GE Global Research

Experimental Design for Combinatorial and High Throughput Materials Development

J.N. Cawse

2002GRC253, November 2002
Class 1

Technical Information Series

GE Global Research

Technical Report Abstract Page

Title	Experimental Design for Combinatorial and High Throughput Materials Development		
Author(s)	J.N. Cawse	Phone	(518)387-6095 8*833-6095
Component	Combinatorial Chemistry and Characterization Technologies Lab, Niskayuna		
Report Number	2002GRC253	**Date**	November 2002
Number of Pages	25	**Class**	1
Key Words	combinatorial; experimental design; DOE; experimental strategy; experimental planning; statistics		

In the past decade, combinatorial and high throughput experimental methods have revolutionized the pharmaceutical industry, allowing researchers to conduct more experiments in a week than was previously possible in a year. Now high throughput experimentation is rapidly spreading from its origins in the pharmaceutical world to larger industrial research establishments such as GE and DuPont, and even to smaller companies and universities. Consequently, researchers need to know the kinds of problems, desired outcomes, and appropriate patterns for these new strategies. This report is an overview of the targets, planning strategies, and statistical issues of this new field.

Manuscript received October 28, 2002

Experimental Design for Combinatorial and High Throughput Materials Development

J.N. Cawse

1.1 Introduction

As the 21st century opens, the opportunities and challenges to materials scientists worldwide have never been greater. Market globalization, environmental performance, customer expectations, population expansion, and resource limitations present a need for materials with new or optimized properties. Although progress has been made in our ability to design or predict the properties of new materials, discovery of most new materials is still solidly based in experimentation.

As chemical and materials knowledge has accumulated over the centuries, the properties of simple materials have become better known and new developments tend to occur in the regions of greater complexity. For example, "The properties of many functional solid-state materials arise from complex interactions involving the host structure, dopants, defects, and interfaces". [1] At this level of complexity the feared "combinatorial explosion" appears. If there are 76 useful, stable elements in the periodic table (Figure 1.1), there are 2850 binary, 70,300 ternary, 1,282,975 quaternary, and $>10^9$ heptanary combinations of these elements. In the case of organic structures, it has been estimated that there are 10^{63} "drug-like" molecules synthesizable from the standard organic toolkit of elements [2] – and these are only a subset of the total number of molecules potentially of interest to materials scientists. Finally, the ultimate properties of useful materials will also be a function of relative compositions and of processing conditions. This further increases the number of possible combinations. Subsets of this enormous number of possibilities form the "experimental space" or "chemical space" of each individual program.

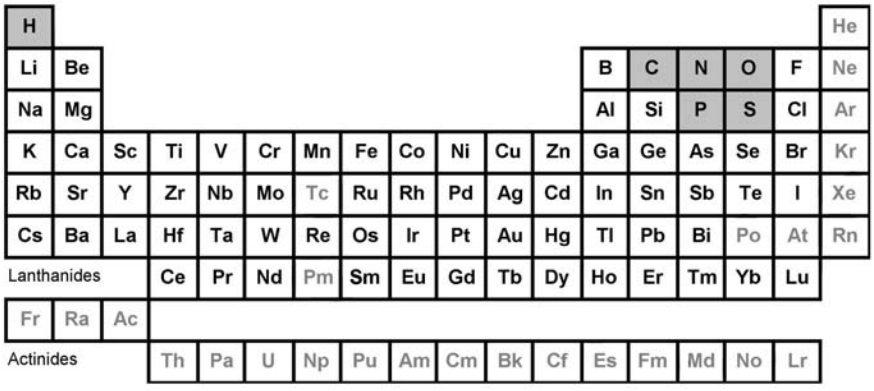

Figure 1.1. Elemental Constituents for Combinatorial Chemistry. Bold: the 76 non-radioactive, non-noble elemental possibilities. Grey background: the 8 elements in the standard organic toolkit.

Combinatorial chemistry and high throughput experimentation have emerged during the last decade as a response to the challenges of materials development in these increasingly

complex experimental spaces. Although its earliest antecedents can be traced to the beginning of this century, combinatorial experimentation really began to take off around 1990 in the pharmaceutical industry [3] and 1995 in materials development. [1,4] This resulted from a convergence of technologies in robots, semiconductor processing, computer software, and analytical capabilities, and from the simple realization that such productivity was possible. The rapid growth of this technology has led to a proliferation of terminology; we will try to use the definitions in Table 1.1. The general term "Combinatorial and High Throughput Experimentation" (CHTE) will be used in this book for the entire field of parallel, miniaturized, high speed experimentation.

Table 1.1 Definitions of CHTE Terms

Experimental space	All possible combinations of composition, formulation and processing variables. in the system under study
Chemical space	All possible molecules within a given set of constraints.
Property space	Multidimensional representation of a set of compounds in which the axes represent quantifiable properties, such as molecular weight, ionization potential, etc., and individual compounds are represented by a vector or set of coordinates.
Combinatorial design	Experimental design whereby relationships between variables can be elucidated through the evaluation of their combinations
Combinatorial library	A set of compounds prepared by combinatorial synthesis or formulations prepared through a combinatorial design.
Combinatorial synthesis	Use of a combinatorial process (e.g. split/pool) to prepare sets of compounds from sets of building blocks.
Parallel Synthesis	Synthetic strategy whereby sets of discrete compounds are prepared simultaneously in arrays of physically separate reaction vessels.
Split/pool Synthesis	Synthetic strategy for assembly of a combinatorial library. A single compound is built onto each single support particle in a series of reactions using a process of splitting, reaction with synthons, pooling, mixing and redividing.
High throughput experimentation	The use of miniaturization, robotics, and parallel techniques to increase the productivity of the research process.
High throughput screening/analysis	Process for rapid assessment of the activity of samples from a combinatorial library or other sample collection, often by running parallel assays.
Factory	Informal term for the entire combinatorial/high throughput process
Hit	Library component whose activity exceeds a predefined, statistically relevant threshold.
Lead	A verified Hit
Descriptor	Numerical representation of a molecular property, including bulk properties (e.g. ionization potential, molecular weight), two-dimensional features (atom connectivities) or three-dimensional features (molecular shape).
Diversity	The 'unrelatedness' of a set of, for example, building blocks or members of acombinatorial library, as measured by their properties (e.g. descriptors).

The development of new experimental tools with awesome levels of productivity does not, however, exempt us from the need to plan our experiments well. Murphy's law continues to rule, and a poorly designed experiment will give us bad information with unprecedented speed and in outstanding quantities. In fact, these new capabilities will be used on harder problems, which require entirely new experimental designs and methods of data analysis and visualization. The field is developing very rapidly and we expect to see many more such design tools during the next few years.

1.2 Historical Basis

A discussion of experimental design for CHTE would be incomplete without at least a brief glance at the historical context. The art and science of experimentation has been evolving since the alchemists, but only in the last 75 years has systematic effort been made to include statistical concepts in the planning stage of experiments, rather than just the data analysis.

1.2.1 Edisonian Experimentation

Science and technology have no lack of examples of extensive – even heroic – experimentation. Table 1.2 gives a few historic examples of systematic and successful experimentation. Even "combinatorial" experimentation has precedents early in the century in the successful efforts to develop an ammonia synthesis catalyst. Mittasch reported the effects of 54 different metals (almost all the metals in the periodic table) on an iron catalyst, and found that "numerous combinations were found to possess catalytic activities." He quotes Willstatter that "multicomponent catalysts can be regarded, in their activities, as equivalent to new specific substances." [5]

Table 1.2. Heroic Experimentation

Experimenter	**Date**	**Goal**	**Approximate**
			number of runs
Thomas Edison	1878-1880	Electric Light	6,000
Alwin Mittasch	1909-1912	Ammonia Catalyst	20,000
Paul Ehrlich	1907-1910	Syphilis drug (Salvarsan)	900

"Edisonian" is a term sometimes used disparagingly in reference to CHTE, implying experimentation that lacks strong scientific underpinnings and is just near-random casting into uncharted waters. This is unfair to both Edison and modern CHTE practitioners. The intellectual effort and scientific rigor used by both to plan and carry out experiments are every bit as creditable as that used by more conventional experimenters.

1.2.2 Design of Experiments: Agricultural antecedents

Classical experimental design strategies grew up in the English agricultural research community in the 1920's, a period of slow, laborious, error-prone experimentation.

The landmark designs developed by Fisher [6] were done when one experiment *per year* was the norm (Figure 1.2). The rationale for designed experiments in that period was

principally one of determination of the main effects of the factors in the presence of enormous natural variation. Replication and blocking patterns were set to minimize and average out random factors such as soil type, wind, sun angle, cloud cover, and slope. Factorial designs and Latin Squares were the preferred arrays.

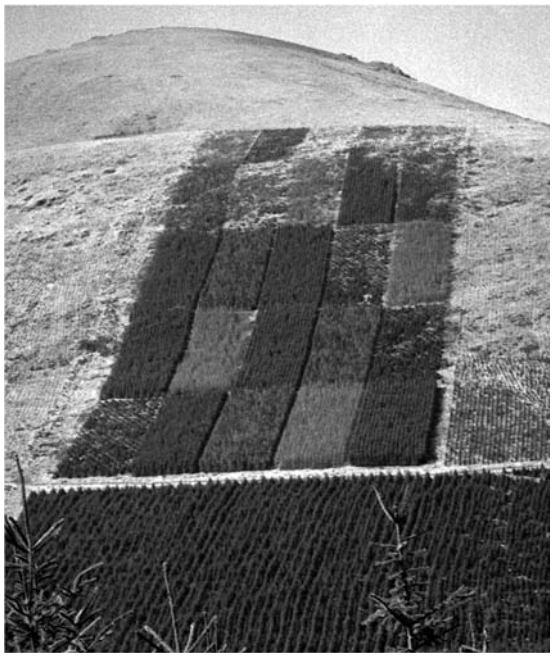

Figure 1.2. Low Throughput Experimental Design. A 5x5 Latin Square planted about 1929. P hotographed in 1945. Reproduced with the kind permission of the Forestry Commission, UK. J.F. Box, "R. A. Fisher, A Life in Science", Plate 6, Copyright ©1978. John Wiley and Sons, NY.

1.2.3 Modern Industrial Design of Experiments

Classic industrial Design of Experiments (DOE's) [7,8] was introduced to the US by Box after World War II and developed in the 1950's and 1960's. Experiments continued to be expensive, since they were typically still done one at a time by highly trained professionals. DOE's are usually attempts to determine the main effects and 2-factor interactions in a minimum number of runs. Factorial and fractional factorial designs were developed and heavily used as simple, effective tools that could be employed by non-statisticians. The mathematical properties of 2-level factorials – from both the statistical standpoint of properties like orthogonality, and the computational properties useful in the pre-computer era – led to them becoming the standard tool. [9]

As statistical sophistication increased and computer use became the norm, more complex DOE's such as central composite designs, mixture designs [10], and optimal designs became common. These allow the experimenter to determine curvature as well as main effects and 2-factor interactions. All of these are now easily generated by standard software (e.g. Design-Expert [11], JMP [12], Minitab [13], and Echip [14]). The popularity of the Taguchi approach to experimentation [15] and the Six Sigma quality initiatives in many major companies [16] has made the use of DOE's increasingly common.

1.2.4 Design of Experiments for Combinatorial and High Throughput Experimentation

In screening for new materials, the emphasis is on the discovery of low probability, high value occurrences ("hits") by searching extensive experimental spaces. These are typically the result of combinations or interactions of three or more important factors, and the system may have dozens of factors, each with multiple levels. The total number of combinations of all of these factors can easily be 10^6, 10^9, or much more. Experimental strategy and tactics must examine this space with a high probability of finding hits, and allow completion within the lifetime of the experimenter (or, more important, the resources of the project).

The basis for Design of Experiments in CHTE must include the following:

- Careful choice of the experimental space with a keen scientific understanding of the system.
- Selection of factors and levels to get broadly diverse view of the possible chemical space that is being studied.
- Estimation of the level of complexity of the combinations and interactions that must be examined to have a reasonable chance of locating some hits.
- An understanding of the sources and amount of variation from the entire CHTE factory and the amount of "difference" which constitutes a hit.
- Recognition of the "Procrustean bed[Φ]" of the (usually fixed) experimental array size and the limitations of the available technology (robots, processing steps, analytical tools).
- Use of the rapid turnaround of CHTE for effective sequential experimentation.

As so often occurs in science, few of the 'new' methods examined here are strictly new. Inspired theft from other disciplines is a recurring theme. The other disciplines clearly include biology, biochemistry and combinatorial mathematics[17] but also physics, geology, computer science, information theory, and mathematical geometry. One observation may be useful to people searching for other methods: *chemistry has now accelerated to approach the speed of the early computers.* Ideas and algorithms that were devised in the 1960's to solve difficult problems on those machines may be adapted to the search problems chemists face now.

1.3 Commercial Targets

The National Institute for Science and Technology (NIST) has been a strong supporter of development of combinatorial methods, and its Vision 2020 Technology Roadmap for Combinatorial Methods [18,19] gave an overview of potential markets and key applications for CHTE. Table 1.3 is adapted from that document, and the following is a more detailed discussion of a few of the major commercial targets for CHTE.

[Φ] NEAR ELEUSIS, in Attica, there lurked a bandit named Damastes, called Procrustes, or "The Stretcher." He had an iron bed on which travelers who fell into his hands were compelled to spend the night. His humor was to stretch the ones who were too short until they died, or, if they were too tall, to cut off as much of their limbs as would make them short enough. (http://www.goines.net/Writing/procrustean_bed.html)

Table 1-3. Markets and Current Key Applications for CHTS

Material Type	Market	Key Discovery Areas
Electro/magnetic/optical materials		
Electronics	Appliances High Speed Circuitry Miniaturization Fiber Optics	Dielectric constant Multifunctional Materials
Photonics	Optical computers Nonlinear optics Optical Networking	Multifunctionality High speed High bandwidth
Magnetic materials	Transformers Power conversion equipment MRI machines	New superconductors Magnetic materials
Scintillators	Medical Imaging	New superconductors
Lighting	Fluorescent lamps Ceramic metal halide lamps Halogen lamps Light-emitting diodes	Phosphors Diode materials
Medical/Biotechnology	Lab-on-a-chip	Proteomics
	Patient-specific materials Biometric materials High-throughput sampling Biodegradable materials	Biomimetics Genetic-based drugs Diagnosis
Chemical		
Catalysis	Industrial chemicals Plastics Alternative fuels	Selectivity Yield Heterogeneous catalysis Homogeneous catalysis
Coatings, Adhesives, Lubricants	Automotive New adhesives Non-solvent based coatings	Multifunctionality
Chemical Process/ Product Design	Routes to commodity chemicals Third-world infrastucture Dream reactions Adsorbtion/separation processes New raw materials	Alternative pathways Process optimization Model refinement Kinetics
Energy Technology	Energy conversion Energy storage	Fuel cell components Battery materials
Other Materials	Non-lead solder Cement Ceramics Hybrid products	Formulation/process interaction Materials interaction Anisotropic materials Nanoscale materials Composites Smart Materials

1.3.1 Catalysts

Benefits

After drugs, the highest potential value for combinatorial chemistry may be the discovery of new catalysts. Catalysts are ubiquitous in industrial chemical processing, and a composition that performs a new reaction or makes an old one more efficient can revolutionize whole areas of the industry. Catalyst-based manufacturing has been estimated to produce over 7000 compounds worth over $3 trillion globally. [20] Improved catalyst effectiveness such as major yield and selectivity improvements will reduce waste and energy consumption and minimize feedstock costs. New routes to currently high cost materials will enable market entry of new feedstocks and raw materials.

Critical technology issues

Industrial catalytic processes have highly important kinetic (time-varying) properties, and are often done in aggressive conditions of temperature, pressure, and chemical environment. [21]

1.3.2 Polymers

Benefits

A significant fraction of new plastic materials are developed through the process of blending known polymers, adding property-modifying chemicals, or performing surface modifications. These processes enhance desired properties such as weatherability or moldability. These blends have become more complex as additional outstanding properties are required to meet the needs of modern products. In addition, the cycle time of the products is decreasing; new computer hardware appears on cycles of 6 months or less. The market for these highly engineered materials in the computer industry alone exceeds $1 Billion/year.

Critical technology issues

Polymer properties are often highly dependent on the details of the mixing, extrusion, and molding processes, and the properties themselves are defined in macroscopic terms (e.g., breaking strength, surface hardness). [22]

1.3.3 Lamp Phosphors

Benefits

Fluorescent lamp phosphors convert ultraviolet emission of a rare-gas/mercury discharge plasma into visible (white) light. General-purpose lighting consumes about one-quarter of all electricity produced in the United States. The incandescent lamp, known for its pleasing appearance to the human eye and its low purchase price, is a very inefficient light source. The replacement of incandescent lamps with compact fluorescent lamps (CFL's) could yield substantial savings in energy and overall lamp life cost with concurrent reduction in greenhouse gas emission from fossil-fuel power plants. CFL's use expensive rare earth based phosphors resulting in a high purchase price. Cheaper and/or

higher performing phosphors will allow CFL's to find higher penetration in residential applications.

Critical technology issues

Phosphors consist of crystal lattices of complex elemental composition plus the presence of doping elements. Defining the constituting elements of a potential host lattice plus potential dopants creates a large number of material libraries from which the most efficient composition would be chosen for further evaluation.

1.3.4 Scintillators

Benefits

Scintillators absorb x-rays and convert them to visible light for electronic detection. Scintillator based detectors are critical components in Computed Tomography (CT) and Digital Radiography (DR) medical imaging equipment, as well as x-ray based systems for non-destructive industrial imaging systems. Computed tomography is a global $1.5B industry. Digital radiography is a new diagnostic imaging modality and is expected to substantially displace x-ray film, which currently is a large fraction of the $3.8B medical x-ray industry.

Critical technology issues

Scintillator quality depends on multiple interlinked properties such as conversion efficiency, linearity, speed, radiation damage, and afterglow. These properties are affected by the scintillator host material and by chemical doping. This doping covers a very wide (5 ppm to 10 mole percent) range, with multiple dopants used to control several properties. Interactions between dopants are complex, and trade-offs between properties are frequently required.

1.3.5 Magnetic Materials

Benefits

A practical superconducting electrical transmission line must be high on the lists of the "holy grails" of materials development. Its benefits, from increased efficiency to improved use of real estate, are in the multi $100 billion dollar range. Modest improvements in magnetic materials would lead to substantial improvements in transformers (already a $20-30B market for superconductors), power conversion equipment, and MRI machines.

Critical technology issues

The highest critical temperature of superconductors has tended to increase with the number of components in the formulation. The most complex of these materials are already 5-component formulations; at 6-8 components the combinatoric explosion is formidable.

1.3.6 Coatings

Benefits

Most finished manufactured products have some sort of coating for protection, decoration, and other property enhancement. These processes are expensive, capital intensive (an automotive painting line may cost as much as $350 million), and often environmentally offensive (US regulatory costs are estimated at $50 million per painting line). Combinatorial experiments in this area can be applied to systems as diverse as flexible flat-panel displays and weather-resistant clearcoats for auto body panels. [23]

Critical technology issues

The system is an experimental space of formulation/process interactions with complex application methodologies. Evaluation of the samples is primarily in terms of macro functionality.

1.4 Capabilities of Combinatorial/HTS equipment

1.4.1 Overview

A wide array of equipment has been developed for CHTE programs during the past decade. Although the details of this hardware depend on the precise problem being studied, there are specific types, and they tend to fall into clearly defined size ranges (Figure 1.3). The limitations in array size, sample size, formulation accuracy, and resolution of analysis equipment all have a direct effect on the choices of experimental designs to use. There is a continual dialog between the hardware for producing sample arrays and the design of the array.

Since comparisons of materials are easiest (and of highest quality) when done within a single array, the hardware acts as a constraint and a goad to the experimentalist. This can sometimes result in a brilliant synthesis of array and experiment, as in the 1024-sample fractal array developed by Xiang et. al. [24] More frequently, it results in a compromise of statistical ideal and practical needs.

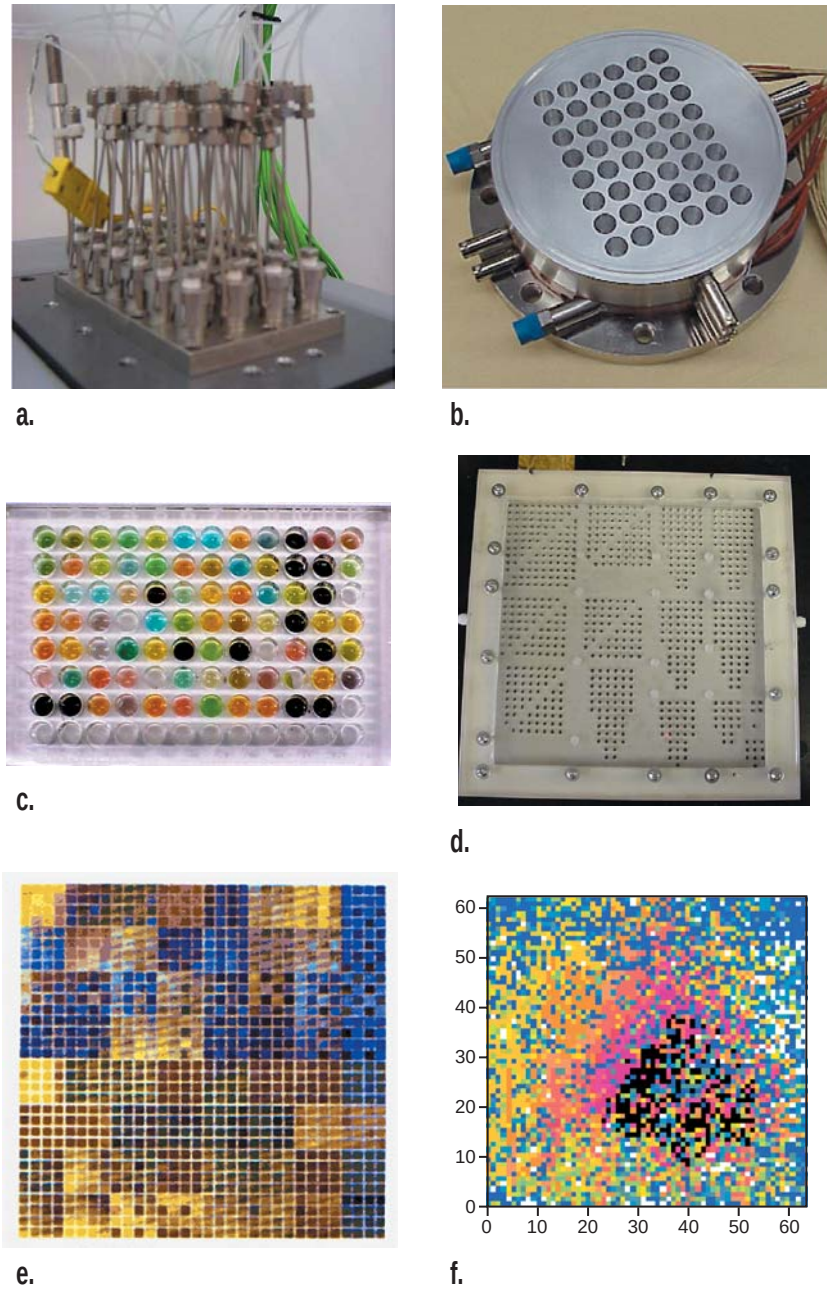

Figure 1.3. Combinatorial Libraries. a. A 6x4 array of microreactors. Photo courtesy of Avantium Technologies BV. b. 6x8 high pressure vial reactor system. Photo courtesy of GE Global Research. c. 96-well plate of catalyst samples. Photo courtesy of GE Global Research. d. Unfolded pentanary catalyst array. Reprinted, with permission, from Mallouk, T.E., and Smotkin, E. Fuel Cell Handbook (forthcoming), Copyright © 2002 John Wiley & Sons. e. 1024-sample fractal design array. Reprinted, with permission, from Sun, T.X. Biotechnology and Bioengineering (Combinatorial Chemistry), **1999**. 61(4), 193-201. Copyright © 1999 John Wiley & Sons. f. 4000-sample continuous gradient. Photo courtesy of Agere Systems.

1.4.2 Microreactors

The smallest array sizes tend to be groups of microreactors. These are constrained by the need to perform physical activities during the processing of the array – gas or liquid addition, mixing, sampling, or analytical measurements. They also tend to be somewhat larger than the following array types, often being in the milliliter rather than microliter scale. For this reason microreactors tend to be used for followup or optimization rather than screening. Arrays of 6, 12, 24, and 48 reactors are typical, and the reaction and analysis times can be relatively long, so the throughput with these arrays is typically 10-100 samples/day.

1.4.3 Robot Mixed Formulations

High-speed materials experimentation has had the advantage of access to a maturing technology of robotic materials handling equipment. This infrastructure, developed for the pharmaceutical industry, includes pipetting robots, handlers, and plate readers. Much of this was designed around the 96-well microtiter plate developed for biochemical screening. This makes the 96-well format a convenient one for materials development, although the robots are quite adaptable to other formats.

Well-based technologies are more flexible in the number and types of combinations possible than the thin film technologies, but they are constrained by the need to physically move and mix components. This requires robotics. The most common robots in CHTE laboratories are pipettors that can accurately transfer 25-1000 microliter quantities using single or multiple pipet heads. The default design of these robots covers the 96 well plate but most are reprogramable to other formats. Smaller quantities of material can be dispensed with very good accuracy using ink-jet type devices. Given the speed of the robots and the need for some manual operations, the throughput of these systems is typically 1-5 well plates (96-480 samples) per day.

1.4.4 In situ Mixing

This technology has primarily been used for the production of gradients or mixtures of thin film materials. The earliest approach in the generation of thin film libraries was the generation of continuous gradients of material on a substrate by sputtering several materials at once on a substrate. The sources are at an angle to the substrate, so the gradient is formed by the decrease of material with distance. This was first suggested by Kennedy in 1965 [25] and Hanak [26] in 1970 but was not widely used for the next 20 years.

In its current form, sputtering has been joined by other methods for generating atomic or molecular vapors such as continuous vapor deposition (CVD) and evaporation. These technologies mix the materials on a molecular scale, but microdroplet deposition and mixing using inkjet or spray methods is also possible. Continuous gradients can also be generated by physically spreading materials such as polymers on a surface. [27] In this approach the number of points is determined more by the resolution of the analysis tools than by the application technologies. Van Dover and Schneemeyer have reported 4,000 points on a 66x63mm rectangle produced by sputtering three materials on a square plate. [28]

A second use of *in situ* mixing has been the generation of thin film libraries using the masking technologies pioneered by the semiconductor industry [29]. Deposition methods can include all the molecular scale methods mentioned above. In this case, the materials are deposited in layers and annealed into a uniform film. The form factor has tended to follow a binary pattern. This began with Xiang and Schultz's 16-sample library, which quickly grew to 128 samples [4] and then to a 1024-sample library. [24]. Addition of a shutter to provide a continuous gradient of material compositions has allowed preparation of ~25,000 distinct compositions on a single 25x25 mm substrate. [30]

1.4.5 Organic Synthesis Technologies

Combinatorial organic synthesis is by far the most developed technology in the CHTE arena. Pharmaceutical development of methods and equipment has evolved extremely rapidly because of the large research budgets of the drug companies. The major method is solid phase synthesis; either parallel synthesis using compounds affixed to pins or the like and "split and mix" synthesis with compounds attached to beads. The first methods typically produce hundreds to thousands of discrete compounds and the second from thousands to hundreds of thousands. Automated liquid-phase parallel synthesizers such as the Chemspeed [http://www.chemspeed.com] or the Myriad (http://www.mtmyriad.com/ps.htm) are more likely to be useful in materials development, where 10 to 1000 mg supplies of compounds are often needed. Liquid-phase synthesizers have 8-64 vessels and can be cycled once or twice a day. The rate-determining step in synthesis technologies is development of robust synthetic methods that can be applied to a wide range of compounds. This takes weeks to months, while the actual synthesis is done in a few days.

This area is thoroughly covered in books [31-34]; journals such as Combinatorial Chemistry (John Wiley and Sons) and the Journal of Combinatorial Chemistry (American Chemical Society); Web sites (www.combinatorial.com, www.combichemlab.com); and by companies such as Pharmacopeia (www.pcop.com), Arqule (www.arqule.com), Argonaut (www.argotech.com), and many others.

1.5 A Systematic Approach to Planning for a Designed Combinatorial Experiment

1.5.1 Background

Regardless of the excitement of using new and powerful technology for conducting an experiment, the "planning activities that precede the actual experiment are critical to successful solution of the experimenters' problem" [35]. If anything, planning must be even more careful since we now have the opportunity of going in the wrong direction faster than ever. Montgomery and Coleman's classic paper on "Planning for a Designed Industrial Experiment" [35] contains a Pre-design Master Guide Sheet (Table 1.4) that is still largely applicable for CHTE work. Their detailed discussion on each of the elements in the guide is also worthwhile reading. I will follow their outline, and then add additional elements unique to CHTE. This guide is useful both on the macro scale, in defining an entire program of experiments that may cover months, and on the micro scale, defining individual experiments.

Table 1.4 Pre-Design Master Guide Sheet

1. Name, Organization, Title
2. Objectives
3. Relevant Background
4. Response Variables
5. Control variables
6. Factors to be "held constant"
7. Nuisance factors
8. Interactions
9. Restrictions
10. Design preferences
11. Analysis & presentations techniques
12. Responsibility for coordination

1.5.2 Objectives

In a typical CHTE program, we know what we are looking for – catalyst activity, phosphor color or luminescence, polymer toughness, electronic properties, and so on. However, it is worthwhile to have an in depth discussion with representatives of the key customers to ensure that all the critical needs are known, both qualitatively and quantitatively. The tools of Customer Needs Mapping and Quality Function Deployment can be very useful for this process. [36] This must be an ongoing process – target creep and mutation are very common. The objectives should also be "(a) unbiased, (b) specific, (c) measurable, and (d) of practical consequence." [35] Once established, the objectives must be prioritized, because it is difficult to analyze for more than one or two critical objectives at maximum throughput. Those "most critical" objectives become the response variables to be measured at the high throughput stage, while less critical objectives become response variables during the secondary and tertiary stages.

1.5.3 Relevant Background

Many of the problems studied using CHTE methods have had years or even decades of study using conventional experimental methods. These will (at least) give some starting places for the study, but the team needs to be careful about being too limited by them. One common phenomenon is the "moth around the flame" effect, in which experimentation consists largely of modest excursions from a known (pretty good) center, because bold experimentation is too expensive. The lively interplay between theory and experiment that marks a good scientific study should be, if anything, accentuated in CHTE because of the rapid feedback from experimentation.

A critical element of the background is the available resources. This is multivariate, with elements of personnel, stakeholders, money, machines, materials, and time.

1.5.4 Response Variables

A high-speed screening program can generally analyze for only one or two critical responses at a rate that matches the high throughput of sample preparation and processing. Definition of the method for performing the high-speed analysis is often the

crucial step in setting up a new CHTE target. Often the analysis cannot measure the customer-critical property directly; instead an easily measurable property that correlates well with the critical property must be chosen. Sometimes a binary response (on or off, present or absent) is sufficient information for identification of a hit; in other cases a quantitative measured response is necessary. This has a profound effect on all aspects of equipment design and experimental design. The remarkable progress that has been made in devising both parallel and rapid serial analytical methods for CHTE has been summarized in a recent review. [37]

1.5.5 Control Variables

This part of the planning process is much more expansive than the factor definition in Montgomery's guidelines. Here we are defining an entire experimental space, which may contain:

- Qualitative formulation factors, such as specifically identified elements or compounds.
- Quantitative formulation factors, including both ranges and intervals.
- Process factors.
- Quantitative or qualitative levels of the process factors.
- Permutations of substituents on chemical structures.

The possible number of combinations of these factors is essentially unlimited. One of the critical intellectual tasks in the planning process is selection of a chemical space that is the "right size". If the space is too restricted, the search will circle around known optima; if too expansive, it will contain too much sterile terrain.

1.5.6 Factors to be "held constant" and "nuisance factors"

A CHTE "factory" usually has a number of process steps, each of which can introduce variation. Controlling this variation so that the output data from the factory has meaning requires intense attention to design and operation. The sources of variation in each process step can be considered as nuisance or noise factors in the statistical sense. A selection of the commoner noise factors is given in Table 1.5. We have found Six Sigma methodology to be a useful framework to design and maintain quality in the factory. [36]

Table 1.5 Nuisance Factors in CHTS

Formulation Experiments	Gradient Experiments
♦Stock Solution	♦Source
– Chemical Source	– Purity
– Amount	– Uniformity
♦Sample	♦Sample
– Amounts	– Annealing
–Total Volume	– Crystallization
♦Reactor	♦Processing
–Temperature Control	– Thickness
– Pressure Control	– System gases
– Reaction Time	
♦Analysis	♦Analysis
– Component Identification	– Positional accuracy
– Quantification	– Quantification

1.5.7 Interactions

CHTE is all about interactions – and usually high order interactions. In many of the targets studied, the main effects and frequently the 2-way interactions of the major process variables have been thoroughly studied. New and substantially enhanced properties will be found in high-order interactions such as:

- a new phase in a 3-way or higher mixture of electronically active ingredients. [28,38]
- synergistic ingredients in a catalyst formulation. [39].

A sense of the degree and type of interactions to be searched is important in deciding on the experimental strategy and designs to be used. This can take the form of an estimate of the domain size of a new phase in a gradient search; the number of interactive components in a reduction/oxidation catalyst cycle; or the number of unique monomer combinations in a polymer.

1.5.8 Restrictions and Design Preferences

Every decision made during the engineering design of the CHTE factory will become a restriction on the factory operation. Therefore there must be an iterative process linking the Customer requirements with the chemical analysis system and the reactor design specifications (Figure 1.4). Simply stated, both the analysis and the reaction system have to be sufficiently capable to allow identification of leads or hits above the composite noise of the systems. However, the ability to perform certain high throughput analyses is a function of the type of reaction and reactor employed and similarly the reaction system can be configured to more easily allow rapid in-line or off-line analysis. Typical restrictions include:

- The standard size of the CHTE array, such as a 96-well plate.
- The sources, masks, and shutters of a system used for producing “landscape” libraries.

- The capabilities of the robot used for charging the ingredients, including:
- Minimum and maximum aliquot size.
- Accuracy of addition (as a function of aliquot size, viscosity, etc.)
- Speed of completion of the array (which may affect sample stability, solvent evaporation, etc).
- Ramp up, ramp down, and uniformity of any processing steps.
- Speed, precision, and resolution of the analysis system.
- Capability of the analysis system to perform *in situ* measurements.

Because most CHTE systems operate with a fixed array size, fitting an "optimized" design to the array is often Procrustean. Therefore issues of array-to-array consistency, blocking, and split plot design must be included in the overall experimental plan.

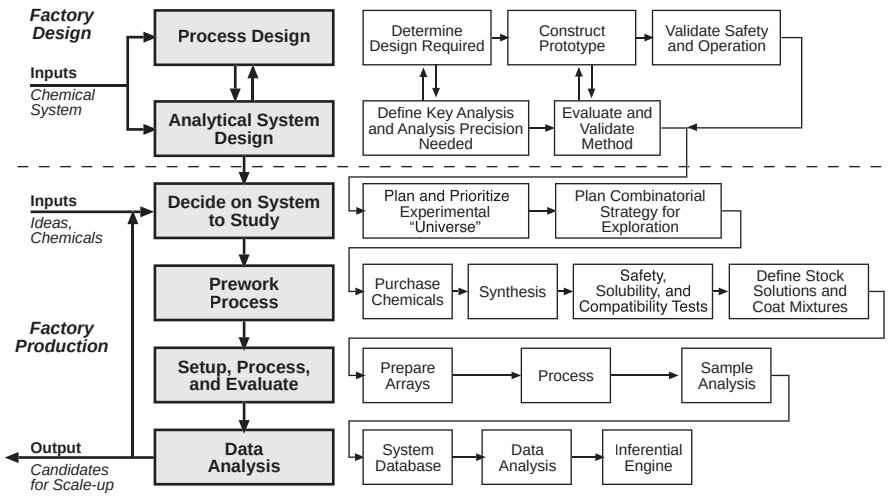

Figure 1.4. Process Map. Inter-relationships between design and construction, and operation of a CHTS system

1.5.9 Analysis and presentation techniques

Analysis of the data from a CHTE study depends very much on the objectives and underlying design. In pure screening mode, a simple "pick the winner" strategy, with confirmation and optimization experiments following, may suffice. More complex strategies will require all the tools of statistical analysis and data mining. Visualization of multidimensional spaces is a major challenge; software such as Matlab [40] and Spotfire [41] is useful.

1.5.10 Responsibility for coordination of the experiment

The structure of a CHTE laboratory could be the subject of at least a full article if not a book. We have found that there are a number of roles that must be allocated:

- Domain experts: understanding of the chemistry and the commercial implications of the target.
- Analytical team: development of high throughput analyses for the critical parameters.
- Synthesis team: generation of non-commercial raw materials.
- Engineering team: development or modification of the reactor and robotic hardware.
- Informatics team: database, experimental design, data analysis, statistics, quality control, and visualization.

One effective structure for the overall laboratory is shown in Figure 1.5. Only the domain experts are completely dedicated to a single target. The other teams are set up to service all the targets. This improves personnel efficiency and improves cross-target communication and learning.

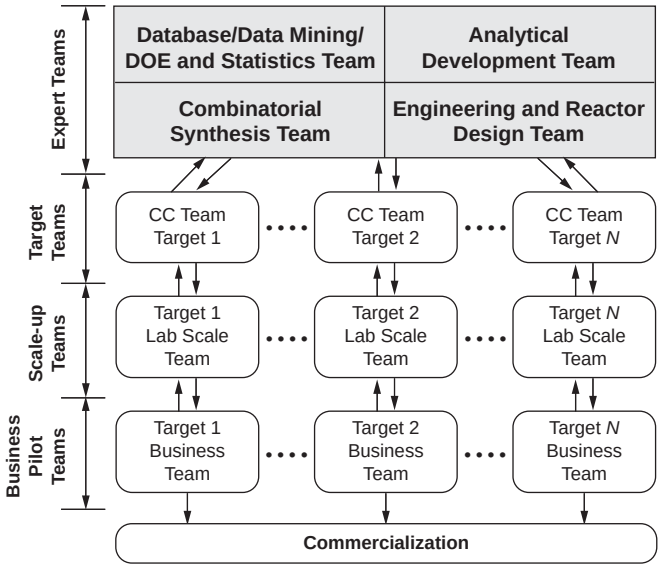

Figure 1.5. Organization of a CHTS Laboratory.

The intellectual effort in planning and analyzing the work in a CHTE system should not be underestimated. In our laboratory we have had instances where the opportunity to increase the throughput of the CHTE "factory" was rejected by the chemists because they could not generate new ideas and plans fast enough.

CHTE systems add areas of planning which were not addressed by Montgomery and Coleman. [35] These include the handling and storage of the data from the experiment; multistage experimentation; and logistics.

1.5.11 Database and Data Handling

A CHTE program in full operation will generate anywhere from 200 to 200,000 experimental samples per week, and each sample will have at least 10 but frequently far more data elements – factor names, factor settings, responses, and associated information. This deluge of data will swamp conventional methods of storage. We have found that a small CHTE program can operate reasonably well with a careful use of spreadsheet methods. Current spreadsheet technology (e.g. Microsoft EXCEL) can contain 230 columns x 65536 (2^{16}) rows, so careful column assignment and the use of filtering functions will enable storage and retrieval of the data from a modest CHTE program. This works best when a single person with a good memory is the experimental planner and data analyst. Larger CHTE programs involving an operating team are better off with a full hierarchical database. Ideally it should include chemical structure searching capability using such software as ISIS [42] or Chemdraw. [43] A database of this magnitude is not a small project and should not be under-resourced.

One class of information that is often left out of database planning is the "meta-information" about the experiment. This includes:

- Connections to preceding experiments, literature, and experience.
- The geometry of the array design.
- Qualitative and anecdotal observations.
- Statistical analysis.
- Conclusions.
- Connections to follow-up experiments.

A well-designed data storage system should include searchable capability for this type of information.

1.5.12 Multistage Experimentation

The high throughput screening process is only the first stage of a development project. It must be intimately tied to subsequent stages of further screening, optimization, lab scale and pilot scale development (Figure 1.6). After a lead is discovered, second stage screening and initial optimization are frequently done using much of the same equipment as the primary screen. Typically, the hit and its immediate neighbors are retested for confirmation and initial establishment of the active region. The optimization stage is a search for main effects, 2-factor interactions, and curvature in the region of the hit. The relatively low cost of high throughput runs allows for less parsimonious experimentation, so we frequently see full factorials or lightly fractionated designs in this stage. The confirmed and optimized hit is then scaled up to conventional laboratory equipment for detailed testing of all the customer requirements. Scale-up tends to become a bottleneck once the CHTE factory is at full production, so the scale-up teams should be in place and prepared for a rapid flow of candidates.

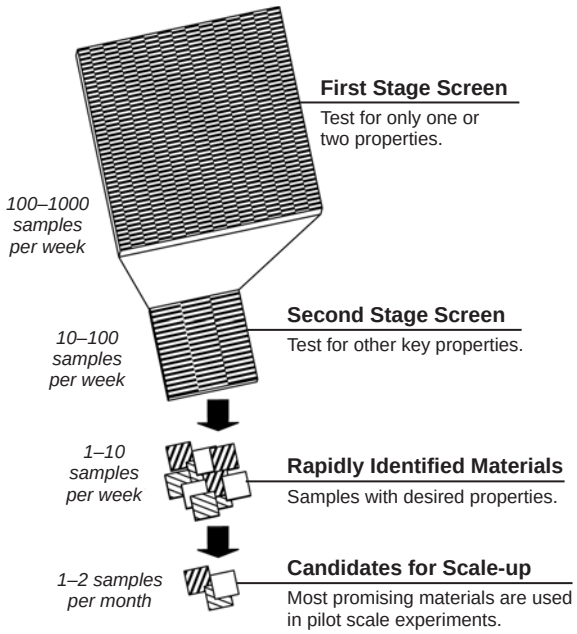

Figure 1.6. Multistage Experimentation. The flow from screening to the pilot plant. Reprinted, with permission, from Cawse, J.N. Accounts of Chemical Research, 2001. **34**: p. 213-221. ©2001 American Chemical Society

1.5.13 Logistics

Just like Henry Ford's assembly line, the CHTE factory must be a marvel of coordinated action. The Symyx mantra is *"Analyze in a day what you make in a day"* [44], and this applies to every other step in the factory as well. A full process map of the steps in the factory should be assembled and studied – both from a day-to-day basis and a longer-term prospect. Cross training of personnel may be required to ensure that the project does not come to a stop when a team member is sick or on vacation. Raw material supply can be an issue; a project can come to a halt when it uses up every commercial variant of a critical component if the synthesis team is not up to speed for generating new materials. Analysis of the data and recommendations for new experiments is often a slow step. Finally, a structure for communication of the progress of the experiment to the team members is essential. This is best tied in with the database.

1.6 Statistical Design Issues

1.6.1 Overview

The field of combinatorial and high throughput materials development is new enough that there has been relatively little examination of the statistical issues in the design of CHTE experiments. One concept that has reached general usage is the idea of *experimental space*, which in this work is the total number of combinations of all the factors and levels that may be varied in the system. It is clear from the discussion above that this space can

be unimaginably large. However, there has been little examination of the geometric and mathematical properties of this space that might help us define search strategies.

1.6.2 The NK Model

Fortunately, there have been important advances recently in a related area – the structure of the genome. The combinatorial issues there are at least as large; a mere decapeptide (with 20 amino acid choices per site) can have $20^{10} \approx 10^{13}$ possibilities, and a minor bacterium with 500 genes can have $2^{500} \approx 10^{150}$ potential genotypes. Kauffman [45,46] has formulated the concept of *correlated fitness landscapes*, which correspond very well with our ideas of experimental space. These landscapes have different degrees of "ruggedness" which bears directly on our ability to search the space and find new optima. Our concept of factors from traditional DOE (which can take on various levels and interact with each other) corresponds with genes (which similarly can take on various properties and interact with each other).

The essence of Kauffman's NK model (which is adapted from the physicist's spin-glass model) is the level of complexity that arises from K interactions of N genes (factors). K can vary from 0 to N-1. "The higher K is – the more interconnected the genes are – the more conflicting constraints exist, so the landscape becomes ever more rugged with ever more local peaks". If K = 0, the "landscape is a 'Fujiyama' landscape, with a single peak falling away along smooth gradual slopes." This is entirely analogous to a *main effects* model (Figure 1.7a) in experimental design. Similarly, a K=1 landscape is one of *2-way interactions* (Figure 1.7b) which can easily be handled by standard 2-level fractional factorial DOE's. I suspect that *the landscape of combinatorial chemistry is one where K=2 or higher* (Figure 1.7c)**.** In fact, I surmise that the landscape of problems which are approachable with our current tools is one where K=2 to (at most) 4 or 5. There are fascinating opportunities for theoretical study in this area.

Some of the properties of these moderately rugged landscapes that may be applicable to our CHTE systems are (from [46]):

- "When K is low, peaks cluster near one another like the high peaks in the Alps… the landscape is *nonisotropic"*.
- As K increases:
- "Landscapes become increasingly rugged and multipeaked, while the peaks become lower."
- "The high peaks spread apart from one another… the landscape is *isotropic*… There is no point in searching far away on an isotropic landscape for regions of good peaks; such regions simply do not exist."
- "Landscapes with moderate degrees of ruggedness share a striking feature: it is the highest peaks that can be scaled from the greatest number of initial positions!"
- In searching a moderately rugged landscape by taking "long" jumps around it, the probability of finding a better region decreases exponentially.

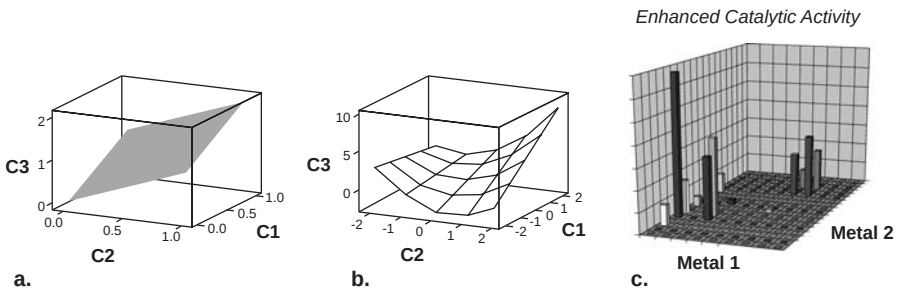

Figure 1.7. Landscapes. a. Main Effects only; b. Two-way interactions and curvature; c. Rugged correlated combinatorial landscape. Figure 1.7c reprinted, by permission, from Cawse, J.N. Accounts of Chemical Research, 2001. **34**: p. 213-221. ©2001 American Chemical Society

This implies that a wise project strategy would therefore be to experiment boldly at the start but switch to optimization of the best regions long before the project resources are exhausted. CHTE has the invaluable property of psychologically energizing the research team to bold experimentation. By continually extracting useful scientific generalizations from the data we obtain in our screening, we can start to predict the large-scale features that will lead us to the optimum regions of our experimental space.

1.6.3 Classical Statistical Issues

Regardless of the landscape we are exploring, the critical issues of statistically based experimental design remain: we are trying to make progress in a system containing experimental error while making efficient use of experimental resources.

Quality

There has been relatively little examination of the quality issues in the generation of combinatorial arrays. Hortatory articles have appeared discussing high error rates in pharmaceutical high throughput efforts. In particular, "…some new users seem to be so mesmerized by the technology's power that they are forgetting basic principles of experimental design." [47] In combinatorial and high throughput work, there are too many samples for each data point to be checked individually, and too many process steps for human oversight. Therefore all the steps of the process must be high quality. A defect minimization process such as Six Sigma must be in place during the design and operation of the factory.

Quality parameters that must be considered in planning the factory include:

- What kinds of improvements are we looking for? Are we looking for big jumps or incremental improvements in the output of our system? How big are the improvements needed relative to the current baseline?
- What risks are we willing to take? What are the consequences of false positives and false negatives? With what certainty do we want to declare a point a "hit"? Since multiple measurements are being made in parallel, care must be taken with simple statistical significance procedures inflating the rate of false positives. A "False Discovery Rate" procedure [48] has been proposed to improve the power of statistical tests in this regard.

- What is a lead? In the words of Gregory Bateson, what will "generate the difference which becomes information by making a difference"? [49]

From these parameters we can estimate an acceptable overall defect level for the process, and what level of replication may be required to deal with the actual defect level. They will also inform the search for the actual experimental designs with the statistical requirements required for successful experimentation.

Efficiency

The efficiency and effectiveness of various approaches to the design of experiments for CHTE will depend on the detailed structure of the space to be examined and the tools available. However, basic statistical principles that apply to any investigation must be considered in the development of an experimental design. These include

- *Resolution*. In classical DOE, Resolution is the ability of a design to evaluate an effect or interaction. [9] In a Resolution III design, for instance, main effects are confounded (mathematically combined so as to be indistinguishable) with 2-factor interactions. There is also the sense of pure physical resolution (as in the resolving power of a camera lens). This applies to gradient systems in which there is continuous variation of composition from point to point. How close can we physically place our sampling points? An understanding of the Resolution of our designs will help us understand what effects we can see (and what we can't).
- *Point Distribution*. Combinatorial space is vast and our resources are limited. We need to find effective ways to distribute our experimental points to sample the space. Fractional Factorial designs and the later D-Optimal designs were revolutionary improvements, allowing complex experiments to be performed with reasonable experimental resources. We are searching for the next generation of fractional and optimal designs.
- *Nesting*. In CHTE we are dealing with different sizes of experimental units [50]. Samples and arrays are formulated and processed in different ways. The error structure often contains sources of variation at the mixing, dosing, formulation, and processing level. If we wish to perform detailed statistical evaluations on these arrays, appropriate assignment of degrees of freedom and sums of squares must be made.
- *Signal/Noise Ratio*. An understanding of the overall quality of our system, and of the degree of improvement that constitutes a hit, will allow us to evaluate the signal/noise ratio and consequently the probability of observing a hit correctly.
- *Blocking*. In our experience, the statistically preferred design generally does not fit in the fixed number of sample points required by the hardware. This will result in either statistical blocking (where the design is larger than the array) or a decision on the best use of the extra space. This space can be used for standards, replicates, comparison points, or mini-designs and "wild hair" experiments.

References

(1) B. Jandeleit; D. J. Schaefer; T. S. Powsers; H. W. Turner; W. H. Weinberg, Angewandte Chemie, International Edition, 38*(*1999), 2494-2532.

(2) R. S. Bohacek; C. McMartin; W. C. Guida, Medicinal Research Reviews, 16*(*1996), 3-50.

(3) S. Borman In *Chemical and Engineering News* 1997; Vol. 1997, p 43-62.

(4) X.-D. Xiang; X. Sun; G. Briceno; Y. Lou; K.-A. Wang; H. Chang; W. G. Wallace-Freedman; S.-W. Chen; P. G. Schultz, Science, 268*(*1995), 1738-1740.

(5) A. Mittasch In Advances in Catalysis; Frankenburg, W. G., Rideal, E. K., Komarewsky, V. I., Eds.; Academic Press: New York, NY, 1950; Vol. 2, p 81-104.

(6) R. A. Fisher The Design of Experiments; Oliver and Boyd: Edinburgh, 1935.

(7) D. M. Steinberg; W. G. Hunter, Technometrics, 26*(*1984), 71-97.

(8) C. Daniel Applications of Statistics to Industrial Experimentation; John Wiley: New York, 1976.

(9) G. E. P. Box; W. G. Hunter; J. S. Hunter Statistics for Experimenters; John Wiley: New York, 1978.

(10) J. A. Cornell Experiments With Mixtures : Designs, Models, and the Analysis of Mixture Data; 2nd ed.; John Wiley and Sons: New York, 1990.

(11) Stat-Ease; 6 ed.; Stat-Ease, Inc: Minneapolis, MN, 2000.

(12) SAS; SAS Institute, Inc: Cary, NC, 2001.

(13) Minitab Inc.; 12 ed. State College, PA, 1999.

(14) I. Echip 2001.

(15) NIST; NIST; Vol. 2001.

(16) M. J. Harry The Vision of Six Sigma: A roadmap for Breakthrough; Sigma Publishing Company: Phoenix, AZ, 1994.

(17) C. J. Colbourn; J. H. Dinitz The CRC Handbook of Combinatorial Designs; CRC Press: New York, 1996.

(18) NIST "Technology Roadmap for Combinatorial Methods," NIST, 2001.

(19) NIST "Combinatorial Methods at NIST," National Institute of Standards and Technology, 2001.

(20) M. McCoy In *Chemical and Engineering News* 1999; Vol. 77, p 17-25.

(21) DOE "Vision 2020 Catalysis Report," Department of Energy, 2001.

(22) DOE "Materials Technology Roadmap," Department of Energy, 2001.

(23) NIST "Combinatorial Methods for Coatings Development," NIST Advanced Technology Program, 1999.

(24) J. Wang; Y. Yoo; C. Gao; I. S. Takeuchi, X.; H. Chang; X.-D. Xiang; P. G. Schultz, Science, 279*(*1998), 1712-1714.

(25) K. Kennedy; T. Stefansky; G. Davy; V. F. Zackay; E. R. Parker, Journal of Applied Physics, 36*(*1965), 3808-3810.

(26) J. J. Hanak, Journal of Materials Science, 5*(*1970), 964-971.

(27) J. C. Meredith; A. Karim; E. J. Amis, Macromolecules, 33*(*2000), 5760-5762.

(28) R. B. Van Dover; L. F. Schneemeyer; R. M. Fleming, Nature (London), 392*(*1998), 162-164.

(29) M. C. Pirrung, Chemical Reviews, 97*(*1997), 473-488.

(30) D. Danielson; M. Devenney; D. M. Giaquinta; J. H. Golden; R. C. Haushalter; E. W. McFarland; D. M. Poojary; C. M. Reaves; H. Weinberg; X. D. Wu, Science, 279*(*1998), 837-839.

(31) *High-Throughput Synthesis*; I. Sucholeiki, Ed.; Marcel Dekker: New York, 2001.

(32) *Combinatorial Chemistry : Synthesis, Analysis, Screening*; G. Jung, Ed.; John Wiley and Sons: New York, 1999.

(33) A. K. Ghose; V. N. Viswanadhan Combinatorial Library Design and Evaluation; Marcel Dekker: New York, 2001.

(34) *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*; E. M. Gordon; J. F. Kerwin, Eds.; Wiley-Liss: New York, 1998.

(35) D. C. Montgomery; D. E. Coleman, Technometrics, 35*(*1993), 1-12.

(36) J. N. Cawse, US Patent WO 01/16785 A2 (2001), to General Electric Co., Inc.

(37) R. A. Potyrailo In Encyclopedia of Materials: Science and Technology; Lifshin, E., Cahn, R., Eds.; Elsevier: New York, 2002.

(38) T. X. Sun; X.-D. Xiang, Applied Physics Letters, 72*(*1998), 525-527.

(39) Y. Sun; H. Buck; T. E. Mallouk, Analytical Chemistry, 73*(*2001), 1599-1604.

(40) Mathworks; Mathworks Inc.: 2001.

(41) Spotfire 2001.

(42) M. I. Systems 2001.

(43) CambridgeSoft 2001.

(44) P. Cohan In *Combi 2000 (Knowledge Foundation)* San Diego CA, 2000.

(45) S. Kauffman The Origins of Order; Oxford University Press: New York, NY, 1993.

(46) S. Kauffman At Home in the Universe; Oxford University Press: New York, NY, 1995.

(47) J. Knight In *Nature* 2001; Vol. 410, p 861-862.

(48) Y. Benjamini; Y. Hochberg, Journal of the Royal Statistical Society B, 57*(*1995), 289-300.

(49) G. Bateson MIND AND NATURE: A Necessary Unity; Bantam Books: New York, 1980.

(50) G. A. Milliken; D. E. Johnson Analysis of Messy Data; Van Nostrand Reinhold: New York, 1984.