Subsurface Digital Twin and Emergence

by

Yushi Zhao

B.A. Computer Science, Goshen College, 2005
M.S. Geological Sciences, Cornell University, 2012
Ph.D. Geological Sciences, Cornell University, 2015

Submitted to the System Design and Management Program
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN ENGINEERING AND MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by: Yushi Zhao
System Design and Management
August 18, 2023

Certified by: Dr. Eric Rebentisch
Research Associate at the Sociotechnical Systems Research Center, MIT
Thesis Supervisor

Certified by: Professor Bradford Hager
Cecil and Ida Green Professor of Earth Sciences, MIT
Thesis Supervisor

Certified by: Professor Ruben Juanes
Professor of Department of Civil and Environmental Engineering, MIT
Thesis Supervisor

Accepted by: Joan S. Rubin
Executive Director
System Design and Management Program, MIT

Subsurface Digital Twin and Emergence

by

Yushi Zhao

Submitted to the System Design and Management Program
on August 18, 2023 in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN ENGINEERING AND MANAGEMENT

ABSTRACT

Subsurface characterization stands at the nexus of humanity's growing demands for materials, energy, and safety amid the burgeoning population and rising living standards. However, challenges in subsurface characterization, rooted in conventional practices, functional silos, limited data density, and technological constraints, impede business efficacy and sustainable development. As societies' expectations shift and industries evolve, a paradigm shift is required in the human-machine relationship and the way we organize work. To meet these challenges and ensure responsible human progress, a systematic solution is needed.

This thesis investigates the concept of a subsurface digital twin as a boundary object that bridges disciplines, scales, and uncertainties, fostering collaboration and real-time informed decision-making. It explores the evolution of subsurface characterization from data-sparse and theory-dependent practices to a holistic digital twin framework. The thesis identifies critical technical and sociotechnical challenges, including data scarcity, over-reliance on empirical relationships, functional silos, and trust. The thesis demonstrates how a subsurface digital twin can enhance cross-functional collaboration and address critical challenges through real-world examples. It highlights the use of geoanalytics and machine learning to predict total organic carbon content and formation brittleness, showcasing the digital twin's power in multidisciplinary workflows. Furthermore, it proposes a solution for uncertainty reduction through integration and laid out future steps for the development of the subsurface digital model, construction of pseudo/surrogate models for probabilistic simulation complex and time-consuming numerical simulations, and use of the digital twin to bridge workflows between data-rich and data-scarce regions across scales.

The thesis outlines the design and value-creating functions of the subsurface digital twin system, facilitating adaptive resolution and agile implementation. It envisions a future where such digital twins revolutionize decision-making, from individual project optimization to enterprise-wide insights. The thesis underscores the importance of strategic investment in digital twins for long-term returns and as a cornerstone of the evolving human-machine relationship and advances the concept of a subsurface digital twin as a transformative approach to subsurface characterization, fostering collaboration, tackling challenges, and paving the way for sustainable progress in a rapidly changing world.

Thesis supervisor: Dr. Eric Rebentisch
Title: Research Associate at the Sociotechnical Systems Research Center, MIT

Thesis supervisor: Professor Bradford Hager
Title: Cecil and Ida Green Professor of Earth Sciences, MIT

Thesis supervisor: Professor Ruben Juanes
Title: Professor of Department of Civil and Environmental Engineering, MIT

Acknowledgments

I would like to take the opportunity to recognize all the support and assistance I received during my time at MIT. First of all, I wish to thank my advisors Eric Rebentisch, Bradford Hager, and Ruben Juanes. Secondly, I wish to thank my program executive director Joan Rubin, and co-director Bryan Moser. Thirdly, I would like to thank all the core lecturers Professor Edward Crawley, Bruce Cameron, and Donna Rhodes. School of Engineering and Sloan Management School lecturers Professor Steven Eppinger, Juanjuan Zhang, Nancy Leveson, David Nino, Richard Neufville, John Hauser, Vivek Farias, Eben Lazarus. Fourthly, I wish to thank Bill Foley, Amal Elalam, Naomi Gutierrez, Jasmine Hicks, Melat Hunde, Ryan Liebel, Jazy Ma, Julie Papp, Ignacio Vazquez Rodarte, Marcelle Durrenberger, Rob Day, Grace Anh, Daniel Richards, Maddie Golison, Anup Sreekumar, Ryan Montvydas, and all the TAs, mentors, sponsors, and support staffs I failed to mention.

Finally, I would like to thank my company Chevron for the corporate sponsorship, management supports, and mentors Fang Lin, Matthew Duke, Jianchang Liu, and Barry Katz for this great development opportunity.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

Subsurface characterization is deeply intertwined with humanity. With the increase in world population and overall standard of living, demand for materials, energy, and safety is ever-increasing[1], [2]. Inattentiveness and delayed decisions in supply and demand while considering environmental requirements could bring tensions to the intricacies of social–economical –environmental balance, stagnated growth, and larger social disparities. Current practices of subsurface characterization have been evolving and improving over the decades resulting from improved solvers, technologies, and computational power. However, the fundamental way of the human-machine relationship and how people perform work has not changed and is growingly challenged by the limitation of the scope of work, functional silos, pressing timelines, and capital constraints. These issues lead to increased uncertainties, increased risks, and overall increased costs which would be a limiting factor or improved efficacy.

Societies' expectations and demands of industry are changing, and so must the relationship between humans and machines and the ways of organizing work[3]. To accommodate responsible and sustainable human progress for generations to come, a more efficient system of subsurface characterization should be implemented. Trial and error have suggested that a holistic subsurface characterization powered subsurface digital twin would be an effective boundary object linking all the puzzles together in meeting the challenges of a new era.

Realizing this vision is challenging. In addition to the challenge of the current way of working, the data density in the field of Earth Science in subsurface characterization and exploration is much lower compared to other industries. These challenges typically result in us being over-reliant on theories, models, and empirical relationships, having inconsistent results, and often depending on a “leap of faith” in addressing the relevancy of our predictions and their embedded uncertainties.

Furthermore, theories are often established and validated under a data-rich system in controlled conditions. Rescaling or analog selection is typically made when one hopes to apply a set of expected relationships or mental models in another geographical area, stratigraphical position, or to address another geological challenge. This is often accompanied by some degree of environmental corrections to account for local variations backed by direct or indirect measurement and data. In most of the exploratory regions, however, we do not enjoy the luxury of great data abundance, completeness, and quality like that from the laboratory or in shallow subsurface fields such as soil science, required to establish a robust analytical relationship. Analog data and model extrapolations and physics-based model simulation approaches are typically used to establish a baseline scenario with a limited amount of understanding of all the local geological controls and justifications for all the input parameters[4]. Uncertainties based on these methods are unfortunately large, and are inversely correlated with data availability. This unfortunately bears significant consequences as it would often in poor economic outcomes and investment decisions, delay timelines, and increase execution risks[5].

In addition to the data limitation, due to the limitation of computational power, the total number of computational model cells typically determines the time required for complex physics-based simulations to run to completion. In a typical basin evaluation, the x-y dimensional grid size is usually in the order of 1km by 1km if a meaning level of details in the z direction is considered. To provide results in a timely manner, practitioners often tradeoff model resolution and simulation time, and the practice of simulating a meaningful number of scenarios to capture and evaluate uncertainties and risks are rarely exercised.

Recently, with the advancement in machine learning and artificial intelligence, there have been proposals and pilots in leveraging ML and AI to help address the limitations of current practice. This is evidenced by many applied fields achieving groundbreaking successes in data-rich industries such as the technology and internet companies, finance banking industry, healthcare and pharmaceutical, and e-commerce and retail industry[6]–[11]. Areas in geoscience where data is relatively rich and where controls are relatively well-understood have also seen new ML-based workflows gaining acceptance by the scientific community and industry. This includes niche disciplines within the domain of petrophysics, geophysics, and geochemistry within development fields with data abundance. Some examples are neural network methods of processing petrophysical logs for rock properties[12], [13], using machine learning to identify faults on seismic images[14], [15], and mineralogic identification and point counting on thin sections[16]–[18], as well as biomarkers and fingerprinting analyses[19]–[21].

The application of ML and AI in subsurface characterization overall remains challenging for a few reasons. Data sampling density is extremely sparse in terms of measurements per rock volume compared to similar metrics in other industries. In order to resolve a subsurface feature, $\frac{1}{4}$ of the wavelet is required[22]. This is translated to a vertical resolution of tens of meters in most 2D seismic data that is gathered as a standard practice. In high-resolution 3D seismic gathers, the vertical resolution can be increased to meters or sub-meters level, but even this level could mean an amalgamated series of geological events and a possible span through millions of years of the geological record. In the medical imaging field, on the contrary, the CT image qualities are high and machine learning and training are on the order of sub 1 millimeter per pixel. Even in a matured development field, the fact that we do not take all the measurements multiple times and at the exact location brings a large amount of data and extrapolation uncertainty. All these issues render an impossible task for a confident data science model. This sometimes begs the question that if ML and AI is even applicable in subsurface characterization if the data availability and density will never be enough. A combined physics-based simulation methods and data science approaches need to be integrated into a system solution.

In addition to data availability and quality, the lack of geological and physical guardrails often leads to overreliance on statistical relationships which often introduces augmented uncertainty due to overfitting and results making no physical sense because of the lack of domain knowledge as guidance[23]. This degrades trust and confidence in practitioners applying this technology in subsurface characterization. There are active debates on the effectiveness of pairing a domain expert with a data scientist in applications versus developing domain experts with sufficient data science knowledge so they can perform advanced analytics on their own. The advantage and disadvantages are pronounced in both options as researchers have pointed out that the trust between two individuals is difficult to establish while the technical debt of retraining the majority of domain experts to be proficient in data science will be a stretch[24]–[26].

The challenge we are facing has always been both a technical and a sociotechnical one. In addition to the trust issue, the nature of decision-making is based on scale. That typical decision-making is basin-scaled, prospect- and play-scaled, and reservoir-scaled. Example decisions are country entry, play risking, prospect evaluation, and reservoir management. These scale-dependent decisions drive how tools are developed and used to answer those questions and are also precisely the reason why the cross-functional collaborations stop there.

Teams are formed to model the subsurface using the same data set but at different scales. As decisions are made at different stages of exploration-development phases, functional silos are formed almost by design. Geoscientists working on basin-scaled models will not need to talk to geoscientists working on modeling the reservoir. Work products generated from one team rarely get shared across the functional boundary. In the rare occasion where works are shared, they would not be directly integrated into other people's workflow. When new data is presented or a new question is raised, a new model is often built as the path of least resistance. This is amplified by pressing timelines business units usually impose on the domain experts whereby they must use simple assumptions or only partially calibrate the model to answer a specific question. Poor documentation leads to the misuse of outputs that are only conditional and not meant to be applicable across the board. As a result, data integration, cross-functional collaboration, and interpretation quality are low, leading to a low overall success rate in exploratory wells. A scale-agnostic solution would not only help to resolve technical challenges but also allow domain experts to collaborate across different scales. This would allow subsurface practitioners to spend more time working on things that create more value than data processing and rework.

Many of these issues have been previously identified. Design thinking brainstorming sessions and Architecting Innovative Enterprise Strategy (ARIES) Framework have been used or proposed in trying to narrow the gap[27], [28]. However, an investment in a holistic solution has not been made also attributed to the pretext of siloed organizational nature and the bounded scope after considering the extent of solutions. Specifically, we tend to favor high impact low effort solutions that we perceive as lower-hanging fruit. These projects often get prioritized, translating to us tackling the symptom instead of the cause by putting off more strategic initiatives for a later time.

In this thesis, we aim to illustrate how these challenges can be addressed using a system approach. We identify the effective boundary object, a subsurface digital twin that is facilitated by a holistic subsurface characterization, in addressing the aforementioned technical and sociotechnical challenges. A subsurface digital model is an indexed and georeferenced subsurface information volume or digital repository. It contains information on an expansive range of subsurface parameters and some through geological time. Some examples are pore pressure, temperature, stresses, thermal maturity, porosity, bulk density, fluid saturation, fluid properties, mineralogy, lithofacies, total organic carbon content, etc. An initial subsurface digital model is produced by the combination of physics-based numerical simulation of a geological model and geoanalytics which is used to calculate geological parameters and their distributions that are not a typical model simulation output. Geoanalytics, or geological data analytics, is a vital component of a subsurface digital model where it acts to reduce subsurface digital model uncertainty through an iterative approach. An example can be that the initial geological model provides the organic carbon development condition such as paleo-water-depth, basin geometry, and sedimentation rate, allowing depositional total organic carbon content and organofacies or hydrogen index and their distributions to be calculated through geoanalytics. This organic richness information is then feedback into the geological model for compaction, hydrocarbon generation, and migration simulations. In addition, geoanalytics also helps with geological data quality controlling (QCing) and validating geological modeling assumptions, especially in paleo events such as exhumation and erosion where data are not typically widely available.

In Chapter 2, we first illustrate the level 0 to level 2 architectural diagrams of an example basin-scaled subsurface digital model, a precursor to a basin-scaled subsurface digital shadow and digital twin. We use a few use cases to demonstrate how a subsurface digital model or digital twin would yield new and improved products with a significant amount of efficiency. The first use case, a geoanalytical approach to formation brittleness distribution prediction using information available from a subsurface digital model, aims to demonstrate how a subsurface digital model can be used to foster cross-functional collaboration, which results in better products quicker.

Understanding the brittleness of rock is important for drilling and completion. In unconventional resource development, the brittleness of rock informs drilling engineers how to most efficiently and safely drill through the rock strata and where to land a well or anchor a directional heel. Completion engineers also need this information for an adequate completion design to maximize economics and minimize risks through hydraulic stimulation. Conventionally, brittleness data are evaluated based on the lithological information from petrophysical logs, thin-section measurements, and geomechanical tests[29], [30]. In Chapter 3, the thesis documents a pilot study, for the first time, using geochemistry and temporal information derived from structural restoration and basin evolution, bringing useful insights on diagenesis to inform rock brittleness prediction, data uncertainty, and extrapolation. In addition to mechanical compaction, metamorphic transformations such as quartz and calcite growth are important contributors to the diagenesis process. Tree-based random forest machine learning algorithms[31] are used to identify key geological controls and data governing brittleness distribution allowing brittleness values to be extrapolated away from well control. This is achieved by applying a satisfactory machine learning model to associated properties extracted from the subsurface digital model. We identified that present-day burial depth is not a correlating factor due to thickness changes resulting from exhumation and erosion. Temporal information such as maximum burial depth and maximum formation temperature is much more correlative to the brittleness distribution. This pilot illustrates for the first time how geochemistry and basin modeling data could be used for petrophysical property prediction through a subsurface digital model and geoanalytics and how information from other disciplines could be valuable if made available to cross-functional teams because this type of temporal information is typically not available to a petrophysicist to conduct a holistic assessment in geomechanical property modeling.

Until now, we have portrayed the subsurface digital model, the subsurface information volume as a deterministic set of numbers. However, in practice, it is just the most likely case of geological properties and their realizations through time and space. Every parameter involved has uncertainties associated with this most likely case. For parameters with greater data abundance, the range of possible outcomes may be smaller than that of model-driven parameters. Ideally, a series of physics-based numerical simulations are required to explore the uncertainties probabilistically. However, this is computationally expensive. This is especially problematic in the case of CO2 sequestration assessment because, in addition to the intricacies of subsurface multiphase fluid flow, reactive-transport effects must be considered to bring pertinency into modeling this type of application through time. Due to the significant risks the lack of uncertainty assessment would bring if reliable simulations cannot be completed, this lack will become a roadblock for subsurface domain experts in assessing the feasibility and viability of CO2 sequestration projects. Recent laboratory-scale CO2 injection and fluid flow experiments have shown challenges in generating satisfactory simulation results in a timely manner in a simplified system[32]–[34]. In the second study, documented in Chapter 4, we illustrate how the emergence occurs when the proposed system produces value greater than the integral of all of its components. This is reflected by examples of cross-functional collaboration and new workflow development documented in Chapters 2 and 3 and a proposal for an ongoing collaboration with Professor Ruben Juanes in Chapter 4. A hypothetical blue hydrogen project and data trades are used as an example to demonstrate how a subsurface digital twin can contribute to Multivariable optimization and real-time decision-making. Chapter 4 then describes how examples show that the holistic approach will lead to improved performance and reduced cost.

The thesis further documents in detail that the subsurface digital twin design has great adaptability when equipped with a flexible resolution data structure. The aim is to be able to take in simulation and analytical results from all resolutions and formats that are available and export insights in any resolution requested and format accepted. In addition to data adaptability, the workflow has to be adaptive in different geological settings and in areas with varied data availability.

The subsurface digital model is also complementary to agile implementation. Significant maintenance cost rework could be eliminated with maintaining one version of the subsurface though the subsurface digital model. The subsurface digital twin also advocates simplicity and avoids unnecessary complexity. By focusing on delivering the minimum viable product (MVP) and iterating based on feedback, teams are less likely to introduce over-engineered and convoluted solutions. This streamlined approach enhances maintainability and resilience.

Chapter 4 also documented meaningful future steps such as characterizing uncertainties into the subsurface digital model for probabilistic bases risks assessments used by decision makers. In addition to the intrinsic ability to reduce uncertainty as a holistic system, an alternative to the resource intensive physics simulations, a machine-learning based surrogate modeling workflow is proposed. The aim is to significantly reduce the simulation time required and arrive at a tolerable uncertainty range compared to that generated by numerous physics-based simulations. This proposal first builds on existing sets of laboratory-scaled experiments and allows us to investigate the feasibility of using machine learning methods to come up with surrogate models where reliable outputs can be achieved with reduced simulation time under a digital model framework and provide a list of important factors we must consider in adopting the combined simulation method.

If successful, this would provide justifications for alternative fluid-flow simulation options that could potentially revolutionize the fluid-flow simulation domain and their impact. This use case also demonstrates how in subsurface modeling space, a subsurface digital twin is an idea integrator for multi-dimensional complex cross-functional investigation where both instantaneous and dynamic effects could be evaluated under one system. We then discuss how this vision is also about bringing the system to the people instead of asking people to catch up to the system. We further articulate how it can be operationalized by adopting an agile framework where we can evolve from a digital model to a digital shadow and eventually toward a digital twin.

The agile implementation allows us to obtain feedback and alignments during implementation while utilizing some of the value-adding functions along the way. Lastly, we describe a future organization powered by subsurface digital twins, how specialists from one function can navigate through a complex organization without experiencing barriers in regard to data and insights, and are encouraged to work with other specialized practitioners by design to avoid siloed interpretations.

A blue-sky vision would be expanding our vision one step further to the scale of an enterprise digital twin, which serves as an effective boundary object to facilitate decisions in real-time. A hypothetical blue hydrogen project could be quickly optimized by pulling data and insights such as available hydrocarbon fluid properties, abundance, cost of operation, CO2 sequestration potentials, H2 storage potentials, mineral rights, access to supply and market, and pipeline and infrastructure availability from the evergreen subsurface digital twin and surface digital twin.

A digital twin is often a strategic investment where the return is higher the longer it is maintained. This type of strategic investment is an important component of a successful digital transformation. It should be invested in parallel with other quick-win solutions that address near-term challenges and be given dedicated long-term support as the human-machine relationship evolves entering industry 5.0.

This thesis work adopts a style that follows a general flow of system architecture and design, use cases, ilities, and future steps. Multiple publication-ready materials are embedded within. Chapter 4 discusses how these components will contribute to the overall theme of the thesis.

Chapter 2

Subsurface Digital Model System Architecture and Design

2.1 Introduction

With the development of data foundation enablers that bring data to our fingertips and the maturation of machine learning workflows, it is still very challenging to deploy a holistic subsurface characterization process, which is based on data and interpretations from all subsurface functions under a common set of geological frameworks. The complication rests with the siloed nature of our industry‘s utilization of multiple data types of different scales, some determined from direct measurements and others derived from indirect measurements. Extrapolations are needed because there is not a complete data set that links any given data type to all other types of measurements at the same location (X, Y, Z). Given the poor data density in our industry, a large degree of uncertainties may impact decision quality if data were extrapolated without a geological context. This unfortunately bears significant consequences as it would often result in poor economic outcomes and investment decisions, delay timelines, and increase execution risks. Furthermore, apart from the constraints imposed by limited data availability, the computational power limitations also play a crucial role in determining the time required for complex physics-based simulations. When conducting a standard basin evaluation, the grid size in the x-y dimensions is typically in the order of 1km by 1km, assuming a meaningful level of detail in the z-direction captured. In order to obtain timely results, practitioners frequently make trade-offs between model resolution and simulation time. Unfortunately, the practice of simulating a significant number of deterministic scenarios to comprehensively assess output uncertainties and risks is rarely exercised[33], [35].

The challenge we are facing has always been both a technical and a sociotechnical one. Establishing trust has been identified as one of the key issues hindering cross-function collaboration. In addition, to trust, silos are also created due to the nature of decision-making based on scale. For example, when decisions are made on a different scale and on a different timetable such as country entry decisions being made years in advance, play risking cycles through a portfolio once every few years, prospect evaluation work done in an exploration phase or contract timeline-driven multi-year scale, and reservoir management on an annual basis, domain-specific tools are developed just to satisfy those needs. Domain experts do not get trained to use different software because they are usually trained to be an expert in one technical domain. All that results in people not having the need, the drive, and the means to collaborate beyond their technical function. Geoscientists working on basin-scaled models do not need to talk to geoscientists working on modeling the reservoir. Work products generated from one team rarely get shared across the functional boundary. In the rare occasion where works are shared, they would not be directly integrated into other people's workflow. When new data is presented or a new question is raised, a new model is often built as the path of least resistance. This is amplified by pressing timelines business units usually impose on the domain experts whereby they must use simple assumptions or only partially calibrate the model to answer a specific question. Poor documentation leads to the misuse of outputs that are only conditional and not meant to be applicable across the board. As a result, data integration, cross-functional collaboration, and interpretation quality are low, leading to a low overall success rate in exploratory wells and lessons if learned, are not shared across functions.

On the flip side, it has been shown that more collaboration, integration, and internally consistent work yields better results as it will cross-constrain input uncertainty[4], [36], [37]. There is a clear value in capturing all the learnings and having only one version of the truth, a holistic subsurface characterization. This would allow subsurface practitioners to spend more time working on things that create more value than data processing and rework.

In this thesis, we aim to illustrate how these challenges can be addressed using a system approach. We identify the effective boundary object, a subsurface digital twin that is facilitated by a holistic subsurface characterization, in addressing the aforementioned technical and sociotechnical challenges. The system problem statement is summarized as follows:

2.1.1 System Problem Statement

To enable more effective decision-making with reduced uncertainty **By** building a multi-dimensional subsurface digital twin integrating models, simulations, and machine learning **Using** cross-disciplinary data sets **While** capturing and minimizing the uncertainties associated with the process.

To keep the thesis focused, the system boundary is limited to the Subsurface Digital Model and the Digital Basin Model, blue shaded boxes in Figure 2.1 below. The Subsurface Digital Model is an indexed and georeferenced subsurface information volume or a digital repository. It contains internally consistent information from different scaled digital models. As the L1 architecture in Figure 2.1 shows, it could contain information from a Digital Basin Model, a Digital Earth Model, and/or a Digital Reservoir Model. The Digital Basin Model, Digital Earth Model, and Digital Reservoir Models are domain-specific and are currently working with a wide range of commercial software packages. It is nearly impossible to align all software developers to be fully integrable with their competitors. The level of complexity in all the interfaces at such a scale would be incomprehensible and an update to one component from one software vendor would potentially cause a "hiccup" resulting in failure in data and information flow to other components. The proposed Subsurface Digital Model, therefore, is a boundary object which allows all domain-specific inputs to be integrated into a central repository. In the case of a component failure, the subsurface digital model will still retain the newest information from the previous iterations of that technical domain. The Subsurface Digital Model is hyper-dimensional and omni-resolution, capable of accepting indexed and georeferenced subsurface information volumes at different resolutions. Effective visualizers would be able to operate on this common data structure to visualize properties at any resolution through time and dimension. More specifically, one can, for example, visualize the subsurface temperature changes through geological time, present day, and dynamically after fluids are injected and extracted from wells. One can also visualize the change in total organic carbon content, hydrogen index, hydrocarbon generation, and migration with an increase in the thermal stress and structural development of the petroleum system. This is the fundamental concept of the Digital Basin Model, which we will touch on in the methodology section of this chapter.

Expanding slightly on Figure 2.1, Subsurface Digital Model, along with Surface Digital Model, Digital Organizational Model, and Digital Facility and Utilization Model yield enterprise-level insights that could be aggregated for enterprise-scaled decisions. In addition to the Subsurface Digital Model, which is shaded in blue, the Surface Digital Model could include information such as land ownership and leasing rights, legal readiness and risk, access to pipelines, facilities, and other infrastructures in the region, demand or customer information and contracts, external facilities and business development, and policy risks. Digital Organizational Model and Digital Facility Utilization Model while not expanded from the L1

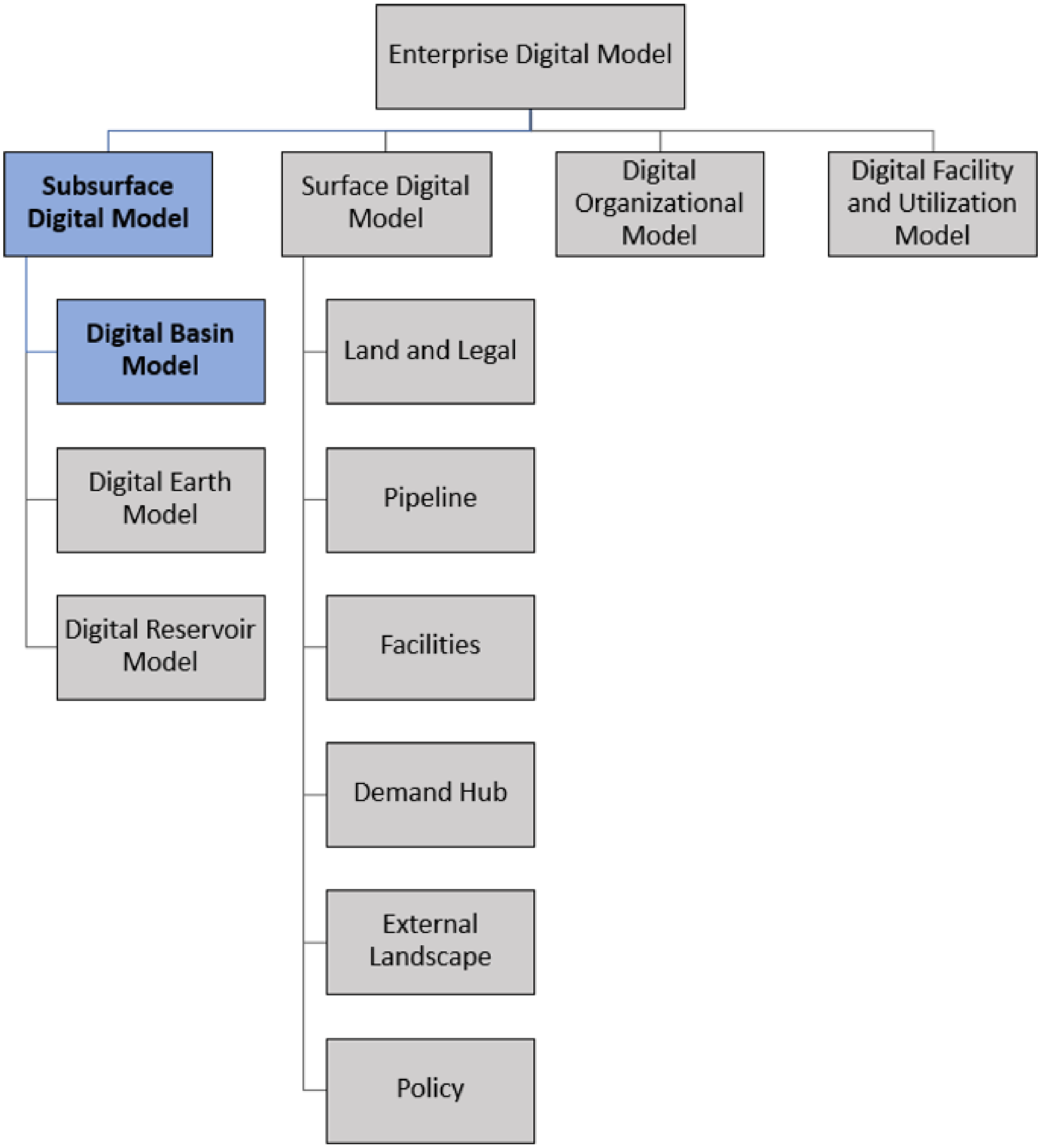


Figure 2.1: L1 Architectural Diagram of Subsurface Digital Twin System. This is for illustration purposes only as it may not include all components and sub-components. Blue-shaded blocks represent the extent of the subject of discussion in the thesis system boundary.

system diagram, it contains people-related information and information from facilities and tangible assets such as office buildings and labs. It is important to note that the diagram does not include an exhaustive list of potential components as this is just an illustration to show how quality data and insights can be leveraged when made available as a system to decision-makers.

Following the “2 down, 1 up” principle[38], the L2 architectural diagram below in Figure 2.2 displayed that the Digital Basin Model consists of a wide range of available data types. This includes geophysical data, petrophysical data, geochemistry data, rock data, production data, surface geology survey, laboratory experimental data, data from collaborators and subscriptions, and other local subsurface data lakes or analog databases. Some of these data are direct measurements such as porosity in thin section samples extracted from a side-wall core, total organic carbon content, vitrinite reflectance, and apatite fission tracks from physical rock samples. Some of these data are indirect measurements such as resistivity and gamma-ray logs, seismic reflections, and density derived from gravity and magnetic data.

These data have different levels of uncertainty and challenges when used alone. For example, high-resolution data such as gamma-ray logs and resistivity logs have a vertical resolution of half of a foot[39], but they are only available where a well is drilled. Seismic data can be collected on a larger scale, but the vertical resolution is generally on the order of meters[22]. To get the best of both worlds, domain experts tie well logs with seismic reflectors and use the reflections to determine the play boundaries[40], [41]. This demonstrates that integrating data across functions enhances product resolution and reduces extrapolation uncertainty from one siloed data.

However, this is just one case of knowledge integration. Moreover, seismic data are not always available where a well is drilled, and in development fields, seismic data are not used as a primary source of data due to their lower resolution nature. In mature unconventional basins such as the Permian Basin, it is common to work with thousands of well tops directly instead of using seismic data to model the subsurface[42], [43]. Challenges in those systems then become the vintage of seismic data, the different time-to-depth conversion velocity models used for different seismic volumes, different stratigraphers picking different formation tops, and the lack of data and interpretation in the geographical area or stratigraphical position lacking prospectivity.

For all of these systems, before siloed data can be integrated together, we must first establish a geological framework. The geological framework provides contexts for data within the same geological and depositional settings to be compared and quality checked. These data then must be weaved into a geological story that follows a structural and stratigraphical framework. Some domains in this field include structural geology, basin framework, clastic-stratigraphy, biostratigraphy, and chronostratigraphy.

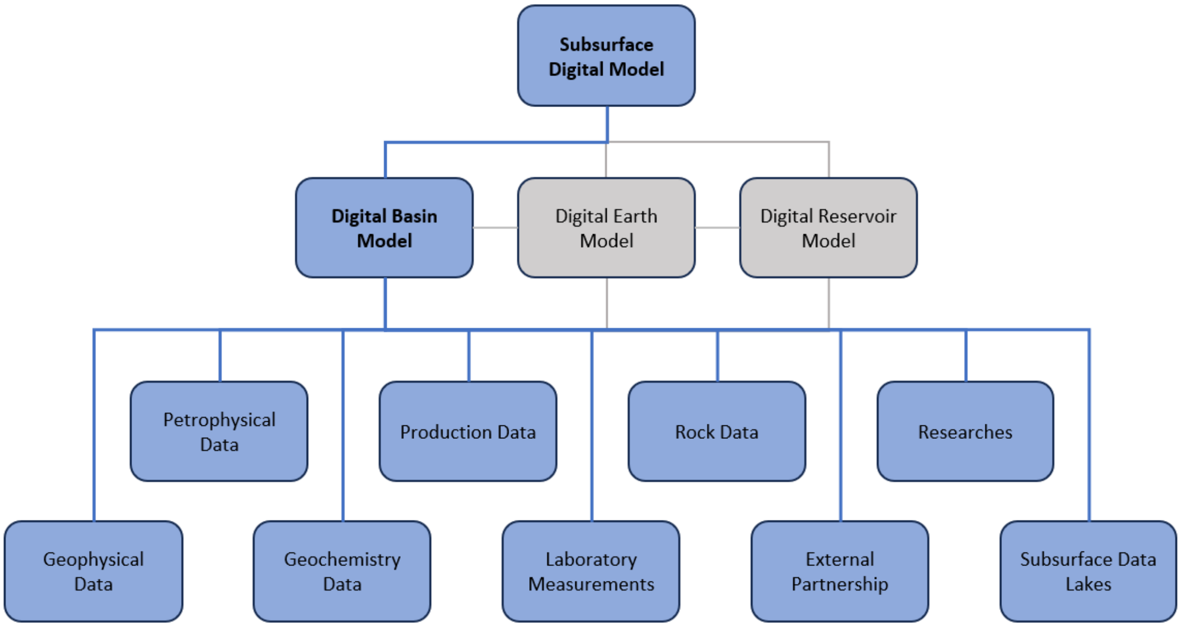


Figure 2.2: L2 Architectural Diagram of Subsurface Digital Twin System. This is for illustration purposes only as it may not include all components and sub-components. Blue-shaded blocks are system components that represent the extent of the subject of discussion in the thesis system boundary.

Stepping out a level of this framework provides us with a higher-level view of how the subsurface digital model could be utilized strategically. As the L0 architecture diagram shows in Figure 2.3, analytics and insights facilitating optimization decisions could be done in real-time when subsurface, surface, organizations, and assets models are utilized at the enterprise level. Financial modeling and external landscape assessment could be further integrated to provide informed pivoting investment decisions. Again this illustration does not include an exhaustive list of all possible components and sub-components as it only serves an illustrative purpose.

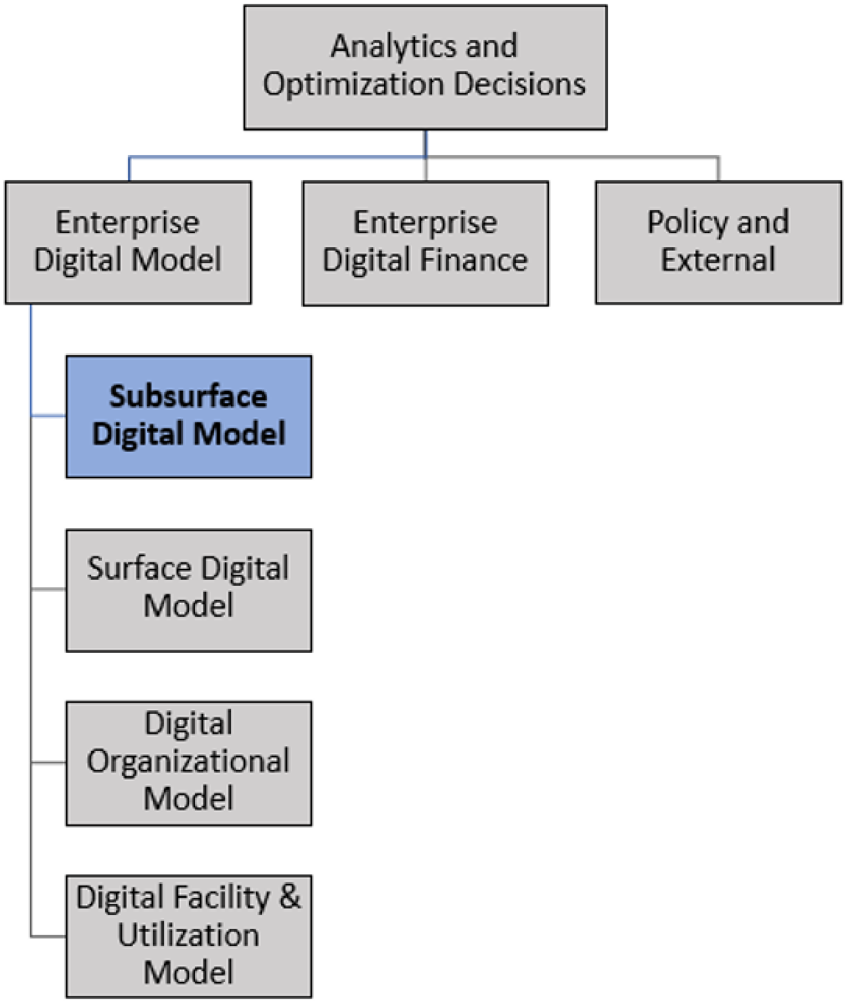
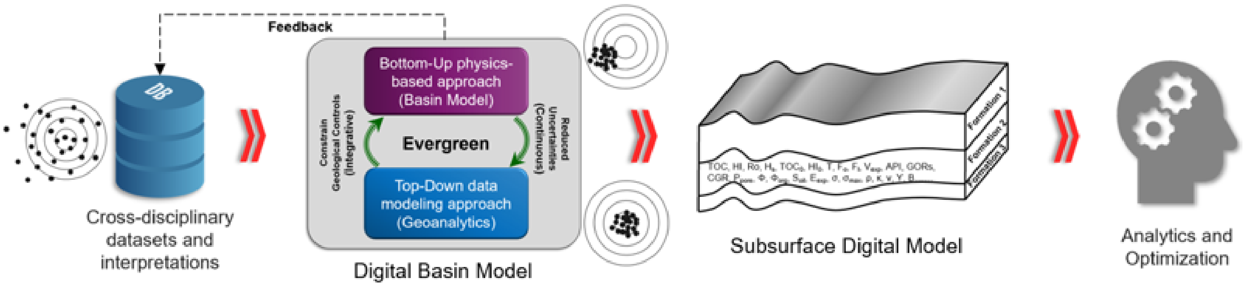

Analytics and Optimization Decisions
Enterprise Digital Model
Enterprise Digital Finance
Policy and External
Subsurface Digital Model
Surface Digital Model
Digital Organizational Model
Digital Facility & Utilization Model

Figure 2.3: L0 Architectural Diagram of Subsurface Digital Twin System. This is for illustration purposes only as it may not include all components and sub-components. Blue-shaded block illustrates the subject of discussion in the system boundary.

To add more functional elements to the L0 diagram, we can see how Digital Basin Model and Subsurface Digital Model contribute to the entire process. As Figure 2.4 shows, cleaned data are fed into the Digital Basin Model, through a combined physics-based, bottom-up basin modeling approach and top-down geoanalytics data-modeling approach. Integrated property volumes are provided in a repository or a Subsurface Digital Model. This internally consistent Subsurface Digital Model is then made available to inform decisions. It is important to note that the integration process of the Digital Basin Model construction itself significantly reduces data and extrapolation uncertainties. This data is then fed back into the raw data source with tags indicating its validation through the Digital Basin Model and deviations from the model trend. Digital Basin Model will be described in greater detail as an example platform for subsurface geological frameworks and interpretations that will be utilized by the Subsurface Digital Model in the methodology section of this thesis chapter.



Feedback
DB
Cross-disciplinary datasets and interpretations
Bottom-Up physics-based approach (Basin Model)
Constrain Geological Controls (Integrative)
Evergreen
Reduced Uncertainties (Continuous)
Top-Down data modeling approach (Geoanalytics)
Digital Basin Model
TOC, HI, Ro, H_s, TOC_o, HI_o, T, F_a, F_t, V_{exp}, API, GORs,
CGR, P_{pore}, Φ, $Φ_{exp}$, S_{at}, E_{exp}, σ, $σ_{max}$, ρ, κ, ν, Y, B
Formation 1
Formation 2
Formation 3
Subsurface Digital Model
Analytics and Optimization

Figure 2.4: Process flow of the Digital Basin Model and Subsurface Digital Model. Every process is iterative and a source of feedback to connected processes

The proposed architecture is built on four distinct layers. Figure 2.5 below shows these layers including the Data Foundational Layer, the Model Engagement Layer, the Aggregated Information Layer, and the Decision and Interactive Layer. This is the minimum number of layers required to yield a stable system architecture for scoped vision.

The Data Foundation Layer at the lower portion of the diagram is mainly responsible for storing data measurements of all sorts and disciplines. These data are typically raw data that are directly sourced from data lakes such as the Open Subsurface Data Universe

(OSDU)[44], internal data lakes, publications, and external database subscriptions. These data usually contain errors and are often incomplete. It requires cleaning procedures and domain expert quality control before entering into the physics-based modeling process.

The Model Engagement Layer above the data foundation is where all the domain experts make sense of data through physics-based model building and numerical simulation. This layer also is responsible for all the cross-functional collaborative workflows where information is shared across the modeling community. There are many software vendors supplying tools and capabilities for subject matter experts in this layer. Domain experts are expected to be the primary stakeholder for this layer.

Above this layer is the Aggregated Information Layer. In this layer, there are only structured, georeferenced, indexed subsurface properties, or information. It is sometimes called 'data,' as the 'data' is often defined loosely. However, it is best to give this type of insight a different name, such as referring to them as 'information.' This would allow the attention of people to distinguish this type of insight from raw data measurements collected from the wellbore or through other indirect means. The 'information' from the Subsurface Digital Model represents the most likely value of that property at a certain location and time. It may be different than the data measurement itself at that location giving competing data inputs and physical relationships to other measurements nearby. The disagreements between data and modeling output reduce through iterations as data are continued to be quality controlled and new geological assumptions or conditions are incorporated into the subsurface models.

Finally, we have on top, the Decision and Interactive Layer. This layer is mainly responsible for information inquiries, data analytics, optimization, and visualization. New data can be easily compared with related properties within the Aggregated Information Layer. Data analytics including machine learning techniques can be utilized in producing new property grids or volumes based on data models. Fit-for-purpose applications can be used to utilize the subsurface digital model. For example, it would be extremely valuable to have the capacity to test new data against the one subsurface interpretation in short-notice opportunities such as data trade and data rooms. Powerful visualization can also be developed to render the right level of resolution for any decision at any time.

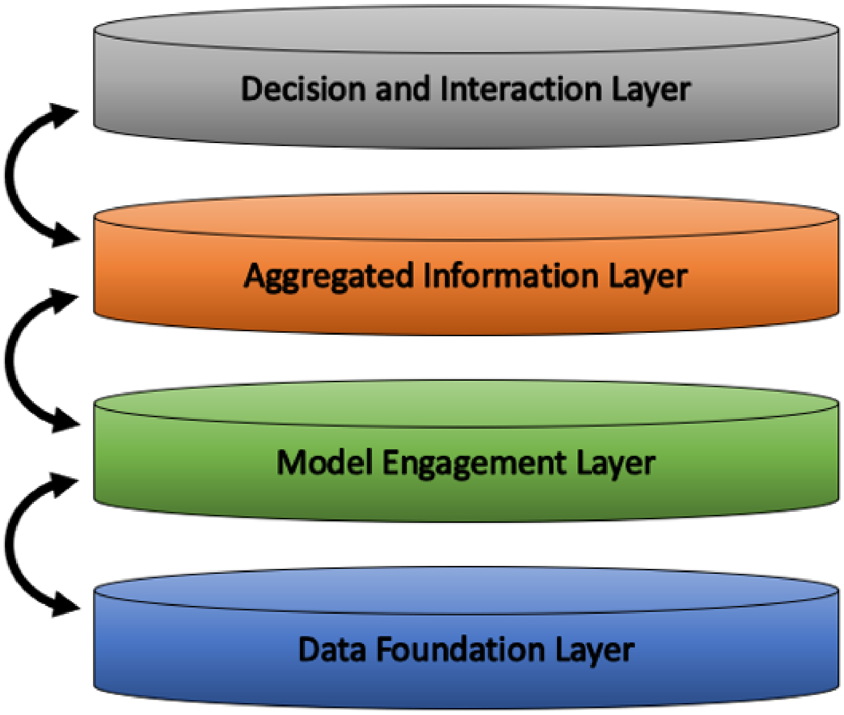


Figure 2.5: System architectural layers enabling data and information flow for the Digital Basin Model and the Subsurface Digital Model.

In general, the upper structures require lower structures and the lower and upper structures operate iteratively and mutually help improve the quality of individual layers. In this illustration, the Digital Basin Model, the Digital Earth Model, and the Digital Reservoir Model from Figure 2.5 all reside in the Model Engagement Layer while the Subsurface Digital Model resides in the Aggregated Information Layer.

2.2 Materials and Methods

2.2.1 Digital Basin Model

To keep the thesis scope narrow, we will be showcasing one type of digital model – the Digital Basin Model as an example modeling workflow in the model and engagement layer. Other types of models such as the Digital Earth Model and Digital Reservoir Model will follow a similar train of thought using similar digital tools and technologies to detail a digital subsurface property volume. Digital Basin Model as Figure 2.6 below shows, has two components, namely Integrated Basin Model and Geoanalytics.

Integrated Basin Model is defined as an extension of the domain of petroleum system analysis or basin modeling. The key difference is the iterative nature and the inclusivity of all different data types that are not typically included in a typical basin model. This could include detailed exhumation and erosion histories, presence and quality of organofacies and organic matter, detailed lithofacies volumes, an excessive number of logs and calibration data points, geophysics data, and paleo-formation-thickness restoration such as resulting from salt tectonics. Boundary conditions such as basal heat flow history (HF), paleobathymetry (PWD), and sediment-water interface temperature (SWIT) are provided as input parameters for a geological model and are then calibrated against common calibration datasets such as temperature, pore pressure, and vitrinite reflectance while at the same time allowing them to be solved as a target variable to best fit the data.

The order of calculation is the reason why it is termed the "Bottom-Up" Physics-based numerical simulation approach. Geoanalytics is developed to complement the lengthy time required to adjust and tune input boundary conditions for model output validation. Data modeling approaches can be adopted to hasten the process needed to get to a suitable boundary condition and input parameter in a short amount of time. Data modeling approaches include classical domain-led analytical relationships, local observation-driven empirical relationships, or machine learning-derived high-dimensional models which are typically difficult to visualize. The physics-based simulation helps to reduce extrapolations and provides a continuous realization of modeling properties that simply extrapolating a raw data set could not provide. On the other hand, data analytics provides useful intel on whether model prediction is in line with observation. In the case where they do not agree, lessons are learned. This typically results in improved data cleaning and labeling and additional model constraints or geological controls being considered, potentially leading to basin-wide implications. The Integrated Basin Modeling forward modeling process works hand-in-hand with the Geoanalytics process in providing an ever-better subsurface characterization.

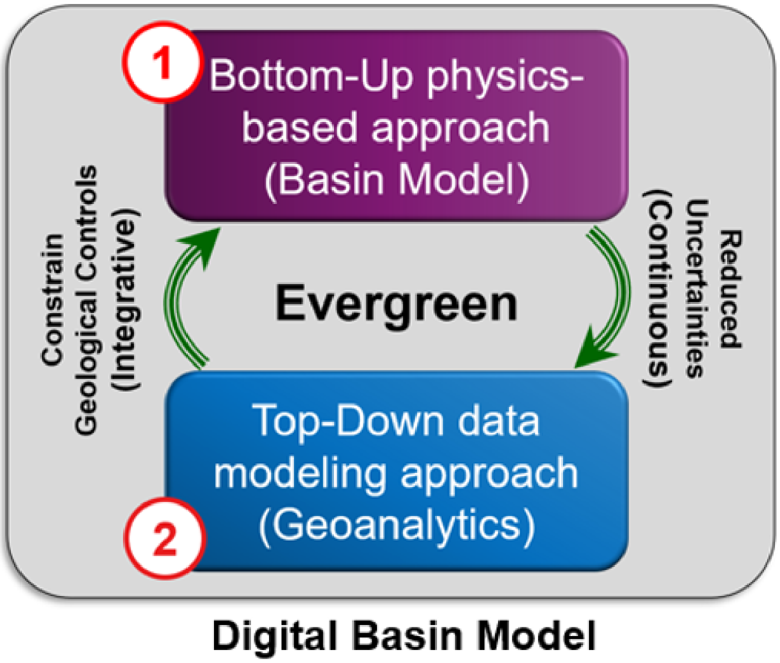

1
Bottom-Up physics-based approach (Basin Model)
Constrain Geological Controls (Integrative)
Evergreen
Reduced Uncertainties (Continuous)
Top-Down data modeling approach (Geoanalytics)
2
Digital Basin Model

Figure 2.6: Functional diagram of the Digital Basin Model. The first component is the bottom-up physics-based simulation approach, and the second component is the top-down data modeling approach or "geoanalytics."

2.2.2 Integrated Basin Model and Geoanalytics

It requires a team of domain experts to assemble all the building blocks of an integrated basin model. As Figure 2.7 shows, it encompasses the extent of an entire basin. It integrates a wide range of data that are used to establish a structural framework and a stratigraphical framework. Surface Geology data such as outcrops information are used to detail the extent of exhumation and erosion and its timing. In the case where outcrops information can be analyzed in labs, we could obtain more information such as maximum thermal exposure and degree of compaction the rock experienced. This would give us direct evidence to model rocks that are still buried deep underground. Gravity and magnetics data help us to understand the general basin geometry deep into the crystalline basement. Although this type of information is in lower resolution, it provides useful information on basin extent, geometry, and justification for the distribution of basal heat flow. Satellite data such as the Shuttle Radar Topography Mission (SRTM) from NASA[45], provides high-resolution topography information that we can integrate to approximate a sediment-water interface temperature (SWIT). Satellite altimetry[46], [47], Light Detection and Ranging (LIDAR)[48], and various Sonar technologies are used to obtain accurate bathymetric depths when the sediment surface is below sea level[4]. Global mean surface temperatures adjusted to paleo-plate-reconstructions are typically used to estimate SWIT through geological time[4]. In the case where the topography is above sea level, the effect of the temperature lapse rate under associated adiabatic conditions may be considered a correction to the sea level temperature[49], [50]. All this information is essential to set up a proper thermal model to capture present-day conditions and is used to simulate temperature through geological time. Together with the basement depths inferred from gravity and magnetic data, it resembles the construction of the 'geological oven.'

To continue detailing the framework of the geological model temporally, key geological events together with regional seismic data and well logs are integrated chronologically into the regional geological framework. Important information such as tectonic events, stress regime, intrusion episodes, fault developments, exhumation, erosion, as well as salt movements gives key information for paleo-structure and bathymetry restoration[51]. Due to stress orientation changes through geological time and differential compaction, depositional centers may shift from one era to another[52]. Paleo-water-depth (PWD), paleo-formation-depths, and paleo-formation-thicknesses are typical output from this exercise and provides a high-order guideline on how and what geological contents or matters went through the 'geological oven' since their deposition. The geological model can then be further detailed by adding vertical resolutions. This is typically achieved by incorporating higher resolution 3D seismic data and a wealth of petrophysical data such as gamma-ray logs, resistivity logs, sonic, and if possible neutron porosity logs and image logs to establish higher order stratigraphic framework. This higher-order stratigraphic framework makes sure bio-stratigraphy, carbonate-stratigraphy, seismic-stratigraphy, and clastic-stratigraphy are internally consistent. In successful cases, this framework can be detailed to the fourth-order system tract level such as in the data-rich Permian Basin[53], [54].

To simulate temperature and pressure through time, we must define the lithology of rocks occupying every model cell in the integrated basin model. The detailed stratigraphical framework also gives guidance on the facies and mineral distribution needed to decorate the lithological model. This process is important because different types of rock will display different physical properties through diagenesis. For example, shaly rocks usually contain a high level of porosity early depositional, but the porosity decreases drastically through compaction. Sandstone on the other hand may have lower initial porosity but will have higher terminal porosity due to stronger structural supports of quartzite minerals[55], [56]. Porosity directly translates into differential bulk density, bulk thermal conductivity, and permeability and pore pressure calculations. Porosity change also impacts the change in geological structure. This has a direct impact on the flow paths of target fluid migration. Hydrocarbon is lighter than water, buoyancy forces drive the hydrocarbon expelled from the source rock interval upward. The change in geological structure may alter fluid flow pathways which may bear significant consequences. In other words, mischaracterized lithology may result in false confidence in prospectivities found in an actual fluid migration shadow. Some clay-rich minerals are also rich in uranium, potassium, thorium, and plutonium. This can also attribute to significant radiogenic heat production[57], [58]. Radiogenic heat generation accounts for half of today's surface heat flux[59], [60]. In data-rich unconventional basins, high-fidelity lithofacies maps are typically produced through geostatistical means using abundant petrophysical logs. There are other forward modeling methods one may adopt that yield alternative extrapolation realizations through the lithofacies modeling exercises such as Compstrat forward modeling and geoanalytics approaches[61], [62]. For consistency purposes, one set of protocols should be followed in generating these lithofacies grids. Some methods may work well in data-rich intervals while performing poorly in intervals with limited well penetration, while others may produce a better overall product balancing between data-rich and data-poor rock strata. These issues illustrate the need to capture input and output uncertainty in our digital basin model and the subsurface digital twin. We will discuss this issue in detail in Chapter 4 of this thesis.

With the 'geological oven' and 'geological matters' defined, present-day temperature, pore pressure, and porosity measurements can be used to tune or calibrate the boundary conditions for the present-day thermal model. Temporal insights captured from information such as vitrinite reflectance, TMax measurements from rock pyrolysis, thin-section quartz and carbonate overgrowth measurements, fluid inclusion analyses, presence of igneous sills, biomarkers, diamondoids, porosity measurements, and apatite fission track measurements, when available can be utilized to validate the geological story while adding timing pertinence in paleo-events[63]–[69]. Timing is critical because it is much riskier to find a commercial accumulation of hydrocarbon when the hydrocarbon generation predates the reservoir geometry and seal formation. Other hydrocarbon migrations delaying mechanisms are required for these types of systems to work such as late structural development or the presence of hoteling mechanisms and interformational seals[57], [70], [71]. For unconventional resources, mechanisms that promote hydrocarbon retention are emphasized instead of expulsion[72], [73]. Regardless of the play type, integrating all different data types into the integrated basin model reduces the uncertainty of our subsurface understanding and opens the door for a high-fidelity digital basin model and a subsurface digital model.

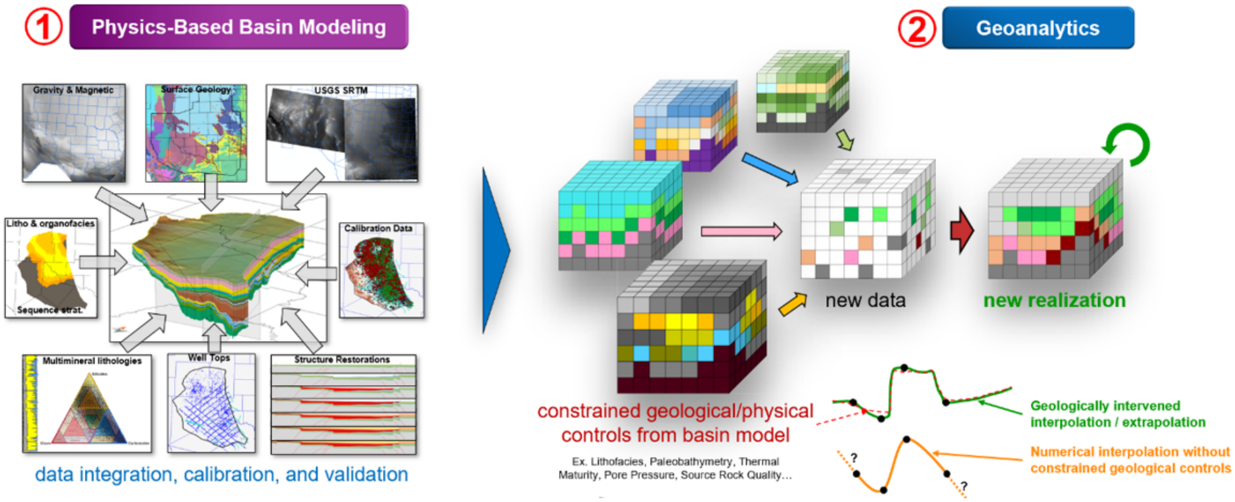

①
Physics-Based Basin Modeling
Gravity & Magnetic
Surface Geology
USGS SRTM
Litho & organofacies
Sequence strat.
Calibration Data
Multimineral lithologies
Well Tops
Structure Restorations
data integration, calibration, and validation
②
Geoanalytics
new data
new realization
constrained geological/physical controls from basin model
Ex. Lithofacies, Paleobathymetry, Thermal Maturity, Pore Pressure, Source Rock Quality...
?
?
Geologically intervened interpolation / extrapolation
Numerical interpolation without constrained geological controls

Figure 2.7: Graphical illustration of the Digital Basin Modeling components. Physics-based modeling approach provides framework information for geoanalytical methods.

Through the integrated basin modeling process, geoanalytics workflows are developed to generate high-fidelity modeling properties. Geoanalytics is data analytics under a geological framework utilizing advanced statistical tools while constrained by domain-specific relationships. As Figure 2.7 shows above, as subsurface information is extracted at the same location (X, Y, Z, T) of a new data type, a data model can be established. Geoanalytics is solution neutral, meaning it can accommodate domain-specific analytical relationships, local empirical observations, or higher dimensional machine learning workflows. After a satisfactory statistical model is constructed to explain all the observations, this model is then applied to all cells in the model to generate an extrapolation of that new data set. This way, geology, domain-specific relationships, and physics are built into the data modeling process. This significantly reduces the extrapolation uncertainty compared to siloed extrapolation work such as inverse distance, seed-based, or geostatistical-based methods that are practiced currently.

2.2.3 Digital Model, Digital Shadow, and Digital Twin

The Digital Model is a broad term that refers to any digital representation of a physical object. They are connected through indirect feedback and manual data flow. A digital model can range from a simple 2D representation to a high-resolution full-physics simulation through time depending on the intended purposes. In this thesis, the subsurface digital model represents a 3D – 4D digital representation of a subsurface system of interest. The content of this subsurface digital model is supplied from physics-based simulations across multiple disciplines and geoanalytics using cross-functional data.

A digital shadow is a dynamic and real-time digital representation of a physical object. Automations are built in so the digital shadow can be updated continuously based on real-time data collected from the physical object. In the field of earth science, this could mean sensors in place to monitor seismic activity, pressure and temperature sensors in the borehole, or flow gauges at the wellhead. The advantage of the digital shadow is to automate data collection and entry so domain experts can be informed in real-time. This is often implemented when strong causal relationships are established, and sensors can be placed to obtain critical information needed.

A digital twin goes one step further in sophistication and complexity. It is typically a digital representation of a physical object, system, or process that is not only monitored in real-time but also includes additional contextual information, historical data, and advanced analytics capabilities such as prediction and prescription. This requires a high degree of understanding of the physical system and controls and is typically the end goal of any digitalization of a physical object or system.

From a system perspective, much more is happening than what is documented in papers and articles[74]. A digital twin is usually perceived as a simple model, but in reality, it is very complicated[75], [76]. Attaining the data, licensing the data and software, and a lot of social system complexities that are not accounted for from the technical complexities of this model/system. Therefore the implementation strategy of the subsurface digital twin is critical as a complex system does not necessarily mean a complicated system. All relationships among components and interactions among stakeholders should be carefully evaluated to help improve decisions.

In implementing the subsurface digital twin vision, the subsurface digital model process can be viewed as a Minimum Viable Product (MVP) of the subsurface digital twin. As Figure 2.8 shows below, when feedback loops between the physical system and the digital system are maintained bidirectionally, a subsurface digital model becomes the subsurface digital twin. For this reason, "Digital Model" and "Digital Twin" terminologies are used interchangeably in this thesis.

Grieves and Vickers[74] also described different types of digital twins from an enterprise perspective such as the digital twin prototype, digital twin instance, digital twin aggregate, and digital twin environment. For the scope of the paper, we limit the digital twin to the system boundary to the system diagrams shown in Figure 2.1 above.

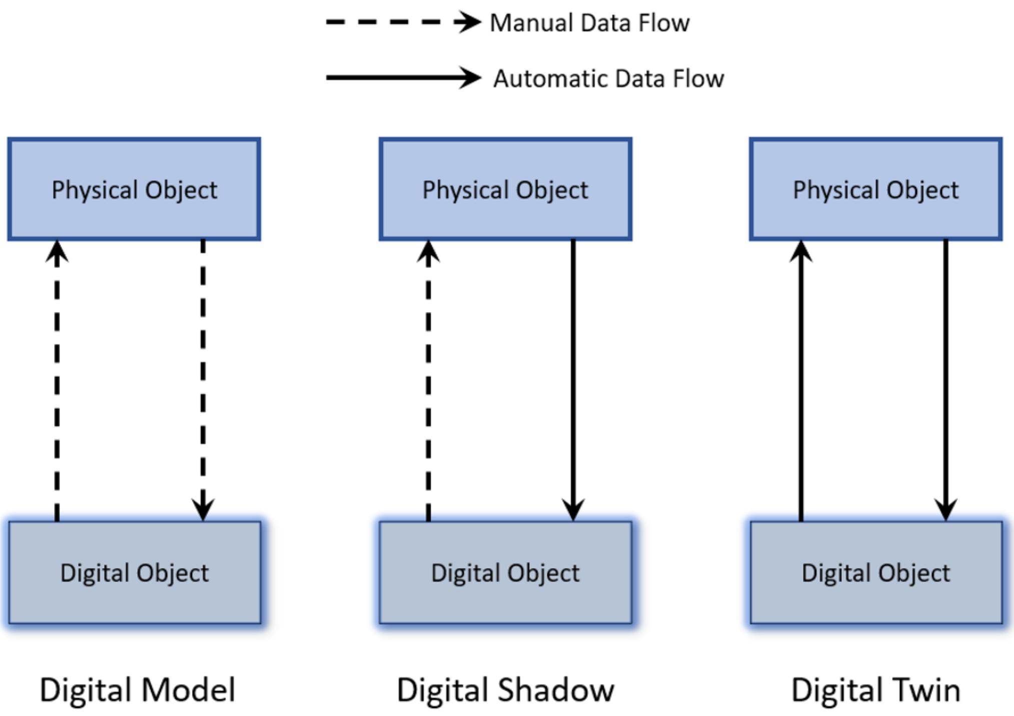


Figure 2.8: Graphical illustration of the technology-focused definition of a digital model, digital shadow, and digital twin. Figure adapted from Kritzinger 2018[77].

2.3 Results and Discussion

The integrated basin model construction process is facilitated by cross-disciplinary collaboration in the form of a Team of Teams organizational model[78], [79]. The modeling process is iterative and learnings are bi-directional. During an internal pilot study, it was identified that there were a lot of geochemistry, well tops, and engineering data that were merged into the database with wrong Xs and Ys, likely resulting from erroneous projection transformation. Many data also have erroneous units and header information such as depth below mud-line and total vertical depth sub-sea assigned. Different software also may denote elevation with a positive number and depth subsea using a negative number while some tools will do the opposite for domain-specific purposes. These findings were shared amongst all project teams, and it was critical to improve everyone else's work product.

Figure 2.9 below shows three different types of data QC classifications categorized through the internal pilot study. The first type is “Automated QC.” This type of data QC does not require a domain expert’s participation. It is generally applying rules to screen out problematic data. Some of these rules include “Is this a text or number?” or “Can this field be a negative number?” Sometimes these even happen behind the scene automatically during a data merge. The second type of data QC is “Advanced QC.” This is the state-of-the-art practice where domain experts will be looking at the analytical correlations, empirical relationships, data distribution profiles, and dynamic measurement profiles to make an interpretation-based decision. Domain experts usually leverage regression, and QC templates on trade-spaces to derive insights. The third type of data QC is system-based data QC. In the subsurface digital model system, all data are talking to each other through physics and first principle-based processes. This process cross-constrains possible data ranges due to the presence of other competing evidence in the same geological location or under the same set of physical laws. Geospatial and temporal relationships are also possible to allow data QC from higher dimensions. This process requires cross-functional collaboration and produces the highest value.

Figure 2.9: Illustration of different data quality control methods (Automated QC, Advanced QC, and QC through system integration and data analytics.

QCed data are generally used with limited validation. For additional quality assurance, data passed through the advanced QC stage can still be validated through geoanalytics.

Geoanalytics not only ensure that all the data involved in the data analytics are from the same location or geological setting, but it also allows users to visualize and operate in the data domain. For example, the total organic carbon content and its distribution, a key contributor to a working petroleum system, is difficult to predict because as organic material matures and expels away from the source rock, the remaining total organic carbon content (TOC) decreases compared to its depositional magnitude[80], [81]. The instantaneous sedimentation influx is also a critical organic matter preservation factor. In the beginning, some level of sedimentation helps to preserve the organic materials by effectively burying the organic-rich sediments through the sediment surface reactive layer, preventing organic material oxidation by the oxygen-rich bottom water and being bioturbated[82]. However, at some point the amount of inorganic influx is going to outpace the additional amount of organic carbon preserved, resulting in the dilution of source rock quality or reduced bulk total organic carbon[83]–[86].

Figure 2.10 below shows the rare opportunity for the observation of total organic carbon content distribution and their relationships with the thermal maturity of the source rock and the instantaneous sedimentation rate during source rock deposition of a world-class source rock as a result of the internal pilot. This system also allows us to explore a P10 and P90 case analytical relationship to establish some level of confidence interval in the final interpretation. Illustrated by the red arrow in Figure 2.10 below, there can be different surfaces that cover the upper and lower envelop of the analytical relationship. Note that the Ro and sedimentation rate axes are anonymized here due to IP concerns but it satisfies the purpose of this section that is to illustrate the utility of such an integrated system can bring to domain experts.

Products generated through geoanalytics are therefore much more geologically sound and are more correlative to other parameters that expect a relationship with the TOC. As

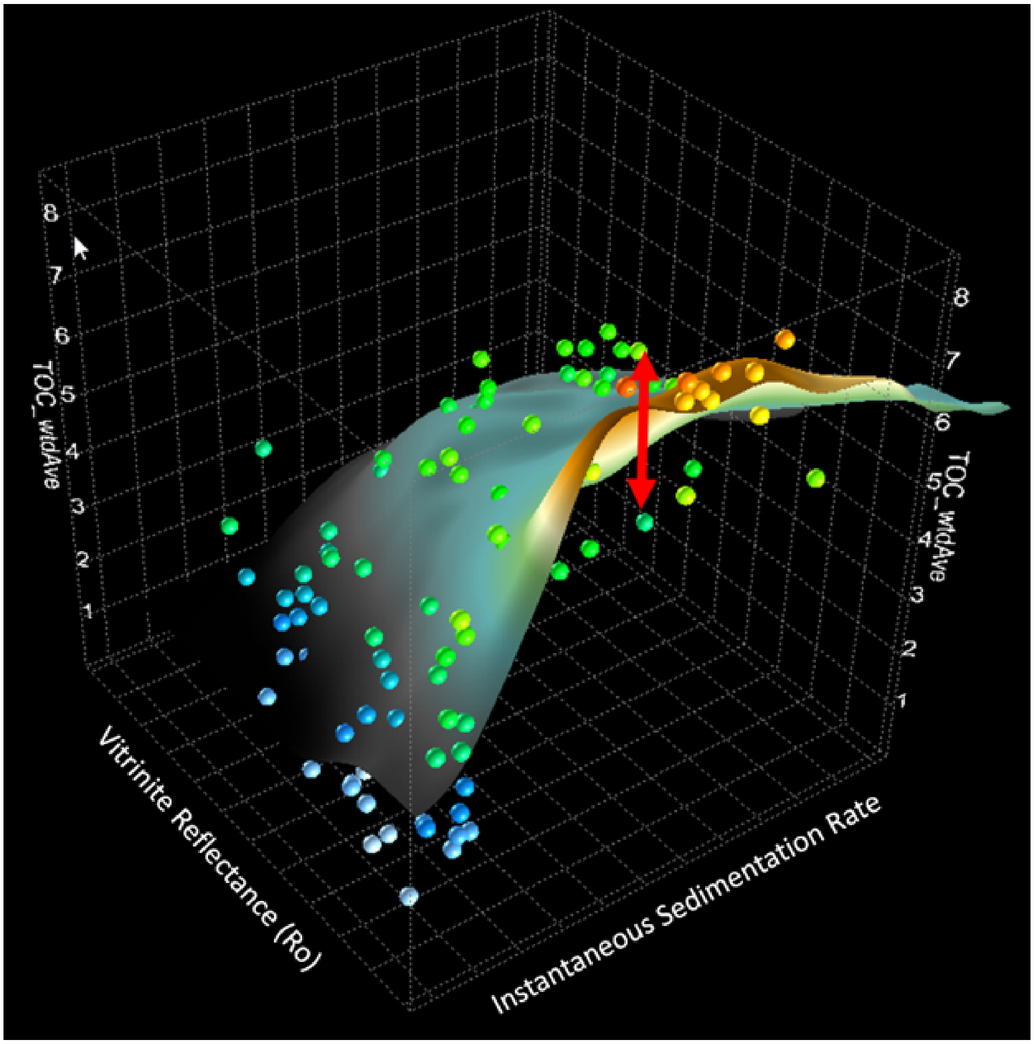


Figure 2.10: Total Organic Carbon content (TOC) as a function of thermal maturity (Ro) of the source rock and the instantaneous sedimentation rate during source rock deposition. The red arrow indicates the degree of data variation and the level of uncertainties in this data model. Some axes are anonymized due to IP concerns.

Figure 2.10 shows, a new TOC realization for the interval of interest is generated by applying this model to all cells with vitrinite reflectance and sedimentation rate values in the same geological settings given that the established analytical solution coverage is expansive through the entire geographical and data domain. Figure 12.11 below shows a histogram of a normalized delta percentage grid generated through geoanalytics and another grid through reverse distance squared extrapolation method. The uncertainties away from well control can easily reach +/- 40% in absolute terms. A use case of higher dimension relationships will be documented in detail in the following chapter (Chapter 3) through machine learning.

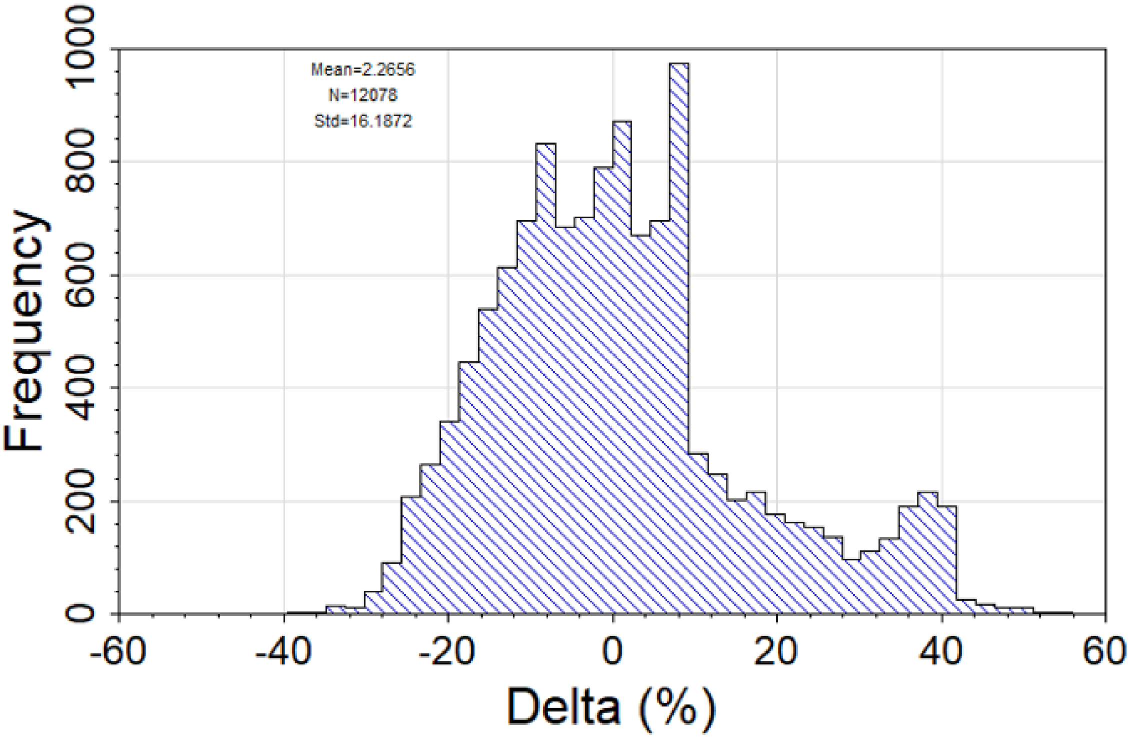


Figure 2.11: Delta histogram showing the uncertainty difference between numerical extrapolation methods and geoanalytical method using the Digital Basin Model.

In summary, the data uncertainties and extrapolation uncertainties can be significantly reduced when a system-based approach is adopted. Figure 2.12 below shows how integrated basin modeling and geoanalytics workflows systematically reduce uncertainty compared to classical basin modeling approaches and siloed interpretations without any type of data integration. When a domain expert uses simple numerical extrapolation methods, the uncertainty is the largest. Basin modeling tools and workflows have been evolving and becoming more inclusive of new data types in a more inclusive manner. The basin modeling process reduces overall uncertainty because different data are integrated with a geological framework. However, in fluid flow simulations, many assumptions are usually made with limited data support. Augmented uncertainty from assumptions such as source rock generation and expulsion kinetics, expulsion efficiency, original TOC and hydrogen index, source rock thickness, and hydrocarbon migration pathway and efficiency results in an amalgamated amount of uncertainties.

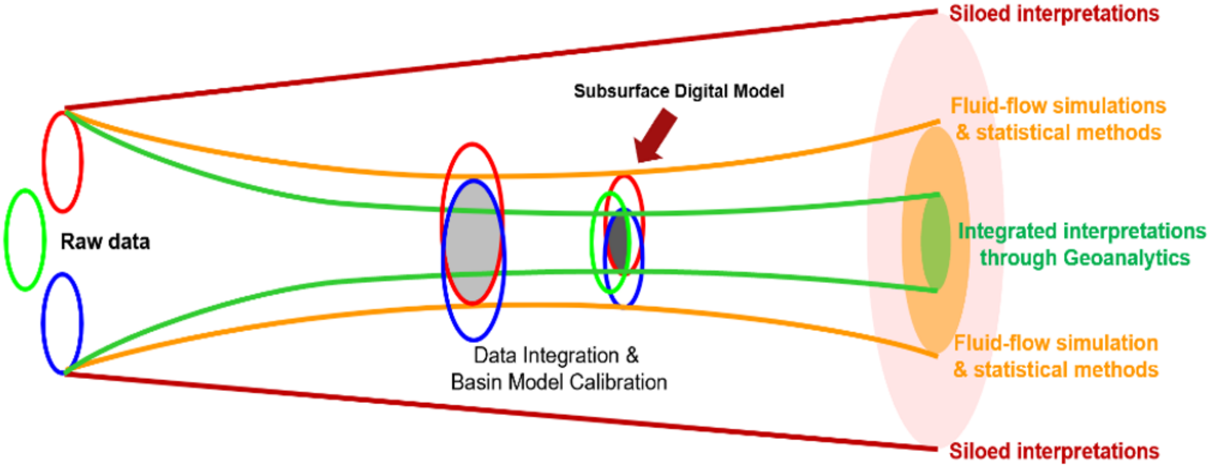


Figure 2.12: Delta histogram showing the uncertainty difference between numerical extrapolation methods and geoanalytical method using the Digital Basin Model.

Integrated basin modeling together with geoanalytics combines the bottom-up physics-based approach and the top-down data modeling approach iteratively. This allows data analytics to be constrained within a meaningful geological context, and products generated can be integrated back into simulation right away to generate a more constrained simulation output. It would be valuable to maintain a digital basin model in every active basin the company operates in.

Chapter 3

Brittleness Prediction Using a Subsurface Digital Twin and Machine Learning Algorithms

3.1 Abstract

Understanding the brittleness of rock is important for drilling and completion. In unconventional resource development, the brittleness of rock informs drilling engineers how to most efficiently and safely drill through the rock strata and where to land a well or anchor a directional heel. Completion engineers also need this information for an adequate completion design to maximize economics and minimize risks through hydraulic stimulation. Conventionally, brittleness data are evaluated based on the lithological information from petrophysical logs, thin-section measurements, and geomechanical tests. In Chapter 3, the thesis documents a pilot study, for the first time, using geochemistry and temporal information derived from structural restoration and basin evolution, bringing useful insights on diagenesis to inform rock brittleness prediction, data uncertainty, and extrapolation. This pilot illustrates for the first time, as a system, how geochemistry and basin modeling data

could be used for petrophysical property prediction through a subsurface digital model and geoanalytics and how information from other disciplines could be valuable if made available to cross-functional teams as this type of temporal information is typically not available to a petrophysicist to conduct a holistic assessment in geomechanical property modeling. Tree-based Random Forest machine learning algorithms are used in geoanalytics to apply key geological controls and data governing brittleness distribution identified to extrapolate brittleness index values away from well controls. A series of data modeling protocols are established in this satisfactory machine learning-based geoanalytics application and its associated properties extraction from the subsurface digital model. As a validation process, we identified that present-day burial depth is not a significant correlating factor due to extensive thickness changes resulting from exhumation and erosion in the area of investigation. Temporal information such as maximum burial depth and maximum formation temperature is much more correlative to the brittleness distribution. It is evident that in addition to mechanical compaction, metamorphic transformations such as quartz and calcite growth are important contributors to the diagenesis process and should be considered for such analyses in the future. A subsurface digital model would facilitate better and quicker products through cross-functional collaboration.

3.2 Introduction

In this chapter, the aim is to demonstrate cross-functional collaboration through a system-based solution, the Digital Basin Model that was introduced in the previous chapter. An internal pilot study assessing rock brittleness distribution is investigated with learnings documented here as a use case.

Understanding the brittleness of rock is important for drilling and completion. In unconventional resource development, the brittleness of rock informs drilling engineers how to most efficiently and safely drill through the rock strata and where to land a well or anchor a directional heel. Completion engineers also need this information for an adequate completion design to maximize economics and minimize risks through hydraulic stimulation[87]. Brittleness estimation away from well control is difficult, however. Conventionally, brittleness data are evaluated based on the lithological information from petrophysical logs, thin-section measurements, and geomechanical tests. These data are collected at the wellbore, extrapolations are typically required to construct a geomechanical model covering the entire reservoir or stimulated rock volume for unconventional systems.

Machine learning methods have been developed in the field of petrophysics so that we can produce a synthetic compressional wave travel time (DTC_SYN) and shear wave travel time (DTS_SYN) sonic log using correlations of other log responses to sonic logs in similar lithologies[88], [89]. However, this only works within a well where log data are being collected. Prediction of geomechanical properties away from wells has been challenging due to limited geological guidelines for geomechanical properties propagation away from well control or to formation without log data/penetration. Overfitting is a common pitfall for high-dimensional machine learning models[90], [91] and even geostatistical methods[92], [93] when used to extrapolate away from well control while domain experts are not always in a good position to interrogate the machine learning (ML) models due to the often-limited transparency of the process[94], [95]. However, there are arguments on both advantages and risks of allowing intervention[96], [97].

The main issue for applying advanced extrapolation methods is that the data density away from well controls is much lower, often in different resolutions and with different modeling assumptions resulting in inconsistencies. A large degree of uncertainties may impact decision quality if data were extrapolated without a geological context. A system-leveled information repository such as the Digital Basin Model would enable machine learning-based geoanalytics workflow which bridges the best of two worlds. The physics and chemistry-based basin model is utilized to reduce geological uncertainties as well as improve interpolation and extrapolation by reducing uncertainties through the honoring of first principles, petroleum system principles, geospatial relationships, temporal relationships, and available data. Products ingested by machine learning/artificial intelligence applications thus result in more informed and pertinent assessments. The system also gives both domain experts and data modelers tools to investigate modeling parameters and intervene as needed. The processes and protocols of modeling and data validation bring credibility to geoanalytical products.

In this manuscript, we demonstrate how cross-disciplinary datasets such as geochemical data as well as simulated basin model temporal information such as maximum burial depths can be used as control of diagenesis which impacted the rock brittleness complimenting conventional lithological data for the first time. Multiple data science QC protocols are established to safeguard geoanalytics model fidelity. Geoanalytics is integrated interpretations and multi-variable data analyses using a cross-constrained physics-based model. While geoanalytics can be applied to other scaled models, in the context of the Digital Basin Model, geoanalytics is being operated on an integrated basin model. Geoanalytics is also solution neutral. It can include classical analytical approaches or machine learning-based analytical approaches. Figure 3.1 below illustrates the high-level flow diagram of how these interactions take place for a machine-learning analytical relationship used for this case study.

3.2.1 System Problem Statement

The SPS of this pilot is **To** reduce exploration and development uncertainty **By** providing a geologically valid mean brittleness index distribution grid for a well-defined rock formation through geoanalytical means **Using** a digital basin model.

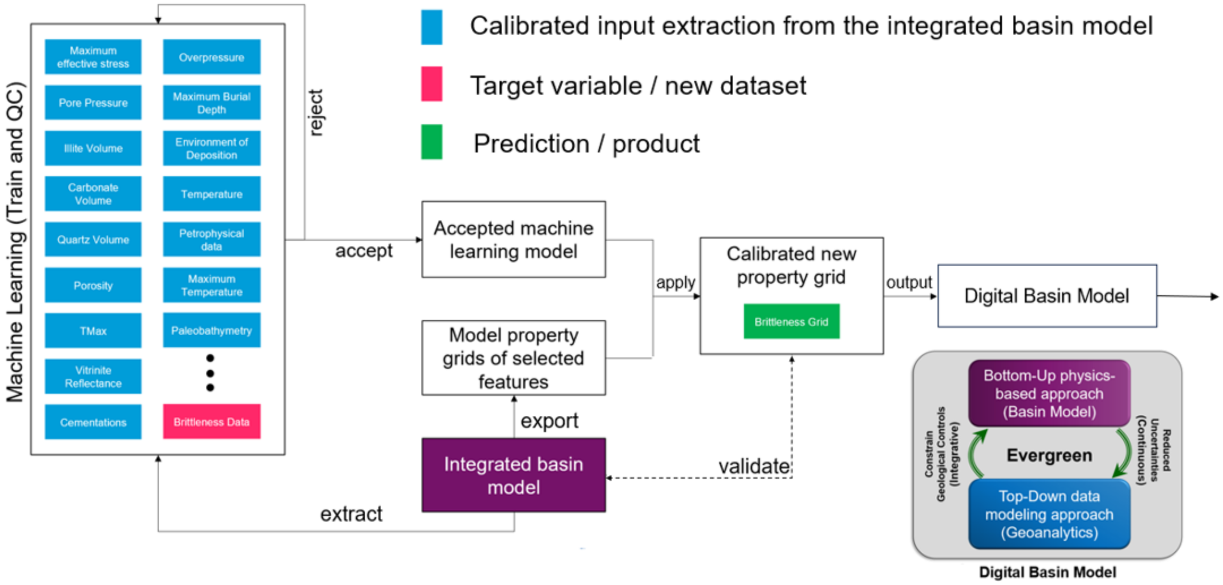

Calibrated input extraction from the integrated basin model
Target variable / new dataset
Prediction / product
Machine Learning (Train and QC)
Maximum effective stress
Overpressure
Pore Pressure
Maximum Burial Depth
Illite Volume
Environment of Deposition
Carbonate Volume
Temperature
Quartz Volume
Petrophysical data
Porosity
Maximum Temperature
TMax
Paleobathymetry
Vitrinite Reflectance
Cementations
Brittleness Data
reject
accept
Accepted machine learning model
apply
Calibrated new property grid
Brittleness Grid
output
Digital Basin Model
Model property grids of selected features
export
Integrated basin model
validate
extract
Bottom-Up physics-based approach (Basin Model)
Constrain Geological Controls (Integrative)
Evergreen
Reduced Uncertainties (Continuous)
Top-Down data modeling approach (Geoanalytics)
Digital Basin Model

Figure 3.1: High-level workflow of the machine learning brittleness index prediction study. Blue boxes are features extracted from the Digital Basin Model, Pink box is the target variable, and the green box is the model output. Bottom right is an illustration of the two main components of the Digital Basin Model.

The pink box from Figure 3.1 above represents the variable of investigation, the brittleness index data. This data has a format of X, Y, Z, and value. Because the brittleness of rock is a function of depositional mineralogy and diagenesis post-deposition, all parameters within the first and second-order relationship to these geological controls are extracted from the integrated basin model or the digital basin model. These datasets are in the same format as the brittleness data, X, Y, Z, and value, represented as blue boxes to the left. Total datasets are then broken into an 80/20 training and testing dataset split. Data analytical modeling is performed on all the extraction training datasets to predict the brittleness training dataset. Upon a satisfactory machine learning model is found and correlating features identified and validated, the model is applied to the entire correlating feature property grids exported from the integrated basin model or the digital basin model. This will result in a digital basin model resolution modeling output of our target variable brittleness index. This is denoted as the green box in the diagram. This product grid is then validated against other independent property grids from the model to validate the consistency and quality of extrapolation. It is then broadcasted to the digital basin model or others to use or incorporated into the integrated basin model for interactive simulation when possible. It is important to note that for the purpose of this pilot test, all the property grids used are basin modeling or geoanalytics products and have already been completed, validated, and integrated into the subsurface digital model prior to this assessment. Some property extrapolation may be required because the digital basin model and subsurface digital model do not permit void spaces in the data volume. Detailed modeling processes and protocols to minimize extrapolation uncertainty are documented in the material and method and result and discussion section in the chapter.

The iterative nature of geoanalytics and integrated basin modeling process reduces the data and extrapolation uncertainty and improves the accuracy of a physics-based basin model through time. The following section will document a series of data modeling protocols which is required to meet our system problem statement requirement.

3.3 Materials and Methods

Neural network machine learning method was used to produce synthetic shear wave travel time (DTS) and compressional wave travel time (DTC) sonic logs when there are data missing in a section of the wells. DTC, gamma ray (GR), neutron porosity (NPHI), density (RHOB), photo-electric factor (PEF), and resistivity (RESD) logs are used to predict a synthetic shear travel time log (DTS_SYN) when DTC data is available. GR, NPHI, RHOB, PEF, and RESD logs are used to predict both the DTS_SYN and synthetic compressional wave travel time log (DTC_SYN) when DTC and DTS data are both absent. Rock brittleness data was then calculated using the Halliburton Brittleness Index method[29]. Brittleness Index data are upscaled to a weighted average value for each well for the formation of interest after standard petrophysical cross-plot data QC are applied. This cleaned dataset and the precise positions (longitudes, latitudes, depths), or (Xs, Ys, Zs) are then used as the position to query information from the integrated basin model.

The petrophysical model and integrated basin model share the same stratigraphical framework. Therefore, all the data extractions happen in the same formation. Parameters that are affiliated with depositional mineralogy composition and diagenesis processes are upscaled and extracted from the integrated basin model. If the extraction position does not lie on top of a grid node, distanced based weighted average method is used for the extraction. This ensures extractions with the same resolution and consistency. This information cannot be used directly before validation. There are typically two issues, 1: the extraction of integrated basin model position is nearing the formation boundary, or the formation is very thin; 2: there is a large sum of model parameters that are extracted, and some may exhibit collinearity. Leaving these issues unchecked would result in higher dimensional overfitted models, problem converging, or a large amount of model uncertainties.

Exploratory data analysis (EDA)[98], as Table 3.1 shows, helps us to explore problematic data and the collinearities of all parameters involved in the analysis[99]. There are data from 489 wells with brittleness index calculations for the formation of interest involved in the study initially, a total of 49 feature parameters are extracted from the integrated basin model from the same 489 brittleness index data positions. This information is placed on a large array grouped by the same Xs, Ys, and Zs. Data engineering workflow under an integrated basin model identifies the reasons for null cells and problematic values. These are typically due to 0 thickness formation and extraction near formation boundaries. After problematic cell removal, a total of 441 quality common datasets are created for every feature extraction plus the brittleness index. Each extracted feature is then plotted against every other feature to determine the collinearity. Features that are controlled by the same factors are grouped with some removed from the modeling process to avoid biasing. Some example parameter

Table 3.1: Summary Table for Exploratory Data Analysis (EDA)

EDA and Data QC Results	Data Amount	%
Brittleness Data	489	100 (Baseline)
Final Data Counts	441	90.2 (Maximize)
Initial features	49	100 (Baseline)
Final features	7	14.3 (Minimize)

that displays great amount of collinearity includes maximum vertical effective stress and maximum burial depth or maximum formation temperature and vitrinite reflectance.

There may be some parameters that demonstrate strong collinearity but at the same time yield new information. This can be for example maximum burial depth and vitrinite reflectance because pressure and stress are the new information the maximum burial depth carries. It may be perceived that the deeper it is the hotter it gets, but this is not always the case when 3D heat-flow is considered with varied rock thermal conductivity and surface temperature difference. In these cases, we aim to preserve these features by investigating similar features within the similarity group to reduce the collinearity and increase the new information added. The guiding principle is to maximize the final data count and minimize the final features involved in the data science model while not compromising the model quality and feature importance. The process of feature engineering is iterative, the number of features or the selected features may still change if the model does not produce consistent outputs and displays unexpected or unjustifiable relationships. 7 final modeling features, quartz volume, maximum burial depths, porosity, calcite volume, TMAX, average TOC, and volume of illite are selected through the entire data modeling process which is the least number of features without compromising model quality and stability.

Figure 3.2 below describes the workflow of the machine learning modeling process. After data assurance and quality checks through the EDA process, the data array extracted is parsed into an 80/20 split training and blind test set. The blind test data set has never been seen by the machine learning model and is used to validate the model's efficacy at the end. The training data set is further divided into small subsets of data chunks and used in different combinations through the cross-validation process[100]–[103].

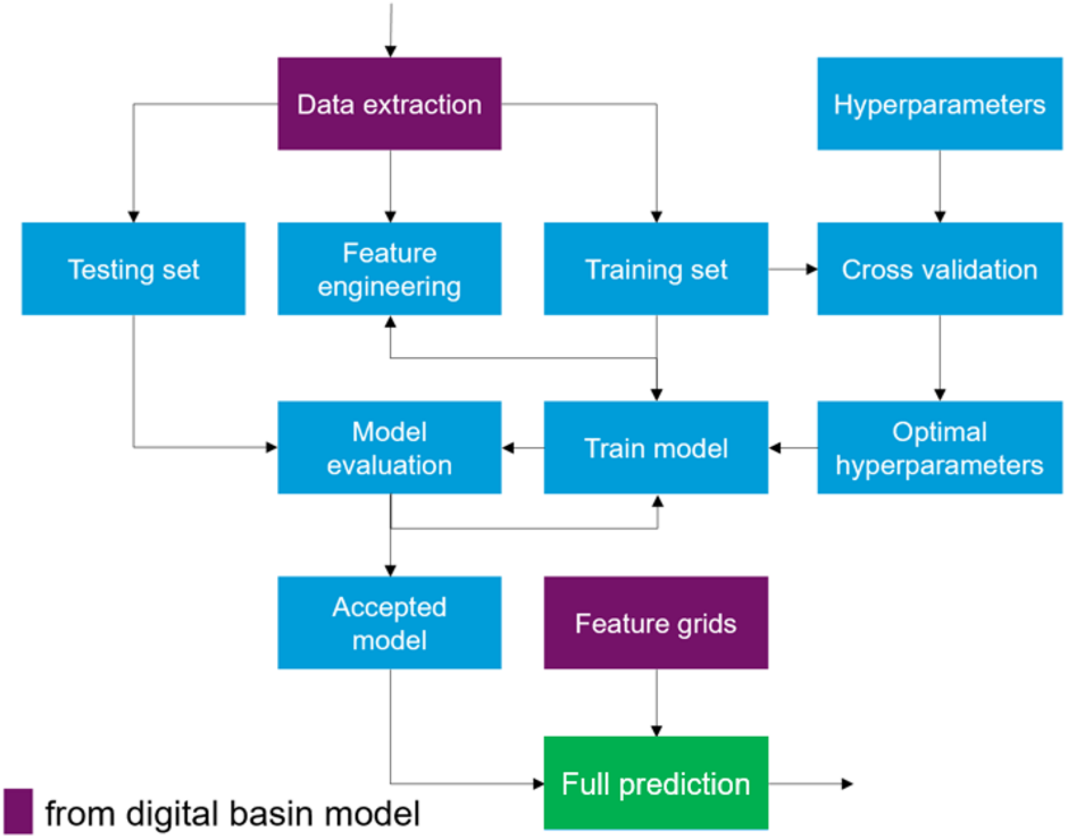


Figure 3.2: Machine learning-based geoanalytics workflow diagram for brittleness index prediction pilot. The purple boxes are parameters extracted from the Digital Basin Model, and the green box is the model output.

The cross-validation process, as Figure 3.3 shows, allows the model to be trained with different combinations of input data. In this pilot study, 10-fold cross-validation approach is used. This means the model breaks the training dataset into 10 random data chunks and then trains and validates the model in 10 different iterations. Model performance is reported as an average of the 10 iterations instead of just one single set training data set. This process helps to minimize biases or to produce a "good" machine-learning model by luck and results in a more stable model.

Extra protocols are implemented for the hyperparameter tuning process which is essential for a set of fair hyperparameters used in the training of the machine learning model. These protocols minimize the risk of overfitting, a common problem in applying machine learning models in lower data density applications. A set of initial model features are selected during the EDA process. Features are iterated, evolved, and finalized with a balance among feature selection, hyperparameter requirement, model performance, and model stability through the feature engineering process. A final set of features are selected at this stage. In this pilot study, we used the tree-based machine learning modeling method random forest[31] due to its advantages such as the reduced overfitting risk, the better handling of biases and uncertainties, and the capability to handle non-linear relationships. However, the machine learning-based geoanalytics using an integrated basin model is machine learning method agnostic. This means it can be used for any type of data modeling approach, not limited to random forest.

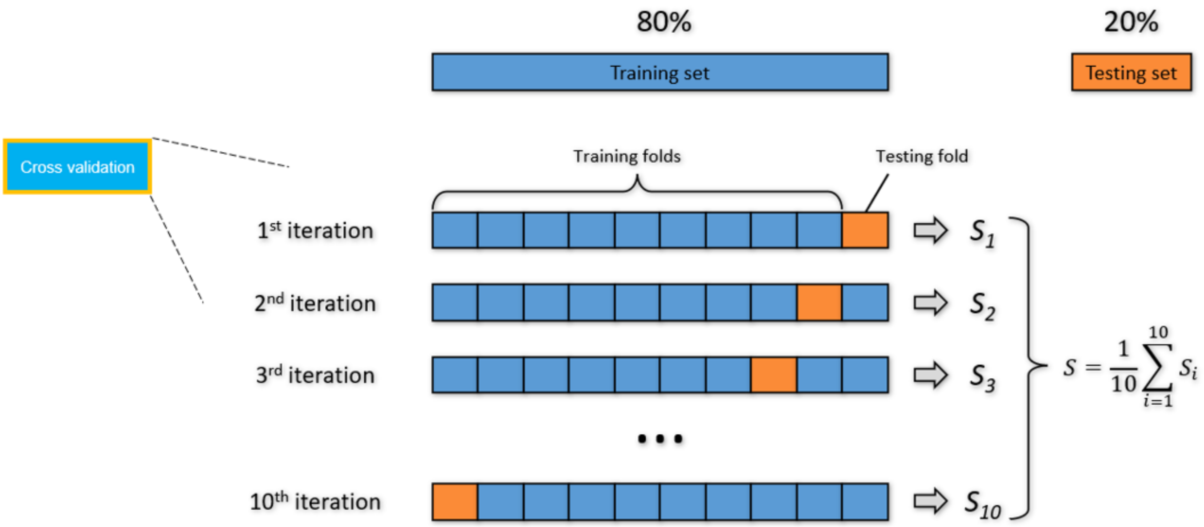


Figure 3.3: Illustration of cross-validation process. 10-fold cross-validation process was adopted in the machine learning brittleness index prediction geoanalytical model.

To further reduce biases, geographical and statistical debiasing protocols are implemented as a condition for fair train-test splits. Figure 3.4 and Figure 3.5 below show an example of post declustering check, what training and testing data split looks like geographically and statistically. The purpose is to ensure training and testing data are not accidentally selected in a clustered geographical area because that could mean a biased sampling of data in similar lithology or geological provenance. Statistically, we need to ensure the models are trained and tested in all data ranges, and not undertrained or undertested in a certain data band.

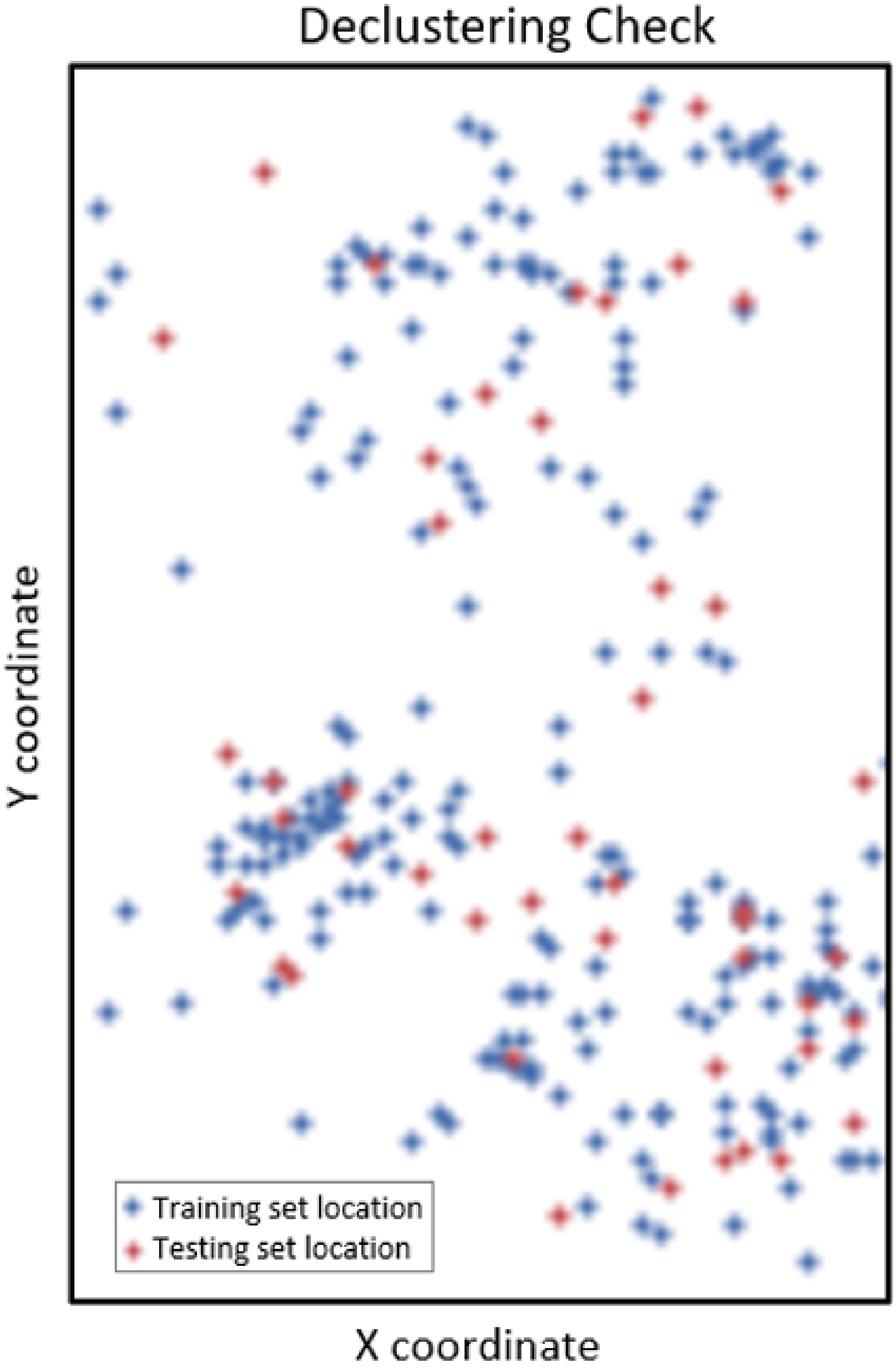


Figure 3.4: Illustration of geographical declustering check for the test and train data splits. Blue crosses are the X-Y location of the training data set and red crosses are the X-Y location of the blind testing data set. Longitude and Latitude information are anonymized and only a subset of the data points are shown due to IP concerns.

Generally, practitioners would like the test and train set displays a similar probability density function across the measurement ranges and no testing data set is outside of training data ranges. Figure 3.5 below shows the histogram of training and testing brittleness index data.

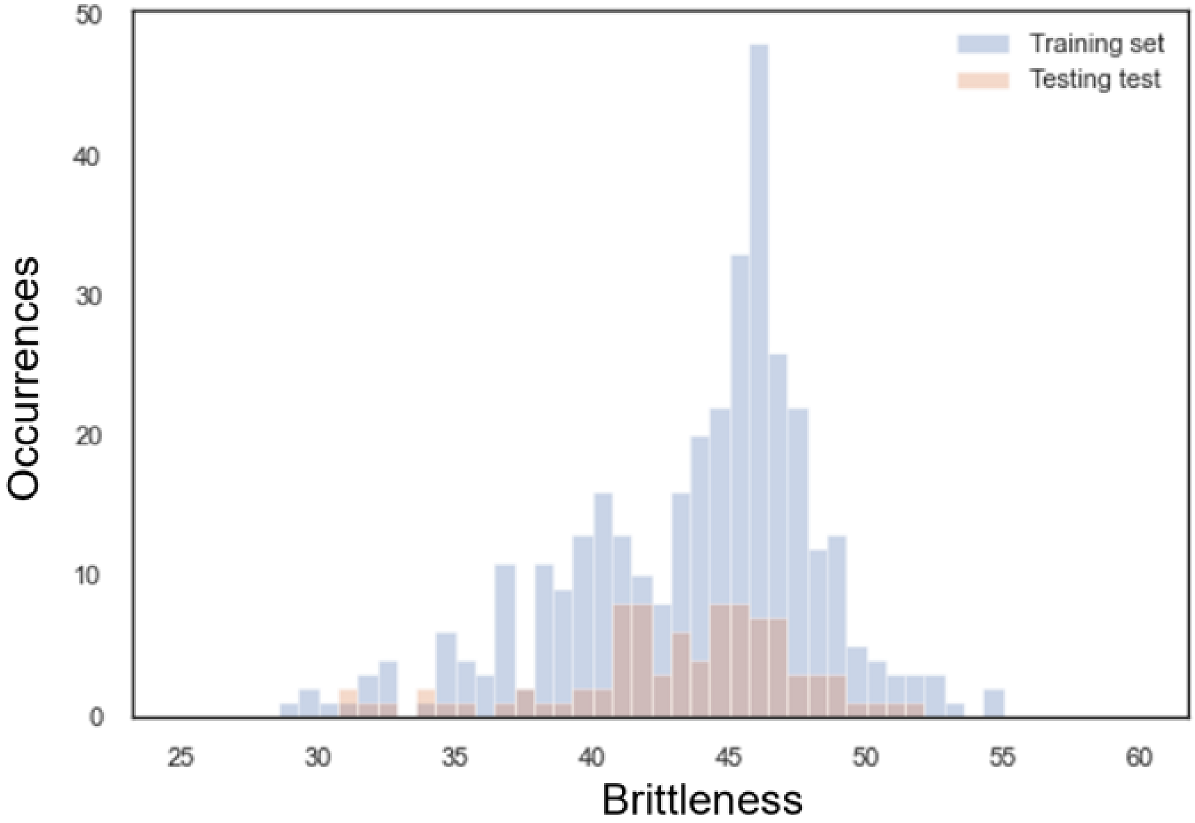
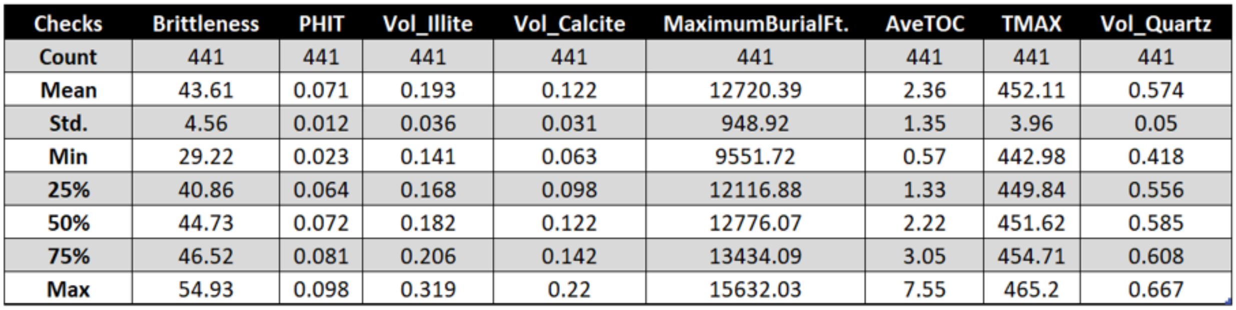


Figure 3.5: Illustration of data statistical distribution and declustering check for the test and train data splits. Blue bars are the histogram of the training data set and red bars are the histograms of the blind testing data set.

Depending on the application of the machine learning model, additional requirements could be implemented if there are additional declustering concerns. For example, if the geology is relatively similar, declustering could require a certain separation of mineralogy or burial depths. For our pilot study, there is enough data coverage and geology variance so that geographical and statistical declustering was sufficient for an unbiased train-test data split.

Multipoint statistical checks are performed on the large data array during the EDA process. Some example checks include valid data counts, mean, standard deviation, minimum, maximum, P25, P50, P75, and the type of data. This helps practitioners to identify inconsistencies and problematic data. Figure 3.6 below shows a summary of this check on the final 7 features selected for the machine learning model after the iterative EDA process. They are target variable Brittleness Index, Porosity from petrophysical logs (PHIT), illite volume (Vol_Illite), calcite volume (Vol_Calcite), quartz volume (Vol_Quartz), average total organic carbon content (AveTOC), pyrolysis measured max temperature (TMAX), and maximum burial depth (MaximumBurialFt).



Checks	Brittleness	PHIT	Vol_Illite	Vol_Calcite	MaximumBurialFt.	AveTOC	TMAX	Vol_Quartz
Count	441	441	441	441	441	441	441	441
Mean	43.61	0.071	0.193	0.122	12720.39	2.36	452.11	0.574
Std.	4.56	0.012	0.036	0.031	948.92	1.35	3.96	0.05
Min	29.22	0.023	0.141	0.063	9551.72	0.57	442.98	0.418
25%	40.86	0.064	0.168	0.098	12116.88	1.33	449.84	0.556
50%	44.73	0.072	0.182	0.122	12776.07	2.22	451.62	0.585
75%	46.52	0.081	0.206	0.142	13434.09	3.05	454.71	0.608
Max	54.93	0.098	0.319	0.22	15632.03	7.55	465.2	0.667

Figure 3.6: 8-point statistical checks for selected features and target variable Brittleness Index. These included total data count, mean, standard deviation, minimum value, 25% percentile value, 50% percentile value, 75% percentile value, and maximum value of the data sets.

Accompanied by the statistical checks are visualizations. This includes boxplots or violin plots that help the practitioners to visualize data distribution and kurtosis. Figure 3.7 below shows the violin chart of the 7 final selected features and the target variable - brittleness index. In this plot, we can see that the white dots and envelopes resemble the data density distribution. They are placed on a normalized vertical scale so they can be displayed side by side. This illustration allows practitioners to quickly identify data issues such as their bimodal or multimodal nature. We can also get a hint on the general correlation or anticorrelation of these features based on their shapes. For example, quartz volume showed a similar shape to that of the porosity and appeared to be opposite in shape to the volume of Illite. This visualization could highlight outliers and apparent correlations between features selected. This is an effective screening tool for domain experts during the EDA process.

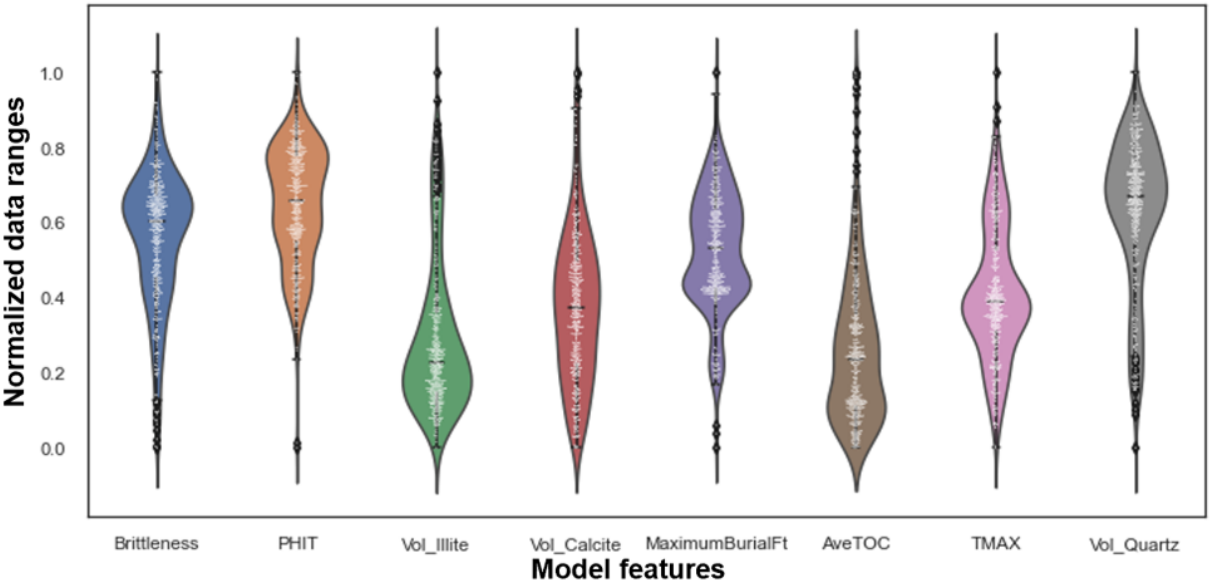


Figure 3.7: Violin plots of the selected machine learning features and target variable, the brittleness index. White dots are a random spread of input data.

Collinearity assessments are essential to maintain the balance of data science model input. As the 2x2 multicollinearity correlation matrix shown in Figure 3.8 below, we observe that Vol_vQuartz to Brittleness Index, TMAX to MaximumBurialFt, MaximumBurialFt to Brittleness Index, TMAX to Brittleness Index, AveTOC to Vol_illite, Vol_Quartz to MaximumBurialFt, Vol_Quartz to TMAX are positively correlated as they show > 0.4 in Pearson correlation coefficient. Vol_Quartz to Vol_illite, Vol_Calcite to PHIT, Vold_Quartz to AveTOC, TMAX to AveTOC, Vol_Illite to Brittleness Index, AveTOC to Brittleness Index, Vol_Quartz to Vol_Calcite, MaximumBurialFt to Vol_Illite, and AveTOC to Maximum-

BurialFt are negatively correlated as they show < -0.4 in Pearson correlation coefficient. Some of these relationships make physical sense such as the correlation between Vol_Quartz to Brittleness Index and Vol_Illite to AveTOC because increased quartz content leads to higher rock brittleness and higher total organic carbon content is largely associated with shaley Illite rocks. Others may be just apparent or carry second-order correlations that may require further investigation.

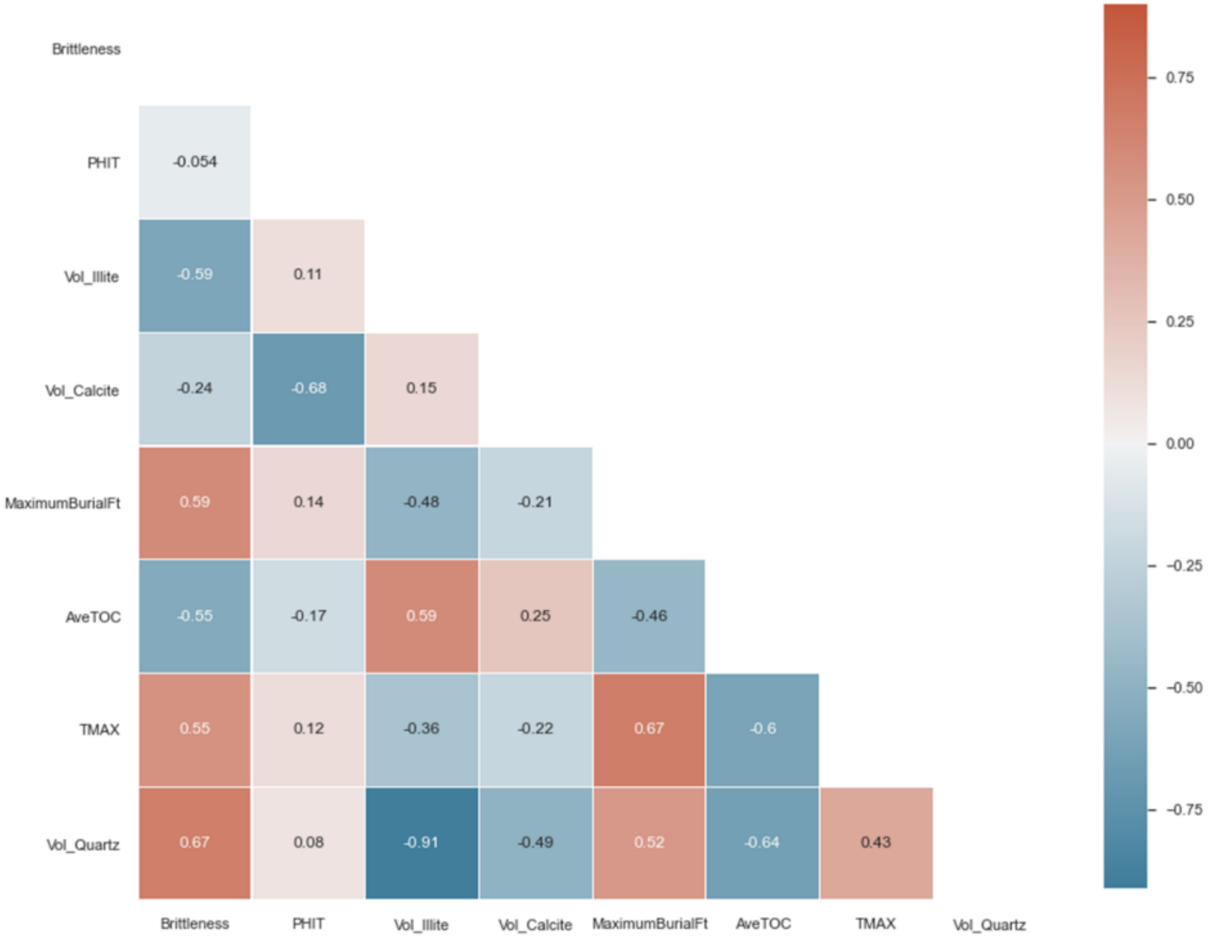


Figure 3.8: Correlation matrix of selected model features during the collinearity assessment. Positively correlated features are shared in orange and negatively correlated features are shared in teal. Higher color saturation represents a higher correlation.

The multicollinearity correlation matrix provides useful information for domain experts to ensure model parameters are selected with justifiable domain evidence during EDA and initial feature selection. However, in addition to knowing feature cross-correlations, understanding the feature importance in predicting brittleness in combined and isolated forms is needed. Additional protocols for domain expert model quality control are described in more detail in the result and discussion section of this thesis chapter that addresses this concern. In particular, this pilot adopted tools such as feature importance score which look at the relative and combined importance of the features in predicting brittleness in a machine-learning model and the partial dependency analysis which looks at the impact of each feature on the prediction of mean brittleness distribution in a machine-learning model.

3.4 Results and Discussion

A feature importance plot is a ranked-based plot that documents in relative terms, how important each of the features used in the machine learning model contribute to the overall model prediction of the target variable. The feature importance plot helps practitioners to visually identify the most important model features that yielded significant influence in the model's predictions. This information can guide the decision-makers in deciding which features to include and exclude. This is also done in tandem with collinearity assessment and model performance evaluation during the feature engineering process. A good sign for a stable model is that the feature incorporates data from different type of measurement sources and location, and the importance ranking and relative importance remains relatively stable. Error bars for each feature are sometimes overlaid when feature importance is run multiple times while the order is still sorted by the main relative importance.

For the pilot presented, the brittleness prediction model is shown in Figure 3.9 below. It is observed that temporal information-bearing data, in red boxes, such as rock pyrolysis data, TOC, and maximum burial depth shared a significant amount of model contribution in addition to mineralogical parameters such as Quartz volume. It is important to point out that these data are not typical data to be integrated into brittleness index calculations. We demonstrate that the integrated basin model and geoanalytics system facilitated cross-function data products and insights integration. And demonstrated for the first time, an integrated workflow incorporating geochemistry and basin modeling insights in brittleness analysis. This integration improves the model's interpretability and reduces overall uncertainties. The scope of the current study is limited to identifying the geological controls for the Brittleness Index occurrences and meaningfully extrapolating away from well control points. Future research could leverage learnings from this assessment in providing additional data for brittleness index calculations.

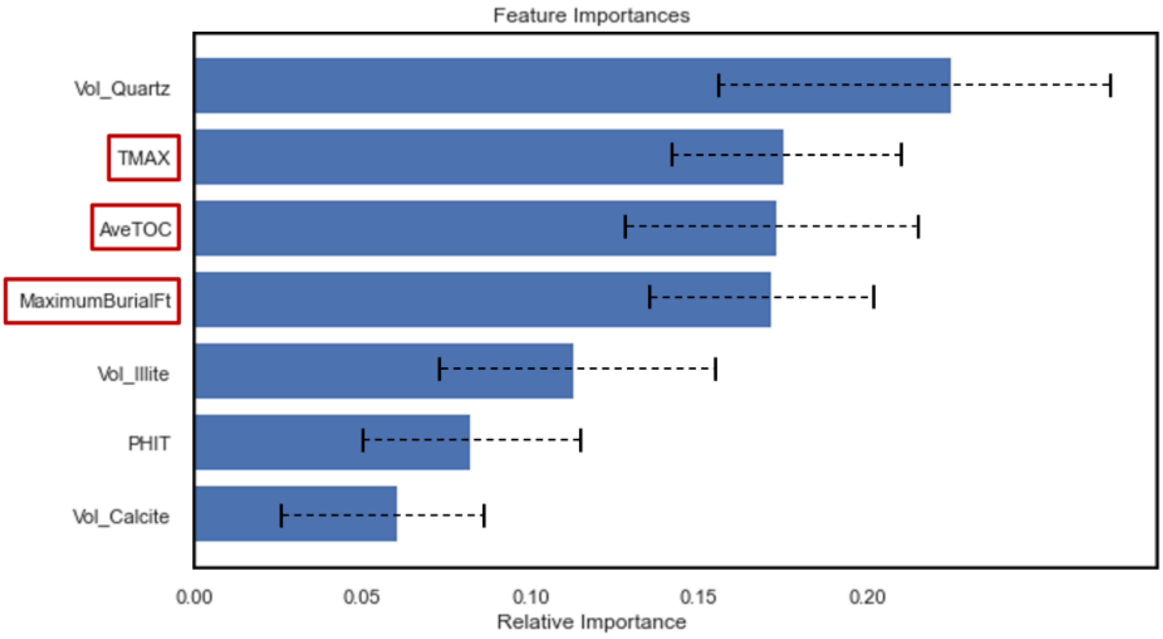


Figure 3.9: Feature important plot for selected machine learning features in predicting brittleness index. Highlighted in red boxes are features from the integrated basin model, not typically available for geomechanical property assessments. Error bars are illustrations of probabilistic variations.

Feature importance analysis gives important insights, but it has limitations in identifying issues due to correlated features, biases, nonlinear relationships, and often limited explanatory power. Additional protocols are required for domain experts to investigate and validate the model parameters. To investigate and validate the features selected for the brittleness prediction model, partial dependency assessments were performed[104]. Partial dependence plots (PDP) shown in Figure 3.10 below offers a visual representation of the relationship between model features and to target variable, the brittleness index. PDP shows how the predicted outcome changes as the feature value varies while holding other features constant, allowing the practitioners to investigate the nature and direction of the correlating relationships. PDP also allows the practitioner to identify nonlinear relationships visually. Monotonic relationships are preferred in the explanatory power of the model[105], [106]. Non-monotonic relationships may indicate indirect relationships or multiple factors at play. This gives the practitioners useful information to further peel back the model and data during the feature engineering process. One most important advantage of the PDP is that it allows domain expert feature validation. Domain experts are able to interrogate the model and assess whether the model aligns with domain expectations. For example, in the machine learning model, lower porosity (PHIT) is correlated with increased brittleness; higher quartz content is associated with higher brittleness; higher maximum burial depth is also associated with higher brittleness. Domain experts can use domain knowledge to gauge whether the machine learning model bears relationships that are consistent with domain expectations. This tool helps to reduce biases and collinearity-driven overfitting.

Another example of the validation outcome is that, in consistency with domain expectation, the brittleness index data distribution is not correlative with the present-day burial depth. This is because there has been a large degree of exhumation and erosion in the basin of investigation. Present-day geometry no longer represents the diagenesis condition the rock experienced. It may be the case for basins with limited structural alteration, however, in the basin of interest, it is observed that the brittleness index data is correlative to the maximum burial depth instead, which is only available post-structural restoration and model simulation. This observation gives additional confidence in the characterization of brittleness distribution in the formation of interest in the study area at the same time validifies the value of assessment through a holistic approach.

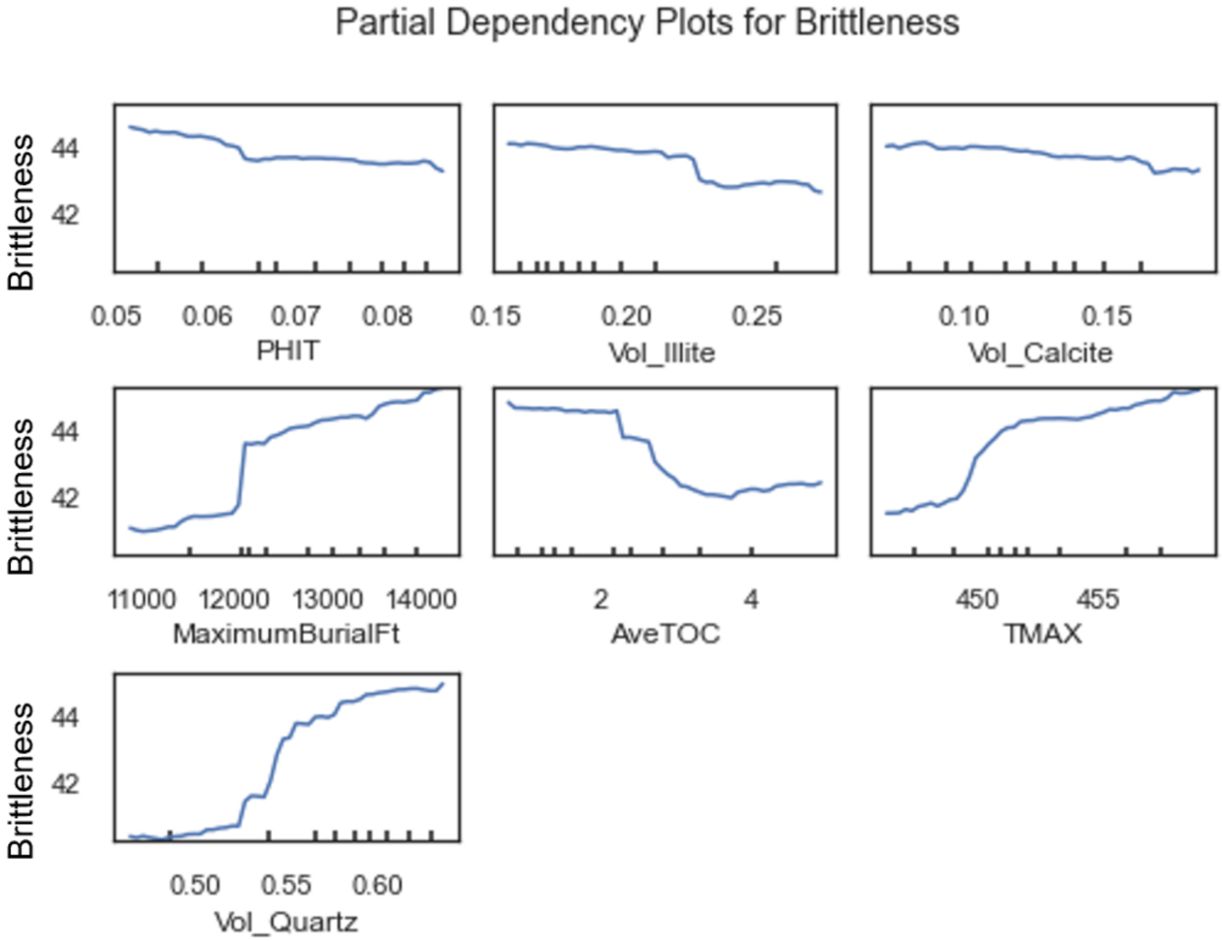


Figure 3.10: Partial dependence plots of selected features for brittleness index. This is an illustration of how the brittleness index changes when changing only one feature while keeping other features constant.

Features finalized should demonstrate a stable partial dependency relationship to the target variable. In addition to verifying the feature importance and the nature of fit against expectations, additional insights are gained from the PDP. For example, it is interesting to observe that there appears to be a sharp change in rock brittleness around 2% TOC value. Additional investigation led to lithofacies interpretation and validated that different depositional settings resulted in different types of rocks being deposited. It is also observed that there appears to be a sharp change in brittleness at around 12000 feet of maximum burial and around 445C of pyrolysis testing temperature. These observations could represent geological conditions in that hydrocarbon started to expel from kerogens and quartz cement is starting to grow where chemistry allowed. Domain-driven relationships integrated assessment and the capability to validate the statistical relationship gives additional confidence in the data model. Diagnostic tools with a high level of transparency warrant domain experts' interpretation integrity.

The blind testing dataset is used to validate the machine learning model. The difference between model prediction and blind test data is documented as modeling errors. Error analysis is another important model QC protocol. Figure 3.11 below shows how we can use error analysis effectively to help ensure robust machine learning models. The figure to the left shows the scatter plot of machine learning predicted brittleness index value against the raw blind test data. The blue line represents the model fit while the red line is the 1:1 line. We observe that the model is not biased or skewed as the model fit aligns with the 1:1 line closely. Uncertainties are shaded by the light blue shade. Data density is reflected by the histogram to the top and the right of the left figure.

The figure in the middle denotes the absolute error between predicted brittleness index values and raw blind test data. The unskewed model exhibits a normal error distribution around "0", which is represented by the dashed black line and histogram to the right of the middle figure. An overly small error may indicate overfitting resulted by hyperparameter tuning and feature engineering processes given the scale of investigation and uncertainties embedded in the data. An overly large amount of error may indicate that the model quality is insufficient to explain all the data variations. This could mean the model is underconstrained, the subsurface is mischaracterized, or more information or features are required.

In addition to absolute errors, relative error as shown in the figure to the right, allows us to investigate the model error distribution across different ranges of brittleness ranges. This is important because a small amount of error on the absolute term may be extensive in relative terms if the measured data is small. Relative error plots allow us to examine the error distribution, in relation to the magnitude of data measurements, across all data ranges so that the model is not biased or exhibit disuniformity in uncertainty distribution. This test provides important feedback to feature engineering and machine learning iterations. Error analysis is an important quality check and assurance step for a robust machine-learning-based geoanalytical model.

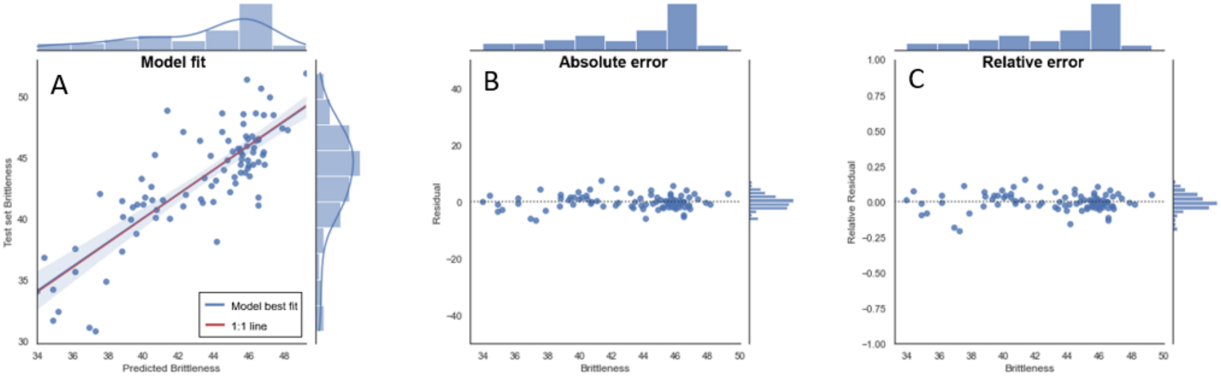


Figure 3.11: Illustrations of modeling error analysis. A: model best fit versus 1:1 data fit line. The red line is the 1:1 line and the blue line is the model fit. B: Absolute error distribution. Histograms are shown at the top and right of the figure. C: Relative error distribution. Histograms are shown at the top and right of the figure.

In addition to error analysis, multiple statistical tests and checks are performed to warrant a robust machine-learning model. There are some common metrics such as Pearson correlation coefficient R and the coefficient of determination R2. Data scientists also often check the performance of errors such as the normalized mean absolute error (NMAE), the normalized root mean square error (NRMSE), the Kendall Tau-b rank correlation coefficient, the Spearman correlation coefficient, or their variants. Different tests may be more diagnostic than others depending on the specific circumstances. Some metrics are more sensitive to outliers while some are more tolerant because some methods are cardinal and some are ordinal, or some methods may limit to linear relationships[106]–[109]. The goal here is not to demonstrate an exhausted list of train-test quality check metrics but to highlight the importance of checking more than one type of model train-test quality check metric. This train-test model QC protocol shows how well the model performs predicting against the training dataset and blind testing dataset. This gives information on the amount of data scatter and how well the model utilizes the training set and how well the model is trained. The process gives practitioners an opportunity to question the validity of the model and results.

Figure 3.12 below indicates that the performance of the testing dataset is lower than the scores for the training datasets. This is because first, there is some degree of data scatter and uncertainties; second, there are compromises taken during feature engineering and hyperparameter tuning process; and lastly, the machine learning model has never seen the blind testing data. Ideally, the testing model performance can be as high as the training dataset. However, this is more achievable in high-fidelity data and data-rich applications.

As Figure 3.12 shows below, the train and test scores for all tests are satisfactory (greater than 0.6) with the exception of the Kendall Tau-b test. This is because the Kendall Tau-b test is very sensitive to outliers. This is identified by running this test multiple times while investigating the training and testing data splits and statistical distribution.

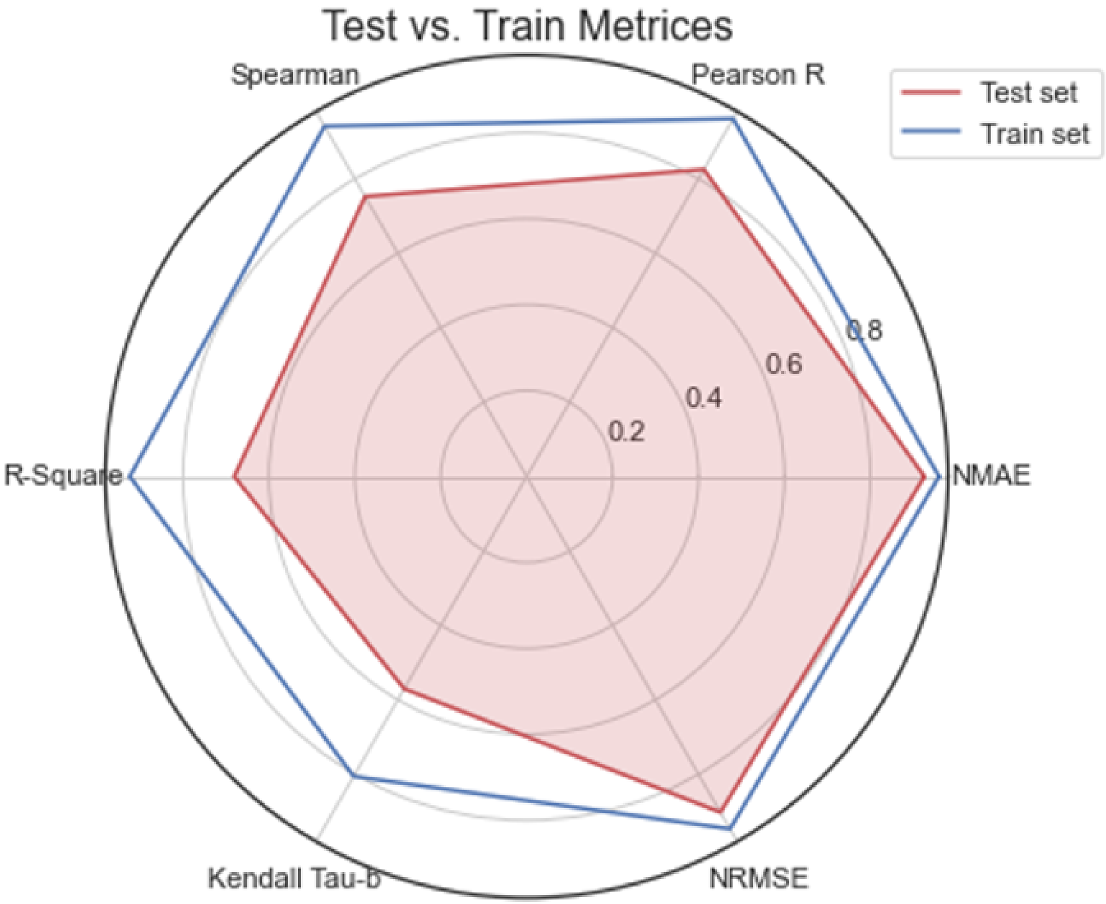


Figure 3.12: Spider diagram of statistical tests for the training and blind testing data set. Blue lines are generated from testing scores for the training set. Red lines are generated from the testing scores of the blind test set.

The model is considered final when all testing protocols are followed and thoroughly investigated. Before applying this data model to the entire area of interest to generate a grid or data volume, high-level domain correlation checks are exercised as due diligence. In this application, the features selected are consistent with the domain expectation that the depositional mineralogy and post-depositional mineralogical alteration are the main drivers for the brittleness index observations. Data visualizations shown in Figure 3.13 below demonstrated that the correlation of the brittleness index could be improved by considering maximum compaction parameters in addition to considering the mineralogy calculations solely. Although the mineralogy data measured present day may include alteration impacts such as quartz and calcite growths, this additional information still gave the model new information for a more inclusive and informed characterization.

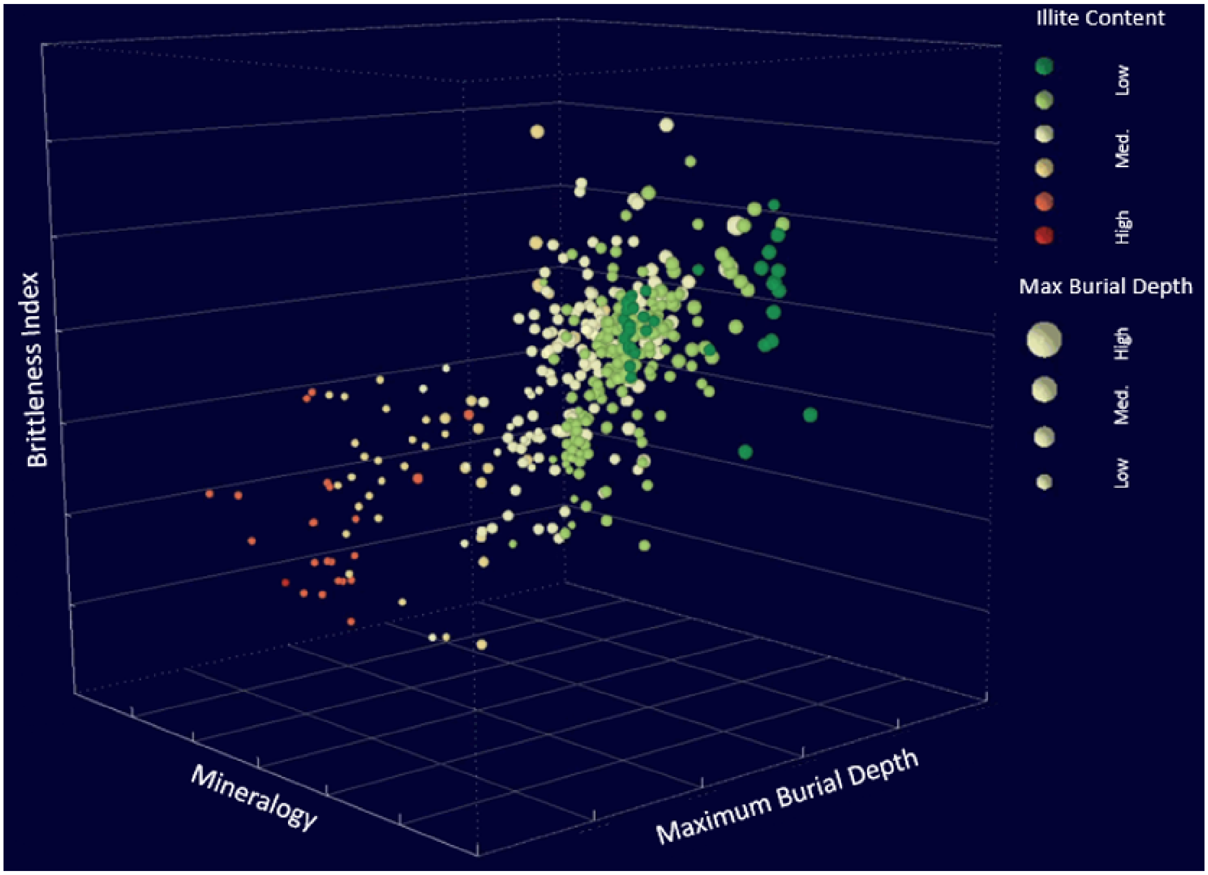


Figure 3.13: Visualization of the Brittleness Index calculations as a function of mineralogy and maximum burial depths. The red color indicates high illite content. Smaller bubble size indicates a lower maximum burial or an indication of a lower degree of diagenesis.

It is challenging to visualize higher dimensional machine learning model fits as there are limits to human visualization and cognition. After applying the machine learning model to the entire area of interest or data volume, the final machine learning-based geoanalytics product is cross-plotted to visualize the improvement instead. Brittleness Index aerial distribution while not shown here, is validated against other interpretations that are not used as the machine learning model features for validation purposes. Figure 3.14 below shows the final brittleness index geoanalytics modeling product using seven selected features previously documented when applied to a formation of interest after thorough validation.

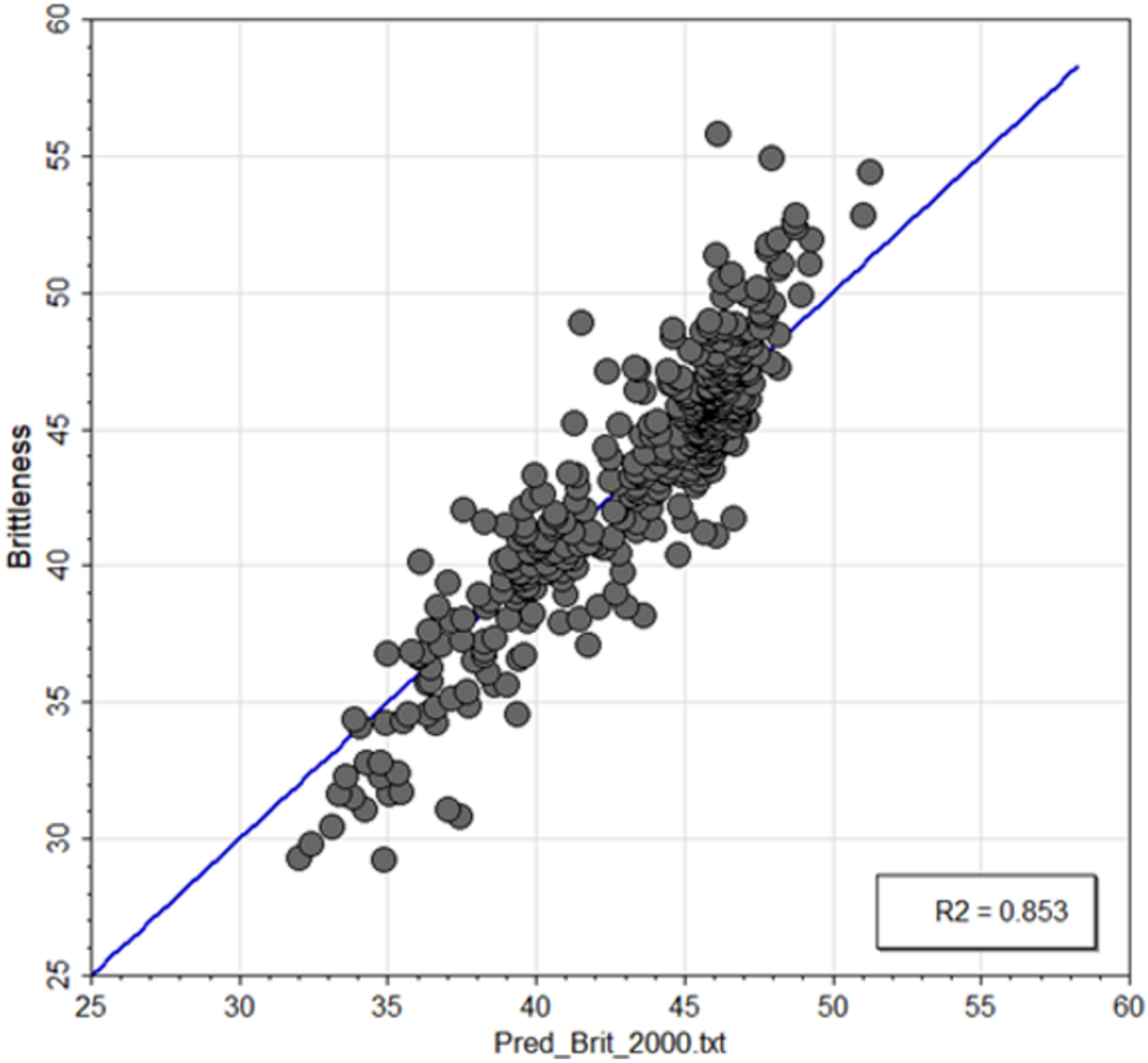


Figure 3.14: Cross-plot of machine-learning geoanalytics model predicted brittleness index and input brittleness index data for the entire dataset. Blue line is the 1:1 line.

It is observed that the brittleness index products are geologically pertinent and demonstrate great characterization fidelity after validating against other independent basin interpretations. Random training and testing split datasets may result in the slight skew of best fit away from the 1:1 line in addition to a tendency to fit data best in the central quantiles or mean using random forest, however. As a result, products sometimes may display narrower ranges than actual measurements. Interval rescaling may be required to achieve a better overall model fit for all data ranges.

This brings data extrapolation uncertainty into the picture. As Figure 3.15 shows, after the data model is applied to the entire area of interest or data volume, the data distribution of the area of interest or the data volume may be very different from the data distribution of the data involved in the machine learning workflow. That is there may be values that are outside of measured ranges for each of the features selected. These data are model extrapolations, which means the model has never been "taught" what to expect for those data ranges. This is because the data measurements may not be evenly sampled within the basin. Denoted under a red bracket in Figure 3.15 below, there are some model outputs that are beyond data ranges of training data set. This means the model is extrapolated to data the model has never seen before. Practitioners need to be cognizant of those areas from each of the input features and exclude them from the model output as necessary to minimize the impact on model extrapolation uncertainty.

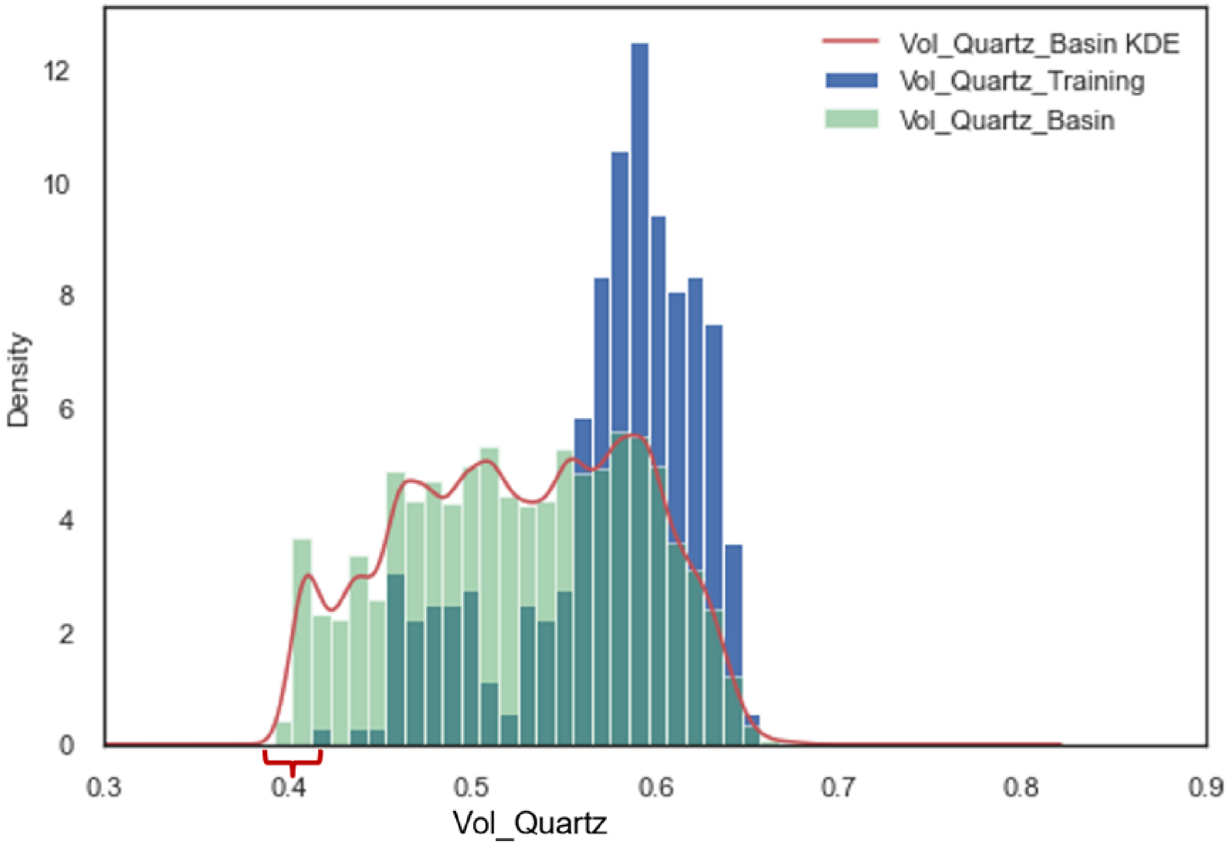


Figure 3.15: Histogram of data ranges the model is trained for and what are actual data ranges extracted from the digital basin model. The lime-green bars are data occurrences for Quartz volume for intervals of interest in the entire area of interest. However, the machine learning model only was trained to a smaller data range shaded by blue bars. Red brackets illustrate the data ranges that the machine learning algorithm has never seen before. Predictions in those data ranges are model extrapolations.

In summary, the machine learning process was successfully applied in geoanalytics. The system of an integrated basin model and geoanalytics is agnostic in the type of analytical relationships, which means it can be domain-driven analytical relationships, empirical relationships, or advanced data modeling methods. The system is also agnostic in the type of machine learning methods, meaning it is not limited to Random Forest[31], Extreme Gradient Boosting (Xgboost)[110], or Neural Network[111]. The system provides a platform that encourages practitioners to bring a balance between domain-based analytical relationships and statistic-based analytical relationships. The fidelity of modeling output hinges on this delicate balance which is facilitated by processes and protocols designed for their purposes. The system design allows domain experts to peel back every layer of embedded insights and investigate each of the modeling processes to allow verification and validation that is otherwise not possible.

The platform also impacts the sociotechnical aspect of a complex organization. Implementation is key for a successful digital transformation. Training employees in masses will be much less efficient in bringing the system to the users. We will discuss the implementation strategy and its impact on the sociotechnical system of a complex organization in Chapter 4.

Chapter 4

Emergences, Uncertainties, and Future Directions

4.1 Introduction

Value is benefit at cost. Delivering maximum value means delivering maximum benefit at minimal cost[38]. Chapter 4 focuses on the benefit delivered by the value-creating functions and the cost minimized through its form and implementation strategy. The proposed system includes a set of components and the relationships among the components. Emergence occurs when the functionality of the system is greater than the sum of the functionalities of the individual components considered separately. Chapter 4 also details why the total functionality is greater than the sum of the individual components for the system proposed through a detailed discussion of emergence and "ilities" as the system operates.

A famous example of an emergent function is the emergent function of sand and a funnel is timekeeping in the form of an hourglass. The value function is a very concentrated summary of the component and stakeholder analysis of the system. A successful enterprise-level digital transformation would be challenging without an accompanying platform tailored to each of future strategies and ways of work. Below are a few emergent functions of the subsurface digital twin.

4.2 Emergent Functions

4.2.1 Cross-function data modeling products

Chapter 3 has demonstrated a successful cross-function pilot. Where with the help of a digital basin model, basin-scaled data modeling exercises could be performed to meet business needs. In addition to what was documented, using suitable model training and testing protocols, cross-functional data can be made into property grids or information volumes. Expanding on what was previously discussed, all the features selected for the machine learning model documented in Chapter 3 were products of the digital basin model and geoanalytics. These data modeling products through geoanalytics bear great geological pertinence. Mineralogical parameters are often correlative to the paleo-water depths with the exception of during some erratic periods of transgressive stands. Geochemical parameters such as TMAX and AveTOC were found to be correlative to thermal and pressure history and organofacies. It is important to note that different basins, or even different stratigraphical units in the same basin may exhibit different correlations to the same or different geological controls that are unique to the system of interest. The subsurface digital model or the digital basin model provides practitioners with a platform for cross-functional modeling work.

Other cross-function data modeling products such as more geochemistry, stratigraphy, geomechanics, geophysics, and even production-related data could be generated through geoanalytics. The more data modeling is done the more complete the subsurface digital model becomes. Due to the reciprocal nature of the system components, the more cross-function data modeling is done the better physics-based simulation results are expected.

4.2.2 Multivariate optimization

Multivariate optimization applications or multi-attribute utility applications in decision theory can be facilitated by this evergreen subsurface digital model that is maintained by a cross-functional team. The subsurface digital model brings a rich suite of insights for decision-makers.

An example optimization problem could be to identify an ideal location or business plan for a blue hydrogen project. Blue hydrogen is categorized as sustainable hydrogen generation through reforming, but instead of emitting the byproduct carbon dioxide, the CO2 is captured and geologically sequestered[112], [113]. A robust business plan for a blue hydrogen plant at scale requires insights not only into market insights, costs and rights for land and facilities, and proximity to existing infrastructures, but it is also ideal to be optimized with subsurface insights. These subsurface parameters could include the availability and quality of hydrocarbon, CO2 sequestration reservoir availability and storage potential, and reservoir storage and contamination risks. If there is a hydrogen storage requirement, subsurface storage could become a viable option compared to surface options. The same concept can be applied to carbon sequestration projects or any other investment decisions. The subsurface digital model should be populated with relevant properties through physics-based simulation and geoanalytical data modeling under one system. The presence of a subsurface digital model would bring more options to decision-makers. This also ties back to the enterprise digital twin concept illustrated as the L0 system architecture diagram (Figure 2.3 discussed in Chapter 2, where an evergreen information system is in place for subsurface and surface insights. This leads to the next emergent properties documented below.

4.2.3 Real-time visualization and decision making

Smart visualization approaches can be implemented on top of the subsurface digital model. As Figure 5 shows above, visualization resides on the top layer - the Decision and Interaction Layer. The Decision and Interaction Layer operates on a structured information volume, which is the outcome of the subsurface digital model or digital twin. This architecture is advantageous because it avoids complex interface interactions with different software vendors. This evergreen subsurface information volume keeps a record of the most recent and most informed subsurface parameters and allows real-time data query, analysis, visualization, and reporting.

Commercial companies from various industries practice data trade and data-sharing sessions on a regular basis with partners and consortiums. When subsurface understanding is required, it is valuable to know the quality and geological fidelity of data presented during those data evaluation sessions. In addition to that, due to the reciprocal nature of the system, it is important to know what data the subsurface digital model is lacking and how the acquisition of a new data type can help improve the subsurface digital model. Acquiring data that will improve the subsurface digital model would bear additional value because the improved model would inform decisions for all applications across the enterprise instead of only serving the occasion and only for a specific project team.

4.2.4 Cross-function collaboration

Companies that identify the importance of human capital put forth policies and processes that would promote an improved employee experience and overall employee well-being. A successful organization often recognizes that humans are at the center of the enterprise and emphasize their needs, capabilities, and experiences while meeting the expectations of shareholders or a healthy bottom line. This is in line with Industry 5.0 – human-centric enterprise design expectations where work is more integrated and at the same time more personalized. The platform proposed facilitates collaboration, integration, personalization, and co-creation. In other words, this meets society's expectations and demands for the energy industry, the evolving relationship between humans and machines, and the ways of organizing work.

The examples documented in this thesis demonstrated how cross-function subject matter experts are able to share data and insights across functional teams effectively. The system utility does not end here. The system also facilitates conversation across the value chain in a business unit, enabling regional teams to connect findings from field geologists and engineers. Teams from similar geological conditions or play types as well as teams from different asset classes could all share learns and workflows openly and securely.

4.2.5 Performance improvement

Enterprise performance is a determent factor for corporate competitiveness. Must win technologies should help us to attain better products quicker. The proposed system allows decisions to be made holistically leveraged by cross-functional integration. At the same time ensuring better product fidelity, cycle time could also be significantly reduced due to reduced rework and recycled material. The cycle time for an exploration decision spans years to decades, in which subsurface characterization takes up a large chunk of that time depending on the complexity and the pace of new data acquisition. Metaphorically speaking, instead of taking multiple "photographs" of an object or multiple "photographs" of different parts of an object without context trying to piece them together and make sense of it, the system promotes a new mindset of being agile. In other words, the insight comes initially from a "blurry photograph." It works toward partial clarity with new data and models in reaching an ever-better understanding of the subsurface. All insights and lessons learned are shared and accumulated with time. At the end state, what is maintained is a multi-resolution "photograph" that has higher resolution where data and interpretation permit. All simulation and data modeling follow the same underlying geological assumptions and a coherent geological story. This holistic approach[114] provides us with an alternative to "the blind men and the elephant" approach in reaching the best approximation to the "truth" quicker. It is recommended that an evergreen subsurface digital model is maintained for every basin.

4.2.6 Cost reduction

Delivering maximum value means delivering maximum benefit at minimal cost. In addition to the cost reduction from improved efficiency and performance, the platform also helps reduce data acquisition and maintenance costs. Data acquisitions and management can be costly for projects. It is a known issue that companies may incur costs in acquiring the same data or different vintage of the same data multiple times with poor data management strategies. This is more pronounced when there are internal organizational barriers and silos where data and insights are often not shared or managed systematically. The subsurface digital model and the supporting data foundational layer could effectively minimize this waste.

There have also been high costs associated with maintenance and rework. Multiple vintages of data and different project teams resulted from multiple versions of interpretations. These reworks could be eliminated when a centralized repository is in place for all. In some extreme cases, internal inconsistencies and conflicting information have been reported. This often causes more rework and delays in a decision. There may be capital costs to the construction of a subsurface digital model or digital twin at first, however, in the long run, significant savings can be realized.

4.3 Ilities

Ilities are systems attributes that emerge as systems operate. Crawley et al.[38] documented 15 “Ilities,” or system attributes related to its implementation and operations compiled over time by tracking mentions in journal articles from the year 1884. These include Quality, Reliability, Safety, Flexibility, Robustness, Durability, Scalability, Adaptability, Usability, Interoperability, Sustainability, Maintainability, Testability, Modularity, and Resilience[115].

While various "ilities" are relevant and crucial in architectural design, the presence of a single "ility" alone is not enough to define the system's architectural uniqueness. The distinctiveness of an architectural project arises from the careful consideration and combination of multiple design elements and features as we documented previously. In addition to that, not every quality or characteristic of the system is unique or exceptional enough to set it apart from other systems. Some "ilities" may be common or shared among many digital systems and may not significantly contribute to the uniqueness or distinctiveness of a particular architectural design. Even if a certain quality or characteristic is considered valuable or important in this architecture, it doesn't guarantee that it will always make the system architecturally distinctive. Different subsurface digital systems might prioritize different "ilities" based on their specific needs and geological characteristics, and the combination of multiple factors contributes to the overall architectural identity of a digital subsurface system.

“Ilities” are also often interconnected[38]. In the context of the system architecture of the subsurface digital twin, higher flexibility also leads to higher usability and ultimately, product quality. Three selected system “ilities” are highlighted below in addition to improved quality that is worth mentioning.

4.3.1 Adaptability

One design objective of the subsurface digital twin is to have a high degree of adaptability. This is to be achieved on two fronts. The first is to be globally inclusive. Global inclusivity means the system should have a great format and error tolerance. Specifically, the system should be able to take in data and insights at all common formats as input. This allows practitioners to focus on developing the model instead of worrying about data formats from different software vendors. There are many reasons for this. There are different data standards currently, and many are pioneered by current software vendors. These resulting output files are problematic to be used directly by tools outside of the recommended operating environment or specific software suites. In order to integrate insights from different software vendors, exporting model output to an open-source or open-domain file format is required. There could also be some resampling and approximation done due to possible different grid resolutions, different grid origins, different definitions of the location of extraction of a grid cell, or different types of grid meshes. This often happens, especially when operations and geoanalytics are required to operate on outputs from software vendors who use triangular or tetrahedral meshes and vendors who use rectangular and cubic meshes. Additional reasons could be as simple as saving, exporting, georeferencing, and coordinate system transformation. This is important because the architectural design of the file system has to recognize all the specific use cases to avoid undesired system emergence or emergency. The system has to be designed in a way where the negative impacts of the capability of integrating all inputs and being globally inclusive are minimized. This calls for the design to be brought to the user, preventing users from deviating from the original data due to repeated operations from the design level.

The second aspect of adaptability is to have the capability to leverage this system to address questions in a wide range of use cases or geological conditions. This system should be applicable not only in regions with different data densities but also has to be translatable in other asset classes or geological conditions. The subsurface digital twin is designed to be inclusive and data-type agnostic. This means that in data-dense plays such as mature development areas in unconventional plays, higher-resolution data types such as petrophysical logs, rock cuttings, and cores can be used to help generate higher-resolution insights. In the data-poor frontier exploration areas, the digital basin model can be constructed with coarser resolution data types. All other plays should be in between these two boundary conditions and adopt a hybrid approach in the construction of the digital basin model. Modeling insights should be provided in a format accepted by all and with a minimal level of alteration.

4.3.2 Maintainability

Currently, physics-based models such as the basin models and earth models are maintained by domain experts or their responsible teams. Documentation of all the models constructed is typically insufficient in capturing all the modeling assumptions and limitations. Although many of the models and outputs are maintained, it is often the case that data and models are recycled and reworked when there are organizational changes to the project team. This leads to increased costs in the form of investment and time. The subsurface digital model harnesses the collective wisdom of a team of domain experts. This reduces the impact of any disruption related to maintaining the understanding of the subsurface.

Maintenance may also have dependencies such as the file format and support from external software vendors. Although it may appear to be the best thing to do at the moment, this is not a long-term maintenance solution to preserve a large sum of models. An unquantifiable amount of value could be lost when models can no longer be maintained due to software and hardware changes. Instead of managing hundreds of models individually, an evergreen subsurface digital model provides the most recent interpretation and understanding of the subsurface. The maintenance complexity would be greatly reduced if the subsurface digital model is the sole objective of maintenance.

Adopting the digital twin and an agile approach also provide benefits in providing data-driven maintenance decisions. The subsurface digital twin provides a wealth of data about the system's behavior and performance. This data-driven approach empowers maintenance personnel to make informed decisions based on actual data, leading to more effective maintenance strategies and resource allocation. In addition to that, the digital twin can also store historical data about the system's operation, maintenance activities, and performance over time. Practitioners can access past records and learn from previous experiences, avoiding repeated mistakes.

Agile implementation of the subsurface digital twin also advocates simplicity and avoids unnecessary complexity. By focusing on delivering the minimum viable product and iterating based on feedback, teams are less likely to introduce over-engineered and convoluted solutions. This streamlined approach enhances maintainability.

4.3.3 Resilience

This architecture is advantageous because it avoids complex and sometimes complicated interface interactions among different software vendors. The proposed architecture simplifies the decision and interaction layer to a single interface which is maintained internally. This reduces implementation risks and maintenance risks. The subsurface digital twin also facilitates the agile implementation of project works. By focusing on delivering the minimum viable product and iterating based on feedback, teams are less likely to introduce over-engineered and convoluted solutions. Projects are thereby completed in a shorter timeline with the right amount of details with improved fidelity. The integration of physics-based simulation also allows practitioners to explore the potential impacts of different events and actions on the system. This proactive approach helps to improve resilience by facilitating the development of effective contingency plans, supply chain planning, and optimizing responses.

The digital twin also improves financial resilience by improving reservoir performance and safety. Through continuous monitoring and analysis, inefficiencies and suboptimal performance in the physical system are identified. By making data-driven adjustments and optimizations, the system's overall performance and efficiency can be enhanced. Improved reservoir management in the form of improved expected ultimate recovery (EUR) and improved carbon geological storage capacity and long-term storage integrity are two examples. Different performance metrics can be optimized under the proposed platform.

The digital twin improves resilience through continuous learning and improvement. Insights gained along the way are not lost, the data collected by the digital twin can be used to learn from past events and performances. Lessons learned are incorporated into system improvements, leading to a continuous learning cycle like through the digital basin model and geoanalytics. As the system evolves and becomes more adaptive, the frequency of learning may be reduced, or the focus may shift through the lifecycle of subsurface assets. This will also help in resource optimizing the allocation and identifying potential areas for redundancy within the system. By understanding resource demands and bottlenecks, the system can be designed and operated with a more robust and resilient configuration.

4.4 Future Steps

There are a few important tasks we opted out from detailing in the previous chapters. These topics are summarized in the Future Steps section in Chapter 4 where they are discussed in more detail. These include the uncertainties associated with this subsurface digital repository or the subsurface digital model, evolution from digital model to digital twin, roadblocks and limitations of physics-based simulation, the continuation of expanding the cross-function integration with an open-source omni-resolution gridding formats, and predictive digital twin maintenance and modeling research.

4.4.1 Incorporation of uncertainties

There are many types of uncertainties associated with the integrated bottom-up simulation and top-down data modeling system. Some examples are data input uncertainty, geological interpretation and interpretability uncertainty, numerical uncertainty, sampling bias and uncertainty, boundary condition uncertainty, time and scale uncertainty, data density bias and uncertainty, interpolation uncertainty, extrapolation uncertainty, geostatistical parameter uncertainty, model structure uncertainty, parameter estimation uncertainty, model selection uncertainty, machine learning data split uncertainty, overfitting and underfitting uncertainty, and prediction confidence uncertainty. What is certain is that there are a lot of uncertainties, perhaps even the uncertainties are uncertain. They are all more or less impacting the overall outcome of the process, the process needs to avoid the accumulation of uncertainties in the final product. The system facilitates the reduction in uncertainty by integrating data from other disciplines and modeling higher-level geological controls such as establishing geological frameworks and the construction of a basin model before diving into high-frequency data. To better analyze and document these uncertainties, they can be first categorized into three groups: Input uncertainty, modeling augmented uncertainty, and interpolation and extrapolation uncertainty.

Input uncertainty is widely speaking about the variation of raw input data, these often contain outliers, inherit problems with data sources, measurement errors, and sample collection biases. Data engineering and domain expert data investigation is often done. The subsurface digital model provides a platform for data QC to be completed in higher dimensions with data and insights from other disciplines. Data through system integration is the most effective way to reduce data input uncertainty. Raw input data that does not agree with a geological story backed by the majority of cross-disciplinary datasets are flagged and evaluated. Some of these processes could be proceduralized and automated or semi-automated by domain experts. Problematic data could be repaired or quarantined for improved data source quality or request for verification. Modeling and augmented uncertainty refers to all the uncertainties that are associated with the modeling process, some are due to the limitation of analytical or numerical solutions while some are due to human choices. These can be related to the process model, domain expectation, analog selection, experience, assumptions, interpretation, and numerical and analytical model selection.

This process is where uncertainties could multiply and get out of hand if left unchecked as the second half of Figure 12 shows above. An example is that instead of simulating hydrocarbon generation, migration, and saturation directly, input parameters such as original total organic carbon content, original hydrogen index, source rock thickness, and thermal maturity could be improved from the data modeling process. Currently, hydrocarbon saturation estimation is heavily based on the migration of hydrocarbon, and the migration of hydrocarbon is based on the generation and expulsion of hydrocarbon out of the source rock. Uncertainties related to source rock richness, availability, and maturity are rolled up into uncertainties of hydrocarbon migration related uncertainties such as lithofacies and structural restoration.

Each of these processes is based on a set of assumptions and analog models and therefore bears a large amount of uncertainties. In addition to that, all these uncertainties are then exacerbated by the seal, reservoir, geometry, and charge timing relationships because if the hydrocarbon generation predates the formation of the reservoir, geometry, and reservoir seal, the expelled hydrocarbon will be retained in the petroleum system without special hydrocarbon retention mechanisms such as hoteling or late structure development[4], [116]. As a result, the cumulative uncertainties in volume assessment can span a few orders of magnitude.

Combined data modeling and simulation help to reduce this uncertainty. Before simulating an outcome, the geological model and assumptions are refined through rounds of geoanalytics. These iterations of the data modeling process provide a better most likely case for input parameters for the simulation. This iterative process limits the possible range of uncertainties introduced by augmented uncertainties of human decisions and numerical models. It is more practical in data-rich systems because some data modeling approaches such as machine learning and inverse regression require target variables to be fed into the model.

Interpolation and extrapolation uncertainties are mainly reflected in two categories. One is the estimation/approximation of values away from data control on a 2D grid or a 3D volume. Simply put, this is how to populate data points into a 2D grid or a 3D volume. Depending on the gridding algorithms used there may be multiple ways to populate the data space. Some examples can be numerical approaches such as inverse distance and seed-based gridding and geostatistical approaches such as kriging. The interpolation and extrapolation uncertainties are huge with this approach, especially in the occasion where data are not as dense or are not evenly distributed geographically. Most of these approaches are bullseye prone without further processing procedures while incapable of extending away from edge data. Simulations, operations, and decisions made based on this information are therefore inheriting a large amount of uncertainty that is purely based on the algorithm used.

Geoanalytics facilitated by a digital basin model as Figure 7 shows above provide a higher dimension geological story to guide the interpolation and extrapolation of these properties. Therefore minimizing the algorithm-generated interpolation and extrapolation uncertainty. The second type of interpolation and extrapolation uncertainty refers to the interpolation and extrapolation of model fits or analytical relationships. Interpolation on a data model is due to a lack of data measurements in certain data ranges, but they are guided by analytical solutions or multiple relationships. The extrapolation on a data model means the extension of curve fitting beyond data measurement and calibration. Because there are no data measurements in those areas, practitioners have to be careful in maintaining geological pertinence in those data ranges because most of the analytical relationships established have a finite range of applicability where the same geological and physical controls are at play.

The subsurface digital model brings cross-disciplinary data into the analytical space for practitioners to find better model fits, potentially improve model fit, fill the interpolatable model ranges, and identify the boundary for data model extrapolation as Figure 10 and Figure 24 show in previous chapters. The subsurface digital model also facilitates using physics such as the relationship among pressure, volume, and temperature (PVT) to limit extrapolation ranges. These processes reduce uncertainties related to data model interpolation and extrapolating.

Although the uncertainties of the subsurface digital model could be reduced through continuous iterations of simulation and data modeling cycles, additional work is required to document a realistic range of uncertainties related to each property stored in the subsurface digital model. Different properties are composed with and are a simulation result or a data modeling product of different constituting input data. There are different data quality and data density among different datasets. Some are modeled based on a model while some are using input data directly. For these reasons a uniform uncertainty on all data types is unrealistic. Due to the variation in geographical distribution of a dataset, there could be lower uncertainty in certain regions compared to others.

A uniform uncertainty distribution on every property grid cell is also not a realistic solution. For this reason, in addition to the most likely value in each of the grid nodes stored in the subsurface digital model, a P10 and P90 value should be modeled and reported as well. P10 and P90 attributes should differ from one property to another based on data availability, convolution of models, data quality, and analytical model fit metrics. Within each P10 and P90 attribute volume, uncertainty ranges on different locations should be different as well. Procedures should be established on how these P10 and P90 properties are modeled and reported. Although it is advantageous to include more intermediate simulation and data modeling realizations in the subsurface digital model, it may not be as practical considering the storage space, data transmission, visualization, and decision requirements. This echoes the principle of not over-engineering a system. This could change when computation and data transmission speed improve in the future, however.

4.4.2 Pseudo/surrogate simulation modeling

Another reason preventing a probabilistic subsurface digital model from being a practical solution is the computation resource requirement for physics-based simulations. Computer clusters and cloud computing are currently used for simulations. Even with the advancement in chip technology and centralized computation architecture, the computational cost associated with simulation in terms of CPU/GPU time is still high. Trade-offs are often made by using lower resolution models, more simulation cores, and waiting for a longer simulation time. A typical basin model or earth model could take days to simulate with clusters while the same tasks could take up months on a local workstation[4], [117]. Institutions operated under capital constraints may have less access to High Performance Computing (HPC) which may fall further behind in research and innovation.

A paradigm shift could be constructing a pseudo-model that does not require a significant amount of simulation time to produce products with a tolerable range of uncertainties. A pseudo-model here refers to a geoanalytical model or a data model coupled with a subsurface digital model that takes in the simulation outputs, inputs, and boundary conditions as input and produces optimized conditions for modeling outputs. This could achieve instant model calibrations, suggest alternative boundary conditions and geological models, and generate numerous simulation outputs within a short period of time. With protocols established, the output uncertainty from the pseudo-modeling process could be managed. This would be a step toward realizing the probabilistic model simulation ambition.

One example of the construction of a pseudo-model construction is through a collaboration with Professor Ruben Juanes at MIT. Recent experiment data has shown encouraging simulation possibilities in the simulation the CO2 migration and storage[34]. A pseudo-model or surrogate model could drastically improve the simulation time for deterministic simulation runs. The validating experiment is to use information collected from the existing laboratory porous medium setup, using a data science model to solve for CO2 migration and saturation. Figure 4.1 below shows how the data science model is designed using a similar combined physics system – geoanalytical approach documented in Chapter 3.

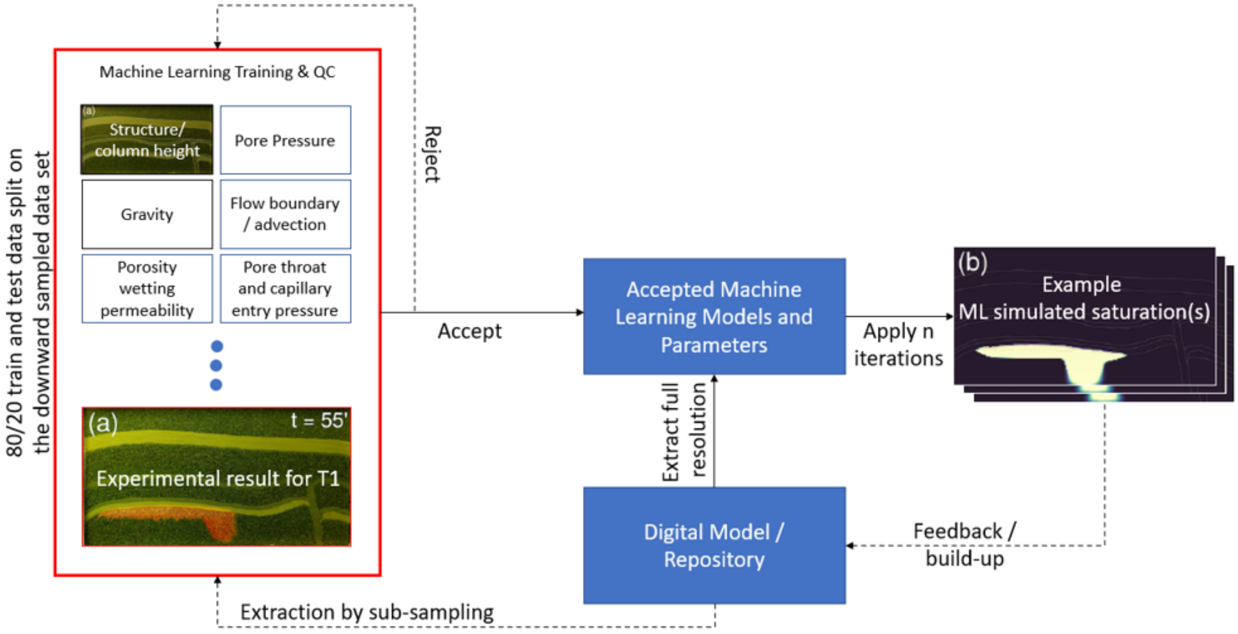

80/20 train and test data split on the downward sampled data set
Machine Learning Training & QC
(a)
Structure/ column height
Pore Pressure
Gravity
Flow boundary / advection
Porosity wetting permeability
Pore throat and capillary entry pressure
(a)
t = 55'
Experimental result for T1
Reject
Accept
Accepted Machine Learning Models and Parameters
Apply n iterations
(b)
Example ML simulated saturation(s)
Extract full resolution
Digital Model / Repository
Feedback / build-up
Extraction by sub-sampling

Figure 4.1: High-level data science modeling design to approximate simulated saturation. This will result in a machine learning model acting as a pseudo/surrogate simulation model. The image of the experimental setup and simulation results are taken from the source publication[32], [33].

The idea is to use additional information such as a high-resolution image of the experimental setup, facies description, gravity drive, flow boundaries, and geometry, in addition to simulated properties such as porosity, wettability, capillary entry pressure, pore pressure, permeability, and rate diffusion to estimate a flow and saturation profile of CO2 post-injection at a given injection rate and duration. High-resolution images may bring additional information such as packing, grain size distribution, and minor heterogeneities in the rock grains that may not be otherwise unavailable for a detailed physics model. Flow breakthroughs could be interfered with by these phenomena left uncharacterized resulting in a large amount of uncertainty in an under-detailed model. The digital data is first subsampled to a manageable matrix size, this could be 20000 evenly extracted data points (200 x 100) from the high-resolution image and physical model or less for example. Both experimental results and model position information are fed into the machine learning model for training. Accepted pseudo-models are used to generate a series of simulation outputs. Learnings are used to improve the physical model to generate better inputs for the machine learning models. After some iteration, a pertinent data science surrogate model may be considered satisfactory and may be validated by a new set of experiment data using the same or a new porous medium. The approach itself is machine-learning language and method agnostic, general statistical guardrails should be implemented during the process as some examples show in Chapter 3.

The process is repeated for different timesteps because there may be different physical controls at different stages of injection. For example, initially, injection rate and advection due to pressure change may be the primary control on fluid flow and saturation. With the flow away from the injection location, gravity and buoyancy may start to dominate the effect on the saturation pattern. Diffusion may become important given saturation some time away from the injection location. Lumping all the physical effects in one data science model may result in difficulties in model diagnostics and repeatability. Multiple stages would give domain experts better tools to investigate the data science model based on domain knowledge. Figure 4.2 below shows a rough sketch of what the proposed multi-step data science model process looks like. Each time step would yield multiple accepted data science models and simulation outputs like the ones shown in Figure 4.1 above.

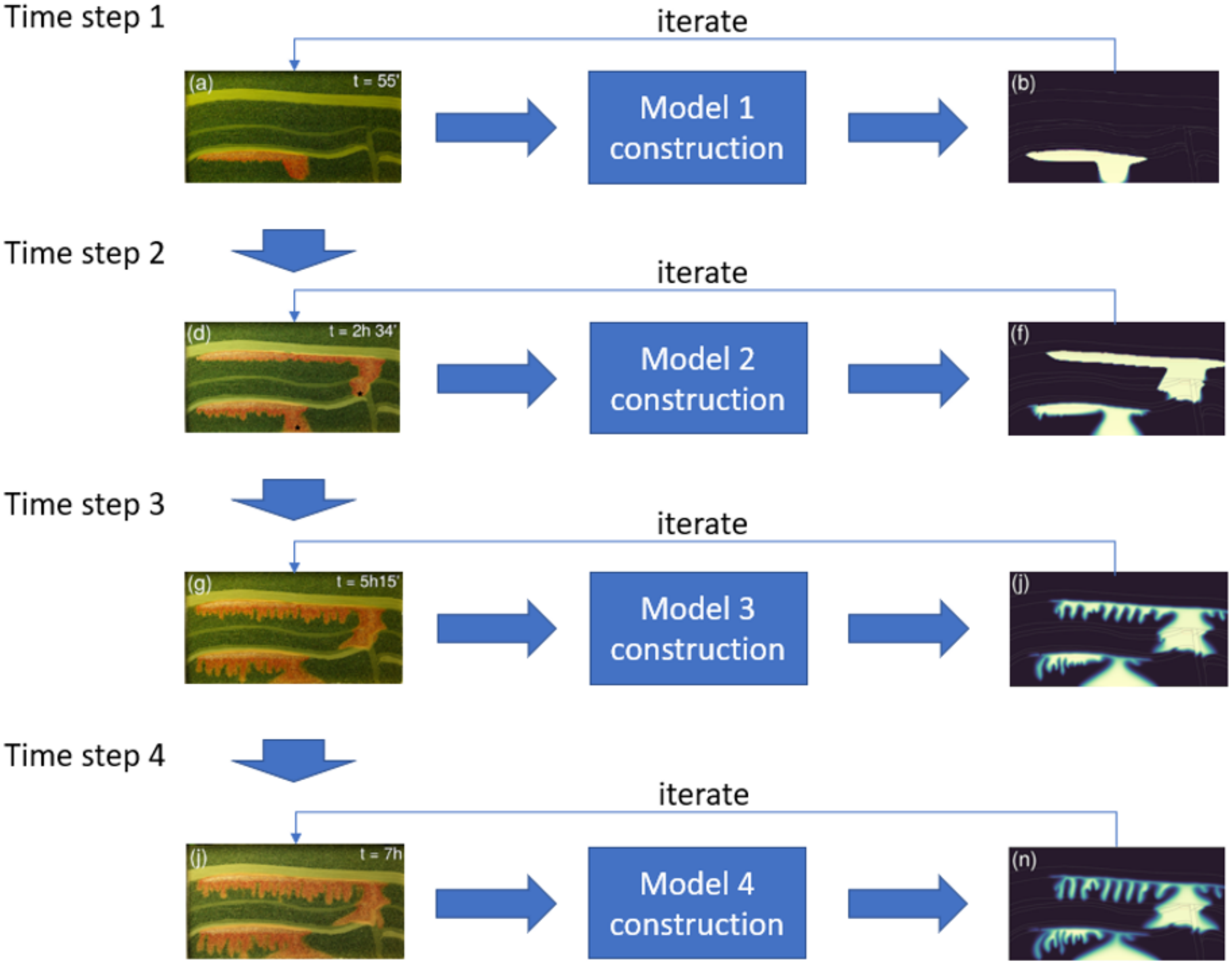


Figure 4.2: High-level data science modeling design to approximate simulated saturation in all time steps. This will result in machine learning models acting as pseudo/surrogate simulation models for all time steps. The image of the experimental setup and simulation results are taken from the source publication[32], [33].

The outcome of this experiment may be unsatisfactory because the data science experiment is done after the design of the physical laboratory-scaled experiment. There is no longer an option to acquire additional digital information such as to image the system using different technologies such as fluorescence and X-Ray. The data science model requires a lot of data and is heavily dependent on the quality of input information. If we were to design a laboratory-scaled experiment to facilitate surrogate model construction, a new porous media fluid flow experiment could be designed to gather more digital information upfront. A multi-permeability column or panel could be first prepared, and digital data of the system saturated with different control fluids at different mixes are collected using different wavelengths of electromagnetic waves such as high-resolution imagery, fluorescence, CT, and MRI scans[118]–[122]. This can be done in combination with tracer flow tests to improve characterization[123]–[125]. This high-resolution digital information could be upscaled and modeled to reflect bulk porosity, pore throat size distribution, and permeability leading to a satisfactory prediction of target fluid saturation in combination with parameters from a physical model simulation. This coupled system reduces simulation time while keeping uncertainties at an acceptable range, which is the goal of the experiment.

Some may argue the necessity of this experiment and the problems of advancing this to the field test even if the lab tests are successful because we will not be able to collect these EM wave-generated data for the subsurface. It would nevertheless provide some theoretical support for the data science approach in predicting fluid flow and saturation. In subsurface characterization, sound waves are typically used. Sound waves generate images at a coarser resolution, but the concept is the same. Subsurface fluid flow and saturation could be enhanced and simulation time could be significantly reduced by integrating seismic data into the physical model simulation. At the same time, it may provide support for other simulation challenges such as reactive transport modeling when chemical reactions present risks to subsurface reservoir performance and integrity.

The coupled physics-based system and geoanalytics under the proposed system is an ideal platform to facilitate this needle shifter, it would also improve the fidelity of the pseudo-model and reduce uncertainty as iteration continues. The subsurface digital model can be viewed as an MVP for the subsurface digital twin. As Figure 8 shows above, when feedback loops between the physical system and the digital system are maintained bidirectionally, a subsurface digital model becomes the subsurface digital twin.

4.4.3 Omni-resolution data architecture

Another requirement for an evergreen subsurface digital model is to be inclusive and widely applicable. This means inputs and outputs at all resolutions. There are multiple ways to design and realize data structure for the Aggregated Information Layer as Figure 5 shows above. For our system, a multiple dimension data storage format should be established with data access speed and visualization as its primary requirement. To reduce external barriers, an industry standard for a multi-resolution subsurface model should be established among all stakeholders. Figure 4.3 below shows an example of multiple-resolution grids during visualization in a top-down 2D view and in a 3D Birds Eye View of example geobodies.

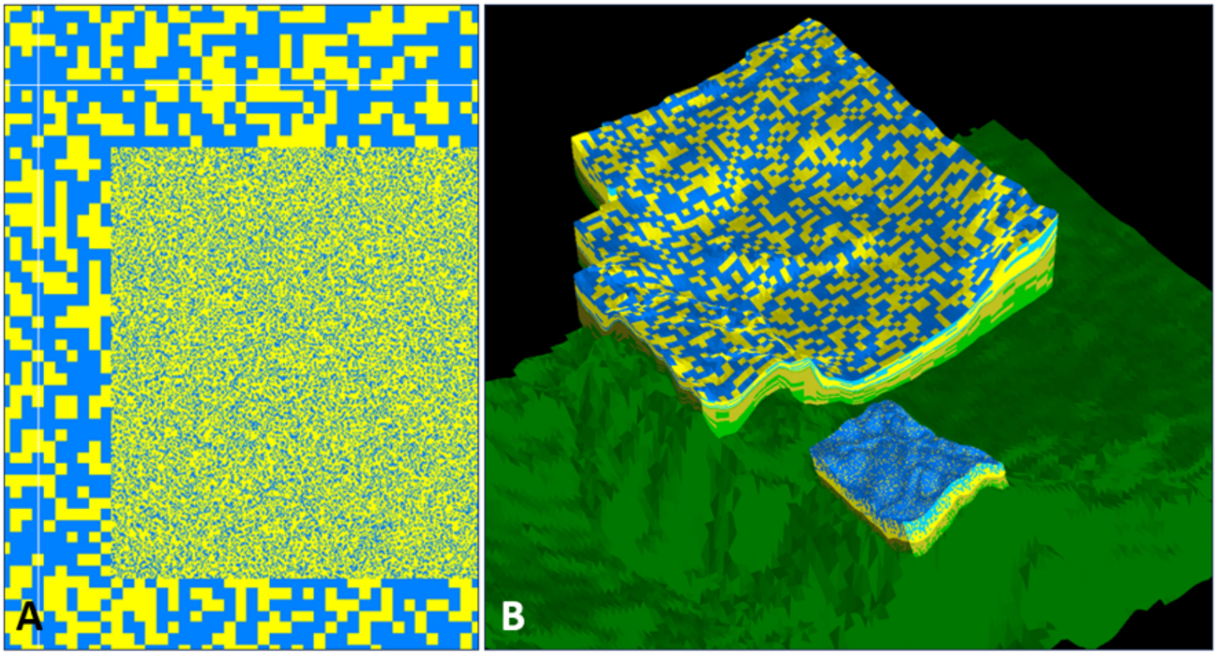


Figure 4.3: 2D and 3D illustration of omni-resolution data grids and volume. A: 2D visualization of a multi-resolution grid (2 resolutions). B: 3D visualization of a multi-resolution data volume (2 resolutions).

4.4.4 Predictive maintenance and prescriptive modeling research

With the advancement in the subsurface digital model and digital twin implementation, additional research in the use case is necessary. This includes developing new workflows and tools to operate on the data architecture of the Aggregated Information Layer. Some examples are developing a digital robot to help improve model conditions by fixing spikes and errors, filling holes and gaps, and looking for patterns and analog information. These would generate recommendations that are feedback to Model Engagement Layer and Data Foundation Layer. Parts of the upkeeping and updates could also become automated as domain experts generate and feed new insights to the Aggregated Information Layer from the Model Engagement Layer.

Research on prescriptive workflows on the Model Engagement Layer and the Aggregated Information Layer is also essential. This shifts us from making predictions to being inquisitive and understanding what happens if something changes. Additional values would be unlocked beyond predictive toward preventative and prescriptive actions. Data acquisition and system maintenance would occur more proactively leading to improved resilience.

In conclusion, the proposed subsurface digital twin is a strategic investment decision. Emerging properties such as cross-functional integration, value chain optimization, real-time decision-making, and uncertainty and risk reduction allow us to do things better and quicker. This type of decision is usually more costly and may take a longer time, but breaking the construction of such a system into multiple milestones MVPs through an agile approach allows operations to take advantage of integration at an earlier time. A subsurface digital model provides a platform for minimizing technical and sociotechnical challenges we face in digital transformation and the expectations placed by society and the industry. It is recommended to construct and maintain a subsurface digital model in every geological region of interest.

References

[1] *Annual Energy Outlook 2023 - U.S. Energy Information Administration (EIA).* [Online]. Available: https://www.eia.gov/outlooks/aeo/index.php (visited on 08/13/2023).

[2] A. Hula, A. Maguire, A. Bunker, T. Rojeck, and S. Harrison, "The 2022 EPA Automotive Trends Report: Greenhouse Gas Emissions, Fuel Economy, and Technology since 1975," Dec. 2022, Number: EPA-420-R-22-029. [Online]. Available: https://trid.trb.org/view/2112905 (visited on 08/13/2023).

[3] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and Industry 5.0—Inception, conception and perception," *Journal of Manufacturing Systems*, vol. 61, pp. 530–535, Oct. 2021, ISSN: 0278-6125. DOI: 10.1016/j.jmsy.2021.10.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278612521002119.

[4] T. Hantschel and A. I. Kauerauf, *Fundamentals of basin and petroleum systems modeling.* Springer Science & Business Media, 2009.

[5] F. O. Hoffman and J. S. Hammonds, "Propagation of uncertainty in risk assessments: The need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability," *Risk analysis*, vol. 14, no. 5, pp. 707–712, 1994, Publisher: Wiley Online Library.

[6] G. Choy, O. Khalilzadeh, M. Michalski, S. Do, A. E. Samir, O. S. Pianykh, J. R. Geis, P. V. Pandharipande, J. A. Brink, and K. J. Dreyer, "Current Applications and Future Impact of Machine Learning in Radiology," *Radiology*, vol. 288, no. 2, pp. 318–328, Aug. 2018, Publisher: Radiological Society of North America, ISSN: 0033-8419. DOI: 10.1148/radiol.2018171820. [Online]. Available: https://pubs.rsna.org/doi/abs/10.1148/radiol.2018171820 (visited on 08/13/2023).

[7] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep Reinforcement Learning from Human Preferences," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[8] M. F. Dixon, I. Halperin, and P. Bilokon, *Machine learning in finance*. Springer, 2020, vol. 1170.

[9] S. Ekins, "The Next Era: Deep Learning in Pharmaceutical Research," *Pharmaceutical Research*, vol. 33, no. 11, pp. 2594–2603, Nov. 2016, ISSN: 1573-904X. DOI: 10.1007/s11095-016-2029-7. [Online]. Available: https://doi.org/10.1007/s11095-016-2029-7.

[10] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: Past, present and future," en, *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, Dec. 2017, ISSN: 2059-8688, 2059-8696. DOI: 10.1136/svn-2017-000101. [Online]. Available: https://svn.bmj.com/lookup/doi/10.1136/svn-2017-000101 (visited on 08/13/2023).

[11] M. R. Kumar, J. Venkatesh, and A. M. J. M. Z. Rahman, "Data mining and machine learning in retail business: Developing efficiencies for better customer retention," *Journal of Ambient Intelligence and Humanized Computing*, Jan. 2021,

ISSN: 1868-5145. DOI: 10.1007/s12652-020-02711-7. [Online]. Available: https://doi.org/10.1007/s12652-020-02711-7.

[12] M. A. Ahmadi and Z. Chen, “Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs,” *Petroleum*, vol. 5, no. 3, pp. 271–284, 2019, ISSN: 2405-6561. DOI: https://doi.org/10.1016/j.petlm.2018.06.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405656117301633.

[13] S. J. Rogers, J. H. Fang, C. L. Karr, and D. A. Stanley, “Determination of Lithology from Well Logs Using a Neural Network1,” *AAPG Bulletin*, vol. 76, no. 5, pp. 731–739, May 1992, ISSN: 0149-1423. DOI: 10.1306/BDFF88BC-1718-11D7-8645000102C1865D. [Online]. Available: https://doi.org/10.1306/BDFF88BC-1718-11D7-8645000102C1865D (visited on 08/13/2023).

[14] L. Huang, X. Dong, and T. E. Clee, “A scalable deep learning platform for identifying geologic features from seismic attributes,” en, *The Leading Edge*, vol. 36, no. 3, pp. 249–256, Mar. 2017, ISSN: 1070-485X, 1938-3789. DOI: 10.1190/tle36030249.1. [Online]. Available: https://library.seg.org/doi/10.1190/tle36030249.1 (visited on 08/13/2023).

[15] W. Xiong, X. Ji, Y. Ma, Y. Wang, N. M. AlBinHassan, M. N. Ali, and Y. Luo, “Seismic fault detection with convolutional neural network,” *Geophysics*, vol. 83, no. 5, O97–O103, 2018, Publisher: Society of Exploration Geophysicists.

[16] P. Bullock and C. P. Murphy, “The microscopic examination of the structure of sub-surface horizons of soils,” *Outlook on Agriculture*, vol. 8, no. 6, pp. 348–354, Jun. 1976, Publisher: SAGE Publications Ltd, ISSN: 0030-7270. DOI: 10.1177/003072707600800607. [Online]. Available: https://doi.org/10.1177/003072707600800607 (visited on 08/13/2023).

[17] P. GAZZI, “Le arenarie del flysch sopracretaceo dell’Appennio modenese : Correlazioni con il flysch di Monghidero,” *Mineralogica et Petrographica Acta*, vol. 12, pp. 69–97, 1966. [Online]. Available: https://cir.nii.ac.jp/crid/1570854175591343232.

[18] J. Maitre, K. Bouchard, and L. P. Bédard, “Mineral grains recognition using computer vision and machine learning,” *Computers & Geosciences*, vol. 130, pp. 84–93, Sep. 2019, ISSN: 0098-3004. DOI: 10.1016/j.cageo.2019.05.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098300419301037.

[19] S. Ge, M. Kintzing, M. Jin, J. Wu, J. Bachleda, and F. Liu, “Time-Lapse Monitoring of Inter-Well Communication and Drainage Frac Height Using Geochemical Fingerprinting Technology with Case Studies in the Midland Basin,” Feb. 2022, D021S006R004. DOI: 10.2118/209145-MS. [Online]. Available: https://doi.org/10.2118/209145-MS (visited on 08/13/2023).

[20] J. Jweda, H. Long, and E. Michael, “Machine-learning assisted production allocation using a 3-D full field geochemical model of produced oils in the Eagle Ford and Austin Chalk of south Texas,” in *Unconventional Resources Technology Conference, 26–28 July 2021*, Unconventional Resources Technology Conference (URTeC), 2021, pp. 2320–2339.

[21] C. Te Stroet, J. Zwaan, G. de Jager, R. Montijn, and F. Schuren, “Predicting Sweet Spots in Shale Plays by DNA Fingerprinting and Machine Learning,” Jul. 2017, URTEC–2671117–MS. DOI: 10.15530/URTEC-2017-2671117. [Online]. Available: https://doi.org/10.15530/URTEC-2017-2671117 (visited on 08/13/2023).

[22] W. M. Telford, L. P. Geldart, and R. E. Sheriff, *Applied geophysics*. Cambridge university press, 1990.

[23] Y. Mao, D. Wang, M. Muller, K. R. Varshney, I. Baldini, C. Dugan, and A. Mojsilović, “How Data ScientistsWork Together With Domain Experts in

Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question?" *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. GROUP, Dec. 2019, Place: New York, NY, USA Publisher: Association for Computing Machinery. DOI: 10.1145/3361118. [Online]. Available: https://doi.org/10.1145/3361118.

[24] S. Park, A. Y. Wang, B. Kawas, Q. V. Liao, D. Piorkowski, and M. Danilevsky, "Facilitating Knowledge Sharing from Domain Experts to Data Scientists for Building NLP Models," in *26th International Conference on Intelligent User Interfaces*, ser. IUI '21, event-place: College Station, TX, USA, New York, NY, USA: Association for Computing Machinery, 2021, pp. 585–596, ISBN: 978-1-4503-8017-1. DOI: 10.1145/3397481.3450637. [Online]. Available: https://doi.org/10.1145/3397481.3450637.

[25] T. Seymoens, F. Ongenae, A. Jacobs, S. Verstichel, and A. Ackaert, "A Methodology to Involve Domain Experts and Machine Learning Techniques in the Design of Human-Centered Algorithms," in *Human Work Interaction Design. Designing Engaging Automation*, B. R. Barricelli, V. Roto, T. Clemmensen, P. Campos, A. Lopes, F. Gonçalves, and J. Abdelnour-Nocera, Eds., Cham: Springer International Publishing, 2019, pp. 200–214, ISBN: 978-3-030-05297-3.

[26] S. Viaene, "Data Scientists Aren't Domain Experts," *IT Professional*, vol. 15, no. 6, pp. 12–17, 2013. DOI: 10.1109/MITP.2013.93.

[27] W. Anderson, J. Dargis, B. DiMartino, Y. Watanabe, A. Yukawa, and Y. Zhao, "Architecting Chevron Shale & Tight Asset Class for Alignment and Innovation," MIT, Course: Systems Architecting Applied to Enterprises, Final Report, May 2023.

[28] J. Dargis, "Architecting an Upstream Oil and Gas Enterprise for Innovation," M.S. thesis, MIT, Sep. 2023.

[29] J. L. Pitcher and J. A. Market, *Azimuthal brittleness logging systems and methods*, Jan. 2016.

[30] D. Zhang, P. Ranjith, and M. Perera, “The brittleness indices used in rock mechanics and their application in shale hydraulic fracturing: A review,” *Journal of Petroleum Science and Engineering*, vol. 143, pp. 158–170, Jul. 2016, ISSN: 0920-4105. DOI: 10.1016/j.petrol.2016.02.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920410516300547.

[31] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[32] B. Flemisch, J. M. Nordbotten, M. Fernø, R. Juanes, H. Class, M. Delshad, F. Doster, J. Ennis-King, J. Franc, S. Geiger, *et al.*, “The FluidFlower international benchmark study: Process, modeling results, and comparison to experimental data,” *arXiv preprint arXiv:2302.10986*, 2023.

[33] J. M. Nordbotten, M. A. Fernø, B. Flemisch, A. R. Kovscek, and K.-A. Lie, “The 11th Society of Petroleum Engineers Comparative Solution Project,” en, 2023.

[34] L. Saló-Salgado, M. Haugen, K. Eikehaug, M. Fernø, J. M. Nordbotten, and R. Juanes, “Direct Comparison of Numerical Simulations and Experiments of $$\hbox {CO}_2$$Injection and Migration in Geologic Media: Value of Local Data and Forecasting Capability,” *Transport in Porous Media*, Jun. 2023, ISSN: 1573-1634. DOI: 10.1007/s11242-023-01972-y. [Online]. Available: https://doi.org/10.1007/s11242-023-01972-y.

[35] Ø. Sylta and W. Krokstad, “Estimation of oil and gas column heights in prospects using probabilistic basin modelling methods,” *Petroleum Geoscience*, vol. 9, no. 3, pp. 243–254, Jul. 2003, ISSN: 1354-0793. DOI: 10.1144/1354-079302-563. [Online]. Available: https://doi.org/10.1144/1354-079302-563 (visited on 08/13/2023).

[36] R. B. Ainsworth, B. K. Vakarelov, and R. A. Nanson, “Dynamic spatial and temporal prediction of changes in depositional processes on clastic shorelines:

Toward improved subsurface uncertainty reduction and management," *AAPG Bulletin*, vol. 95, no. 2, pp. 267–297, Feb. 2011, ISSN: 0149-1423. DOI: 10.1306/06301010036. [Online]. Available: https://doi.org/10.1306/06301010036 (visited on 08/13/2023).

[37] C. R. Berger and R. J. Calabrese, "Some Explorations in Initial Interaction and Beyond: Toward a Developmental Theory of Interpersonal Communication," *Human Communication Research*, vol. 1, no. 2, pp. 99–112, Dec. 1975, ISSN: 0360-3989. DOI: 10.1111/j.1468-2958.1975.tb00258.x. [Online]. Available: https://doi.org/10.1111/j.1468-2958.1975.tb00258.x (visited on 08/13/2023).

[38] E. Crawley, B. Cameron, and D. Selva, *System architecture: strategy and product development for complex systems*. Prentice Hall Press, 2015.

[39] J. H. Schön, *Physical properties of rocks: Fundamentals and principles of petrophysics*. Elsevier, 2015.

[40] R. H. Herrera and M. van der Baan, "A semiautomatic method to tie well logs to seismic data," *Geophysics*, vol. 79, no. 3, pp. V47–V54, 2014, Publisher: Society of Exploration Geophysicists.

[41] A. Munoz and D. Hale, "Automatically tying well logs to seismic data," *CWP-725*, 2012, Publisher: Citeseer.

[42] J. E. Adams, "Stratigraphic-Tectonic Development of Delaware Basin1," *AAPG Bulletin*, vol. 49, no. 11, pp. 2140–2148, Nov. 1965, ISSN: 0149-1423. DOI: 10.1306/A6633888-16C0-11D7-8645000102C1865D. [Online]. Available: https://doi.org/10.1306/A6633888-16C0-11D7-8645000102C1865D (visited on 08/13/2023).

[43] T. Hoak, K. Sundberg, and P. Ortoleva, "Overview of the structural geology and tectonics of the Central Basin Platform, Delaware Basin, and Midland Basin, West

Texas and New Mexico," Dec. 1998. DOI: 10.2172/307858. [Online]. Available: https://www.osti.gov/biblio/307858.

[44] P. Jong and J. Krebbers, "Transforming the subsurface data and application landscape with the open subsurface data universe forum," en, in *SEG Technical Program Expanded Abstracts 2019*, San Antonio, Texas: Society of Exploration Geophysicists, Aug. 2019, pp. 4987–4989. DOI: 10.1190/segam2019-3216470.1. [Online]. Available: https://library.seg.org/doi/10.1190/segam2019-3216470.1 (visited on 08/13/2023).

[45] T. G. Farr, P. A. Rosen, E. Caro, *et al.*, "The Shuttle Radar Topography Mission," en, *Reviews of Geophysics*, vol. 45, no. 2, 2007, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2005RG000183, ISSN: 1944-9208. DOI: 10.1029/2005RG000183. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1029/2005RG000183 (visited on 08/13/2023).

[46] D. B. Chelton, J. C. Ries, B. J. Haines, L.-L. Fu, and P. S. Callahan, "Chapter 1 Satellite Altimetry," in *International Geophysics*, L.-L. Fu and A. Cazenave, Eds., vol. 69, Academic Press, Jan. 2001, pp. 1–ii, ISBN: 0074-6142. DOI: 10.1016/S0074-6142(01)80146-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0074614201801467.

[47] L.-L. Fu and A. Cazenave, *Satellite altimetry and earth sciences: a handbook of techniques and applications*. Elsevier, 2000.

[48] S. E. Reutebuch, H.-E. Andersen, and R. J. McGaughey, "Light Detection and Ranging (LIDAR): An Emerging Tool for Multiple Resource Inventory," *Journal of Forestry*, vol. 103, no. 6, pp. 286–292, Sep. 2005, ISSN: 0022-1201. DOI: 10.1093/jof/103.6.286. [Online]. Available: https://doi.org/10.1093/jof/103.6.286 (visited on 08/13/2023).

[49] M. Z. Jacobson, *Fundamentals of atmospheric modeling*. Cambridge university press, 1999.

[50] P. Sheridan, S. Smith, A. Brown, and S. Vosper, “A simple height-based correction for temperature downscaling in complex terrain,” en, *Meteorological Applications*, vol. 17, no. 3, pp. 329–339, 2010, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/met.177, ISSN: 1469-8080. DOI: 10.1002/met.177. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/met.177 (visited on 08/13/2023).

[51] M. P. Jackson and M. R. Hudec, *Salt tectonics: Principles and practice*. Cambridge University Press, 2017.

[52] D. Forand, V. Heesakkers, and K. Schwartz, “Constraints on Natural Fracture and In-situ Stress Trends of Unconventional Reservoirs in the Permian Basin, USA,” Jul. 2017, URTEC–2669208–MS. DOI: 10.15530/URTEC-2017-2669208. [Online]. Available: https://doi.org/10.15530/URTEC-2017-2669208 (visited on 08/13/2023).

[53] R. Barnaby and W. Ward, “Outcrop Analog for Mixed Siliciclastic–Carbonate Ramp Reservoirs—Stratigraphic Hierarchy, Facies Architecture, and Geologic Heterogeneity: Grayburg Formation, Permian Basin, U.S.A.,” *Journal of Sedimentary Research*, vol. 77, no. 1, pp. 34–58, Jan. 2007, ISSN: 1527-1404. DOI: 10.2110/jsr.2007.007. [Online]. Available: https://doi.org/10.2110/jsr.2007.007 (visited on 08/13/2023).

[54] R. D. Wilson, J. Chitale, K. Huffman, P. Montgomery, and S. J. Prochnow, “Evaluating the depositional environment, lithofacies variation, and diagenetic processes of the Wolfcamp B and lower Spraberry intervals in the Midland Basin: Implications for reservoir quality and distribution,” *AAPG Bulletin*, vol. 104, no. 6, pp. 1287–1321, Jun. 2020, ISSN: 0149-1423. DOI: 10.1306/12031917358. [Online]. Available: https://doi.org/10.1306/12031917358 (visited on 08/13/2023).

[55] Bjørlykke K., “Principal aspects of compaction and fluid flow in mudstones,” *Geological Society, London, Special Publications*, vol. 158, no. 1, pp. 73–78, Jan. 1999, Publisher: The Geological Society of London. DOI: 10.1144/GSL.SP.1999.158.01.06. [Online]. Available: https://doi.org/10.1144/GSL.SP.1999.158.01.06 (visited on 08/13/2023).

[56] S. Ehrenberg, “Assessing the Relative Importance of Compaction Processes and Cementation to Reduction of Porosity in Sandstones: Discussion; Compaction and Porosity Evolution of Pliocene Sandstones, Ventura Basin, California: DISCUSSION1,” *AAPG Bulletin*, vol. 73, no. 10, pp. 1274–1276, Oct. 1989, ISSN: 0149-1423. DOI: 10.1306/44B4AA1E-170A-11D7-8645000102C1865D. [Online]. Available: https://doi.org/10.1306/44B4AA1E-170A-11D7-8645000102C1865D (visited on 08/13/2023).

[57] M. W. Downey, “Evaluating Seals for Hydrocarbon Accumulations1,” *AAPG Bulletin*, vol. 68, no. 11, pp. 1752–1763, Nov. 1984, ISSN: 0149-1423. DOI: 10.1306/AD461994-16F7-11D7-8645000102C1865D. [Online]. Available: https://doi.org/10.1306/AD461994-16F7-11D7-8645000102C1865D (visited on 08/13/2023).

[58] D. W. Waples, “A New Model for Heat Flow in Extensional Basins: Radiogenic Heat, Asthenospheric Heat, and the McKenzie Model,” *Natural Resources Research*, vol. 10, no. 3, pp. 227–238, Sep. 2001, ISSN: 1573-8981. DOI: 10.1023/A:1012521309181. [Online]. Available: https://doi.org/10.1023/A:1012521309181.

[59] N. H. Sleep, “Evolution of the mode of convection within terrestrial planets,” *Journal of Geophysical Research: Planets*, vol. 105, no. E7, pp. 17 563–17 578, 2000, Publisher: Wiley Online Library.

[60] D. L. Turcotte and G. Schubert, *Geodynamics*. Cambridge university press, 2002.

[61] T. Sun, F. J. Laugier, R. Caldwell, A. Harris, and M. Sullivan, "Computational Stratigraphy of Deepwater Sedimentary Systems," vol. 2018, OS12A–02, Dec. 2018, Conference Name: AGU Fall Meeting Abstracts ADS Bibcode: 2018AGUFMOS12A..02S. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018AGUFMOS12A..02S (visited on 08/13/2023).

[62] B. Willis and T. Sun, "Relating depositional processes of river-dominated deltas to reservoir behavior using computational stratigraphy," *Journal of Sedimentary Research*, vol. 89, no. 12, pp. 1250–1276, Dec. 2019, ISSN: 1527-1404. DOI: 10.2110/jsr.2019.59. [Online]. Available: https://doi.org/10.2110/jsr.2019.59 (visited on 08/13/2023).

[63] K. Bjørlykke and P. K. Egeberg, "Quartz Cementation in Sedimentary Basins1," *AAPG Bulletin*, vol. 77, no. 9, pp. 1538–1548, Sep. 1993, _eprint: https://pubs.geoscienceworld.org/aapgbull/article-pdf/77/9/1538/4509377/aapg_1993_0077_0009_1538.pdf, ISSN: 0149-1423. DOI: 10.1306/BDFF8EE8-1718-11D7-8645000102C1865D. [Online]. Available: https://doi.org/10.1306/BDFF8EE8-1718-11D7-8645000102C1865D.

[64] K. Gallagher, R. Brown, and C. Johnson, "FISSION TRACK ANALYSIS AND ITS APPLICATIONS TO GEOLOGICAL PROBLEMS," *Annual Review of Earth and Planetary Sciences*, vol. 26, no. 1, pp. 519–572, May 1998, Publisher: Annual Reviews, ISSN: 0084-6597. DOI: 10.1146/annurev.earth.26.1.519. [Online]. Available: https://doi.org/10.1146/annurev.earth.26.1.519 (visited on 08/13/2023).

[65] R. H. Goldstein, "Fluid inclusions in sedimentary and diagenetic systems," *Fluid Inclusions: Phase Relationships - Methods - Applications. A Special Issue in honour of Jacques Touret*, vol. 55, no. 1, pp. 159–193, Jan. 2001, ISSN: 0024-4937. DOI:

10.1016/S0024-4937(00)00044-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002449370000044X.

[66] S. Morad, "Carbonate Cementation in Sandstones: Distribution Patterns and Geochemical Evolution," in *Carbonate Cementation in Sandstones*, May 1998, pp. 1–26, ISBN: 978-1-4443-0489-3. DOI: 10.1002/9781444304893.ch1. [Online]. Available: https://doi.org/10.1002/9781444304893.ch1 (visited on 08/13/2023).

[67] K. Peters, J. Moldowan, and P. Sundararaman, "Effects of hydrous pyrolysis on biomarker thermal maturity parameters: Monterey Phosphatic and Siliceous members," *Organic Geochemistry*, vol. 15, no. 3, pp. 249–265, Jan. 1990, ISSN: 0146-6380. DOI: 10.1016/0146-6380(90)90003-I. [Online]. Available: https://www.sciencedirect.com/science/article/pii/014663809090003I.

[68] E. Roedder, "Volume 12: Fluid inclusions," *Reviews in mineralogy*, vol. 12, p. 644, 1984.

[69] Z. Wei, J. M. Moldowan, S. Zhang, R. Hill, D. M. Jarvie, H. Wang, F. Song, and F. Fago, "Diamondoid hydrocarbons as a molecular proxy for thermal maturity and oil cracking: Geochemical models from hydrous pyrolysis," *Organic Geochemistry*, vol. 38, no. 2, pp. 227–249, Feb. 2007, ISSN: 0146-6380. DOI: 10.1016/j.orggeochem.2006.09.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0146638006002488.

[70] R. Rateau, N. Schofield, and M. Smith, "The potential role of igneous intrusions on hydrocarbon migration, West of Shetland," *Petroleum Geoscience*, vol. 19, no. 3, pp. 259–272, Jul. 2013, ISSN: 1354-0793. DOI: 10.1144/petgeo2012-035. [Online]. Available: https://doi.org/10.1144/petgeo2012-035 (visited on 08/13/2023).

[71] T. T. Schowalter, "Mechanics of Secondary Hydrocarbon Migration and Entrapment1," *AAPG Bulletin*, vol. 63, no. 5, pp. 723–760, May 1979, ISSN: 0149-1423. DOI: 10.1306/2F9182CA-16CE-11D7-8645000102C1865D. [Online].

Available: https://doi.org/10.1306/2F9182CA-16CE-11D7-8645000102C1865D (visited on 08/13/2023).

[72] D. M. Jarvie, R. J. Hill, T. E. Ruble, and R. M. Pollastro, "Unconventional shale-gas systems: The Mississippian Barnett Shale of north-central Texas as one model for thermogenic shale-gas assessment," *AAPG Bulletin*, vol. 91, no. 4, pp. 475–499, Apr. 2007, ISSN: 0149-1423. DOI: 10.1306/12190606068. [Online]. Available: https://doi.org/10.1306/12190606068 (visited on 08/13/2023).

[73] E. Sandvik, W. Young, and D. Curry, "Expulsion from hydrocarbon sources: The role of organic absorption," *Proceedings of the 15th International Meeting on Organic Geochemistry*, vol. 19, no. 1, pp. 77–87, Dec. 1992, ISSN: 0146-6380. DOI: 10.1016/0146-6380(92)90028-V. [Online]. Available: https://www.sciencedirect.com/science/article/pii/014663809290028V.

[74] M. Grieves and J. Vickers, "Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems," in *Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches*, F.-J. Kahlen, S. Flumerfelt, and A. Alves, Eds., Cham: Springer International Publishing, 2017, pp. 85–113, ISBN: 978-3-319-38756-7. DOI: 10.1007/978-3-319-38756-7_4. [Online]. Available: https://doi.org/10.1007/978-3-319-38756-7_4.

[75] E. Rebentisch, D. H. Rhodes, A. L. Soares, R. Zimmerman, and S. Tavares, "The digital twin as an enabler of digital transformation: A sociotechnical perspective," in *2021 IEEE 19th International Conference on Industrial Informatics (INDIN)*, Journal Abbreviation: 2021 IEEE 19th International Conference on Industrial Informatics (INDIN), Jul. 2021, pp. 1–6. DOI: 10.1109/INDIN45523.2021.9557455.

[76] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the Digital Twin: A systematic literature review," *CIRP Journal of Manufacturing Science and Technology*, vol. 29, pp. 36–52, May 2020, ISSN: 1755-5817. DOI:

10.1016/j.cirpj.2020.02.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1755581720300110.

[77] W. Kritzinger, M. Karner, G. Traar, J. Henjes, and W. Sihn, “Digital Twin in manufacturing: A categorical literature review and classification,” *16th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2018*, vol. 51, no. 11, pp. 1016–1022, Jan. 2018, ISSN: 2405-8963. DOI: 10.1016/j.ifacol.2018.08.474. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896318316021.

[78] G. S. McChrystal, T. Collins, D. Silverman, and C. Fussell, *Team of teams: New rules of engagement for a complex world*. Penguin, 2015.

[79] R. C. Ford and W. A. Randolph, “Cross-Functional Structures: A Review and Integration of Matrix Organization and Project Management,” *Journal of Management*, vol. 18, no. 2, pp. 267–294, Jun. 1992, Publisher: SAGE Publications Inc, ISSN: 0149-2063. DOI: 10.1177/014920639201800204. [Online]. Available: https://doi.org/10.1177/014920639201800204 (visited on 08/14/2023).

[80] D. M. Jarvie, “Total organic carbon (TOC) analysis: Chapter 11: Geochemical methods and exploration,” 1991, Publisher: AAPG Special Volumes.

[81] J. Qin, L. Zheng, and Tenger, “Study on the restitution coefficient of original total organic carbon for high mature marine source rocks,” *Frontiers of Earth Science in China*, vol. 1, no. 4, pp. 482–490, Oct. 2007, ISSN: 1673-7490. DOI: 10.1007/s11707-007-0059-5. [Online]. Available: https://doi.org/10.1007/s11707-007-0059-5.

[82] S. M. Henrichs and W. S. Reeburgh, “Anaerobic mineralization of marine sediment organic matter: Rates and the role of anaerobic processes in the oceanic carbon economy,” *Geomicrobiology Journal*, vol. 5, no. 3-4, pp. 191–237, Jan. 1987,

Publisher: Taylor & Francis, ISSN: 0149-0451. DOI: 10.1080/01490458709385971. [Online]. Available: https://doi.org/10.1080/01490458709385971.

[83] D. J. Burdige, "Preservation of Organic Matter in Marine Sediments: Controls, Mechanisms, and an Imbalance in Sediment Organic Carbon Budgets?" *Chemical Reviews*, vol. 107, no. 2, pp. 467–485, Feb. 2007, Publisher: American Chemical Society, ISSN: 0009-2665. DOI: 10.1021/cr050347q. [Online]. Available: https://doi.org/10.1021/cr050347q.

[84] J. I. Hedges and R. G. Keil, "Sedimentary organic matter preservation: An assessment and speculative synthesis," *Marine Chemistry*, vol. 49, no. 2, pp. 81–115, Apr. 1995, ISSN: 0304-4203. DOI: 10.1016/0304-4203(95)00008-F. [Online]. Available: https://www.sciencedirect.com/science/article/pii/030442039500008F.

[85] B. J. Katz, "Controlling Factors on Source Rock Development—A Review of Productivity, Preservation, and Sedimentation Rate," in *The Deposition of Organic-Carbon-Rich Sediments: Models, Mechanisms, and Consequences*, N. B. Harris, Ed., vol. 82, SEPM Society for Sedimentary Geology, Jan. 2005, p. 0, ISBN: 978-1-56576-218-3. DOI: 10.2110/pec.05.82.0007. [Online]. Available: https://doi.org/10.2110/pec.05.82.0007 (visited on 08/13/2023).

[86] R. Tyson, "Sedimentation rate, dilution, preservation and total organic carbon: Some results of a modelling study," *Organic Geochemistry*, vol. 32, no. 2, pp. 333–339, Feb. 2001, ISSN: 0146-6380. DOI: 10.1016/S0146-6380(00)00161-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0146638000001613.

[87] K. S. Mews, M. M. Alhubail, and R. G. Barati, "A Review of Brittleness Index Correlations for Unconventional Tight and Ultra-Tight Reservoirs," *Geosciences*, vol. 9, no. 7, 2019, ISSN: 2076-3263. DOI: 10.3390/geosciences9070319. [Online]. Available: https://www.mdpi.com/2076-3263/9/7/319.

[88] M. Mullen, R. Roundtree, and B. Barree, "A Composite Determination of Mechanical Rock Properties for Stimulation Design (What To Do When You Don't Have a Sonic Log)," Apr. 2007, SPE–108139–MS. DOI: 10.2118/108139-MS. [Online]. Available: https://doi.org/10.2118/108139-MS (visited on 08/13/2023).

[89] Y. Zhang, H.-R. Zhong, Z.-Y. Wu, H. Zhou, and Q.-Y. Ma, "Improvement of petrophysical workflow for shear wave velocity prediction based on machine learning methods for complex carbonate reservoirs," *Journal of Petroleum Science and Engineering*, vol. 192, p. 107 234, Sep. 2020, ISSN: 0920-4105. DOI: 10.1016/j.petrol.2020.107234. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920410520303193.

[90] T. Dietterich, "Overfitting and undercomputing in machine learning," en, *ACM Computing Surveys*, vol. 27, no. 3, pp. 326–327, Sep. 1995, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/212094.212114. [Online]. Available: https://dl.acm.org/doi/10.1145/212094.212114 (visited on 08/13/2023).

[91] X. Ying, "An Overview of Overfitting and its Solutions," *Journal of Physics: Conference Series*, vol. 1168, no. 2, p. 022 022, Feb. 2019, Publisher: IOP Publishing, ISSN: 1742-6596. DOI: 10.1088/1742-6596/1168/2/022022. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1168/2/022022.

[92] D. Nerini, P. Monestiez, and C. Manté, "Cokriging for spatial functional data," *Statistical Methods and Problems in Infinite-dimensional Spaces*, vol. 101, no. 2, pp. 409–418, Feb. 2010, ISSN: 0047-259X. DOI: 10.1016/j.jmva.2009.03.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X0900061X.

[93] M. A. OLIVER and R. WEBSTER, "Kriging: A method of interpolation for geographical information systems," *International Journal of Geographical Information Systems*, vol. 4, no. 3, pp. 313–332, Jul. 1990, Publisher: Taylor &

Francis, ISSN: 0269-3798. DOI: 10.1080/02693799008941549. [Online]. Available: https://doi.org/10.1080/02693799008941549.

[94] A. Holzinger, "From Machine Learning to Explainable AI," in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, Journal Abbreviation: 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Aug. 2018, pp. 55–66. DOI: 10.1109/DISA.2018.8490530.

[95] N. Savage, "Breaking into the black box of artificial intelligence.," eng, *Nature*, Mar. 2022, Place: England, ISSN: 1476-4687 0028-0836. DOI: 10.1038/d41586-022-00858-1.

[96] O. Loyola-González, "Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View," *IEEE Access*, vol. 7, pp. 154 096–154 113, 2019, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2949286.

[97] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x. [Online]. Available: https://doi.org/10.1038/s42256-019-0048-x.

[98] J. W. Tukey *et al.*, *Exploratory data analysis*. Reading, MA, 1977, vol. 2.

[99] J. Fox and G. Monette, "Generalized Collinearity Diagnostics," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 178–183, Mar. 1992, Publisher: Taylor & Francis, ISSN: 0162-1459. DOI: 10.1080/01621459.1992.10475190. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475190.

[100] D. M. Allen, "The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction," *Technometrics*, vol. 16, no. 1, pp. 125–127, Feb. 1974, Publisher: Taylor & Francis, ISSN: 0040-1706. DOI: 10.1080/00401706.1974.10489157. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1974.10489157.

[101] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella, “The ‘K’ in K-fold Cross Validation,” en, *Computational Intelligence*, 2012.

[102] T. Fushiki, “Estimation of prediction error by using K-fold cross-validation,” *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, Apr. 2011, ISSN: 1573-1375. DOI: 10.1007/s11222-009-9153-8. [Online]. Available: https://doi.org/10.1007/s11222-009-9153-8.

[103] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *Journal of the royal statistical society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974, Publisher: Wiley Online Library.

[104] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation,” *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, Jan. 2015, Publisher: Taylor & Francis, ISSN: 1061-8600. DOI: 10.1080/10618600.2014.907095. [Online]. Available: https://doi.org/10.1080/10618600.2014.907095.

[105] N. S. Chok, “Pearson’s Versus Spearman’s and Kendall’s Correlation Coefficients for Continuous Data,” Sep. 2010. [Online]. Available: http://d-scholarship.pitt.edu/8056/.

[106] Y. Dodge, *The concise encyclopedia of statistics*. Springer Science & Business Media, 2008.

[107] M. G. Kendall, “A New Measure of Rank Correlation,” *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938, Publisher: [Oxford University Press, Biometrika Trust], ISSN: 0006-3444. DOI: 10.2307/2332226. [Online]. Available: https://www.jstor.org/stable/2332226 (visited on 08/13/2023).

[108] R. G. Pontius, O. Thontteh, and H. Chen, “Components of information for multiple resolution comparison between maps that share a real variable,” *Environmental and*

Ecological Statistics, vol. 15, no. 2, pp. 111–142, Jun. 2008, ISSN: 1573-3009. DOI: 10.1007/s10651-007-0043-y. [Online]. Available: https://doi.org/10.1007/s10651-007-0043-y.

[109] C. J. Willmott and K. Matsuura, "On the use of dimensioned measures of error to evaluate the performance of spatial interpolators," *International Journal of Geographical Information Science*, vol. 20, no. 1, pp. 89–102, Jan. 2006, Publisher: Taylor & Francis, ISSN: 1365-8816. DOI: 10.1080/13658810500286976. [Online]. Available: https://doi.org/10.1080/13658810500286976.

[110] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, *et al.*, "Xgboost: Extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[111] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities.," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, Apr. 1982, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.79.8.2554. [Online]. Available: https://doi.org/10.1073/pnas.79.8.2554 (visited on 08/13/2023).

[112] G. He, D. S. Mallapragada, A. Bose, C. F. Heuberger-Austin, and E. Gençer, "Sector coupling *via* hydrogen to lower the cost of energy system decarbonization," en, *Energy & Environmental Science*, vol. 14, no. 9, pp. 4635–4646, 2021, ISSN: 1754-5692, 1754-5706. DOI: 10.1039/D1EE00627D. [Online]. Available: http://xlink.rsc.org/?DOI=D1EE00627D (visited on 08/14/2023).

[113] M. Noussan, P. P. Raimondi, R. Scita, and M. Hafner, "The Role of Green and Blue Hydrogen in the Energy Transition—A Technological and Geopolitical Perspective," en, *Sustainability*, vol. 13, no. 1, p. 298, Jan. 2021, Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2071-1050. DOI:

10.3390/su13010298. [Online]. Available: https://www.mdpi.com/2071-1050/13/1/298 (visited on 08/14/2023).

[114] E. Rebentisch and L. Prusak, *Integrating program management and systems engineering: Methods, tools, and organizational systems for improving performance*. John Wiley & Sons, 2017.

[115] O. L. De Weck, D. Roos, and C. L. Magee, *Engineering systems: Meeting human needs in a complex technological world*. Mit Press, 2011.

[116] K. E. Peters, D. J. Curry, and M. Kacewicz, "An overview of basin and petroleum system modeling: Definitions and concepts," 2012, Publisher: AAPG Special Volumes.

[117] Matthäi S. K., Geiger S., Roberts S. G., *et al.*, "Numerical simulation of multi-phase fluid flow in structurally complex reservoirs," *Geological Society, London, Special Publications*, vol. 292, no. 1, pp. 405–429, Jan. 2007, Publisher: The Geological Society of London. DOI: 10.1144/SP292.22. [Online]. Available: https://doi.org/10.1144/SP292.22 (visited on 08/14/2023).

[118] P. Dijk, B. Berkowitz, and P. Bendel, "Investigation of flow in water-saturated rock fractures using nuclear magnetic resonance imaging (NMRI)," en, *Water Resources Research*, vol. 35, no. 2, pp. 347–360, 1999, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/1998WR900044, ISSN: 1944-7973. DOI: 10.1029/1998WR900044. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1029/1998WR900044 (visited on 08/14/2023).

[119] J. Mason, M. McGlue, and T. Jordan, "Unification of Compositional Data and High-Resolution Facies Analysis in the Union Springs Formation of New York," in *AAPG Eastern Section Meeting*, Sep. 2016.

[120] S. Saraf and A. Bera, "A review on pore-scale modeling and CT scan technique to characterize the trapped carbon dioxide in impermeable reservoir rocks during sequestration," *Renewable and Sustainable Energy Reviews*, vol. 144, p. 110 986, Jul. 2021, ISSN: 1364-0321. DOI: 10.1016/j.rser.2021.110986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032121002781.

[121] K. E. Washburn and G. Madelin, "Imaging of multiphase fluid saturation within a porous material via sodium NMR," *Journal of Magnetic Resonance*, vol. 202, no. 1, pp. 122–126, Jan. 2010, ISSN: 1090-7807. DOI: 10.1016/j.jmr.2009.10.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1090780709002948.

[122] W. P. Węglarz, A. Krzyżak, G. Machowski, and M. Stefaniuk, "ZTE MRI in high magnetic field as a time effective 3D imaging technique for monitoring water ingress in porous rocks at sub-millimetre resolution," *Magnetic Resonance Imaging*, vol. 47, pp. 54–59, Apr. 2018, ISSN: 0730-725X. DOI: 10.1016/j.mri.2017.11.015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0730725X1730262X.

[123] Y. Xu, L. Chen, Y. Zhao, L. M. Cathles, and C. K. Ober, "Supercritical CO 2 -philic nanoparticles suitable for determining the viability of carbon sequestration in shale," en, *Environmental Science: Nano*, vol. 2, no. 3, pp. 288–296, 2015, Publisher: Royal Society of Chemistry. DOI: 10.1039/C5EN00003C. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2015/en/c5en00003c (visited on 08/14/2023).

[124] C. Yao, Y. Zhao, G. Lei, T. S. Steenhuis, and L. M. Cathles, "Inert Carbon Nanoparticles for the Assessment of Preferential Flow in Saturated Dual-Permeability Porous Media," *Industrial & Engineering Chemistry Research*, vol. 56, no. 25, pp. 7365–7374, Jun. 2017, Publisher: American Chemical Society, ISSN: 0888-5885. DOI: 10.1021/acs.iecr.7b00194. [Online]. Available: https://doi.org/10.1021/acs.iecr.7b00194.

[125] Y. Zhao, "The use of nanoparticles to assess subsurface flow heterogeneity," *Cornell University*, no. Cornell Thesis and Dissertations, 2015. [Online]. Available: https://hdl.handle.net/1813/39481.