From experimental data to mechanistic hypotheses: analysis of proteomic data using a very large-scale causal model

Dexter Pratt, Genstruct Inc, One Alewife Center, Cambridge MA 02140

Abstract

High-throughput proteomic analyses of tissue and bio-fluid samples can yield datasets comprising measured differences in hundreds - or even thousands - of proteins. In principle, this rich source of data can provide a systems-level view of the biological processes in an experiment, leading to testable hypotheses describing the mechanisms that led to the observed changes. But typically, the integration of hundreds of observations to infer the active biological networks is an unmanageable task, limiting the analysis to categorization of the changed proteins by annotations and by patterns of modulation. To identify disease mechanisms, compound mechanisms and biomarkers from proteomic and systems biology experiments requires the development of a model of biology. Using a mental model, a scientist can reason about hundreds of distinct molecules present within a cell, but reasoning over tens of thousands of molecules and their interrelationships is impossible. We describe the development and application of a very large-scale causal, computable model of biology which has been used to identify molecular cause and effect hypotheses consistent with data from proteomic experiments. Automated causal analysis can be used to define upstream networks of molecular events which could result in experimentally observed protein changes. It can be used to identify possible causal pathways linking initial experimental perturbations to observed protein or phenotypic changes. Large-scale causal analysis is a powerful new systems-based approach for the interpretation of molecular state measurements in drug discovery.

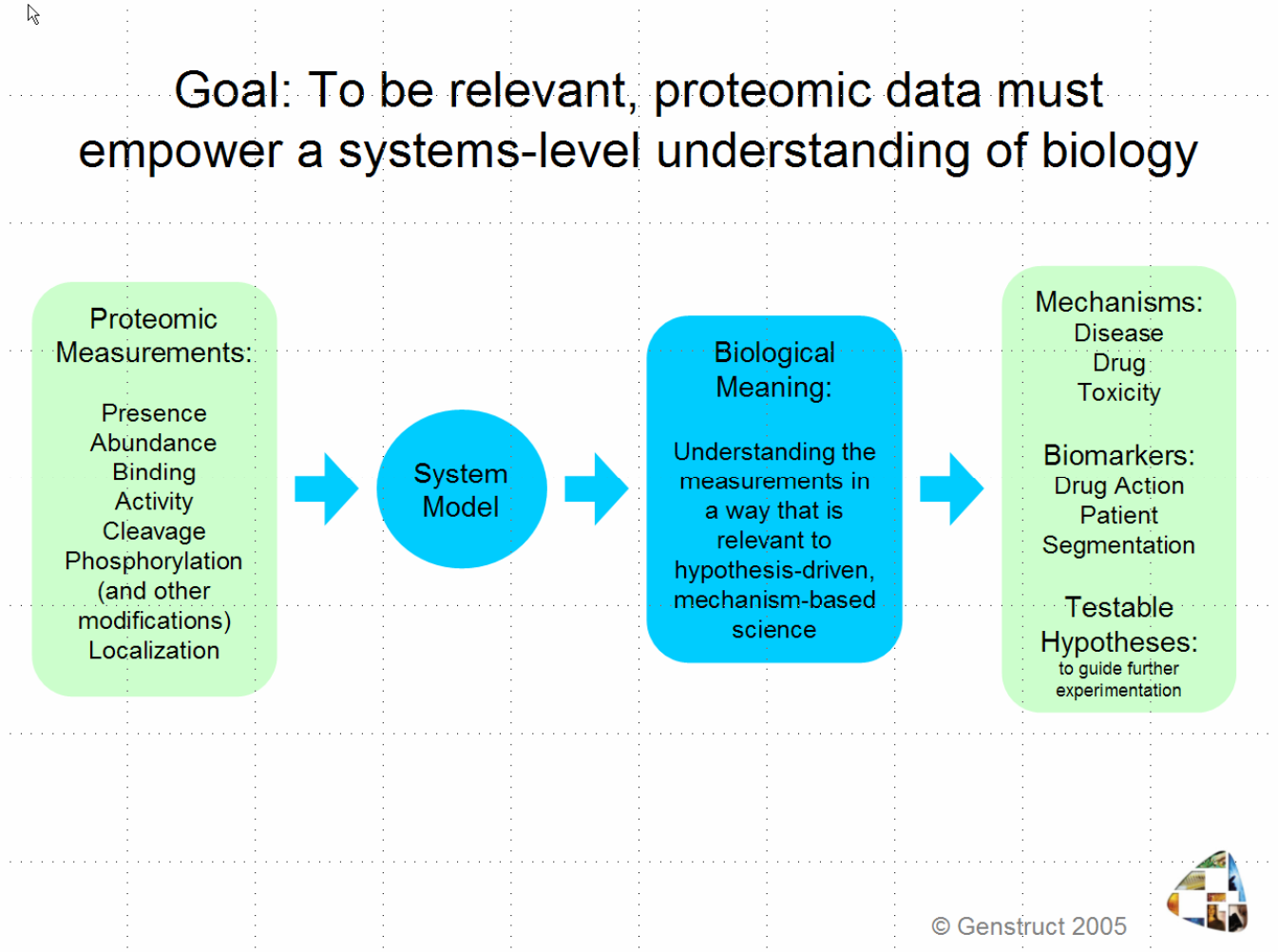

Figure 1. Proteomic data and systems level understanding

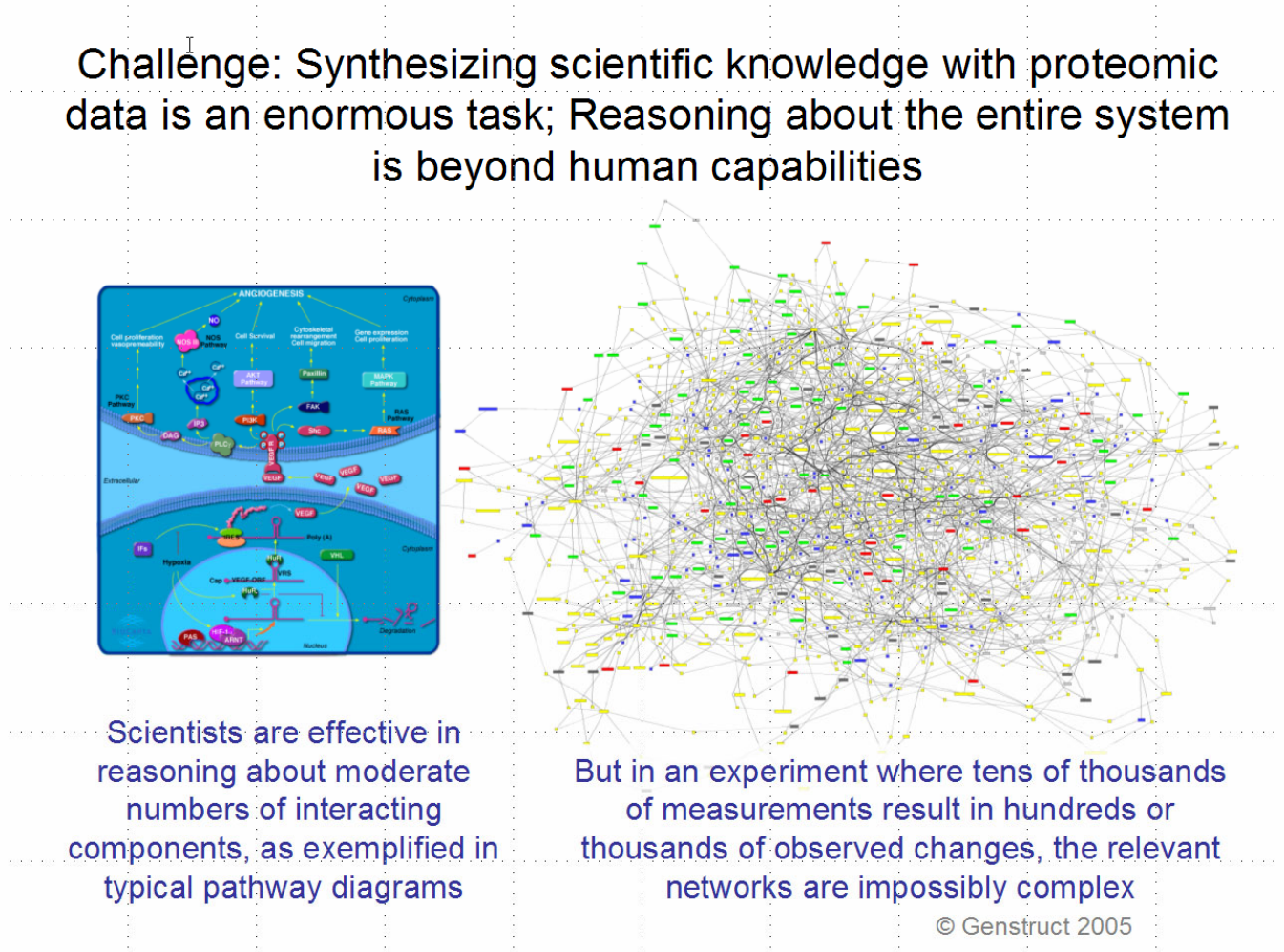

Figure 2. The conceptual challenge of high-throughput data

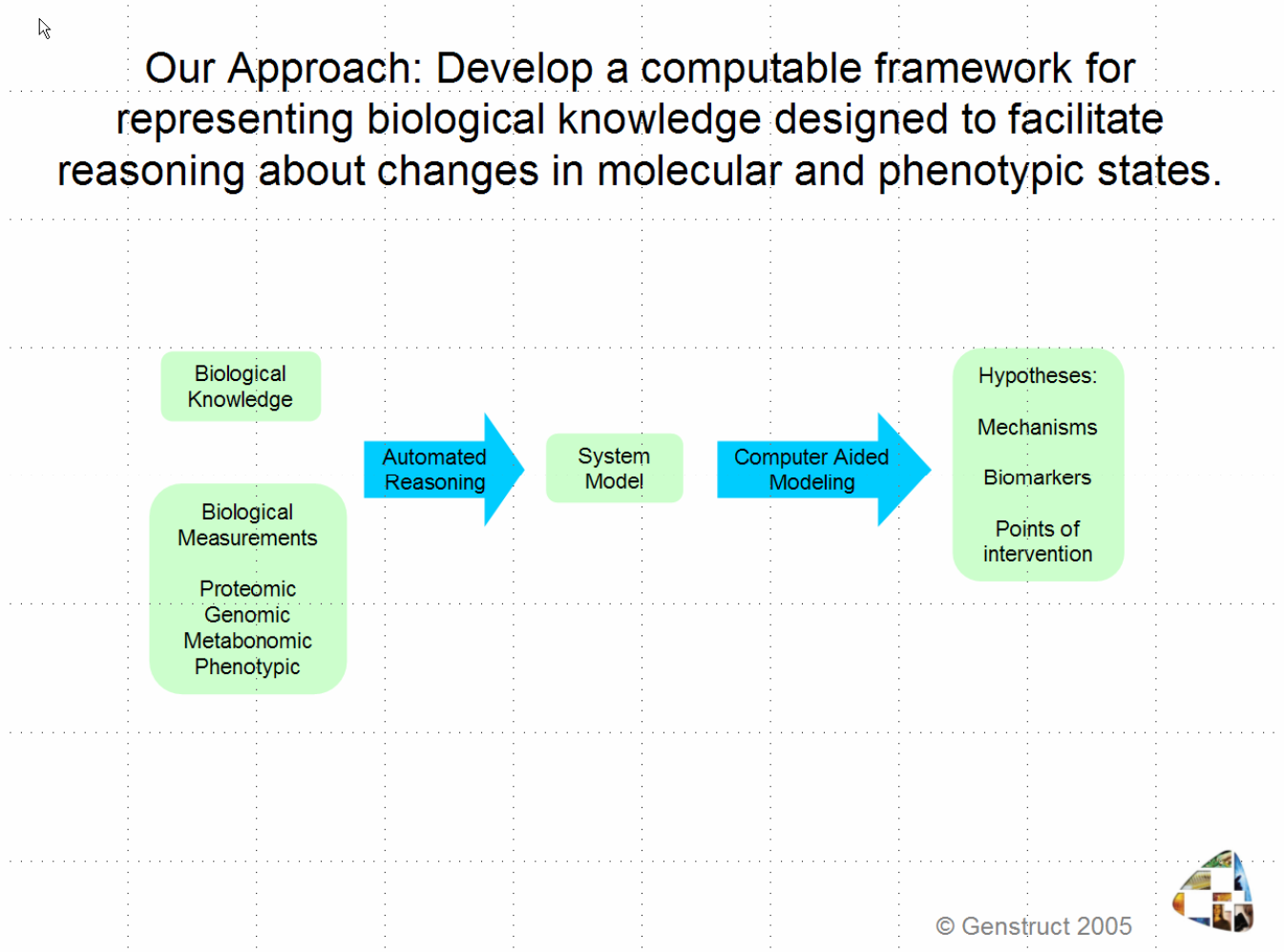

Figure 3. Computable framework for reasoning about biological state changes

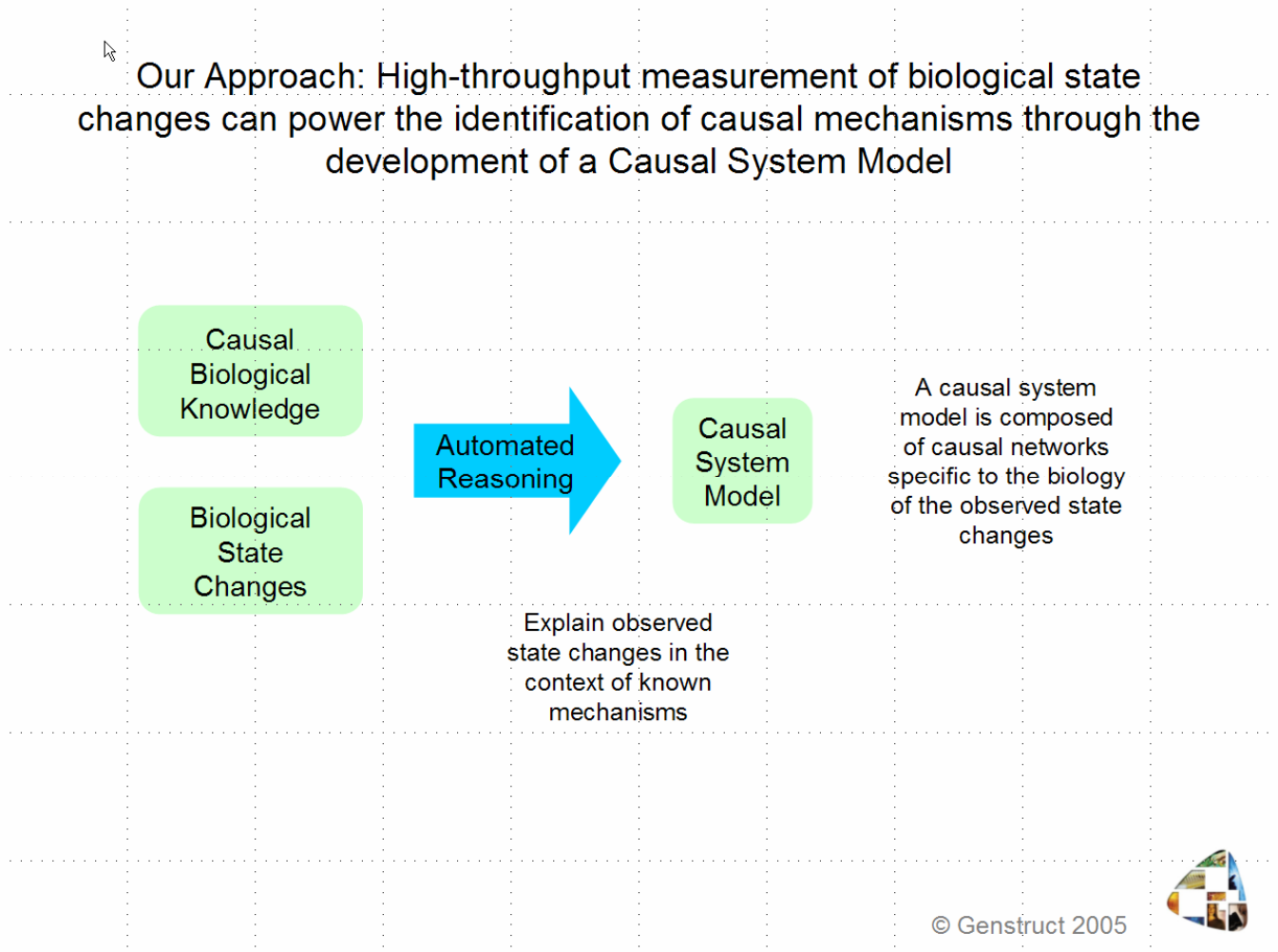

Figure 4. Causal System Models

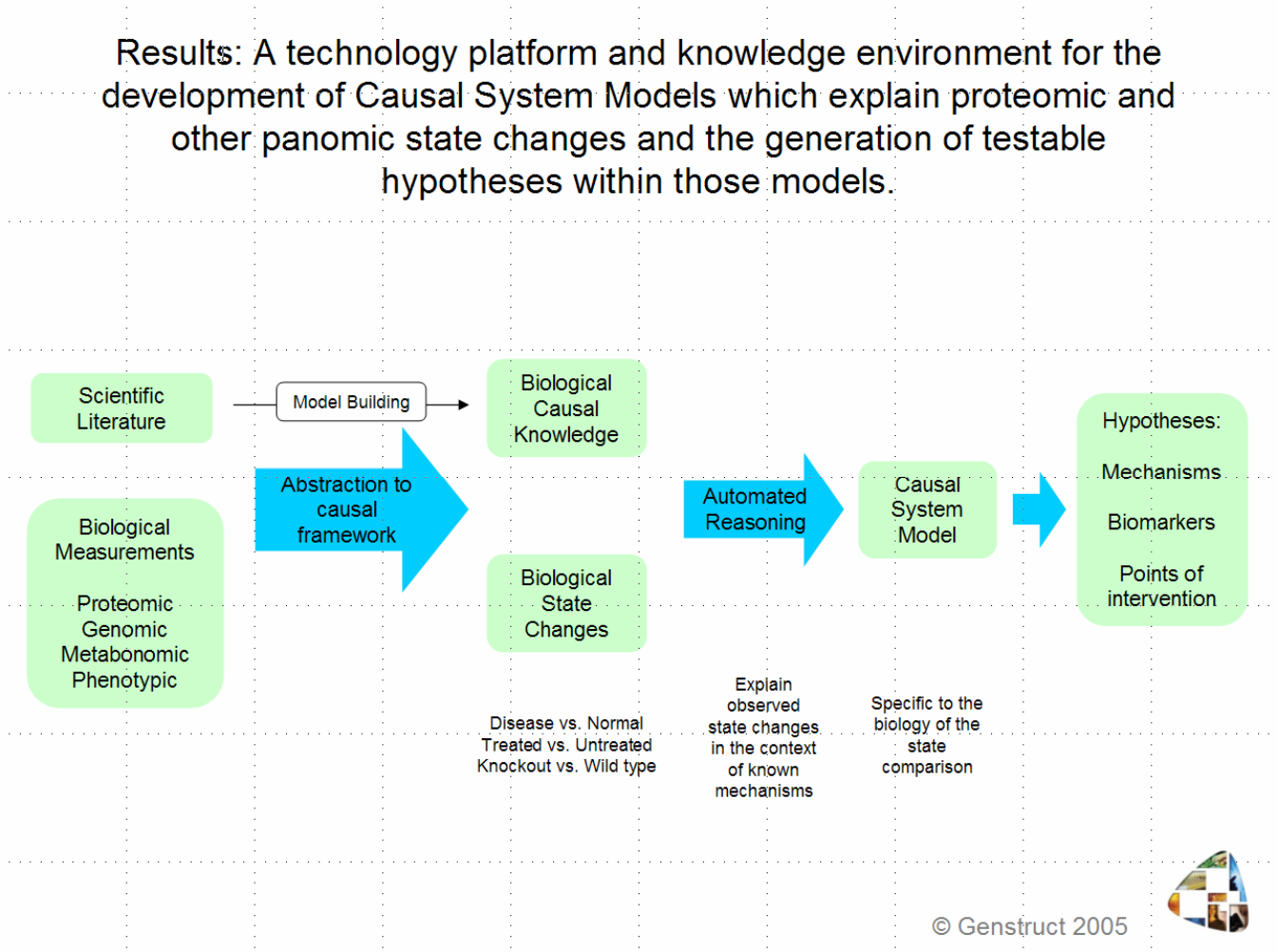

Figure 5. An integrated technology platform and knowledge environment

Framework: A compact, precise ontology of molecular entities, their activities and modifications, biological processes and locations

Entity	Meaning	Measurement
X	Abundance of X	Proteomics, Metabonomics
exp(X)	Gene expression of X	RNA profiling
catof(X)	Catalytic activity of X	Activity assays
kaof(X)	Kinase activity of X	Kinase activity assay
X {P@Y}	X phosphorylated at tyrosine	Phosphoprotein-mapping
taof(X)	Transcriptional activity of X	Promoter-binding assays

Examples of entities used in the Genstruct knowledge representation

Figure 6. Examples of entities used in the Genstruct knowledge representation

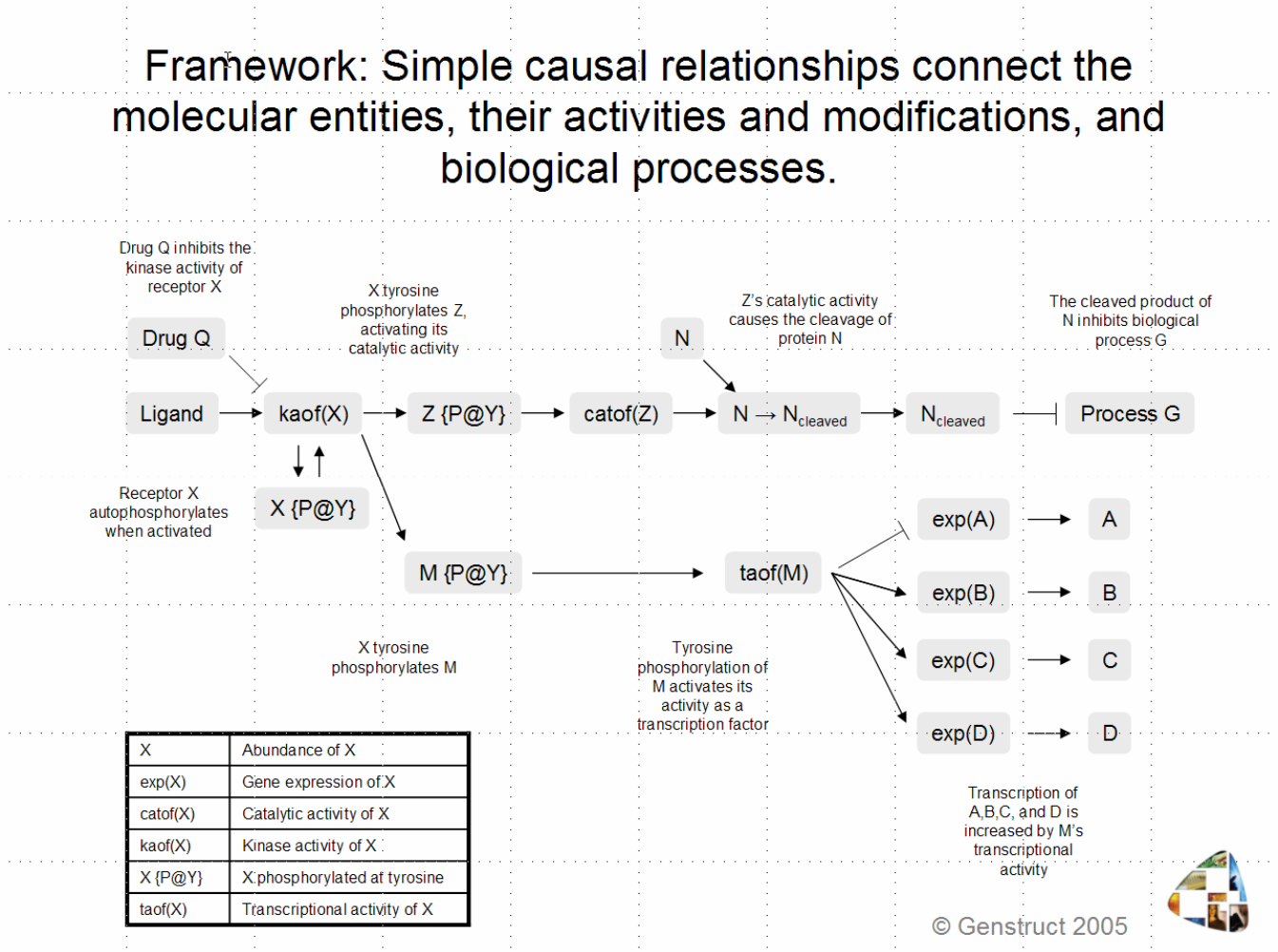

Figure 7. Example of representation of a simple causal network

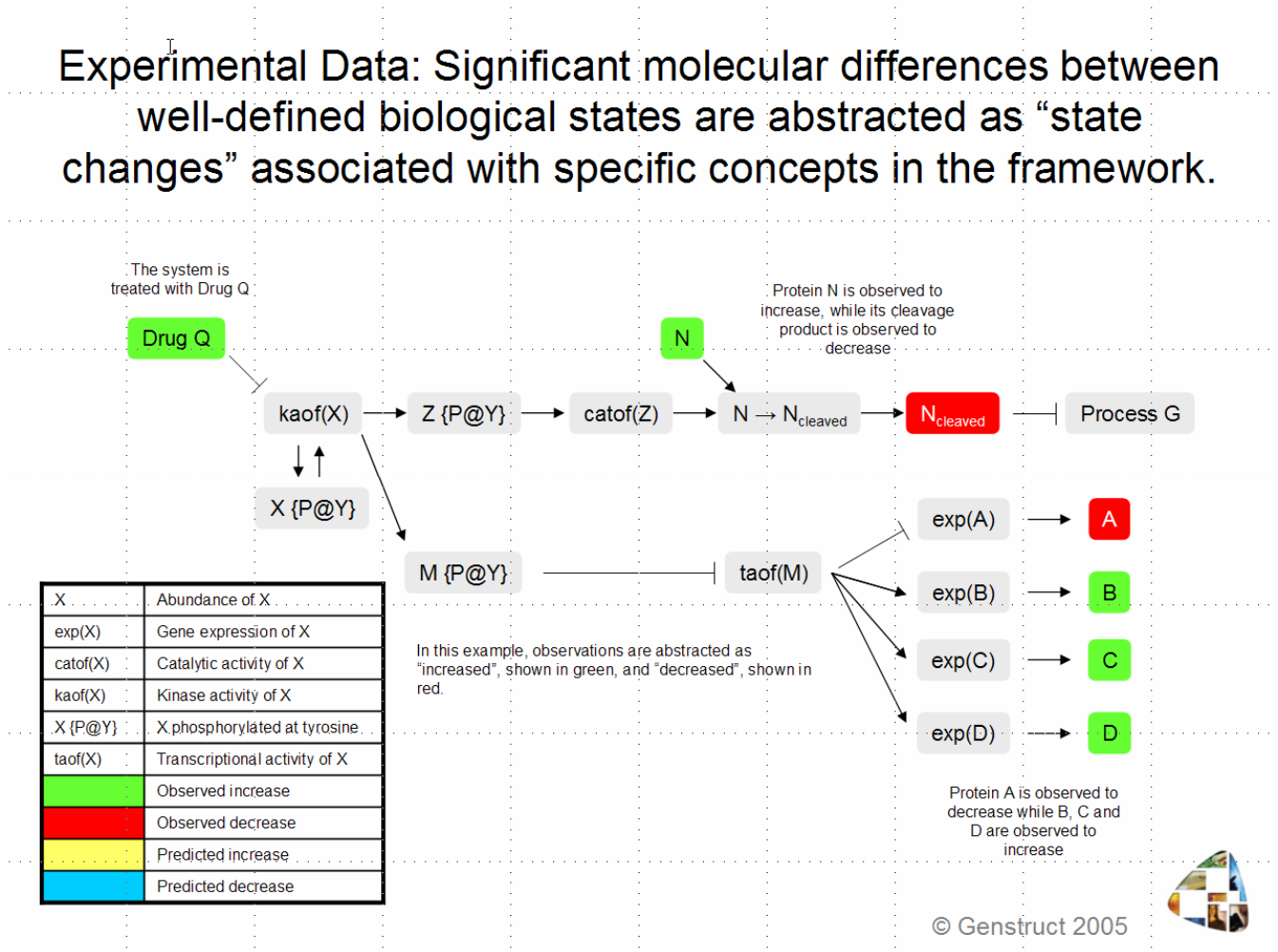

Figure 8. Biological state changes associated with specific entities

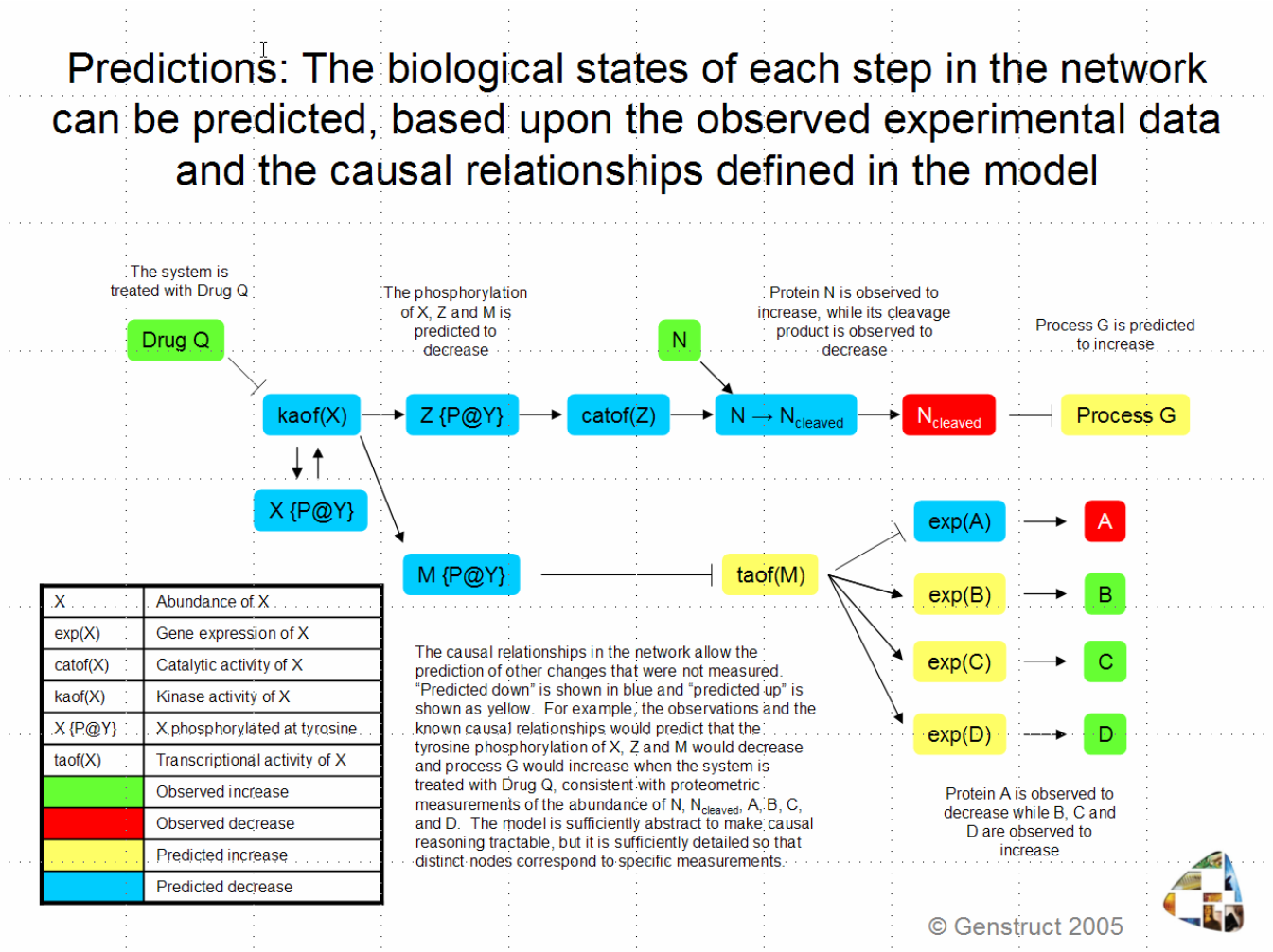

Figure 9. Predictions of state changes based on observed changes and causal relationships

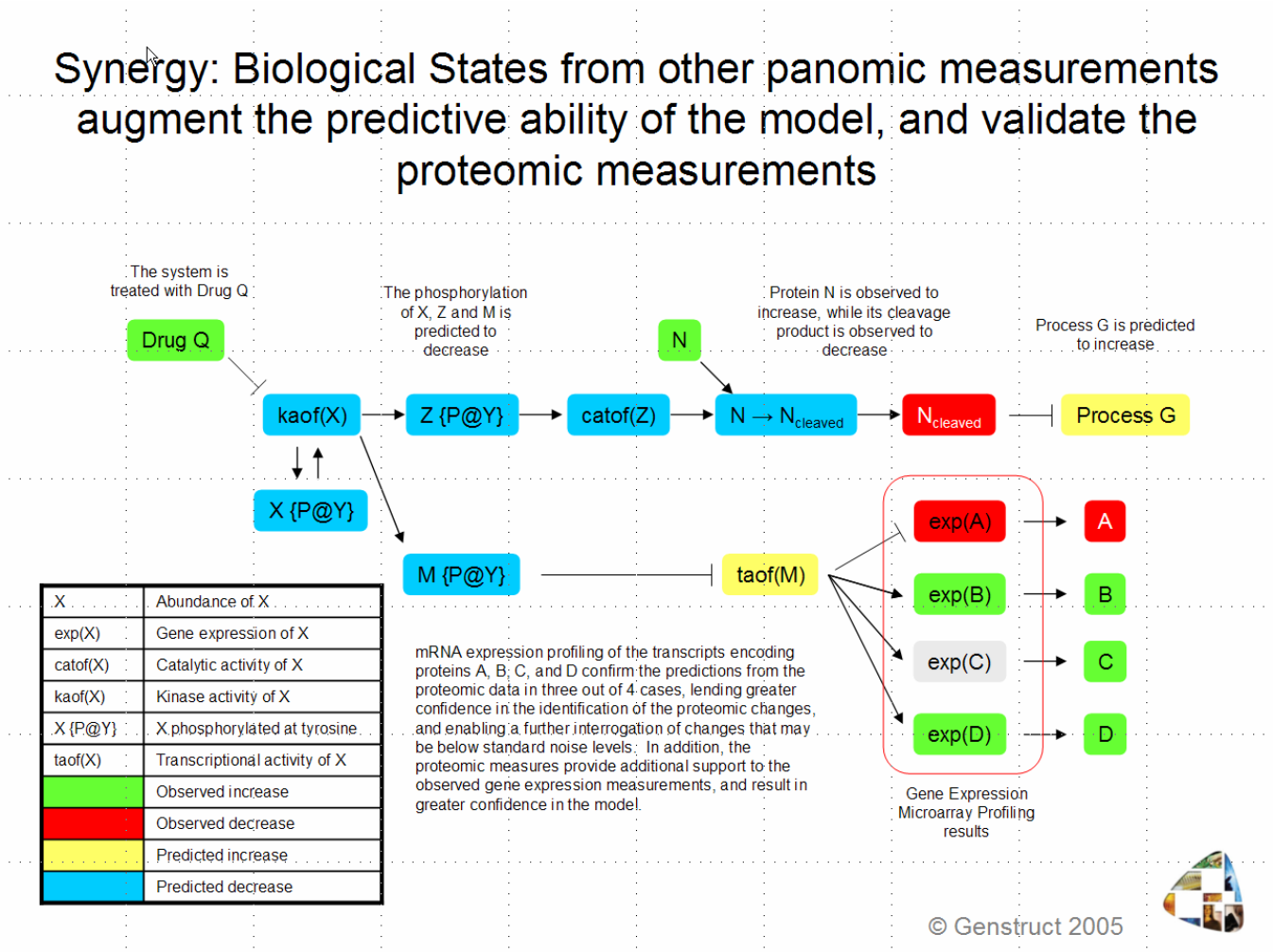

Figure 10. Integration of data from multiple measurement modalities in a causal framework

Reasoning: Causal System Models are created by identifying the most explanatory biological networks and merging them into a single, internally consistent model

- All possible explanations for each observed change are explored
- Each explanation is evaluated as a hypothesis:
  - How many observed changes can it's predictions explain?
  - How consistent are it's predictions with the observed changes?
  - Is the support for the hypothesis by the observed data significantly better than could occur by chance?
- The highest ranking biological networks are assessed for mutual compatibility and their relationships to relevant phenotypes and experimental perturbations.
- An internally consistent set of networks are selected and merged into a Causal System Model of the mechanisms which differ between the biological states compared in the experiment.

Figure 11. Causal reasoning to generate and select hypotheses

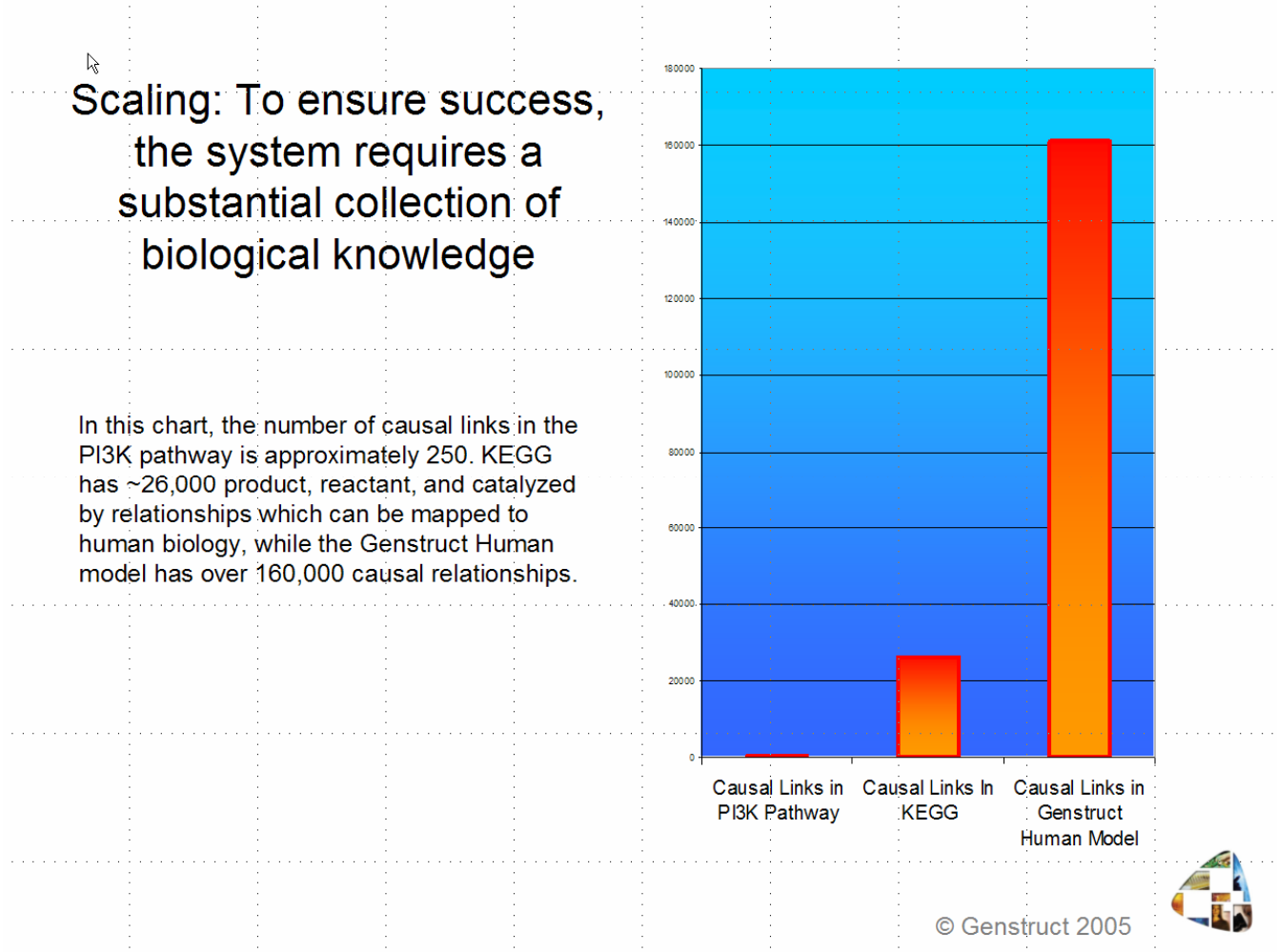

Figure 12. Scale of the system

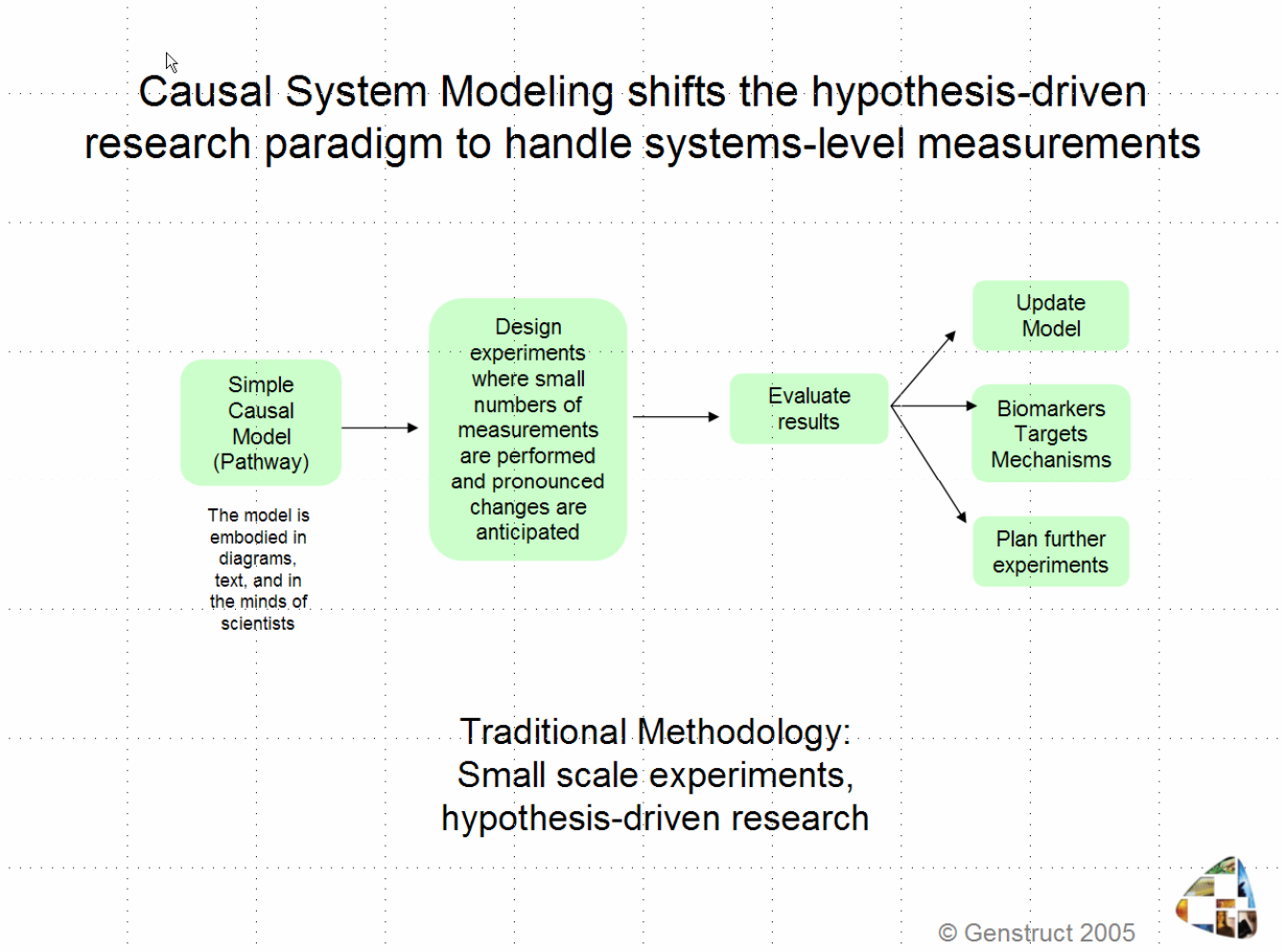

Figure 13. Small scale experiments and hypothesis-driven research

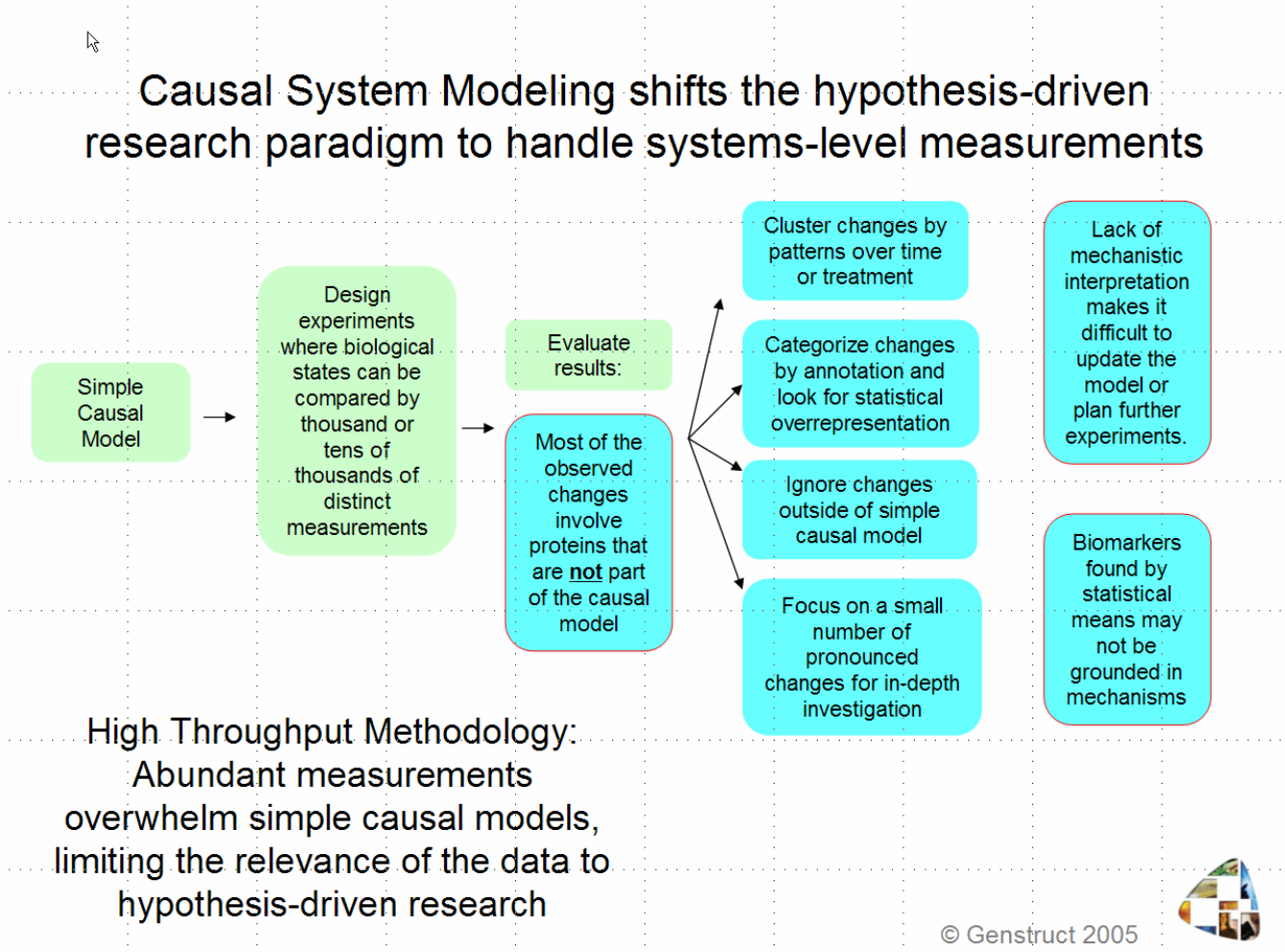

Figure 14. High throughput methodology and hypothesis-driven research

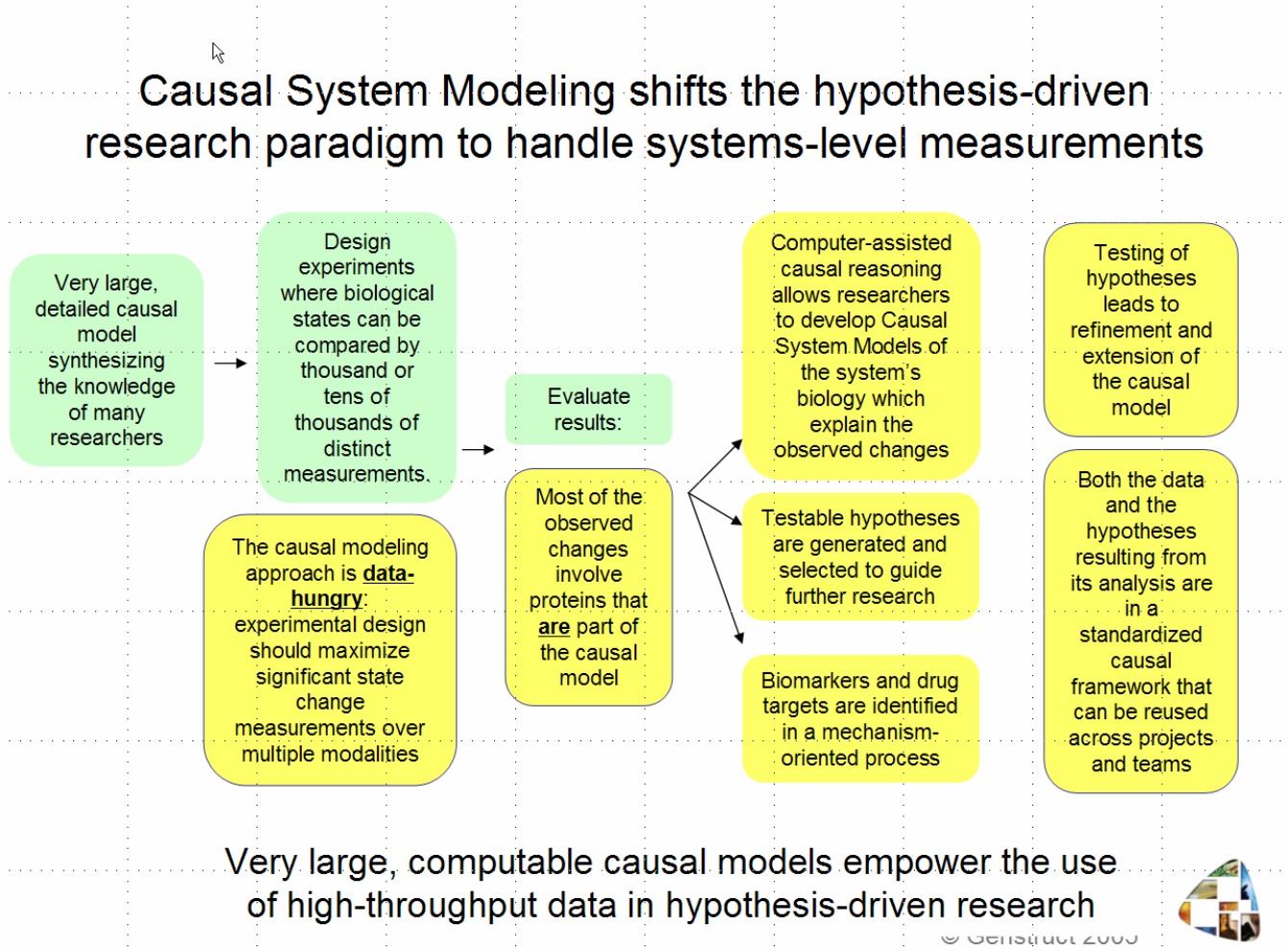

Figure 15. Causal System Modeling and hypothesis-driven research

Causal System Models enhance the value and utility of proteomic measurements

- Because the framework integrates measurements from multiple modalities (proteomic, genomic, metabonomic), each gains value by synergy with the others.
  - RNA transcript abundance can be compared with protein abundance
  - Changes in enzyme abundance can be compared to changes in substrates
  - Ambiguous downstream abundance changes can be resolved to specific upstream signaling pathways by phosphorylation assays
- Specific identification of protein species becomes more valuable when the Causal System Model uses their changes to distinguish between competing hypotheses
  - Full-length proteins vs. cleavage products
  - Identifying changes in protein modification
  - Distinguishing isoforms and splice variants
  - Verifying species origin in xenograft models

Figure 16. Causal System Models enhance the value and utility of proteomic measurements

Summary

- Large scale causal models are a practical means to incorporate high-throughput proteomics data into hypothesis-driven, mechanism-based research
  - Generation of testable hypotheses
  - Identification of biomarkers
  - Selection of novel drug targets
- Large scale Causal Models can integrate and simultaneously exploit measurements from multiple 'omics technologies.
- Causal System Modeling shifts the research paradigm to make high-throughput measurements a critical part of hypothesis driven research.
- Causal system modeling has been successfully applied to a diverse range of biology and model systems
  - Breast cancer
  - Prostate cancer
  - Muscle hypertrophy and atrophy
  - Type 2 diabetes
  - Vascular inflammation
  - Dyslipidemia
- Construction and effective use of large scale causal models depends upon a compact, precise ontology focused on simple causal relationships.
- Due to the conserved nature of biology, once biological knowledge is encoded, it can be transparently reused, both within a disease area and across mammalian biology
- Our methodology generates Causal System Models that characterize disease states or other biological phenomena. Because they are fully supported by the literature references that underlie each causal connection, they are both a mechanism for "what if" predictions and a dynamic knowledge resource.

Figure 17. Summary