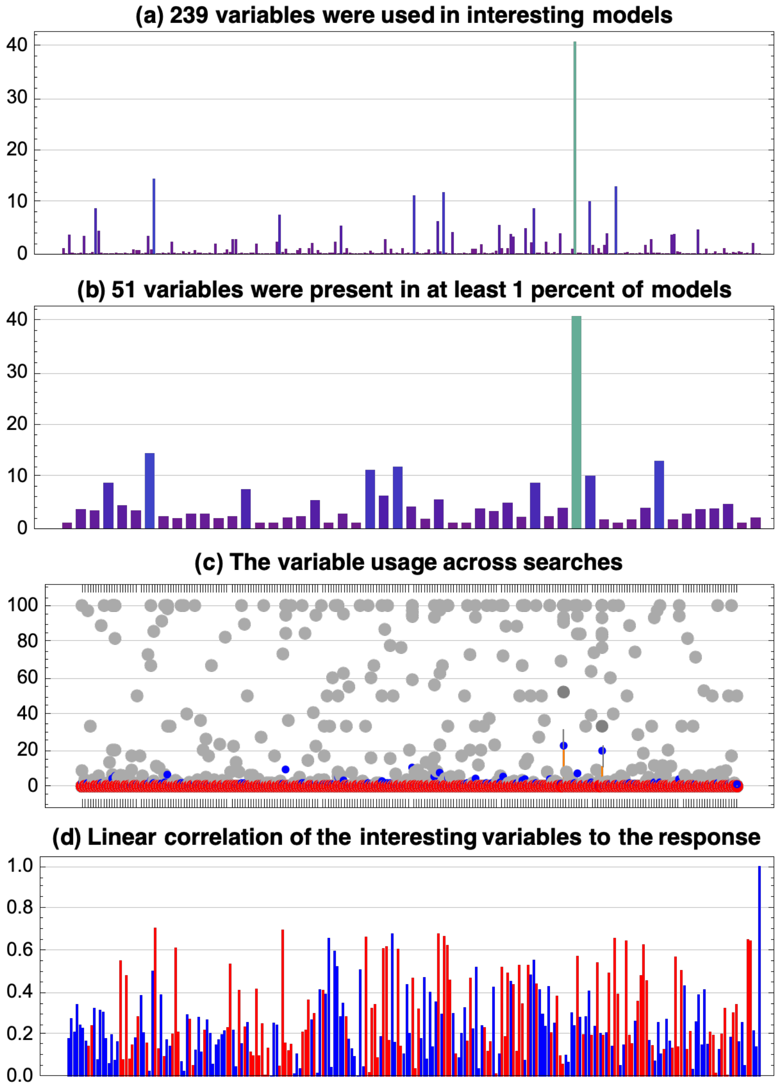


Figure 6: The model set used in Figure 4 is the result of 64 independent evolutions each of approximately 200 generations. Hence approximately 4×10^6 models (not variable combinations) have been evaluated. There are 5×10^{11}unique triplets of the variables. We are truly looking at a needle-in-a-haystack problem. In (a) we see that 3% of the available variables appeared in any of the interesting models. In (b) we see only 20% of those were in 1% (~10) of the selected models. In (c) we look at the fractional presence of variables across the 52 searches which contributed to the interesting models. Despite the support of novelty and innovation, the founders effect coupled with different initial random models allow some variables to emerge and dominate the quality model results of a particular search. In (d) we show the selected variables were not chosen based upon linear correlation to the response. In fact, m5098 which was in over 11% of the models only has a 3.7% linear correlation to the target.

Variable Combination Table

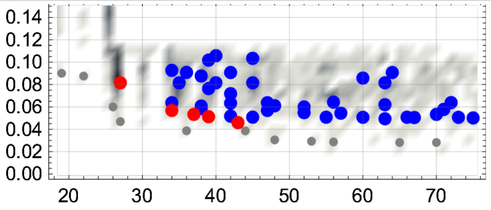
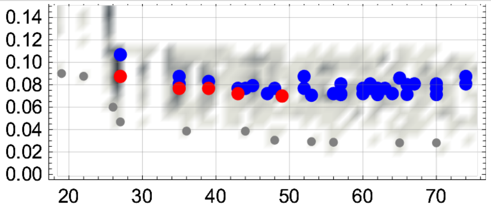
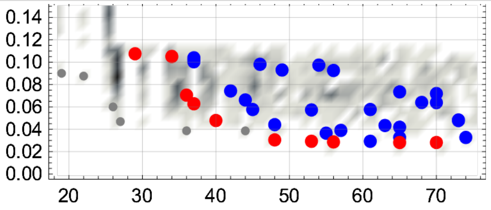
	num ⟹ %	Variables Used	ParetoFrontPlot
1	50 ⟹ 5.4 %	m4284 m5477 m6536	
2	46 ⟹ 5.0 %	m5889 m6072 m6162	
3	36 ⟹ 3.9 %	m5477 m6670 m7431	

Figure 7: Continuing to look at the identified interesting models, we can examine which variable sets are the most popular to see if such jump out in terms of intuition and performance to be worthy of focused development. In this example we show the three most popular of the 243 combinations present in the data. Alternately, we could explore the quality landscape and select variables from interesting individual models.

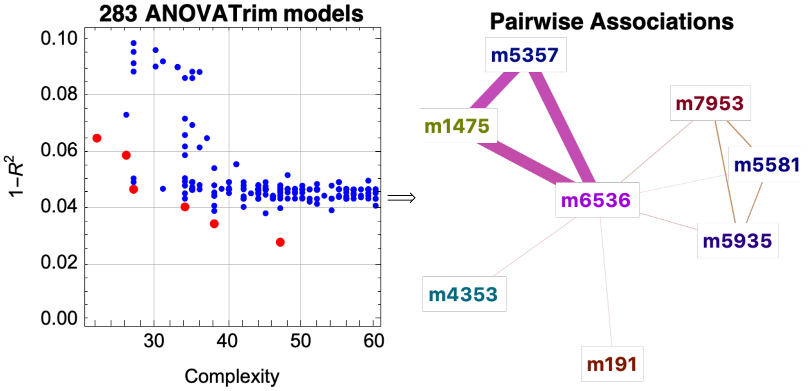


Figure 8: Here we have built 3-variable models from ten high-performing variables selected through multiple rounds of variable focusing from the original 8,282, selected the knee of the Pareto front and applied ANOVATrim to the result. The strength of variable associations is denoted by the line thickness. Despite the strong preference for one trio, a totally distinct set also yields quality models

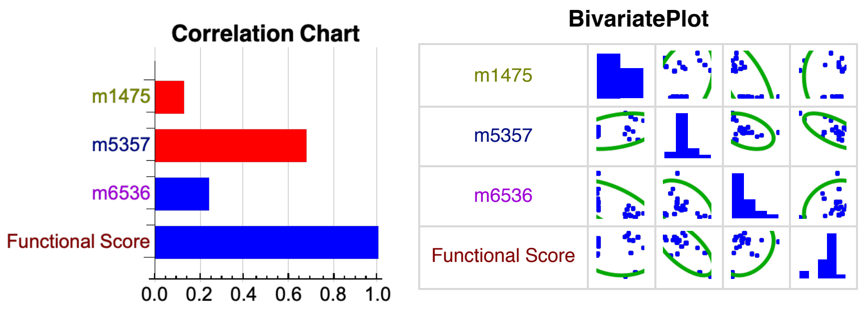


Figure 9: If we look at the most popular variable combination from Figure 8, we see that the variables were not chosen based upon their linear correlation to the response. Additionally, from the bivariate plot we can see that the chosen variables cover the design space pretty well.

MetaVariables & Basis Functions

The explicit algebraic expressions developed during symbolic regression can be mined for MetaVariables — i.e., the small building blocks that are discovered and, presumably due to their usefulness, propagate through the evolved population (Figure 10). Discovered relationships can provide mechanistic insight as well as reused as pre-defined building blocks for subsequent rounds of model development.

Closely related to MetaVariables are the top-level additive terms used if basis search has been enabled. These can be viewed as linearizing transforms since the resulting models can easily be dropped into a linear statistics framework. This attribute is leveraged when ANOVATrim is applied as a model filtering step.

CMaT MetaVariables from 29 Model Searches

Rank	⌗ models	MetaVariable	⌗ Evolutions	% Evolutions	Max Count	Max %	Mean %
1	71	Ethanol 1.1795 D6 IL15 − 6 LIF − 6	16	55.2	16	100.0	37.3
2	48	$\frac{1}{\text{OSM} - 6}$	14	48.3	16	100.0	34.8
3	46	$\frac{\text{IL2 Conc}}{\text{OSM} - 6}$	12	41.4	16	100.0	29.6
4	32	IL2 Conc OSM − 6	8	27.6	15	100.0	15.3
5	55	Ethanol 1.1795 D6 IL15 − 6	13	44.8	14	100.0	31.8
6	15	$\frac{1}{\text{IL2R} - 6}$	4	13.8	11	100.0	5.4

Figure 10: MetaVariables are functional building blocks in developed models. Here we look for those simple structures which have been discovered and reused during the evolutionary search. These relationships can provide insight into mechanisms.

Model Selection and Ensemble Definition

To this point in the analysis workflow, we have looked at populations of models — and, presumably, iteratively focused the allowed variable set to a small number of meaningful and useful inputs. Although the insight gained in this process can be useful, capturing the full value of predictive modeling requires a predictive model. The problem with any empirical model is one of trust that the proper variables have been chosen and the model form adequately captures the targeted system behavior. As mentioned earlier, we prefer ensembles as a way to hedge our bets. The other recommendation is to not chase accuracy by adding variables and complexity. To illustrate the latter point, consider models developed against the bone regeneration data using week 9 information to predict the final bone volume shown in Figure 11. Any model is limited to no more than four since we only have twenty records available; however, we can also achieve reasonable accuracy with fewer variables — albeit, at a slightly lower accuracy.

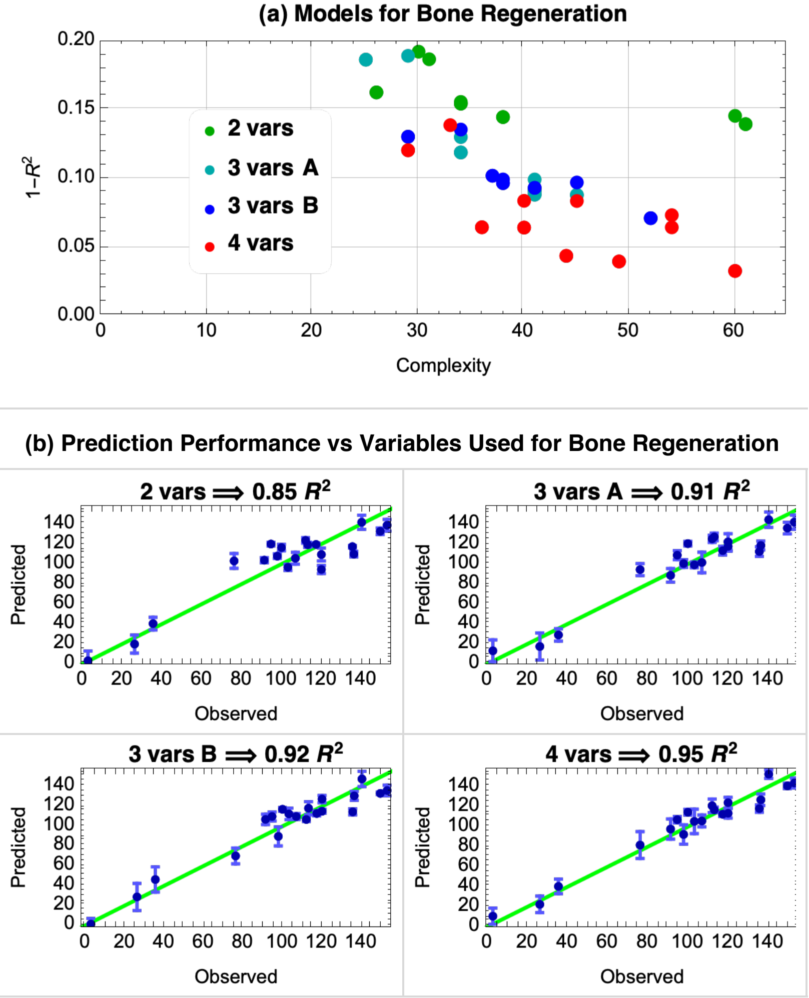


Figure 11: In (a) show the accuracy and complexity of models selected for four ensembles predicting bone regeneration with the models in each ensemble color-coded. Given that we only have twenty observations available for our modeling, is there a practical difference between the ensemble performance shown in (b)? Adding complexity and variables in the pursuit of accuracy does not, necessarily, provide a commensurate improvement in understanding or practical performance. In all cases, we have used interval arithmetic and ANOVATrim in our quest to be conservative in the model development.

In Figure 12, we look at the response plots of the bone regeneration ensembles at a specified reference point (in this case, one of the data records). By moving around parameter space, we can visually explore the impact of modifying each variable — which is important if the parameters are controllable. If we have included a diverse set of models in our ensemble (aka, a loose definition of the knee of the ParetoFront so some are under-fitted and others are over-fitted) then the constituent models will diverge when asked to extrapolate into new regions of parameter space — effectively providing a trust metric on the ensemble prediction.

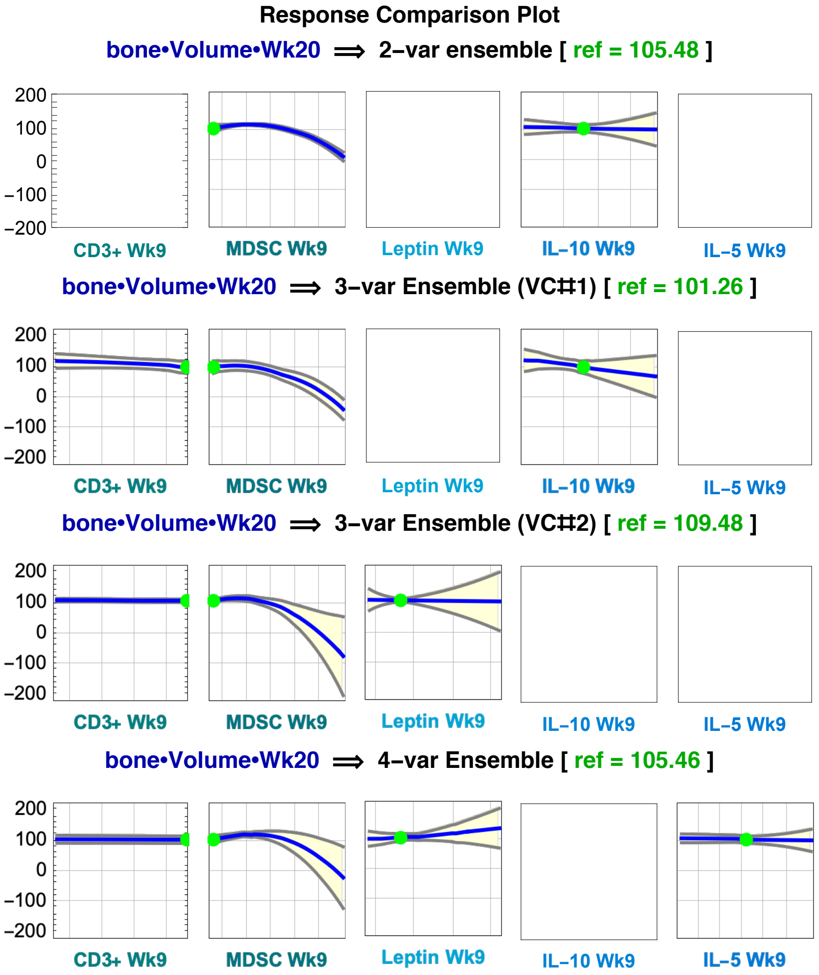


Figure 12: Here we look at the bone regeneration behavior with each plot showing the effect of changing the associated variable while holding the others at the reference point (green dot). The key take-away should be that the common variables across models shift similarly but the constituent models diverge (yellow envelope) when asked to operate in unexplored regions of parameter space.

ParetoGP ensembles tend to extrapolate reasonably well. To illustrate this, consider Figure 13 where we have built CMaT models using the original 18 point data set using only the process parameters and, then applied those models to the expanded data. Moving outside the nominal data range by +50% in two of the three variables is extremely aggressive but the ensemble does a unreasonably good job of extrapolation while also flagging the predictions as suspect due to the constituent model divergence.

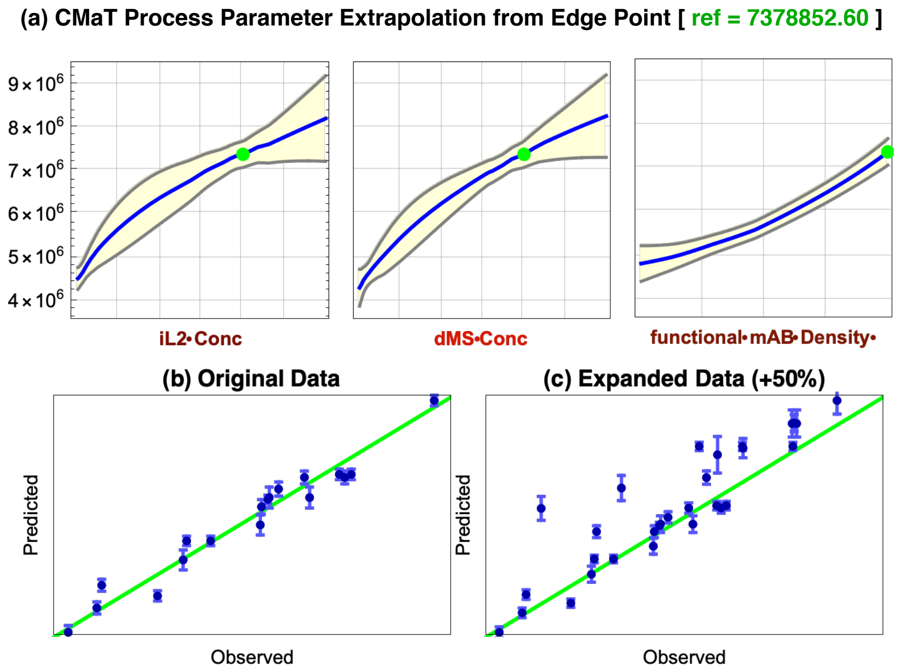


Figure 13: Initial analysis of the original DOE data for the CMaT data set indicated that regions outside the nominal data range would be desirable and, as a result, additional data was collected. (a) shows an extrapolation of original data set models using only the process parameters from the boundary point. As expected, the constituent models diverge when asked to extrapolate. In (b) we see that the process parameter models performed well against the original data set while in (c) we see that the original model does remarkably well when asked to extrapolate +50% relative to the observed data range in two of the three variables. Although anecdotal, the graceful degradation of ParetoGP ensembles seems to be endemic.

Data Cubes & Summary Statistics

The data sets considered for predictive analytics are, typically, record-oriented with each record corresponding to an observed target response. In some situations, the data can more accurately be considered to be three-dimensional with a multiple values corresponding from a single measurement. To illustrate, we might have a time-series associated with a given sensor or, in the case of the the single-cell multiomics, RNA measurements from individual cells of 23 patients so the nominal data set size of 7,219 records by 1,333 columns is deceptive. The RNA expression is highly variable so, as illustrated in Figure 14, the nominal model accuracy is not very high — it is a very noisy measurement. To take out the noise we deleted the entries for each variable which were not expressed and calculate summary statistics for the ~300 measurements associated with each of the 23 patients.

The ability to handle wide data sets is very powerful. The example used in Figure 14 also illustrates the risk of chasing model accuracy since a three-variable model would not have had the proper level of data support despite passing a variety of criteria. Examining the data of the selected variables is always appropriate.

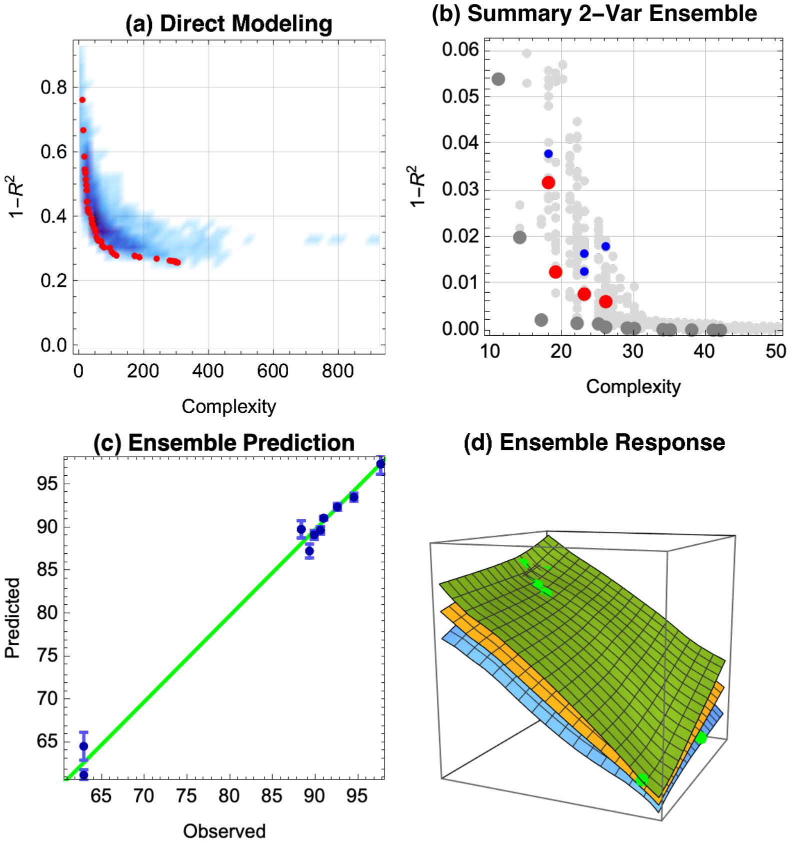


Figure 14: The original data set had ~300 measurements of each of RNA expressions for each of 23 patients. The noisy nature of the measurements made modeling difficult (a). In (b) we use the median, interquartile range and max for each variable to produce a 23×1,941 input matrix and get extremely accurate models despite restricting to a limit of three variables and three basis sets and requiring each model pass both ANOVA and interval arithmetic criteria. An ensemble was formed from a 2-variable subset of the models — the individual performance is shown in (b) — and prediction performance shown in (c) with the ensemble prediction and divergence shown in (d) along with the data records. Notice that the ensemble divergence increases away from the observed data points. The key observation here is that the 23 data records only had 10 unique records so it would have been inappropriate to use more than two variables. The best practice is that, if a model is unreasonably accurate, it and the data should be scrutinized harshly.

Trade-Off Analysis

To this point, we have focused on a single target variable. Often, however, we have multiple objectives which must be balanced for system optimization. Although the most predictive variables may differ, we can explore the modeling options for each of the targets and exploit possible variable substitutions to identify an acceptable common set as illustrated in Figure 15. Ideally, the chosen set will be controllable parameters rather than observed responses.

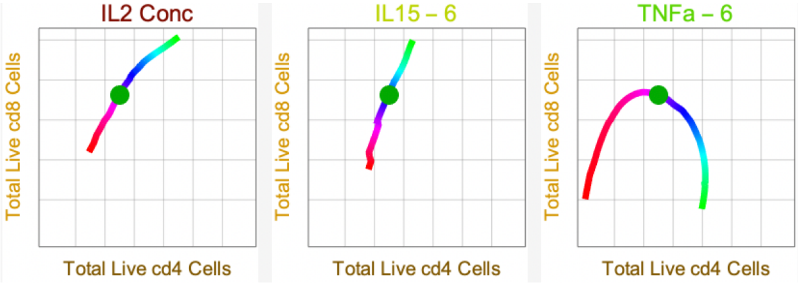


Figure 15: CD4 and CD8 T-Cells need to be balanced. Hence, here we have modeled both and selected a common set of variables which provided acceptable model accuracy for both (out of 53 day six variables). The plots explore the effect of changing model variables from the reference point (green dot) from the minimum observed value (red) to the maximum (green). The best variables for each response did not overlap so we had to look for a common set which would allow exploring the trade-off behavior.

Competing Technologies

Unfortunately, for the very wide data sets used here, the classic machine learning techniques struggle with the problem of too many variables as does statistical fit-regularization approaches such as LASSO. Iterative approaches to Partial Least Squares Regression (PLSR) is generally the most robust; however, the presumption of linearity generally precludes the variable focusing coming from ParetoGP. The required hyperparameter tuning for the others should inspire caution.

ParetoGP also naturally explores alternative variable combinations rather than being greedy in variable selection. The ability to develop diverse models and combine them into ensembles to effectively provide a predictive model with a trust metric is also critical.

Conclusions

Herein we have described a workflow appropriate for biological data analysis and illustrated the associated thought process and analysis tools using a variety of real biological data sets.

Multiomics biological data is intrinsically ill-conditioned with wide data sets comprised of correlated variables. ParetoGP’s exploration of explicit algebraic models rewarding simplicity as well as accuracy provides a mechanism to identify driving factors and insights into biological pathways and couplings. Combining diverse simple-but-accurate models enables robust prediction as well as a trust metric on those predictions to detect extrapolation into new regions of parameter space.

From an analyst viewpoint, the developed whitebox models in just a few variables provide clarity without requiring extensive hyperparameter tuning. Domain knowledge is important, however, in choosing the most desirable variables for inclusion in the deployed models since multiple variable combinations are often effective predictors and the risk of spurious relationships when working with very limited data sets is always a possibility.

Research Opportunities

Coevolution of variable sets effective against multiple targets or a viable alternative is something that is needed since the current approach is too manual and labor-intensive. Quantized inputs and responses is also a risk which we attempt to mitigate via the use of interval arithmetic; however, response behavior and stability in the interstitial spaces remains a concern.

Since fully exploring the search space is not practical, it may be desirable to use mutual information metrics on discovered variable sets to suggest alternatives which might also be explored for their viability.

References

[Maughon, et al] — Metabolomics and cytokine pro!ling of mesenchymal stromal cells identify markers predictive of T-cell suppression, Cytotherapy, 24 (2022) 137–148

[Facundo] — private communication providing data set

[Pradhan, et al] — Single-Cell Transcriptomic Attributes and Unbiased Computational Modeling for the Prediction of Immunomodulatory Potency of Mesenchymal Stromal Cells, bioRxiv preprint

[Cheng, et al] — Early systemic immune biomarkers predict bone regeneration after trauma, PNAS 2021 Vol. 118 No. 8 e2017889118

[Odeh-Couvertier, et al] — Predicting T-cell quality during manufacturing through an artificial intelligence-based integrative multiomics analytical platform, Bioeng Transl Med. 2021;e10282

{Ghassemi, et al] — Rapid manufacturing of non-activated potent CAR T cells, NATURE BIOMEDICAL ENGINEERING,

[BioInsights] — Expert Roundtable, Cell & Gene Therapy Insights, 2021: 7(4), 503-518

[SystemsThinking] — National Academies of Sciences, Engineering, and Medicine. 2021. Applying systems thinking to regenerative medicine: Proceedings of a workshop. Washington, DC: The National Academies Press. https://doi.org/10.17226/26025.

[OrdinalGP] — M. Kotanchek & N. Haut, Back To The Future: Revisiting OrdinalGP & Trustable Models After a Decade, GPTP 2021

[Keijzer] — Dissertation (Interval Arithmetic, use of R2)

[Kotanchek/Smits, et al] — appropriate ParetoGP reference