English AI Library Edition

Artificial Intelligence for New Materials Design and Rocket Propulsion Engineering

AI Potentials, Self-Driving Laboratories, and Physics-Informed Machine Learning (PIML)

Kim Kyung-jin,
Attorney at Law

kimkj.com · AI Library · 2026

Artificial Intelligence for New Materials Design and Rocket Propulsion Engineering

AI Potentials, Self-Driving Laboratories, and Physics-Informed Machine Learning (PIML)

Kim Kyung-jin, Attorney at Law

Starting with machine learning potentials that learn the movement of a single atom, this book covers ten chapters: generative models that first set desired properties and then create crystal structures in reverse; self-driving laboratories where robots repeat experiments on their own; high-entropy alloys and ceramic composites that can protect the hottest parts of rockets; defect monitoring and fatigue life assurance in metal 3D printing; physics-informed machine learning that puts physical laws into the loss function to calculate combustion; cooling channel design; and real-time diagnosis and reinforcement learning control of liquid rocket engines. It is written without equations for general readers who like science and technology. Each chapter includes verified supporting materials. This is a reviewed and revised version that reflects the advisor-style criticism of two artificial intelligence reviewers.

Public Research Materials Repository

This book was compiled and produced with artificial intelligence. It synthesizes materials collected in major languages from around the world. NotebookLM does not limit the language of questions based on the language in which the materials were written. Readers can ask questions in their own language in the same research materials repository used to prepare this book. The same applies even when some materials are in Korean, English, or other languages. If you open the settings in NotebookLM and set the output language to the language you want, the answers and generated results will appear in that language.

The public NotebookLM repository used to prepare this book (AI for Materials Science and Rocket Engine Engineering, 129 materials): https://notebook.google.com/notebook/ac39e3ce-77a9-4e78-9aaf-b52c45c2934d

The original Korean edition of this book is available in the kimkj.com AI Library: https://kimkj.com/ai-library/?mod=document&uid=12453

Table of Contents

Click each chapter title to go to the corresponding text. To read from the beginning, select Chapter 1.

Chapter 1. Foundations of Materials Informatics and Machine Learning Interatomic Potential (MLIP)

Introduction

Every object we use is made from materials. A spoon is made from metal, and a cup is made from glass. The nozzle of a rocket engine is also made from special metals and ceramics. When the material changes, the properties of the object also change.

The properties of a material begin with tiny particles that we cannot see. Those particles are atoms. Atoms are the basic units that make up matter. Depending on how atoms bond and move, metal can become hard or soft.

People who build rockets want to know these atomic movements in advance. Inside a rocket engine, the temperature rises to several thousand degrees. The pressure is also very high. We must know in advance whether a metal can withstand such conditions without melting. Making and testing real parts takes a great deal of money and time. So scientists move atoms inside a computer. This is atomic simulation.

To move atoms with a computer, rules are needed. This is because we must know in which direction and with how much force each atom is pushed. For a long time, the accurate way to make these rules and the fast way to make them were different. The accurate method was slow, and the fast method was inaccurate. A new path that breaks down the wall between the two is machine learning interatomic potentials (Machine Learning Interatomic Potentials, MLIP). From here on, this technology will be called MLIP.

This chapter explains step by step what MLIP is and how it has developed. First, it looks at two long-standing methods of atomic calculation. Then it shows how machine learning connected the two. It follows the path from early methods to the latest methods. Finally, it points out what to watch for when using this technology in rocket materials and new materials research. It does not use equations. It explains the ideas in words that readers can picture.

The term materials informatics also appears here for the first time. It means materials information science. It is the field that gathers large amounts of data about materials and analyzes them with computers to find new materials. MLIP is a core tool within that field. It handles calculations at the atomic level.

It is worth saying in advance why this story connects to rockets. Rocket development is hard to separate from the search for new materials. If there is a metal that is lighter and can withstand higher temperatures, rocket performance improves. If each such material is made and tested one by one in a laboratory, it can take years. If candidates are first screened through atomic calculations inside a computer, this time can be reduced. MLIP is the key to making those calculations fast and accurate. After reading this chapter, readers will understand the broad picture of how new materials are born inside a computer.

1. Introduction and Background: Combining Density Functional Theory (DFT) and Machine Learning

What This Section Covers

To handle atoms with a computer, we must know the energy and forces between atoms. There have been two main ways to find these values. One is an accurate method that calculates even the movement

of electrons. The other is a fast method that uses simple formulas set in advance. This section explains what each method does well and what it cannot do well.

It also examines the old problem between the two methods. The accurate method is too slow, and the fast method is often wrong. Machine learning appeared to solve this problem. This section first draws the broad picture.

Several terms in this section will be used throughout the chapter. They are energy, force, and potential energy surface. If these meanings are clear now, the next section becomes easier. There are no difficult equations. The explanation uses only words that readers can picture.

The two key values in calculations about atoms are energy and force. Energy shows how stable an atomic arrangement is. The lower the energy, the more comfortable the arrangement is. Force shows in which direction each atom is pushed. If we know the force, we can calculate where an atom will move in the next moment.

Energy and force are not separate values. Force comes from how sharply energy changes with position. If energy is the height of the ground, force is the slope of that ground. A ball rolls down in the direction of a steep slope. Atoms are also pushed toward lower energy. So if we know the energy map, we can obtain force from the slope of that map. This relationship becomes important again later.

Density functional theory (Density Functional Theory, DFT) is a method that calculates the distribution of electrons to find energy and force. From here on, it will be called DFT. Electrons are tiny charged particles that move around atoms. Where electrons gather, and how much they gather there, determines chemical bonds. DFT calculates this electron distribution using the principles of quantum mechanics. Quantum mechanics is the law of physics for the very small world. For this reason, DFT draws chemical bonds accurately.

The values produced by DFT are not absolute truth. DFT handles the push and pull between electrons through approximation formulas. The results differ slightly depending on how these approximation formulas are chosen. So DFT values are reference values obtained under fixed settings, rather than true values. Machine learning learns these reference values. When DFT is called the answer later, it should be understood in this sense.

The problem is speed. In DFT, the amount of calculation grows quickly as the number of atoms increases. If the number of atoms doubles, the calculation time increases by much more than double. Even with a large supercomputer, the number of atoms that can usually be handled is only in the hundreds. The time that can be calculated is also limited to a very short moment. This is not suitable for watching a large mass for a long time, as in rocket materials.

On the other side is the empirical force field (Empirical Force Field). This method finds forces between atoms using simple formulas set in advance. It does not calculate electrons one by one. Instead, it determines forces using only a few values, such as the distance between atoms. Because the calculation is light, it can handle millions of atoms. It can also observe them over much longer times.

The empirical force field is fast, but it has weaknesses. A force field that assumes fixed bonds cannot describe well the moment when bonds break and new ones form. Reactive force fields and force fields for metals that appeared later can handle bond changes and defects to some extent within a narrow range. Even so, it is difficult to maintain accuracy across broad conditions. Phase transitions, in which atoms change their positions greatly, are also difficult. In hot and harsh environments such as rocket engines, such events occur often. For this reason, errors become large when only a simple force field is used.

The difference between the two methods can be compared to everyday life. DFT is like grading one student's homework carefully. It is accurate, but grading the homework of every student in the school takes a long time. An empirical force field is like quickly checking only whether the answers are right. It is fast, but it makes mistakes on difficult problems. For a long time, scientists had to choose either accuracy or speed.

A machine learning interatomic potential is an attempt to break down this wall of choice. A potential is like a map that tells us the energy of each atomic arrangement. This map is called a potential energy surface. MLIP uses machine learning to learn the reference values calculated by DFT. Machine learning is a computational method that finds rules on its own from many examples. If it learns enough examples, it quickly predicts the energy and forces of new atomic arrangements.

Let us look a little more closely at how MLIP learns. First, DFT calculates the energy and forces of many atomic arrangements. These calculation results are like a workbook with the answers written in it. The machine learning model studies this workbook and learns the rules. It develops its own sense of why a certain arrangement has a certain energy. After training, it can also solve new arrangements that were not in the workbook.

The goal of this method is clear. It is to create calculations that are as accurate as DFT and as fast as an empirical force field. If this goal is achieved, rocket materials can be tested inside a computer on a large scale and over a long time. We can see at the atomic level how metal vibrates at high temperature and how cracks spread. Candidate materials can be screened before real tests. The rest of this chapter is the story of several technologies that aim to achieve this goal.

One point should be made in advance. MLIP does not completely replace DFT. It can learn only when reference values made by DFT exist. The two methods are partners, not competitors. DFT creates reference values, and MLIP spreads those values quickly. This cooperation makes large calculations possible.

2. Classical Descriptor-Based Machine Learning Models

What This Section Covers

To learn atomic energy with machine learning, the shape around an atom must first be converted into numbers. This is because computers handle numbers, not pictures. This list of numbers is called a descriptor (Descriptor). This section looks at three descriptor methods used by early MLIP.

The three methods each have different features. One uses a neural network, one compares how similar two environments are, and one uses deep learning to learn even the descriptors. All three produced good results, but they also revealed limits. Those limits led to the next generation of technologies.

A good descriptor must follow a rule. The same material must always be expressed with the same numbers. A desk remains a desk even if it is turned around. If an atomic structure is rotated or moved as a whole, its position changes. Even so, the material remains the same. A descriptor must not be shaken by such changes. This property is called invariance (Invariance).

Problems arise when invariance is not preserved. The model mistakes the same material for several different materials. Then it must learn the same thing separately many times. Learning becomes slow, and prediction becomes unstable. Early MLIP research focused on putting this invariance into descriptors by hand.

2.1 Neural Networks That Learn Energy Atom by Atom (BPNN)

Behler and Parrinello presented an important idea in 2007. The idea was to divide the total energy into the sum of atomic energies. This method is called atomic energy decomposition. Each atom looks only at its own surroundings and produces a small energy value. When the values of all atoms are added, they become the energy of the whole structure. This neural network is called the Behler-Parrinello Neural Network (Behler-Parrinello Neural Network, BPNN).

A neural network is a computational structure modeled on the connections between cells in the human brain. It receives input and produces an answer through several stages. When trained with many examples, it learns the rules between inputs and answers. BPNN takes the shape around an atom as input and produces the energy of that atom as the answer.

The tool that converts the surroundings of an atom into numbers at this point is the atom-centered symmetry function (Atom-Centered Symmetry Functions, ACSF). This function places one atom at the center. Then it examines neighboring atoms within a certain distance. It records the distances to neighbors and the angles formed by the neighbors as numbers. It is divided into a part that records only distances and a part that also records angles. The numbers made this way do not change even when the structure is rotated or moved.

This method also has meaning for rocket materials research. When a metal becomes hot, atoms vibrate greatly around their positions. The distances and angles to neighbors continue to change. ACSF captures these changes as numbers. So it became a first step in learning the movement of materials according to temperature.

However, the limits were clear. When the number of element types increases, the number of required values grows quickly. Rocket alloys often mix several elements. If there are five elements, the combinations of element pairs increase. Three-atom combinations that handle angles grow even faster. The problem in which the computational burden grows as the number of types increases is called the curse of dimensionality.

Another limit is that descriptors must be designed by hand. A person decides which distances and angles to record and how densely to record them. If this choice is wrong, important information is missed. Creating good descriptors was itself a difficult assignment. Efforts to reduce this burden led to the next methods.

Even so, the meaning left by this method is great. The idea of dividing the total energy by atom still remains alive. Most latest models also follow this decomposition principle. Each atom looks only at its surroundings, produces a value, and then all values are added. Thanks to this method, calculations can be divided and handled even when there are many atoms. The idea of Behler and Parrinello became a foundation stone of MLIP.

2.2 A Potential That Measures the Similarity Between Two Environments (GAP)

Bartok, Csanyi, and others chose a different path in 2010. Instead of a neural network, they directly compared how similar two environments were. This method is called the Gaussian Approximation Potential (Gaussian Approximation Potential, GAP). When it encounters a new atomic environment, it compares it with environments already learned. If there are many similar environments, it uses their energies as references to produce an answer.

The key tool for this comparison is Smooth Overlap of Atomic Positions (SOAP). SOAP draws the area around an atom as a soft cloud, not as hard points. It is like placing one blurred ball around each neighboring atom. If the clouds of two atomic environments overlap a great deal, they are treated as

similar environments. This method is designed to remain stable even under rotation.

There is a reason for drawing atoms as soft clouds. Atoms always vibrate a little. If positions are treated only as hard points, even small vibrations can make the value jump sharply. If they are drawn as clouds, small vibrations pass smoothly. This is why SOAP was strong in handling defects in crystals and disordered materials.

The Gaussian approximation potential achieved high accuracy even with a small amount of data. This is a great help when DFT calculations are expensive. It has been widely used in studies of crystal defects, liquid states, and disordered amorphous materials. Defects and disorder are important topics in rocket materials as well. For that reason, this method became a long-cited benchmark.

Its limits appear when the data grows large. This method compares each new environment with all environments it has learned. As the training data grows, the number of targets for comparison also grows. Then the burden of calculation and storage grows together. It requires too much computation to handle large datasets of hundreds of thousands of items or more, as is common today.

For this reason, this method gave way to other structures in large-scale problems. Deep learning neural networks were more suitable for learning quickly from large data. This is because deep learning becomes more powerful as data increases. Still, the accuracy standard set by the Gaussian approximation potential remains alive today. It is often used as a comparison target when measuring the performance of new models.

The idea left by SOAP also continues. It is the idea of viewing the area around an atom as a smooth density. This idea entered many later models. Treating atoms as soft distributions instead of hard points makes calculation stable. This is because values do not jump from small vibrations. This stability helps in high-temperature calculations, where atoms vibrate strongly, as in rocket materials.

2.3 Deep Learning That Learns Even the Descriptors by Itself (DeePMD)

Zhang, Car, and others brought deep learning fully into atomic calculations in 2018. This method is called Deep Potential Molecular Dynamics (DeePMD). Molecular dynamics is a simulation that follows the movement of atoms over time. DeePMD calculates the forces used in this simulation with deep learning.

Earlier methods used descriptors designed by people. DeePMD greatly reduced this burden. A neural network changes the coordinates around an atom into a useful form by itself. A person does not have to choose distances and angles one by one. At the same time, it preserves properties that remain stable under rotation and translation.

The term deep learning also needs a little more explanation here. Deep learning means a neural network built with many layers. When layers are stacked deeply, the network can learn complex rules. A shallow neural network learns only simple rules. A deep neural network captures tangled relationships between atoms. In return, it takes a great deal of data and computation to train. The word deep comes from the depth of the layers.

The strength of this method is scale. In deep learning, the computational burden does not grow sharply even when there are many atoms. For this reason, DeePMD can handle the movement of millions of atoms. It has been used for many materials, such as water, metals, and semiconductors. It was powerful in tasks that observe large systems for a long time.

DeePMD is also meaningful for rocket materials research. To see how a mass of metal changes at high temperature, many atoms are needed. If only a small piece is examined, the overall behavior can be

missed. DeePMD can handle large pieces, so it fit this kind of research.

Still, a limit shared by the early descriptor-based methods remained. These methods look separately only at neighbors within a fixed distance. To include the influence of distant atoms, the viewing distance must be increased. Then the number of neighbors grows greatly, and the computational burden increases. The structure made it difficult to include long-range interactions.

Another door opened by DeePMD is its combination with large computing tools. This method led to studies that handle tens of billions of atoms on large supercomputers. It is also used to calculate alloys made by mixing several elements at large scale. Research on high-temperature alloys for rockets requires this kind of large calculation. This is because behavior missed in small pieces can be seen in large systems.

This limit called for a new idea. The idea was to move beyond placing a single atom at the center and looking around it. What if atoms could exchange information with one another? If information is exchanged several times, the influence of distant atoms can be passed along. This idea leads to the graph neural networks in the next section.

3. Modern Equivariant Graph Neural Networks (GNNs)

This section covers the method that has become the mainstream of MLIP today. It is the Graph Neural Network (GNN). This method treats atoms as points and neighbor relationships between atoms as lines. Here, a line is not the chemical bond itself. It means that neighbors within a certain distance have been connected. These lines can also change during calculation. It is similar to a subway map that draws stations as points and the connections between stations as lines. The model calculates energy by exchanging information along points and lines.

The key word in this section is equivariance. It is a property paired with the invariance discussed in the previous section. First, I explain what equivariance is and why it is needed in atomic calculations. Then I review representative models that include this property, from early models to recent ones. Finally, I cover large models that have already learned from broad data.

A graph neural network calculates by exchanging information. Each atom sends messages to its neighbors. It gathers messages from its neighbors and updates its own state. This process is repeated several times. After one exchange, it knows the atoms right next to it. After two exchanges, it knows the neighbors of its neighbors. As the process repeats, the influence of more distant atoms is passed along. This method is called message passing.

This structure solves some of the limits of the previous section. There is no need to force the viewing distance to grow. This is because information beyond immediate neighbors arrives through repeated message exchanges. However, if the number of layers is fixed, the range that information can reach is also fixed. Effects such as very long-range electrostatics or dispersion forces are hard to capture fully with this method alone. Such effects must be added as separately designed long-range terms. Even so, graph neural networks were a major advance in efficiently capturing a wide range of interactions.

3.1 Symmetry That Works Even When Rotated or Reflected

There are symmetries that hold in the atomic world. Even if a structure is moved or rotated as a whole, the laws of physics remain the same. The collection of translations, rotations, and reflections is called E(3). The collection of rotations among them is called SO(3). The names are difficult, but the meaning is simple. They are groups of translations, rotations, and mirror-like reflections. Rotation and reflection are different motions. Rotation turns the shape as it is, while reflection creates a mirror image with left

and right reversed.

The invariance discussed earlier is the property that a value does not change under these motions. Energy is an invariant quantity. Even if a structure is rotated, the energy of that structure must be the same. But not every value is invariant. Force is a value with direction. If the structure rotates, the direction of the force must rotate with it.

This property is equivariance. Equivariance means that when the input rotates, the output rotates in the same way. Force is an equivariant quantity. If a structure rotates to the right, the force on each atom also rotates to the right. Atomic calculations match physical laws only when both invariance and equivariance are preserved.

We can separate invariant and equivariant quantities in one sentence. If a value stays the same after rotation, it is an invariant quantity. If the value rotates together with the input, it is an equivariant quantity. Energy is a single number without direction, so it is invariant. Values such as temperature and density also have no direction, so they are invariant. Values with direction, such as force and velocity, are equivariant. A model must know which side each value belongs to and handle it accordingly.

Here I continue the point left aside in the previous section. I said that force is the slope of the energy map. A good model does not fit force separately. It first finds the energy and then obtains the force from the slope of that energy map. There is a reason for keeping this order. If the model outputs force as it pleases, it can produce forces that do not conserve energy. If a long simulation is run with such forces, energy that was not there may appear or disappear by itself. Then the simulation gradually inflates or collapses. If force is obtained from the slope of energy, this mismatch does not occur. In physics, a force that comes from energy in this way is called a conservative force.

Putting these two properties into the model from the beginning brings a major benefit. The model does not have to learn the same material separately from many directions. A rule learned once automatically works in every direction. So the model learns stably even with less data. Since DFT calculations are expensive, this saving of data is a major advantage.

In rocket materials, many events depend on direction. Cracks spread in specific directions. Deformation that slips along crystal planes also has direction. Oxidation reactions also occur differently depending on the direction of the surface. If the model cannot handle force directions accurately, it draws these events incorrectly. Models that preserve equivariance have an advantage here.

Let us move this symmetry into a simpler picture. Suppose we take several pictures of the same toy block structure. A picture taken from the front and a picture taken from the side look different. Still, the block structure itself is the same. People understand this quickly. But to a computer, the two pictures are different inputs. A model without symmetry may treat the two pictures as different materials. A model with symmetry treats the two pictures as the same material. This difference greatly divides learning efficiency.

In short, symmetry is not decoration but structure. A model that preserves symmetry saves data and stabilizes predictions. A model that breaks symmetry can produce physically strange answers. This is why modern MLIP research focuses on how to include this property well.

3.2 From Methods That Saw Only Distance to Methods That Exchange Direction Too

SchNet is an early graph neural network that appeared in 2018. This model exchanges only distances between atoms as messages. Distance is a value without direction. So SchNet is an invariant graph neural network that handles only invariant information. Its structure is clean, and calculation is fast.

A method that uses only distance has a weakness. It is hard to directly capture bonds where direction matters. Even at the same distance, a bond differs if the angle differs. If this difference is missed, errors appear. DimeNet addressed this point by adding angles between atoms to the messages. This continued the effort to include directional information.

Batzner, Kozinsky, and others took a large step in 2022. Their model is called NequIP. This model exchanges direction itself as messages, beyond distance and angle. It passes and mixes values that have direction as they are. This is equivariant message passing. It captures the direction of force and the three-dimensional shape around atoms much more faithfully.

A major advantage of NequIP is data efficiency. Because it includes directional information according to principle, it learns well even from few examples. The published paper states that it achieved high accuracy with far less training data than existing models. In research where each DFT calculation is expensive, this saving is a practical benefit.

Let us note again why data efficiency matters. A single DFT calculation takes a great deal of computer time. If tens of thousands of calculations are needed for training, the computational cost grows sharply. NequIP can reach similar accuracy with far fewer cases. Then more difficult materials can be studied with the same budget. New materials for rockets often have little data. A model with high data efficiency is useful in such situations.

This advance is directly meaningful for rocket material calculations. Cracks, defects, and phase transitions are all tied to direction. Equivariant models draw these directional events more accurately. They can learn behavior in extreme environments even from limited DFT data. For this reason, after NequIP, equivariant methods came close to becoming the standard in this field.

Around the same time as NequIP, research to improve efficiency also continued. These were attempts to handle directional information while making calculation lighter. A recent study proposed a more efficient equivariant model and was published in a journal in 2025. The direction is to reduce computational burden while preserving accuracy. As these efforts accumulate, equivariant methods have become practical for real research.

There is, however, a cost. Mixing directional information according to the principle requires heavy mathematical operations. This operation is called a tensor product. It is accurate, but it brings a large burden in computation and memory. This burden later becomes a problem again. This discussion continues in Section 5.

3.3 A Method for Capturing Relationships Among Many Atoms at Once (MACE)

Batatia, Csanyi, and others introduced MACE in 2022. This model combines an equivariant graph neural network with an old physical theory called many-body expansion. Many-body means effects created by several atoms together. In a material, the relationship between one atom and another is not the only important thing. Arrangements formed by three atoms or four atoms together also change properties.

Earlier graph neural networks exchanged messages many times to capture these many-body effects. If the number of exchanges grows, the computational burden increases and training slows down. MACE found another path. It multiplies neighbor information collected once to create the effects of many atoms at the same time. So it captures effects involving four or more atoms with only a small number of message-passing steps.

The benefit of this design is that it achieves both accuracy and speed. Fewer message-passing steps make the calculation lighter. At the same time, it preserves accuracy because it does not miss many-body

effects. This advantage grows in large crystals or long molecular dynamics simulations. For this reason, MACE is used as a strong benchmark model in many studies.

MACE is also used in research on rocket materials. A recent research team studied a high-entropy alloy made by mixing aluminum, titanium, chromium, molybdenum, and tungsten. This alloy is a material that mixes several elements in similar amounts. The researchers used MACE to build an interatomic potential for this alloy. They then examined phase stability as temperature changed. This study was published in a physics journal in 2026. It is an example of MACE being used to design alloys for high temperatures. However, what this study showed was a calculation of phase stability. It did not verify cracks or lifetime in real parts through this one study.

We must also see the limits. MACE also uses tensor products when mixing directional information. So if the precision of direction, called angular momentum, is increased, the computational burden grows quickly. In very large systems or very long simulations, this burden remains a problem. Later studies try to reduce this burden.

Let us move the many-body effect into a simple example. If we look at only two students in a classroom, we know only whether they get along. When three students gather, the mood changes. When four students gather, another flow appears. Atoms are the same. The relationship between two atoms alone cannot capture all the properties of a material. The arrangement created by several atoms together changes the properties. MACE captures these relationships among many atoms in a small number of steps.

Another task is situations outside the learned range. No matter how accurate MACE is, it can be wrong for unfamiliar arrangements that were not in training. Rocket environments are always extreme and contain many unfamiliar events. So a more accurate model needs a device that lets it report moments when it does not know. That device is active learning, discussed in Section 4.

3.4 Large Models That Learn Broad Materials in Advance (CHGNet, MatterSim)

The models discussed above are usually trained for one material or a narrow range. When they meet a new material, they must be trained again with data suited to it. The movement to reduce this burden is foundation potentials. A foundation potential is a large model that has learned very broad materials data in advance. It is like a cook who has learned many kinds of dishes and can respond well to new ingredients.

CHGNet is a representative foundation potential released in 2023. This model was trained on a large computational database called the Materials Project. The Materials Project is a public resource that gathers DFT calculation results for many materials. CHGNet also learns clues about charge and magnetic states. So it tracks lithium movement in battery materials and phase changes in transition metal oxides well.

There is a reason CHGNet learns charge and magnetic states. In battery materials, lithium moves in and out, and charge moves with it. In transition metal oxides, the magnetic state of atoms changes properties. If these changes are missed, the behavior of the material is drawn incorrectly. CHGNet includes these clues in its representation of the atomic environment and covers a wider range of materials. However, it is not an electrochemical model that tracks charge one by one as a conserved value. It is better to understand it as a model that captures clues about magnetic states through learned representations. This ability is useful because oxides and transition metals are often used in rocket materials.

MatterSim is another foundation potential released in 2024. This model targets a very wide range of temperatures and pressures. The public paper states that it covers from 0 kelvin to 5000 kelvin and from 0 gigapascals to 1000 gigapascals. Kelvin is a unit of absolute temperature, and gigapascal is a unit of pressure. The hot conditions of rocket engines fall within this range. However, overlap with the training range and actual accuracy are different matters. 1000 gigapascals is far higher than the pressure in a rocket combustion chamber. Rocket-specific situations, such as combustion gases or oxidized surfaces, must be verified separately. For this reason, this model is drawing interest as a starting point for extreme-materials research. The public paper states that the model saved a great deal of data. Because it learned broad conditions in advance, it can respond quickly to new materials.

Recently, a new foundation potential has also appeared that improves both accuracy and speed. It is a model called GRACE. This model was built with a framework called graph atomic cluster expansion. The researchers trained this model on part of one of the largest materials datasets. The published study states that this model established a new balance between accuracy and efficiency. This study was published in a journal on computational materials in 2026. It showed that a large model can be refined well and adapted to a specific task while preserving accuracy.

Foundation potentials are changing the way research is done. In the past, models were built from the beginning for each material. Now, researchers first use a large model and refine only the necessary parts. This way of making small adjustments is called fine-tuning. Even when a new alloy appears, the model can respond quickly with little data.

Efforts to compare these large models fairly are also growing. As many foundation potentials have appeared, it has become hard to judge which one is better. So public evaluation arenas have emerged to measure performance under the same conditions. These are benchmark platforms built by several research teams. Such arenas compare accuracy, speed, and stability together. Users can choose a model suited to their own problem.

Still, there are points to be careful about. A foundation potential is not automatically correct in every situation. Errors can grow in alloys or surface reactions that were rare in the training data. So a good model must be accompanied by verification. Unfamiliar structures must be calculated again with DFT and checked. This checking procedure is the subject of the next section.

4. Active Learning Loop and Error and Uncertainty Evaluation

No matter how good a model is, it can be wrong in situations it has not learned. The problem arises when the model is wrong but still gives a confident answer. Then it is hard for people to notice the error. This section deals with methods that let a model report by itself when it does not know.

That method is active learning. Active learning is a circular procedure in which the model selects structures it is not confident about and calculates them accurately again. We will examine how this loop works and how lack of confidence is measured as a number. We will also see why this procedure is a safety device in extreme and unfamiliar environments such as rocket materials.

Let us move interpolation and extrapolation into an everyday example. If we know the temperatures from the past several years, we can estimate today’s temperature. Estimating within a range we have already experienced is interpolation. But if an unprecedented heat wave arrives, the estimate goes wrong. Moving into a range we have not experienced is extrapolation. Models are the same. They work well within the range they have experienced and become unstable outside it.

The place where a model predicts safely is inside the training data. This area is called the interpolation region. Inside the area surrounded by the training data, predictions are reliable. But the situation

changes outside it. This outside area is called the extrapolation region. In the extrapolation region, the model can give an answer that looks plausible but is wrong.

Errors in the extrapolation region are dangerous. During a simulation, atoms may overlap or a structure may collapse strangely. If this happens quietly, the whole result becomes unreliable. Rocket environments are always extreme, so extrapolation happens often. High temperature, strong collisions, oxidation, and cracks are rare events in training data. So a device that catches these moments is essential.

The active learning loop runs through four steps. First, the model is trained on a small DFT dataset. Next, this model is used to calculate atomic motion quickly. During the calculation, structures that make the model uncertain are marked. Only those structures are selected and calculated accurately again with DFT. The new results are added to the data, and the model is corrected. If these steps are repeated, the model gradually learns a wider range of situations. It is like checking a map again at every fork on an unfamiliar road.

The core of this loop is measuring lack of confidence as a number. This number is called uncertainty. There are several ways to measure uncertainty. The following subsections examine representative methods.

4.1 Measuring the Moment When a Model Does Not Know as a Number (UQ)

The first method for measuring uncertainty is to use an extrapolation grade. A model called the Moment Tensor Potential (MTP) uses this method. This model measures, as a number, how different the current atomic arrangement is from the training data. This number is called the extrapolation grade. If the extrapolation grade is low, the arrangement is familiar and the prediction is safe. If the extrapolation grade is high, the arrangement is unfamiliar and must be calculated again.

The second method is to compare the answers of several models. Several models are trained slightly differently on the same data. This group is called an ensemble. If the answers of these models are similar for a new structure, they are generally reliable. If the answers differ greatly, that structure is a warning sign. However, models learned from the same data may all be wrong together for an unfamiliar input. So agreement within an ensemble is only a reference indicator, not a complete guarantee. It is like asking several people the same question and seeing a hard problem when their answers differ.

The third method is to make the model itself provide an error range. Recent research has made progress in this direction. One research team attached a method called evidential deep learning to an interatomic potential. This method outputs both the predicted value and how much that value can be trusted. It is computationally lighter because several models do not need to be run separately. This study was published in a journal in 2025.

Research is also emerging that connects uncertainty more deeply with active learning. One method makes the simulation seek out uncertain places by itself. It guides atomic motion toward arrangements about which the model is uncertain. Then it collects structures with high learning value more quickly. This is a way to fill unfamiliar situations efficiently.

This procedure is a framework for maintaining safety in rocket materials research. In extreme environments, unexpected arrangements always appear. Without an uncertainty device, there is a risk of believing a wrong result as true. With an uncertainty device, the calculation stops at dangerous moments and checks them. That makes the results more trustworthy.

Methods for handling uncertainty are still emerging. Recently, some studies have even addressed cases where the model itself is designed incorrectly. If the model's basic assumptions are wrong, the uncertainty also comes out wrong. A study pointing out this problem was published in a journal in 2025. This means that measuring uncertainty is a difficult task. Even so, progress in this direction supports the safety of extreme-materials calculations.

Another benefit of this circular procedure is that it saves data. DFT calculations are expensive. So calculating every structure without selection causes great waste. Active learning selects only structures that have high learning value and calculates them. With the same budget, it gathers more useful data. As a result, a better model can be obtained more quickly.

This method can be compared to studying for a test. Solving problems you already know again wastes time. If you select only the problems you got wrong and solve them, your skill improves quickly. Active learning also selects and fills only the model's weak parts. It is a way to gain large progress with limited computation. In rocket materials research, the computational budget is always tight. So this saving becomes a real strength in development.

However, accurately measuring uncertainty is itself a difficult task. Some methods carry a heavy computational burden, while others have lower accuracy. If the model is poorly designed from the start, even its uncertainty estimates can be wrong. Research on this problem continues today. Uncertainty quantification is not yet a finished technology. It is still a developing field.

5. Geometric Neural Networks That Remove Tensor Products: Computational Efficiency and Smooth Curves

Equivariant models are accurate, but they carry a heavy computational load. A major source of that load is the tensor product discussed earlier. This section explains why tensor products are burdensome and reviews recent research that aims to reduce or remove them.

Another important issue here is smoothness. The result of an atomic calculation is a potential energy surface, an energy map. If this surface is rough, force calculations jump and the simulation breaks down. This section explains how new methods try to achieve both speed and smoothness.

A tensor product is a mathematical operation that mixes directional values according to principle. Equivariant models use this operation to handle directional information accurately. The problem is that the burden of this operation grows quickly as directional precision increases. Computation time and memory use both grow. In simulations of large systems or long time periods, this burden becomes a wall.

We can picture tensor products in simpler terms. When mixing information that has direction, the rules are strict. Two directional values cannot simply be added or multiplied. They must be paired and mixed in a way that follows physical laws. The tensor product performs this difficult pairing accurately. It is accurate, but the number of pairs grows quickly with directional precision. So even a small increase in precision increases the computational burden.

Research to overcome this wall is moving in several directions. One direction is to make special programs that compute tensor products faster. A recent study greatly accelerated molecular dynamics calculations in this way. This study was published in a journal in 2026. This path keeps the tensor product itself but optimizes the computation.

5.1 A New Design That Does Not Use Tensor Products at All

Another direction is to avoid tensor products altogether. A study called Geodite took this path. Its title is Tensor-Product-Free Equivariant Interatomic Potential. It replaces the costly operation of mixing directions with another method. At the same time, it preserves rotational symmetry according to principle.

Let us briefly review how this method replaces tensor products. Instead of using the costly pairing operation, it builds local coordinate systems. It sets up a reference frame of directions around each atom. When information is handled on this frame, direction can be preserved with lighter computation. Symmetry is maintained, while the burden is reduced.

The study reports two achievements. One is speed. The public abstract says it is three to five times faster than models with similar accuracy. The other is the breadth of accuracy. It achieved accuracy comparable to leading models in materials stability prediction, thermal conductivity, lattice vibrations, and nanosecond-scale molecular dynamics. In other words, it greatly increased speed while preserving accuracy. This study is early research that has not undergone peer review. So it should be viewed not as a settled standard, but as a promising candidate technology.

Removing tensor products improves more than speed. It can also help make curves smoother. When directional precision is truncated in tensor products, fine ripples can appear on the surface. These ripples can make force calculations jump. Designs that do not use tensor products are made to avoid such ripples. The curve remains smooth even in regions where bonds form and break.

This smoothness is important in rocket materials calculations. At high temperatures, atoms vibrate strongly and bonds often break. If the curve is rough, forces jump in these regions and the simulation collapses. A smooth curve stabilizes the calculation even under extreme conditions. That is why smoothness is as important as accuracy.

Other studies pursue similar goals. One study spreads directional information over a spherical grid and mixes it with a lightweight neural network. This method is also faster than tensor products while preserving expressive power. Another study smoothly changes the cutoff distance according to the situation. These studies are all still at an early stage. Several teams are testing different paths toward the same goal.

Let us look a little further at why smoothness matters so much. Forces come from the slope of the energy surface. If the surface is smooth, the slope also changes smoothly. If the surface has ripples, the slope jumps suddenly. A jumping force pushes atoms in the wrong direction and ruins the simulation. So a smooth surface is needed as much as accurate energy. Even when the visible error is small, a rough surface can create major problems in long calculations.

The meaning of this trend is clear. The goal is to keep the accuracy of equivariant models while reducing their computational burden. If computation becomes lighter, larger systems can be observed for longer periods. This directly helps when studying large materials such as rocket parts over long times. However, evidence is still accumulating, so validation must be examined together.

6. Structural Pruning for Electron Density Modeling and Lightweight Hardware Operation

Large models are accurate, but they are burdensome and require large computing equipment. This section discusses methods for making large models lighter while preserving physical laws. It examines two ideas. One is pruning, which removes less important computational pathways. The other is adding clues about the distribution of electrons to the model. The key point of this section is to make the model lighter without cutting carelessly. Atomic calculations must preserve many physical properties. These

include symmetry, force direction, and consistency between energy and forces. This section explains how to reduce a model while preserving these properties.

Pruning is a method that reduces less important pathways in a neural network. In a neural network, pathways that carry information are called channels. Not all channels are equally important. Reducing less used channels makes computation lighter. It is like taking rarely used items out of a bag to make the bag lighter.

The problem is that cutting arbitrary channels can disturb physical laws. In an equivariant model, channels are tied to direction. If channels in the same directional group are cut apart, the direction of force becomes strange. One possible misunderstanding should be corrected here. If a model first produces energy and then obtains force from its gradient, reducing channels still preserves the relationship between energy and force itself. Pruning does not immediately break this relationship. Instead, it can degrade the precision of directional handling and prediction accuracy. That is why pruning must be done carefully.

6.1 Structural Pruning That Preserves Physical Consistency

Structured pruning is a method that cuts channels by directional group. It does not cut them one by one in a scattered way, but handles them as groups. This is necessary to keep the balance of directions intact. The ratio between scalar channels and vector channels is also managed. A scalar is a value without direction, and a vector is a value with direction. This ratio must be preserved for symmetry and force direction to remain intact.

We can picture why channels must be cut as groups. A compass needle only makes sense when it can indicate north, south, east, and west together. If only north and south remain while east and west are cut away, direction becomes distorted. Directional channels in an equivariant model are the same. The values within one group must stay together to indicate direction correctly. If only some of them are cut apart, the direction of force becomes misaligned. That is why pruning must always use groups as the unit.

Which channels to cut is decided by importance. We measure how much each channel contributes to the result. Channels with small contributions are reduced first. Even then, the directional group is the unit. The values within one group are assessed together. In this way, the framework that carries information is preserved while only excess weight is removed.

A recent research team applied this method to foundation potentials. The study is titled Dense SO3 Equivariant Atom Models via Structured Pruning. The public abstract says it reduced the number of parameters by several times. Parameters are the values that the model learns. The computational cost of pretraining was also reduced by several times. At the same time, the abstract says that after fine-tuning, errors in energy and forces actually decreased. This is early research released in 2026. Because it has not undergone peer review, it should not be treated as a settled result.

The goal of this research is to run precise calculations even on lightweight equipment. Large models usually require several pieces of high-performance equipment. If a model is reduced well, similar calculations can be performed on smaller equipment. If the barrier to research is lowered, more researchers can take part in materials calculations. However, which types of equipment will be sufficient still needs further confirmation.

This direction also has meaning for rocket materials research. To simulate extreme environments over long periods, calculations must be repeated many times. If the model is lightweight, the same equipment can run more experiments. This helps screen candidate materials quickly. The core task is to gain speed

while preserving accuracy.

There is another method with a goal similar to pruning. It transfers the knowledge of a large model to a small model. This method is called knowledge distillation. A well-trained large model becomes the teacher for the small model. The small model learns by imitating the teacher's answers. Then it can achieve performance close to the large model despite its smaller size. Pruning and knowledge distillation are both ways to move a large model into a lighter form.

The point to watch is that physical validation must not be skipped. A lightweight model obtained through pruning must also be checked. Consistency between energy and forces, force direction, symmetry, and uncertainty indicators must be examined together. Only when these properties are preserved can a lightweight model be trusted. A model that is only fast is risky in extreme calculations.

6.2 Methods for Adding Electron Density Clues

When a model is reduced, its expressive power may also decrease. One idea for filling this loss is to add electron density. Electron density is a map that shows how electrons are spread through space. It is the basic physical quantity originally handled by DFT. The idea is to add this clue directly into the model to supplement missing expressive power. A method that mixes different kinds of information in this way is called a hybrid approach.

Let us briefly review the types of bonds here. An ionic bond is a bond formed by giving and receiving electrons. Salt is made of this kind of bond. A metallic bond is a bond in which electrons flow freely among many atoms. This bond allows iron to conduct electricity well. A covalent bond is a bond in which atoms share electrons. In this way, each type of bond has a different electron shape.

Electron density information helps distinguish bond types. There are many kinds of bonds in materials. Ionic bonds, metallic bonds, and covalent bonds are different from one another. By seeing where electrons gather and how much they gather, we can identify these differences. Adding this clue to the model can help capture the nature of bonds more accurately.

This method is also expected to help recover long-range effects. The influence of electrons extends beyond close neighbors. When a model is made lightweight, it is easy to miss these distant effects. The expectation is that electron density clues can capture some of these distant influences if used well. However, adding electron density does not automatically solve long-range electrostatics and dispersion forces. Research is still accumulating on what density to add, how to add it, and how to validate it. So this direction is better understood as a proposal, not as a settled solution.

Rocket materials often contain a complex mixture of bond types. Metals and ceramics are used together, and oxide layers form. To handle such materials, the nature of the bonds must be distinguished accurately. Methods that add electron density can help with this distinction and strengthen extreme materials calculations.

The possibility of large-scale calculations also opens up. A recent research team simulated an alloy made from multiple elements at the scale of one million atoms. It handled a large system using an interatomic potential for multiple components. This is early research released in 2026. High-entropy alloys are important candidates for high-temperature rocket materials. The larger the system that can be handled, the closer the observed behavior becomes to that of real parts.

Let us view the discussion in this section against the reality of rocket development. Many rocket companies and research institutes do not have abundant computing resources. If large models can run only on several pieces of equipment, the barrier is high. If models become lighter, more teams can begin atomic calculations. They can quickly screen candidate materials and reduce the number of targets for

real testing. This cycle between computation and experiment shortens the development period.

The two ideas in this section point toward the same goal. The goal is to make computation lighter while preserving accuracy. Pruning removes excess weight, and electron density clues fill missing expressive power. Used together, they can make models useful even on smaller equipment. This direction is still developing, and validation is still accumulating.

Summary

This chapter was about the rules for handling atoms on a computer. These rules are interatomic potentials. For a long time, scientists had to choose between an accurate method and a fast method. DFT was accurate but slow, while empirical force fields were fast but inaccurate. A machine learning interatomic potential (MLIP) is a tool that tries to break down the wall between the two.

Let us briefly look back on this journey. At first, people converted the surroundings of atoms into numbers. Next, models learned those numbers by themselves. Then atoms were made to exchange information with one another. Now large models that have already learned broad classes of materials have appeared. At each step, less depended on human hands and more depended on the power of data. This flow is the main line of MLIP development.

Early MLIPs converted the shapes around atoms into numbers designed by people. This descriptor approach included BPNN, Gaussian approximation potentials, and DeePMD. They achieved major results, but they also revealed limits. When the number of elements increased, the computational burden grew, and it was hard to include long-range effects.

The next generation is equivariant graph neural networks. This method sees atoms as points and neighbor relationships as lines, then exchanges information. It saves data by placing symmetry for rotation and translation inside the model. From SchNet to NequIP and MACE, both accuracy and efficiency improved. Foundation potentials such as CHGNet, MatterSim, and GRACE have become large models that have already learned broad classes of materials.

Even accurate models have weaknesses. They can be wrong in unfamiliar situations they have not learned. That is why active learning and uncertainty quantification are needed, so the model can signal when it does not know. Rocket environments are always extreme and often involve unfamiliar events. With this safeguard, calculations for extreme materials can be trusted.

Recent research pursues speed, smoothness, and lightness at the same time. It reduces or removes tensor products to make computation faster. It shrinks models through structural pruning and adds expressive power with clues from electron density. Many of these methods are still early studies that have not yet undergone peer review. They are promising, but it is too early to treat them as established standards.

For general readers of this chapter, five points are worth remembering. Atomic calculations have long involved a trade-off between accuracy and speed. Machine learning has built a bridge that eases this trade-off. Preserving symmetry is central to a good model. No matter how good a model is, it must tell us when it does not know. And much of the latest technology is still early research under validation.

These five points also apply to the chapters that follow. Creating crystal structures, designing alloys, and working with ceramics all stand on these atomic calculations. Without accurate atomic calculations, the designs built on top of them also become unstable. That is why the content of this chapter becomes the foundation stone of the whole book.

One balanced view should also be kept in mind. MLIP is a powerful tool, but it is not universal. If the training data is poor, the predictions also become poor. If validation is neglected, plausible-looking errors can mislead us. So this technology does not replace human judgment or experiments. Computation narrows the candidates, and experiments make the final confirmation. This cooperation is the way to bring new materials safely into the world.

All these technologies converge on one goal. It is to test rocket materials and new materials inside computers at large scale, over long periods, accurately and safely. If candidates are screened before actual parts are made, money and time can be saved. From the next chapter, we move from these prepared atomic calculations to the story of directly creating new materials.

Related Supporting Materials

Kohn, W., & Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. Physical Review, 140, A1133-A1138. https://doi.org/10.1103/PhysRev.140.A1133

Behler, J., & Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. Physical Review Letters, 98, 146401. https://doi.org/10.1103/PhysRevLett.98.146401

Bartók, A. P., Payne, M. C., Kondor, R., & Csányi, G. (2010). Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Physical Review Letters, 104, 136403. https://doi.org/10.1103/PhysRevLett.104.136403

Bartók, A. P., Kondor, R., & Csányi, G. (2013). On representing chemical environments. Physical Review B, 87, 184115. https://doi.org/10.1103/PhysRevB.87.184115

Zhang, L., Han, J., Wang, H., Car, R., & E, W. (2018). Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. Physical Review Letters, 120, 143001. https://doi.org/10.1103/PhysRevLett.120.143001

Schütt, K. T., Sauceda, H. E., Kindermans, P. J., Tkatchenko, A., & Müller, K. R. (2018). SchNet: A deep learning architecture for molecules and materials. Journal of Chemical Physics, 148, 241722. https://doi.org/10.1063/1.5019779

Gasteiger, J., Groß, J., & Günnemann, S. (2020). Directional message passing for molecular graphs (DimeNet). International Conference on Learning Representations. https://arxiv.org/abs/2003.03123

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., & Kozinsky, B. (2022). E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials (NequIP). Nature Communications, 13, 2453. https://doi.org/10.1038/s41467-022-29939-5

Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C., & Csányi, G. (2022). MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. Advances in Neural Information Processing Systems, 35, 11423-11436. https://arxiv.org/abs/2206.07697

Batatia, I., Batzner, S., Kovács, D. P., Musaelian, A., Simm, G. N. C., Drautz, R., Ortner, C., Kozinsky, B., & Csányi, G. (2025). The design space of E(3)-equivariant atom-centred interatomic potentials. Nature Machine Intelligence, 7, 56-67. https://doi.org/10.1038/s42256-024-00956-x

Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C. J., & Ceder, G. (2023). CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. Nature Machine Intelligence, 5, 1031-1041. https://doi.org/10.1038/s42256-023-00716-3

Yang, H., Hu, C., Zhou, Y., Liu, X., Shi, Y., Li, J., et al. (2024). MatterSim: A deep learning atomistic model across elements, temperatures and pressures. arXiv:2405.04967. https://doi.org/10.48550/arXiv.2405.04967

Deringer, V. L., Caro, M. A., & Csányi, G. (2021). Machine learning interatomic potentials as emerging tools for materials science. Advanced Materials, 33, 1902765. https://doi.org/10.1002/adma.201902765

Shapeev, A. V. (2016). Moment tensor potentials: A class of systematically improvable interatomic potentials. Multiscale Modeling and Simulation, 14, 1153-1173. https://doi.org/10.1137/15M1054183

Podryabinkin, E. V., & Shapeev, A. V. (2017). Active learning of linearly parametrized interatomic potentials. Computational Materials Science, 140, 171-180. https://doi.org/10.1016/j.commatsci.2017.08.031

Lysogorskiy, Y., Bochkarev, A., & Drautz, R. (2026). Graph atomic cluster expansion for foundational machine learning interatomic potentials (GRACE). npj Computational Materials, 12, 114. https://doi.org/10.1038/s41524-026-01979-1

Reschützegger, T., Aykent, S., Perin, G. J., Nunes, B. H., Cipcigan, F., Ferreira, R. N. B., Steiner, M., & Thiemann, F. L. (2026). Equivariant interatomic potentials without tensor products (Geodite). arXiv:2601.15492. Preprint, not yet peer reviewed. https://arxiv.org/abs/2601.15492

Wang, C., Hu, S., Tan, G., & Jia, W. (2026). Compact SO(3) equivariant atomistic foundation models via structural pruning. arXiv:2605.08885. Preprint, not yet peer reviewed. https://arxiv.org/abs/2605.08885

Tan, C. W., Descoteaux, M. L., Kotak, M., de Miranda Nascimento, G., Kavanagh, S. R., Zichi, L., Wang, M., Saluja, A., Hu, Y. R., Smidt, T. E., Johansson, A., Witt, W. C., Kozinsky, B., & Musaelian, A. (2026). High-performance training and inference for deep equivariant interatomic potentials. Digital Discovery, 5, 1558-1567. https://doi.org/10.1039/d5dd00423c

Miklaucic, N., Wei, L., Dong, R., Fu, N., Omee, S. S., Li, Q., Dey, S., Fung, V., & Hu, J. (2025). Facet: Highly efficient E(3)-equivariant networks for interatomic potentials. arXiv:2509.08418. Preprint, not yet peer reviewed. https://arxiv.org/abs/2509.08418

Han, K., Cong, H., Deng, B., & Barati Farimani, A. (2026). Smooth dynamic cutoffs for machine learning interatomic potentials. arXiv:2601.21147. Preprint, not yet peer reviewed. https://arxiv.org/abs/2601.21147

Xu, H., Cui, T., Tang, C., Ma, J., Zhou, D., Li, Y., Gao, X., Gong, X., Ouyang, W., Zhang, S., & Su, M. (2025). Evidential deep learning for interatomic potentials. Nature Communications, 16, 937. https://doi.org/10.1038/s41467-025-67663-y

Kulichenko, M., Barros, K., Lubbers, N., Li, Y. W., Messerly, R., Tretiak, S., Smith, J. S., & Nebgen, B. (2023). Uncertainty-driven dynamics for active learning of interatomic potentials. Nature Computational Science, 3, 230-239. https://doi.org/10.1038/s43588-023-00406-5

Perez, D., Subramanyam, A. P. A., Maliyov, I., & Swinburne, T. D. (2025). Uncertainty quantification for misspecified machine learned interatomic potentials. npj Computational Materials, 11, 263. https://doi.org/10.1038/s41524-025-01758-4

Yang, Z., Wang, X., Li, Y., Lv, Q., Chen, C. Y.-C., & Shen, L. (2025). Efficient equivariant model for machine learning interatomic potentials. npj Computational Materials, 11, 49. https://doi.org/10.1038/s41524-025-01535-3

Shuang, F., Ying, P., Liu, K., Wei, Z., Liu, F., Fan, Z., Jiang, M., & Dey, P. (2026). Benchmarking chemically scalable machine-learning interatomic potentials for large-scale simulations of multicomponent alloys. arXiv:2604.01642. Preprint, not yet peer reviewed. https://arxiv.org/abs/2604.01642

Arnold, J. E., Woodgate, C. D., Hafizi, R., Harris, M. J., Mottura, A., García Fuentes, G., & Gurrutxaga-Lerma, B. (2026). Thermodynamics and phase stability of the AlTiCrMoW high-entropy alloy simulated using machine-learned interatomic potentials. Physical Review Materials, 10, 073605. https://doi.org/10.1103/5tz1-zg9z

Chiang, Y., Kreiman, T., Zhang, C., Kuner, M. C., Weaver, E. J., Amin, I., Park, H., Lim, Y., Kim, J., Chrzan, D., Walsh, A., Blau, S. M., & Asta, M. (2025). MLIP Arena: Advancing fairness and transparency in machine learning interatomic potentials via an open, accessible benchmark platform. Advances in Neural Information Processing Systems, 38. https://arxiv.org/abs/2509.20630

Chapter 2 Generative AI and Diffusion Models for Crystal Structure Prediction (CSP) and Inverse Materials Design

Introduction

When we make an object, we usually choose the material first. Suppose we are making a pot. First, we choose steel or aluminum. Then we decide its thickness and shape. Last, we put it on heat and check what properties appear.

The old way of finding new materials is similar. First, we choose one material we already know. Then we measure its properties by computer or experiment. If it is good, we keep it. If it is bad, we discard it. This method is reliable. But it is slow because each candidate must be checked one by one.

The problem is that the world of materials is too large. There are many kinds of elements. The number of atoms can also be changed. The positions of atoms and the size of the lattice can also be changed. When all of these are combined, the number of cases grows like stars in the sky. People cannot possibly reach the end by testing them one by one.

This chapter is about turning the direction of thought around. We do not choose the material first. We decide the desired property first. We decide that we want a solid electrolyte that conducts electricity well. We decide that we want a catalyst that survives in a hot rocket nozzle. Then AI draws a crystal structure that fits those properties on its own. This method is called inverse design.

The importance of inverse design can be compared to cooking. The old method is cooking with ingredients already in the refrigerator. Because the ingredients are fixed, the dishes that can be made are also fixed. Inverse design starts by deciding what food we want to eat. Then it works backward to find the ingredients needed for that food. It even obtains ingredients that were not in the refrigerator. In this way, we are not trapped by the ingredients in the refrigerator. We move closer to the food we imagined.

The main character here is generative AI. It is easy to think of an AI that draws images. It starts from a cloud of blurry dots. It removes noise little by little. At the end, a human face or landscape appears. Inverse materials design works on the same principle. It starts from a blurry group of atoms. It removes noise and forms a regular crystal. This noise-removal method is called a Diffusion Model.

In this chapter, we look step by step at how diffusion models make crystals. We look at how they follow symmetry rules. We look at Flow Matching, a technique that speeds up calculation. We also look at new ideas such as inpainting, which fills empty spaces, and mirage atoms. We look at how reinforcement learning guides models toward good structures. Finally, we examine real cases where this method has been used for battery solid electrolytes, catalysts, and aerospace alloys. We explain step by step why rockets and new materials need this technology.

1. From Choosing Materials First to Inverse Design That Starts with Properties

This section deals with how the direction of materials search has changed. The old method chooses materials first and checks properties later. The new method decides properties first and makes materials later. We examine why this shift is needed and what it means for rockets and new materials.

The old method is called Forward Property Screening. Forward means moving from materials to properties. Screening means filtering many candidates. Researchers take candidates from known crystal databases. There are two representative repositories. They are the Materials Project and the Inorganic Crystal Structure Database (ICSD).

After taking out candidates, researchers calculate their properties. For this, they use a physics calculation called Density Functional Theory (DFT). This calculation solves the motion of electrons inside atoms to obtain energy and properties. The results are accurate. But the amount of calculation is large. Computers must run for a long time. So only a small number of candidates can be examined in a day.

The basic limit of this method is that it is trapped in a narrow cage. Testing is done only on materials that are already known. New materials that are not in the database cannot become candidates in the first place. Even if an unknown material has excellent properties, we cannot reach it. In this way, the old method has weak power to create something new.

Inverse materials design breaks this cage. It reverses the order. First, it sets the target properties. It decides that ionic conductivity must be high. It decides that the material must be thermodynamically stable. Then AI directly draws a structure that satisfies those conditions. It can also draw structures that were not in the database. The center of materials search moves from testing to creation.

Let us consider how large the number of cases is. About 90 elements can be used in nature. Suppose we place only ten atoms in a box. For each position, we choose which element to place. The atomic positions can also be shifted little by little. Even this combination alone becomes an astronomical number. More candidates appear than grains of sand on a beach. People could spend a lifetime calculating them one by one and still not finish.

The time required for one calculation is also not small. DFT calculations take longer as the number of atoms increases. If the number of atoms doubles, the calculation grows several times larger. A computer may run for days to solve one large structure. So we cannot blindly calculate many candidates. We must choose candidates well in advance. AI gives us the wisdom to select which candidates to calculate first.

Drawing one crystal means deciding three things at the same time. First, we decide the lattice. The lattice is the shape and size of the box that contains the atoms. Next, we decide the atomic positions. This means where atoms are placed inside the box. Finally, we decide the atomic species. This means which element sits at each position. All three must fit together to become a real crystal. If even one is wrong, atoms collide or the structure collapses.

It is difficult to handle these three things at once. The lattice is made of continuous values with lengths and angles. Atomic positions are also continuous values. But atomic species are discrete values. After chlorine, there is no element that is half like chlorine. Values with different properties must be handled together in this way. Diffusion models do this difficult work. They recover continuous values and discrete values side by side from noise.

Let us explain the principle of diffusion models a little more. Training has two directions. First, in the forward direction, noise is gradually added to a clean crystal. After many steps, the crystal turns into a disordered group of points. AI learns how to reverse this process. It practices guessing the original crystal from a noisy state. After this practice, it forms new crystals from disordered points.

The power of this method lies in how well it learns from data. Thousands of known crystals are shown to the model. The model learns a sense of which arrangements are plausible. It learns the distances atoms must keep from one another. It learns which elements sit well in which positions. With this learned sense, it proposes new crystals. It learns on its own without people writing every rule by hand.

Inverse design is not limited to regular crystals. Some materials have disordered atoms, like glass. These materials are called amorphous materials. A study published in 2025 made amorphous structures with a diffusion model. It made them thousands of times faster than older calculation methods. It produced various structures by changing composition and processing conditions. In this sense, inverse design reaches beyond crystals.

Inverse design models also learn chemical rules. Each atom has a number of bonds it can form with its neighbors. This number is called valence. If valence is violated, an impossible structure appears. A study published in 2026 put this valence rule into the generation process. As a result, it selected only materials that can actually exist. The trend of embedding common knowledge from physics and chemistry into models continues.

This shift has great meaning for rockets and new materials. Rockets require materials that can withstand extreme environments. The combustion chamber is hot. The nozzle is eroded by fast gas. Batteries must be light and safe. Materials that satisfy all these conditions are not common in nature. That is why inverse design is needed to draw what does not yet exist. AI walks through material space where people have not been able to go.

2. Generative Models That Follow the Symmetry Rules of Crystals

This section deals with the symmetry rules of crystals. A crystal is not a pile of atoms thrown anywhere. It is an ordered structure in which atoms repeat according to rules. If AI does not know these rules, it draws fake crystals. We examine how symmetry is put into models by dividing the models into two generations.

To understand crystals, we must first grasp the word symmetry. Symmetry is the property of keeping the same shape even after a certain operation. If both sides overlap when a sheet of paper is folded in half, it has left-right symmetry. In crystals, several layers of such symmetry are stacked. They remain the same when shifted sideways, rotated, or reflected.

The collection and classification of all these symmetry rules is called a Space Group. In three-dimensional crystals, there are exactly 230 space groups. Each crystal belongs to one of these 230 groups. This classification fits ideal three-dimensional periodic crystals. Amorphous materials such as glass and quasicrystals fall outside it. If we know the space group, we can be said to know all the repetition and rotation rules of that crystal.

Inside a space group, there are fixed places where atoms can sit. These groups of places are called Wyckoff Positions. A Wyckoff position is not one site but a group of sites tied by symmetry. If an atom is placed at one site, the same atom is also placed at the matching sites under the symmetry rules. It is like placing a ball in front of a mirror and seeing a ball appear inside the mirror. So when one site is set, the positions of several atoms are set at once. The number of sites attached to this group is called multiplicity. If a generative model receives this rule, it can avoid placing atoms on top of one another.

The power of a space group lies in drawing the whole with little information. For a crystal with high symmetry, only a few atoms need to be set. The rest are filled automatically by the symmetry rules. It is easy to think of a snow crystal. The shapes extending in six directions are the same. If one direction is drawn, the other five are obtained by rotation. A crystal also determines the whole from a small piece in this way.

Problems arise if a generative model does not know this rule. Suppose AI places atoms one by one freely. Even a small amount of computational noise breaks symmetry. A crystal with almost no symmetry is called P1. P1 is also a proper official space group. Some real materials do belong to P1. So a structure

cannot be called fake just because it is P1. The problem arises when a crystal that should have high symmetry collapses into P1. In that case, interatomic distances, energy, and stability must be checked separately.

For this reason, recent research tries to put symmetry rules into the model in advance. Instead of placing atoms one by one, it follows a different order. First, it creates a small basic unit that contains all the symmetry. Then it applies the symmetry rules to unfold the full crystal. This makes the generated result more likely to keep symmetry.

A good way to handle symmetry is a property called equivariance. Equivariance means that if the input rotates, the output rotates with it. Gloves make this easy to understand. If a right-hand glove is turned inside out, it becomes a left-hand glove. The hand and glove are flipped together. If a crystal model has equivariance, its properties remain the same even when the structure is rotated. Symmetry is preserved naturally without being forced.

Using a neural network with this equivariance makes learning easier. This is because the model knows symmetry on its own. The same structure does not need to be shown from many angles. It learns well even with less data. That is why recent crystal generative models use equivariance as an important tool. This property will also become central in the Flow Matching models discussed later.

A comprehensive review of crystal generative models published in 2026 summarizes this trend well. The review points out that the ability to preserve symmetry is a core condition for a good crystal generative model. It sees symmetry-preserving models as producing more stable candidates. This review compares diffusion models and Flow Matching models side by side.

2.1 The Starting Point of Crystal Generative Diffusion (CDVAE)

Research on crystal generative diffusion began with an early model. The name of this model is CDVAE. We examine one by one what it did well and what it lacked.

CDVAE stands for Crystal Diffusion Variational Autoencoder. This model was published in 2022 and opened the field. It established the first diffusion framework that handled crystal lattices, atomic positions, and atomic species together.

CDVAE's major contribution is that it considers periodic boundary conditions. A periodic boundary condition means that a crystal repeats endlessly in the same shape. Wallpaper patterns make this easy to understand. One piece continues up, down, left, and right to cover the whole wall. A crystal also forms a bulk material through repeated small units.

This model moves atomic positions little by little using a method called Langevin Dynamics. The direction that pushes the atoms is the gradient of the probability learned by the model. It is different from the actual force between atoms or from free energy. So this process alone does not guarantee a stable structure. After generation sets the framework, DFT refines the structure again. These two steps must be viewed separately.

However, CDVAE has a clear limit. It cannot directly fix space group symmetry. Because it moves atoms freely, symmetry is easily broken. As a result, the generated structure often collapsed into a P1 state with no symmetry. Crystals that require symmetry needed extra verification.

Another limit is that the number of atoms must be set in advance. This model first predicts the number of atoms and then builds a structure to match that number. Once the number is fixed, flexibility falls. It becomes hard to handle cases where atoms are missing or added. This limit is later solved by the mirage atom method.

These shortcomings are not flaws but stepping stones. The problems raised by CDVAE became the goals of the next generation of models. How can symmetry be preserved? How can the number of atoms be handled freely? These two questions guided later research. Knowing the limits of an early model is the starting point for progress.

2.2 Symmetry-Aware Conditional Diffusion Models

The next generation of models tried to follow symmetry rules from the start. I will separate two studies with similar names so they are not confused. Performance figures should be handled carefully, only within verified bounds.

The core idea of a symmetry-aware model is simple. It does not place atoms freely one by one. It first creates an asymmetric unit. An asymmetric unit is the smallest group of atoms that can create the whole crystal through symmetry rules. The model applies the symmetry operations of the space group to this small seed. Through these operations, the seed unfolds into the whole crystal.

The strength of this method is that symmetry is preserved by itself. Because the seed is unfolded by symmetry rules, symmetry is built into the result. There is no need to adjust atoms separately. Paper folding makes this easy to understand. If paper is folded once and cut, the left and right sides match when it is unfolded. A symmetry-aware model cuts out a crystal in this way.

Here, two studies with similar names must be distinguished. One is SymmCD. It is a symmetry-preserving diffusion model for crystal generation. Levy and others presented it at the International Conference on Learning Representations (ICLR) in 2025. It is a formal peer-reviewed paper. The identifier of this paper is arXiv:2502.03638.

The other is a separate study that handles Wyckoff positions. Ishii and others released it as a preprint in 2026. A preprint is early research that has not yet gone through peer review. This paper proposes a method called WyckoffDiff-Adaptor. Its identifier is arXiv:2601.08115. The two papers have similar goals, but their methods and authors are different. This book presents the two studies separately and does not mix them together.

There is also an intermediate bridge between CDVAE and symmetry-aware models. One example is DiffCSP. This model generates the lattice and fractional coordinates together through diffusion. Fractional coordinates are relative positions when the box size is treated as 1. DiffCSP preserved symmetry better than CDVAE. It paved the way toward later models that directly include symmetry. In this way, models move forward one generation at a time by filling earlier gaps.

Performance figures must be treated carefully. There is a figure saying that the symmetry preservation rate exceeds 98 percent, but the original text does not confirm it as final. So this book does not present a specific number as confirmed performance. Instead, it understands the result as a direction. Symmetry-aware models preserve symmetry better than older models. As a result, they produce more candidates close to real crystals.

Symmetry-aware models also handle conditional generation well. Conditional Generation is a method that gives the desired property as a condition. Suppose a value called the band gap is specified. The band gap is an energy threshold that shapes the properties of a semiconductor. The model draws only crystals that match that band gap. When conditions are added in this way, only candidates that fit the target are collected. This greatly reduces the waste of drawing useless candidates.

One preprint released in 2026 pushed this conditional generation further. It generated crystal structures and electronic properties together. By learning electronic properties at the same time, it matched target properties more closely. In tests that used band gap and formation energy as conditions, the success rate

rose. The share of stable and novel candidates also increased. The trend of handling symmetry and conditions together is continuing.

It is worth asking why this technology is needed for new materials. Symmetry is deeply tied to the properties of materials. A fake structure with broken symmetry also gives wrong property predictions. Aerospace alloys and battery materials matter only if they can lead to real synthesis. Generative models that preserve symmetry reduce useless candidates. This raises the quality of candidates that move on to experiments.

3. Flow Matching for Faster Crystal Generation

Diffusion models are accurate, but they are slow because they take many steps. Flow matching increases speed by reducing the number of steps. We also look at how it handles the curved space where crystals exist.

First, we need to understand why diffusion models are slow. A diffusion model creates a structure by removing noise little by little. It removes only a small amount at a time. So it needs hundreds to thousands of steps to finish. Many steps mean long computation. It takes time to create one candidate. In materials search, where many candidates must be generated, this time becomes a burden.

Flow Matching solves this problem from a different angle. Instead of removing noise little by little, it learns a shortcut. It learns a path that goes directly from the starting state to the target state. A river current makes this easy to understand. Once the waterway is set, a leaf flows along that path. Flow matching learns this waterway itself.

When this waterway is written as a formula, it becomes an ordinary differential equation. An Ordinary Differential Equation is an equation that describes how a value changes. It shows how a value changes over time. Flow matching follows this equation from the starting point to the ending point. Diffusion shakes randomly at every step, but this path does not. Once the starting point is set, it follows a fixed path to the end. So the path is straight and the number of steps is small. However, the starting point itself is still chosen randomly. A fixed path does not automatically reduce distortions in lattice angles and volume.

The difference between diffusion and flow matching can be compared to coming down a mountain. Diffusion feels its way down one step at a time in the fog. It is careful but slow. Flow matching comes down along a trail it has learned in advance. Because it knows the path, it moves quickly with long steps. Both methods reach the destination at the foot of the mountain, but the time they take is different.

When dealing with crystals, there is a difficulty: the space is not flat. A lattice has lengths and angles. Atomic coordinates are positions inside a repeating box. This kind of space is not flat like paper but curved like a ball. A curved space is called a Riemannian Manifold.

If curved space is treated like flat paper, errors occur. It is like flattening a globe into a map and distorting area. The polar regions are drawn larger than they really are. Crystals are the same. The angles and volume of the lattice become distorted. So flow matching uses a ruler for curved space from the start. This ruler is called a Riemannian Metric. Because curved space is handled as curved, distortion is reduced.

3.1 Finding Shortcuts on Curved Space

We examine how flow matching handles curved space through representative models. The discussion centers on verified papers.

Finding a shortcut in curved space is like flying over Earth. The shortcut from Seoul to New York is not a straight line on a map. This is because Earth is round. The real shortcut is a curved line along Earth's surface. This kind of curve is called a Geodesic. Crystal generation also moves along geodesics on curved space.

FlowMM is the model that first applied this idea well to crystals. FlowMM extends Riemannian flow matching for crystals. It also considers the symmetry unique to crystals. It handles translation, rotation, atom-order permutation, and periodic boundary conditions. This model greatly reduced the number of steps needed to find stable materials. The published paper reports that it found stable candidates in about three times fewer steps than existing methods.

FlowMM does two tasks. One is to predict a stable structure when the composition is fixed. The other is to propose a new composition and its structure together. The first task is called Crystal Structure Prediction. The second task is true discovery of new materials. Flow matching performs both tasks quickly. So it is used widely across many parts of materials search.

Another verified case is CrystalFlow. This study was published in Nature Communications in 2025. It is a formal peer-reviewed paper. CrystalFlow creates crystalline materials with a flow-based generative model. It generates the lattice and atomic arrangement together. It showed that flow matching makes training easier than diffusion models.

Flow matching is also used for organic crystals. One preprint released in 2026 used Unit Cell Flow Matching. With this method, it quickly predicted organic crystal structures. This study is not a model only for inorganic crystals. So care is needed not to extend its results to inorganic materials. Still, it shows that flow matching can be used across different kinds of crystals.

Let us connect speed to rockets. Materials inverse design must generate and filter many candidates. Even if ten thousand are generated, only a few may be useful. If it takes too long to create one candidate, the whole search is blocked. Flow matching solves this bottleneck. It generates more candidates in the same amount of time. This speed is valuable in fields with strict requirements, such as aerospace materials.

How much does flow matching reduce the number of steps? The FlowMM paper reports that it reduced the steps needed to find stable candidates by about three times. It travels the path taken by diffusion models in fewer steps. The time needed to create one candidate becomes shorter by that amount. Large claims such as a hundredfold reduction or shortening a day to a few minutes have no basis. Even if we look only at reported values, the saved time accumulates as more candidates are generated.

Flow matching still has tasks left. To learn a shortcut, it must know the target distribution well. If training data is lacking, it learns the wrong path. Computing on curved space also remains difficult. So flow matching does not replace all diffusion models. The two methods are chosen according to the problem. Use diffusion when accuracy is urgent, and use flow matching when speed is urgent.

4. Redrawing Only Part of a Crystal and Letting the Model Decide the Number of Atoms

A crystal does not always need to be drawn from a blank page. There is also a way to change only part of an existing structure. Inpainting, which fills vacancies and defects, is one such method. There is also an idea called mirage atoms, which handles problems where the number of atoms is not fixed.

First, we need to define Inpainting. Inpainting is a technology that naturally fills empty parts. It is often seen in photo editing. If a utility pole is removed from a photo, that place becomes empty. Artificial intelligence looks at the surrounding sky and fills the empty space. It does the same thing in crystal

structures. In crystals, it naturally fills empty atomic sites and defect sites.

Drawing a crystal from a blank page is different from inpainting. Blank-page generation starts without placing any atoms. Inpainting starts from an existing structure. It empties only part of the structure and redraws only that part. The rest is left unchanged. The two methods have different uses. Blank-page generation is used to find completely new materials. Inpainting is used to improve a good material.

There is a reason this technology is needed in materials research. In practice, researchers rarely draw an entire crystal from scratch. In many cases, they already have a good host matrix. They change its properties by adding or removing a small amount of another element. Replacing some atoms with other atoms in this way is called doping. A state in which an atom is missing or has entered another site is called a defect.

Doping is a method the semiconductor industry has used for a long time. Pure silicon does not conduct electricity well. A very small amount of phosphorus or boron is added to it. Then its electrical properties change greatly. The chips inside smartphones run on this principle. Crystal inpainting entrusts this doping design to artificial intelligence. The model proposes which element to place at which site.

This small change can greatly change the properties. It is like how the taste changes when even a small amount of another substance is mixed into salt. If one atom in a crystal is replaced, electrical conductivity changes. Catalytic performance changes. That is why researchers must precisely design what to place and where. Inpainting helps with this precise design. It keeps the host matrix as it is and redraws only the sites to be changed.

There is a reason inpainting is more practical than generation from a blank slate. Finding a good host crystal itself takes a long time. If a verified host matrix already exists, it is faster to improve it. It is like repairing one room instead of building a sturdy house from scratch. The foundation does not shake, so the risk is lower. That is why working researchers often use inpainting.

To fill an empty site, the surrounding atoms must be read carefully. If any element is inserted at random, the structure collapses. It must match the size of neighboring atoms. The electrical balance must also fit. Artificial intelligence learns the rules of the host matrix and satisfies these conditions. It uses a method called masking to hide only the sites to be changed. It redraws only the hidden sites and fixes the rest.

4.1 Mirage Atom Diffusion (MiAD), which determines the number of atoms by itself

There is a preprint model called MiAD as a method for handling the number of atoms freely. Let us examine why the idea of a mirage atom is useful.

MiAD stands for Mirage Atom Diffusion. It is a preprint released in late 2025. It is early-stage research before peer review. Its formal title means a diffusion model that creates new crystals from scratch. Its identifier is arXiv:2511.14426.

The core idea of this model is the mirage atom. Mirage means an optical illusion. A mirage atom is a virtual site that may become a real atom or may disappear. The model first fills a box with plenty of virtual sites. Then, during the diffusion process, it decides whether each site is a real atom or an empty site. An element sits at a site that remains real, and sites that do not remain disappear.

The advantage of this method is that it does not fix the number of atoms in advance. Earlier models first set the number of atoms and then made a structure. MiAD turns virtual sites on and off and determines the number by itself. What this paper actually showed was generation that determines the number of

atoms by itself. It is not a study that verified defect concentration, doping sites, or non-stoichiometry. Still, there is room to attach this method to materials whose number of atoms is flexible.

A material whose elemental ratio is not exact is called a non-stoichiometric compound. Such materials are common in real materials. An empty site where one atom is missing changes the properties. Clusters of several defects can also form. There are also interstitial defects, in which atoms enter leftover sites. These small disturbances determine conductivity and strength. A generation method that handles the number of atoms freely offers a clue for expressing such disturbances.

MiAD reported results in benchmark tests. It used a widely used test set called MP-20. A metric that shows the ratio of stable, unique, and novel crystals is called the SUN rate. The paper reported certain results on this metric. However, these values are numbers from the preprint stage. They should therefore be treated as early results, not confirmed performance.

Let us consider what this technology means for rockets and new materials. Real aerospace materials are not perfectly regular. There are always defects where some atoms are missing or have changed places. These defects determine strength and conductivity. A model that handles the number of atoms freely expresses such real structures better. It therefore produces candidates closer to laboratory results. However, this method is still at the preprint stage, so its defect design performance needs further verification.

5. Guiding Candidates Toward Better Ones with Reinforcement Learning

A generative model produces many candidates. But not all of them are useful. Reinforcement learning sets the direction toward better candidates. Let us examine the principle by which reinforcement learning guides the model and the safeguards around it.

Reinforcement learning is learning that changes behavior by looking at rewards. It is easy to understand if you think of dog training. If the dog does well, it gets a treat. If it does not, it does not get one. The dog increases the actions that earn treats. Artificial intelligence also increases actions that earn high rewards.

In materials generation, good structures receive high rewards. Stable crystals with good properties receive high scores. Structures in which atoms collide or are unstable receive low scores. The model changes its generation habits toward higher scores. As tests are repeated, the number of usable candidates increases.

Generative models and reinforcement learning have different roles. The generative model spreads plausible candidates broadly. Reinforcement learning drives those candidates toward the target. It is like one hand scattering seeds and another hand opening a path for water. The two hands must move together to bear good fruit. Generation alone is not enough, and reinforcement learning alone is not enough. When the two are connected, the model approaches the desired properties.

There is a reason this method is needed. A generative model is good at making plausible candidates. But plausibility and actual stability are different. Many structures look good on the surface but collapse thermodynamically. So generation alone is not enough. Stability, novelty, and synthesizability must be considered together. Reinforcement learning puts these three factors into the reward and guides the model.

Synthesizability is a difficult target to handle. A material that is stable on paper may not be made in the laboratory. The path to making it may be too difficult, or the raw materials may be unavailable. So looking only at stability is not enough. Researchers must also see whether it can actually be made. The reward in reinforcement learning tries to include these real-world conditions. Only then do candidates

emerge that can move to the laboratory.

There are several ways to attach reinforcement learning to materials generation. A representative method is called Proximal Policy Optimization. This method changes the policy only a little at a time. This is because learning becomes unstable if it changes suddenly and greatly. It moves forward carefully, little by little, and increases the reward. In this way, it improves performance stably. This method is also widely used to refine language models.

A preprint published in 2025 shows this direction well. It fine-tuned a diffusion model with reinforcement learning and raised the success rate of inverse design. It produced more structures that matched the target properties. It also increased the ratio satisfying the conditions of stability, uniqueness, and novelty. This research is at an early stage. Therefore, DFT calculations and experimental synthesis are still needed for industrial application.

5.1 Safeguards and Performance Metrics in Reinforcement Learning

Reinforcement learning has two safeguards. One prevents the model from straying too far. The other stores good candidates and uses them again. We also look at metrics for measuring performance.

Reinforcement learning has a hidden risk. If the model pursues only rewards, it can become biased toward strange structures. It gets trapped in a narrow region that gives high scores. It loses diverse candidates and repeats only one answer. This phenomenon is called mode collapse. When collapse occurs, exploration stops.

The device that prevents this is KL regularization. KL is a measure of how different two distributions are. This device keeps the fine-tuned model from moving too far away from the distribution it originally learned. It is like being tied to a rubber band. The model moves in a new direction, but if it strays too far from its roots, it is pulled back. In this way, it finds good candidates while preserving diversity.

Another device is experience replay. Good candidates are stored instead of being thrown away. The stored candidates are brought back and used for learning. It is like collecting problems you solved well in a study notebook and solving them again. This prevents the model from forgetting good structures that were hard to find. Learning continues more stably.

A metric often used to measure performance is the SUN rate. SUN means crystals that are stable, unique, and novel. It is the ratio of candidates that satisfy all three conditions. Stable means that the material does not easily separate into other materials. Unique means that generated materials do not overlap with one another. Novel means that the material was not in the database.

To measure stability, researchers look at a value called energy above hull. This value is calculated with DFT at absolute zero. The lower the value, the more stable the material is in calculation. If the value is close to 0, it is maintained well within the calculation. However, actual temperature and pressure, water and oxygen, and reaction rates are not included in this value. There are also calculation errors. Therefore, this value should be read as an indicator that points to computationally stable candidates. The reward in reinforcement learning uses this value as an important ingredient.

The reward puts several goals on the scale together. If it pursues only stability, the result is a material with bland properties. If it pursues only properties, the result is a material that is hard to synthesize. So weights are assigned to stability, properties, and novelty. Which goal matters more is adjusted by these weights. This weight adjustment is the designer's responsibility. When the scale is adjusted to match the target, candidates gather in the desired direction.

Let us consider what these safeguards mean for rocket material design. Aerospace materials have meaning only when they can actually be synthesized. A structure that is good only on paper is useless. Reinforcement learning includes synthesizability in the reward. KL regularization and experience replay keep exploration stable. As a result, the ratio of candidates worth moving to the laboratory increases.

6. Practical Applications of the Inverse Design Pipeline

How have the technologies discussed above been used in real research? We look at solid electrolytes for batteries, catalysts, and aerospace alloys. We use verified recent studies as the basis. We examine how generation, computation, and experiment work together.

To use the technologies discussed above in practice, several tools must be connected. Diffusion models and flow matching create candidates. Symmetry awareness and inpainting improve candidate quality. Reinforcement learning guides candidates toward the target. These tools show their strength when they are connected into one flow. This flow is called an inverse design pipeline. A pipeline means connecting several processes through pipes.

Practical inverse design usually follows three steps. First, in the generation step, artificial intelligence draws candidate structures that match the target properties. Next, in the computational validation step, DFT and machine learning interatomic potential (MLIP) calculate stability and properties. Finally, in the experimental synthesis step, a self-driving laboratory or researcher actually makes and verifies the material.

These three steps do not flow in only one direction. Results obtained from experiments are returned to the generative model. The model uses this feedback to draw the next candidates better. This feedback loop is called active learning. A study published in 2025 greatly increased inverse design efficiency through active learning. It repeated candidate generation and validation to approach the target properties.

Let us look more closely at the second step, computational validation. A candidate made by a generative model is still a drawing on paper. Computation checks whether this drawing is truly stable. DFT is accurate but slow. So researchers use the machine learning interatomic potential discussed in the previous chapter together with it. This method produces accuracy close to DFT much faster. It quickly filters many candidates. Only a small number of promising ones are checked again with DFT.

This collaboration determines the speed of exploration. No matter how fast generation is, it is useless if validation is slow. If validation becomes a bottleneck, candidates only pile up. Fast generation and fast validation must work together. That is why recent research connects the three steps into one automated flow. The self-driving laboratory discussed in the next chapter handles the end of this flow.

6.1 Designing Halide Solid Electrolytes for All-Solid-State Batteries

The first case is the design of solid electrolytes for batteries. We look at why this material is important and how AI helps. We focus on verified numbers.

An all-solid-state battery uses a solid electrolyte. In ordinary batteries, the electrolyte is liquid. Liquid electrolytes can catch fire. Solid electrolytes reduce this risk. But solid does not always mean safe. They may react with lithium metal, moisture, or electrodes. Sulfide-based electrolytes and some halide-based electrolytes can produce harmful decomposition products, or resistance and cracks can form at interfaces. Still, interest is high because they offer a major benefit: lower risks of leakage and fire.

The properties required of a solid electrolyte are demanding. Lithium ions must pass quickly through the solid. How well ions pass through is called ionic conductivity. A higher value is better. At the same time, the material must withstand high voltage. It must not be brittle, and it must adhere well under pressure. It is difficult to meet all these conditions.

Halide solid electrolytes are considered promising candidates. Halides refer to elements such as chlorine, bromine, and iodine. This family is relatively stable at high voltage. It also has a fair balance between softness and conductivity. So research is active. A representative material is a compound made of lithium, yttrium, and chlorine.

Artificial intelligence helps design these materials in several ways. One 2025 study combined machine learning with DFT calculations. It predicted several properties at once and screened good halide candidates. It predicted ionic conductivity, elastic modulus, and the electrochemical stability window together. This approach, which looks at several properties at the same time, made the results more practical.

Another study published in 2026 used a method that extracts rules from data. It identified key chemical features that control ionic conductivity. Based on these features, it designed new compositions. The paper reported that one designed material showed ionic conductivity of about 12 millisiemens per centimeter at room temperature. This value is high for this family. It shows a path from extracting rules from data to using them in design.

It is difficult to calculate how ions pass through a solid. Lithium ions move by jumping from one site to another. Conductivity is high when the barrier for this jump is low. To calculate this barrier, one must track atomic motion for a long time. Machine learning interatomic potential (MLIP) is used here as well. A 2025 study reported that this method is useful for calculations of solid ion conductors. Candidates are generated first, and this method then checks ion transport.

Now consider what this material means for aerospace and rockets. Satellites and probes depend heavily on batteries. When there is no sunlight, they rely on stored electricity. If the battery is unstable, the whole mission is at risk. Safe and light all-solid-state batteries reduce this risk. AI inverse design helps find such battery materials faster.

6.2 Designing Catalysts That Split Water

The next case is catalyst design. We will see what a catalyst is and how AI finds good catalysts. We will look at a recent case linked with a robotic laboratory.

A catalyst is a material that helps a chemical reaction proceed faster. It changes very little itself. But it increases only the reaction rate. A catalyst is essential in devices that make hydrogen and oxygen from water. This device is called a water electrolysis device. With a good catalyst, more hydrogen can be obtained using the same electricity.

Catalyst performance depends on how atoms are arranged on the surface. Reactions occur on the catalyst surface. Reactants attach to the surface, detach, and change. The process in which a substance attaches to a surface is called adsorption. If it attaches too strongly, that is a problem. It does not detach well, and the reaction is blocked. If it attaches too weakly, that is also a problem. It detaches before the reaction occurs. A good catalyst holds it with the right strength.

Adsorption strength is determined by the arrangement of surface atoms. Even a small change in the distance between atoms changes the strength. So how the surface is shaped matters. Artificial intelligence adjusts adsorption strength by changing atomic positions on the surface. It finds an arrangement that holds reaction intermediates with the right strength. This arrangement is the key to a

good catalyst.

The reaction that makes oxygen is difficult. This reaction is called the oxygen evolution reaction. To drive this reaction, a voltage higher than the theoretically required voltage must be applied. This extra voltage is called overpotential. Overpotential is a different concept from activation energy, which is the barrier to the reaction rate. A good catalyst has a low overpotential. A low overpotential means less electricity is used. So lowering overpotential is a major goal of catalyst design.

One study published in 2026 used generative AI to inversely design catalysts. The study dealt with high-entropy catalysts, which mix several elements. It connected spectroscopic information, a generative model, and robotic experimental equipment into one system. The robot made candidates, measured their performance, and returned the results to the model. This feedback loop ran quickly.

The numbers reported by this study show its practical value. It greatly reduced the time needed to make and measure one sample. Work that used to take about 20 hours was reduced to 78 minutes. The best candidate found by the system also lowered the overpotential by another 32 millivolts. This case is academic research, but it also shows robotic automation. It shows how much speed can increase when a generative model and experiments work together.

This trend changes human work. In the past, a researcher spent all day making one sample. Now a robot does that work day and night. The person thinks about what goal to set. The person judges which results have meaning. The work moves from using the hands to using thought. The human role does not disappear. It moves upward. The next chapter examines this self-driving laboratory in detail.

There is a reason high-entropy catalysts that mix several elements are gaining attention. Mixing many elements creates diverse sites on the surface. This raises the chance that a site suitable for the reaction will appear. But the number of combinations is so large that people cannot test them all. A generative model scans this vast space instead. It narrows good compositions and reduces the number of experiments. A robot quickly makes and checks those candidates.

Now consider what catalysts mean for rockets and space exploration. Long space missions must produce water and air on their own. Catalysts are used to obtain oxygen by splitting water. Catalysts are also used to make fuel. A good catalyst makes more oxygen and fuel with the same electricity. That reduces the load a spacecraft must carry. AI inverse design brings such catalysts closer.

Hydrogen is also used as rocket fuel. Liquid hydrogen produces strong thrust. If technology for obtaining hydrogen by splitting water advances, fuel production becomes easier. Here too, catalysts hold the key to the reaction. If good catalysts can be made from cheap elements, costs fall. The goal is to use less precious platinum. Generative AI is powerful in the search for these non-precious-metal catalysts.

6.3 Inverse Design of Aerospace Alloys and Superalloys

The final case is alloy design for rocket engines. We will see why new alloys are needed. We will examine how generative models find alloy compositions. The discussion is based on verified recent studies.

Rocket engines must withstand extreme heat. The inside of the combustion chamber is thousands of degrees hot. The nozzle is eroded by fast gas. For many years, nickel-based superalloys have held this role. A superalloy is an alloy that does not lose strength even at high temperature. But nickel superalloys also have temperature limits. In hotter environments, they melt or soften.

High-entropy alloys have emerged as an alternative. Ordinary alloys mix small amounts of other elements into one main element. High-entropy alloys mix several elements in similar amounts. When mixed this way, the structure becomes unexpectedly stable. Those made from elements with high melting points are called refractory high-entropy alloys, or refractory HEAs. These alloys are considered candidates to replace nickel superalloys in environments above 1100 degrees.

The reason high-entropy alloys are stable lies in disorder. When several elements are evenly mixed, there are many possible arrangements. These many possibilities hold the structure together. This property is called high configurational disorder. In this case, disorder preserves order. Because of this property, the alloy holds up well even in extreme environments. Rocket nozzles and combustion chambers are typical places that require this durability.

The problem is that the number of combinations is again as vast as the stars in the sky. Even mixing only five elements creates countless possible ratios. It is difficult to make and test them one by one. This is where a generative model steps in. When target properties are set, it searches backward for a matching composition. Strength and toughness values are given as conditions. The model proposes element ratios likely to produce those values.

Several studies published in 2026 show this direction. One study reviewed several generative methods for making alloy compositions. It examined variational autoencoders, generative adversarial networks, and diffusion models together. Another study collected literature and data and showed a process for narrowing candidates for high-temperature high-entropy superalloys. It selected compositions that matched target properties and reduced the number of candidates to test. Both studies are at the stage of narrowing candidates, not making real parts.

The properties of an alloy are not determined by composition alone. Even the same composition changes depending on how it is made. How heat is applied and how the alloy is cooled changes the microstructure. Microstructure means the size and arrangement of crystal grains. So alloy design must consider both composition and process. Generative models are also moving toward handling both together. Later chapters discuss these processes in more detail.

However, caution is needed. Refractory high-entropy alloys still face barriers. They still tend to fracture easily at room temperature. At high temperature, they may react with oxygen and suffer surface damage. So there is a gap between the promise in the laboratory and real parts. Even if a generative model narrows the candidates, these two barriers must be overcome through experiments. Controlling room-temperature brittleness and high-temperature oxidation is the next task.

6.4 Comparing Generative Models and the Importance of Verification

How should different generative models be compared? We will see which metrics are used. We will examine why verification is more important than results. We will also explain why unverified numbers require caution.

Several criteria are considered together when comparing generative models. One is stability, which checks whether the generated crystal can maintain itself. Novelty is also checked. This asks whether the material was absent from databases. The validity of atomic distances is measured too. It checks whether atoms are not too close to one another. Symmetry preservation is also examined. It checks whether crystal symmetry is maintained. The final criterion is speed, which measures how long it takes to make one candidate.

A model is not good just because one metric is good. Even if it is fast, it has little use if stability is low. Even if it preserves symmetry well, it has limited meaning if it lacks novelty. So the balance among

several metrics matters. The important metric changes depending on the problem. For battery materials, stability and conductivity come first. If rapid search is urgent, speed comes first.

A fair comparison requires the same test paper. In the field of crystal generation, there are widely used test sets. MP-20 is a representative example. It is a collection of about 20,000 crystals drawn from the Materials Project. There is also a set containing only perovskite crystals. All models are tested on the same set. Only then can we compare which model is better. If the test papers differ, the scores cannot be placed side by side.

There is a point to be careful about here. Numbers that list the performance of several models side by side are often not verified in the original sources. So this book does not use numbers from unverified tables. Unverified numbers can mislead readers.

Instead, we can get a sense from one verified result. It is a generative model study published by Microsoft in Nature in 2025. The model is called MatterGen. This model generates inorganic materials across the periodic table. It can also be fine-tuned for target properties. The paper reported that this model’s outputs were stable and novel twice as often as those of older models. It also reported that they were ten times closer to the energy minimum.

This model can learn again for several conditions. If it learns for elastic values, it produces hard materials. If it learns for magnetic values, it produces magnetic materials. It can also learn to use only desired elements. This is a method of adjusting one base model for several goals. This method is called fine-tuning. It changes the goal without rebuilding the foundation.

This study went as far as experimental validation. The target bulk modulus was set at 200 gigapascals, and a new material structure was generated. The material was then synthesized in the laboratory. However, what was synthesized was a solid solution in which the atoms were randomly mixed. The researchers measured Young's modulus by nanoindentation and used calculated values to estimate the bulk modulus. The estimated values reached a maximum of 169 gigapascals in four trials, with an average of 158±11 gigapascals. The measured values and converted values must be read separately. The calculated structure and the synthesized structure must also be read separately. Even so, this is a rare demonstration that connects generation to synthesis.

Extra care is needed when dealing with preprints. Several models introduced in this chapter are still at the preprint stage. A preprint has not yet gone through peer review. Its results may change later. For this reason, this book identifies preprints as early-stage research when introducing them. Even strong numbers are not treated as settled before validation. This attitude protects readers from mistaken beliefs.

Let us look at why validation matters more than results. A generative model can create an unlimited number of candidates. But creation does not make all of them real. They must be filtered by computation. They must be confirmed by experiment. Materials inverse design has meaning only when generation, computation, and experiment run together. This principle must be stricter in fields where safety is at stake, such as aerospace materials. A candidate can be trusted only after computational validation and experimental synthesis.

Summary

This chapter was about reversing the direction of materials discovery. The old method chose a material first and measured its properties later. The new method sets the properties first and draws the material later. This new method is called materials inverse design. Changing just one direction greatly changes the power of exploration. It opens a path to drawing materials that did not exist before.

The main actor in inverse design is generative artificial intelligence. A diffusion model shapes a crystal by removing noise from a blurred group of atoms. Its principle resembles image-generating AI. Crystal generation is harder. This is because the lattice, atomic positions, and atomic species must all be matched at the same time.

Crystals have a strict rule called symmetry. A three-dimensional crystal belongs to one of 230 space groups. CDVAE, an early model, could not directly fix symmetry, so structures often collapsed. Symmetry-aware conditional diffusion models reduced this problem by making a small seed and expanding it through symmetry. It is important to distinguish studies with similar names instead of mixing them together.

Efforts to increase speed have also continued. Flow matching learns a shortcut instead of removing noise little by little. FlowMM and CrystalFlow are representative examples. They quickly generate crystals along geodesics on curved spaces. This speed is highly useful in materials discovery, where many candidates must be sampled.

Techniques that change only part of a structure have also advanced. Inpainting fills vacancies and defects. Mirage atom diffusion handles the number of atoms freely. It places virtual sites that may or may not exist and decides the count on its own. This better expresses defect structures in real materials.

Reinforcement learning guides generative models toward better candidates. Stability, novelty, and synthesizability are included in the reward. KL regularization and experience replay keep exploration stable. Performance is measured with the SUN rate and energy above the hull.

Practical cases show that this technology is not just a story on paper. In the design of halide solid electrolytes, data analysis and design were connected. In the design of high-entropy catalysts, robotic experiments and generative models worked together and greatly reduced the time required. Microsoft's generative model synthesized a designed material in the laboratory and confirmed its properties.

One idea runs through this chapter. Artificial intelligence is a guide that walks the wide space of materials on our behalf. It scans paths in a few days that a person could not visit in a lifetime. It selects candidates that look promising and presents them. But computation and experiment decide whether those candidates can be trusted. The guide narrows the path, and people open the final door.

Finally, let us return to the principle of validation. A generative model produces an unlimited number of candidates. But creation does not make all of them real. Unverified numbers should not be used. Candidates must be filtered by computation and confirmed by experiment. Materials inverse design has meaning only when generation, computation, and experiment run together. In fields where safety is at stake, such as rockets and space exploration, this principle must be stricter. A structure drawn by AI is not the final answer. It is a carefully chosen experimental candidate.

Related Supporting Materials

Xie, T., Fu, X., Ganea, O. E., Barzilay, R., & Jaakkola, T. (2022). Crystal diffusion variational autoencoder for periodic material generation. International Conference on Learning Representations. https://arxiv.org/abs/2110.06197

Jiao, R., Huang, W., Lin, P., Han, J., Chen, P., Lu, Y., & Liu, Y. (2023). Crystal structure prediction by joint equivariant diffusion. Advances in Neural Information Processing Systems. https://openreview.net/forum?id=DNdN26m2Jk

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. Nature, 624, 80-85. https://doi.org/10.1038/s41586-023-06735-9

Zeni, C., Pinsler, R., Zugner, D., Fowler, A., Horton, M., Fu, X., et al. (2025). A generative model for inorganic materials design. Nature, 639, 624-632. https://doi.org/10.1038/s41586-025-08628-5

Levy, D., Panigrahi, S. S., Kaba, S. O., Zhu, Q., Lee, K. L. K., Galkin, M., Miret, S., & Ravanbakhsh, S. (2025). SymmCD: Symmetry-preserving crystal generation with diffusion models. International Conference on Learning Representations. https://arxiv.org/abs/2502.03638

Ishii, T., Hisama, K., & Shinohara, K. (2026). Symmetry-aware conditional generation of crystal structures using diffusion models. arXiv. https://arxiv.org/abs/2601.08115

Okhotin, A., Nakhodnov, M., Kazeev, N., Lazarev, M., Ustyuzhanin, A. E., & Vetrov, D. (2025). MiAD: Mirage atom diffusion for de novo crystal generation. arXiv. https://arxiv.org/abs/2511.14426

Chen, J., Guo, J., Fako, E., & Schwaller, P. (2025). Accelerating inverse materials design using generative diffusion models with reinforcement learning. arXiv. https://arxiv.org/abs/2511.03112

Miller, B. K., Chen, R. T. Q., Sriram, A., & Wood, B. M. (2024). FlowMM: Generating materials with Riemannian flow matching. Proceedings of Machine Learning Research, 235. https://arxiv.org/abs/2406.04713

Luo, X., Wang, Z., Wang, Q., Shao, X., Lv, J., Wang, L., Wang, Y., & Ma, Y. (2025). CrystalFlow: A flow-based generative model for crystalline materials. Nature Communications, 16, 9267. https://doi.org/10.1038/s41467-025-64364-4

Lo, A., Mucko, L., Cheng, A. H., Cai, A., Price, A. J. A., & Matusik, W. (2026). Fast organic crystal structure prediction with unit cell flow matching. arXiv. https://arxiv.org/abs/2606.03199

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nicklas, M., & Le, M. (2022). Flow matching for generative modeling. International Conference on Learning Representations. https://openreview.net/forum?id=PqvMRDCJT9t

Hoogeboom, E., Satorras, V. G., Vignac, C., & Welling, M. (2022). Equivariant diffusion for molecule generation in 3D. International Conference on Machine Learning, 8867-8887. https://arxiv.org/abs/2203.17003

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv. https://arxiv.org/abs/1707.06347

Metni, H., Ruple, L., Walters, L. N., Torresi, L., Teufel, J., Schopmans, H., Ceder, G., & Friederich, P. (2026). Generative models for crystalline materials. Advanced Materials, 38(18), e23620. https://doi.org/10.1002/adma.202523620

Mal, S., Ahmed, N., Jami, J., Mishra, S., & Sen, P. (2025). DiffCrysGen: A generative diffusion model for accelerated design of inorganic crystalline materials. arXiv:2510.12329. https://arxiv.org/abs/2510.12329

Okuda, I., Takahara, I., & Mizoguchi, T. (2026). Inverse materials design via joint generation of crystal structures and local electronic descriptors. arXiv:2605.01286. https://arxiv.org/abs/2605.01286

Yang, K., & Schwalbe-Koda, D. (2025). A generative diffusion model for amorphous materials. npj Computational Materials, 12(1), 29. https://doi.org/10.1038/s41524-025-01901-1

Han, X.-Q., Guo, P.-J., Gao, Z.-F., Sun, H., & Lu, Z.-Y. (2025). InvDesFlow-AL: Active learning-based workflow for inverse design of functional materials. npj Computational Materials, 11(1), 364. https://doi.org/10.1038/s41524-025-01830-z

Cheng, M., Luo, W., Tang, H., Yu, B., Cheng, Y., Xie, W., Li, J., Kulik, H. J., & Li, M. (2026). Enhancing materials discovery with valence-constrained design in generative modeling. Nature Computational Science. https://doi.org/10.1038/s43588-026-01037-2

Choong, L. Y. A., Chen, Z., & Ng, M.-F. (2025). Discovery of effective halide solid electrolytes for solid-state rechargeable batteries via machine learning and DFT calculations. ACS Applied Energy Materials, 9(1), 507-520. https://doi.org/10.1021/acsaem.5c03277

Li, D., Lin, Z., Luo, F., Xiao, C., Dai, Z., Ye, Z., Wang, Q., Guo, Y.-G., & Yang, C. (2026). Data-driven chemical feature extraction and material design for halide solid-state electrolytes. eScience, 100590. https://doi.org/10.1016/j.esci.2026.100590

Zhou, D., Yang, R., Jia, Z., Cai, Y., Zhao, L., Guo, L., Ye, G., Wang, S., Chen, L., Liu, D., Smith, P. E. S., Huang, Y., Zhu, Q., & Jiang, J. (2026). A practical inverse design approach for high-entropy catalysts using generative AI. Nature Synthesis, 5(5), 730-739. https://doi.org/10.1038/s44160-025-00983-5

Li, C., Xian, Y., Zhou, Y., Ding, X., Sun, J., & Xue, D. (2026). Generative design for alloys: Harnessing generative models for faster discovery. Advanced Materials, 38(29), e20478. https://doi.org/10.1002/adma.202520478

Rousseau, F., Belmonte, T., Sur, F., & Nominé, A. (2026). Inverse design of high-entropy superalloys using machine learning and generative artificial intelligence. Materials & Design, 266, 116097. https://doi.org/10.1016/j.matdes.2026.116097

Du, H., Hui, J., Zhang, L., & Wang, H. (2025). Universal machine learning interatomic potentials are ready for solid ion conductors. arXiv:2502.09970. https://arxiv.org/abs/2502.09970

Chapter 3 Self-Driving Laboratories and Scientific Multi-Agent Systems

Introduction

Imagine baking bread at home. You decide how much flour and water to use. You mix the dough and set the oven temperature. You slice the finished bread and taste it. If it is too hard, you add more water and lower the temperature. The next loaf comes out a little better. You repeat the work of choosing ingredients, making, tasting, and adjusting many times.

Research that finds new materials is similar. A researcher decides the ratio of ingredients. The researcher mixes powders and bakes them at a high temperature. The structure and properties of the material are measured. If the result is poor, the conditions are changed. This process must be repeated hundreds of times to produce one good material.

The problem is time. If a person does this by hand, it can be done only a few times a day. Experiments stop at night. It can take years to search a wide range while changing ingredients little by little. Materials for rockets must withstand difficult conditions. So there are very many candidate materials, and it is hard to test them all by hand.

A self-driving laboratory is a research laboratory that gives this repetition to machines. Artificial intelligence (AI) chooses the conditions to test next. Robots weigh, mix, and bake powders. Analytical equipment measures the results. Those results go back to the AI. Experiments continue on their own without a person giving instructions each time. Like a robot vacuum that cleans by itself, the laboratory runs by itself through the night.

This structure, in which results return to the next choice, is called a closed loop. A closed loop means a circuit that is closed like a ring. Experiment, measurement, learning, and the next experiment keep turning without a break. This chapter explains how this self-driving laboratory works. It examines how AI chooses the next experiment. It looks at how machines read sentences in papers. It explains how robots and AI talk safely. Finally, it looks at real self-driving laboratories around the world.

This chapter also explains why self-driving laboratories are rising now. Recent AI quickly predicts the properties of materials. Computation can produce thousands of good candidates in a day. But actually making and checking them is slow. Prediction has become fast, but verification has not kept up. A self-driving laboratory fills this gap. Robots quickly make and check the candidates produced by computation. It is a bridge that aligns the speed of prediction with the speed of verification.

Many systems in this chapter are still at an early research stage. Some are preprints just presented to academic communities. A preprint is early research that has not gone through peer review. So this chapter looks at both principles and limits, rather than boasting about performance numbers. The goal of a self-driving laboratory is not fast experiments. It is trustworthy experiments.

3.1 What Parts Make a Self-Driving Laboratory Run?

What this section covers

This section looks at the parts that make up a whole self-driving laboratory. There are four main parts. They are the head that decides the next experiment, the hands that actually make it, the eyes that measure the result, and the memory that updates what has been learned. These four parts are linked like a loop. They pass experimental results to the next experiment and keep running.

First, it reviews the meaning of the term closed loop. Then it examines what each part does. Finally, it explains why this structure is needed for rocket materials research.

A self-driving laboratory runs through four connected parts. In the planning layer, AI chooses the composition and process conditions to test next. Here, process conditions mean baking temperature, holding time, heating rate, and atmospheric gas. In the execution layer, robots weigh and mix powders, then press them into pellets. Then they place them in an electric furnace and bake them. In the measurement layer, tools such as X-ray diffraction equipment measure the structure of the material that has been made. In the learning layer, measurement results enter the AI model and update its knowledge.

What links these four parts is the flow of information. The planning layer sends a work specification to the execution layer. A work specification is an instruction sheet that says what to do, how much to use, and how to do it. The execution layer sends the prepared sample to the measurement layer. The measurement layer sends values such as phase purity and lattice constant to the learning layer. Phase purity means how purely the desired crystal was obtained. The learning layer uses those values to refine the next plan. When information completes one full cycle, one experiment is finished.

The strength of a closed loop lies in fast repetition. Even if one experiment fails, the data remain. Failed conditions also show what does not work. AI learns from both success and failure. So the next conditions improve little by little. A person checks the larger direction once every few days, while the machine handles the detailed repetition.

Now consider why rocket materials need this structure. Rocket nozzles and combustion chambers withstand extreme conditions. Temperatures rise to thousands of degrees. Pressure and vibration are also high. Materials used in such places must meet very demanding conditions. A small change in composition can greatly change properties. So there are tens of thousands of candidates to test. Human hands cannot search this wide space completely. A self-driving laboratory can quickly narrow this wide space.

In 2026, research continued to raise self-driving laboratories to the next level. The Royal Society of Chemistry in the United Kingdom organized the concept of self-driving laboratories 2.0. The first generation of self-driving laboratories solved only one fixed problem. The next generation aims to handle many kinds of equipment and problems flexibly. A 2025 review paper that broadly examined self-driving laboratories was also published. This paper discusses both the maturity of the technology and the barriers that remain. A publication of the American Chemical Society reports that self-driving laboratories are changing how chemists work.

This structure also has clear limits. Equipment costs a great deal of money. Connecting robots and analytical instruments is also difficult. Even weighing powders accurately is not easy. Automatically detecting failed experiments is difficult too. So today's self-driving laboratories fit only certain material groups well. A universal laboratory that can handle every material does not yet exist. The remaining sections of this chapter examine how these barriers are being overcome one by one.

3.1.1 Bayesian Active Learning for Choosing the Next Experiment

This subsection covers the part that serves as the head of a self-driving laboratory. The key question is how AI chooses the next experiment. The method used here is Bayesian active learning. This subsection explains in simple terms what this method is and why it reduces the number of experiments.

A hypothesis is an idea about why the next experiment will be good. A good self-driving laboratory forms good hypotheses. Testing random conditions wastes time. It must choose and test places that are likely to work. Bayesian active learning helps with this selection. Active learning means a method in

which a machine chooses the data it will learn from by itself.

This method first builds a surrogate model. A surrogate model is a small computational model that imitates an expensive experiment. It is built from experimental results obtained so far. The surrogate model predicts the properties of conditions that have not yet been tested. Two values come out with each prediction. One is the expected performance. The other is how uncertain that prediction is.

The key is to look at these two values together. A place with high expected performance is worth testing. A place with high uncertainty is also worth testing. That is because an unexpectedly good material may be hidden in a poorly known area. If we chase only performance, we circle around places we already know. If we chase only uncertainty, we wander through useless places. A good method balances these two aims. This is called the balance between exploration and exploitation. Here, exploration means examining unknown places, and exploitation means digging deeper into known places.

An acquisition function measures this balance as a number. The acquisition function gives each candidate condition a score. If the candidates are lined up by score, the one at the front becomes the next experiment. A representative acquisition function is Expected Improvement. Expected Improvement measures how much the next experiment is expected to improve on the best result so far. Another method is Upper Confidence Bound. This method makes the system consider both high performance and unknown regions. A self-driving laboratory uses these scores to decide the next composition, temperature, and time.

Now consider why this method matters for rocket materials. Ultra-high-temperature ceramics and special alloys cost a great deal for each experiment. They use expensive raw materials and require long baking times. So reducing the number of experiments directly saves money and time. Bayesian active learning selects and tests only conditions that are likely to work. It greatly reduces the number of experiments compared with testing everything blindly. Some self-driving laboratories report that they reduced the number of required experiments by dozens of times.

Real cases show this effect. In 2026, one research team built a laboratory that autonomously grows thin films. A thin film is made by stacking atoms layer by layer. This laboratory read electron diffraction patterns with a real-time camera. A diffraction pattern is a signal that shows whether atoms are stacking well. The AI watched this pattern and chose the next conditions. As a result, the number of required experiments fell sharply. The team said it was reduced by more than thirty times compared with checking conditions one by one. Good conditions were found much faster. This research is also at an early stage, but it clearly shows the direction.

In 2026, research actively refined this method for self-driving laboratories. One preprint organized Bayesian optimization for autonomous materials laboratories. It explains the path from algorithms to real experimental workflows. Another preprint proposed a method that narrows the search space by itself. It reduced the candidate space by about half through active learning while preserving most of the good solutions. This helps researchers find good materials faster. A practical guide to Bayesian optimization for experimental science was also published in 2026.

This method also has limits. At first, the surrogate model has little data, so its predictions are rough. The first few experiments may point in the wrong direction. If the composition space is very wide, the amount of computation grows. If measurements contain a lot of noise, predictions become unstable. So people do not simply trust the predictions of the surrogate model. They use predictions as a guide, but always check them with real measurements. Human supervision remains even in a self-driving laboratory.

3.2 Extracting Synthesis Methods from Papers and Turning Them into Safe Plans

What this section covers

This section covers two things. One is making machines read synthesis methods in papers. The other is turning those methods into plans that robots can execute safely.

People read papers and picture the experimental steps in their minds. Machines cannot do that. So sentences must be changed into a form that machines can handle. This section examines that translation process and safe planning.

In papers, synthesis methods are written in sentences. They state which powders are mixed and in what amounts. They state at what temperature and for how many hours the material is baked. They state what gas the material is treated in. Papers accumulated over decades contain millions of such methods. This knowledge is a major foundation for new materials research. But most of it is scattered in human language.

The problem is that these sentences differ from person to person. Some papers say the material was baked overnight. Some only say it was treated at a high temperature. They do not state exactly how many hours overnight means or how many degrees high temperature means. People fill these gaps with experience. Machines cannot do that. So scattered sentences must be converted into exact numbers.

This work is valuable in rocket materials research. Synthesis methods for ultra-high-temperature ceramics and special alloys are scattered across old papers. It would take a long time for people to read and organize all of them. If machines automatically extract and organize the sentences, they save a great deal of time. A self-driving laboratory turns the collected methods directly into robot instructions, shortening the distance from paper to experiment.

When this knowledge is scattered, repeated research occurs. Another researcher tests conditions that have already failed. Researchers start from the beginning because they do not know a method that has already succeeded. If methods in papers are gathered in one place, this waste decreases. It is difficult for people to read millions of papers. So it is valuable for machines to extract and organize sentences. Well-organized methods become a good starting point for a self-driving laboratory.

There are two steps here. One is extracting methods from sentences. A Large Language Model does this work. A Large Language Model is AI that learns and handles human writing. The other is turning the extracted methods into safe plans. A planning language does this work. The next two subsections examine each step in detail.

3.2.1 Language Models That Build Synthesis Plans from Papers (MSP-LLM)

This subsection explains how to extract synthesis methods from papers. It examines the MSP-LLM study as a representative case. It looks at what problem this study tries to solve and what steps it follows.

MSP-LLM is a study on a large language model that plans materials synthesis. Its name comes from the title, an integrated large language model framework for complete materials synthesis planning. It is an early study released in 2026. The researchers are Heewoong Noh, Kyoungsoo Na, Namkyung Lee, and Chanyoung Park. This study is a preprint presented at a conference. So it should be seen as an early attempt, not as a field standard.

The problem this study addresses is synthesis planning. Synthesis planning means setting the full sequence for making a target material. This task is divided into two smaller problems. One is precursor prediction. A precursor is a raw material added at the beginning to make the target material. The other

is synthesis operation prediction. A synthesis operation is the sequence of actions such as mixing, heating, and cooling. A complete plan comes out only when both problems are solved together.

MSP-LLM connects these two problems into one flow. First, it decides what type of material the target material is. This type becomes an intermediate basis for judgment. Then it selects precursors that fit that type. Finally, it builds an operation sequence suited to those precursors. In this way, it creates a chain from type to raw materials, and from raw materials to sequence. When linked as a chain, the plan becomes chemically coherent from beginning to end.

A simpler example helps explain this flow. Deciding what dish to make corresponds to the target material type. It is like deciding to make kimchi stew. Then deciding what to buy corresponds to precursor prediction. It is like choosing kimchi, pork, and tofu. Finally, deciding the cooking order corresponds to operation prediction. It is like setting the order of stir-frying, adding water, and boiling. MSP-LLM completes the plan by connecting these three steps.

This kind of automatic planning is valuable for rocket materials. To make a new alloy or ceramic, even choosing the raw materials is difficult. Even for the same target, the result changes if the raw materials change. The result also changes if the heating sequence changes. It takes a long time for people to examine all these combinations. An automatic planner quickly provides a good starting point based on knowledge from papers. A self-driving laboratory can begin experiments from this starting point.

This method also has clear limits. A large language model sometimes invents facts that do not exist. This is called hallucination. If invented raw materials or conditions are used directly in experiments, resources are wasted. It may also produce dangerous combinations. So an extracted plan must always be checked by a person or another verification tool. First, one must examine whether it follows chemical rules and whether it is safe. Automatic planning is a tool that helps people, not a tool that pushes people out.

3.2.2 Preventing Robot Accidents with a Planning Language (PDDL)

This subsection explains how to turn an extracted synthesis method into a safe robot plan. The tool used here is the planning language PDDL. It explains in simple terms how this language prevents robot accidents.

An experimental plan is not just a sequence chart. Each action has conditions that must be met in advance. A sample is placed in an electric furnace only when the furnace is empty. A sample is taken out only after the furnace has cooled. Mixing is done only after powder weighing is finished. If these before-and-after conditions are violated, accidents occur. It is dangerous if a robot opens the door of a hot electric furnace carelessly.

PDDL (Planning Domain Definition Language) is a language for writing such plans. In Korean, it means a language for defining a planning domain. This language records two things for each action. One is the precondition. It states what must be true before this action can be performed. The other is the effect. It states what changes when this action is performed. For example, the preconditions for a heating action are that the electric furnace is in standby mode and that its temperature is low. The effect after completing this action is that the sample has been heat-treated.

When conditions are written in this way, the machine finds a safe sequence by itself. An action whose conditions are not met is not executed. An access command to a hot electric furnace is blocked. A sample that has not been weighed does not move on to mixing. These rules become a fence that prevents robot accidents in advance. They translate laboratory safety rules followed by people into machine language.

However, this fence does not guarantee safety by itself. Preconditions work only when people enter each one accurately. If a condition is omitted or written incorrectly, a hole appears in the fence. So PDDL is not a device that guarantees safe execution. It is a tool that lets people anticipate risks and turn them into rules. The density of the rules is only as dense as human judgment.

This plan can be shown as a diagram. Each action is placed as a node, and before-and-after relationships are connected with arrows. This kind of diagram is called a directed acyclic graph. It means that the arrows flow in one direction and do not come back. This diagram shows which tasks can be done at the same time. For example, while one sample is being heated, the powder for another sample is weighed. If tasks that do not interfere with each other run side by side, the whole laboratory operates that much faster.

This safety planning is valuable in rocket materials experiments. An ultra-high-temperature electric furnace is very dangerous if handled incorrectly. Highly reactive raw materials must not come into contact with air. Many pieces of equipment move at the same time in one laboratory. In a laboratory that runs overnight without people, rules to prevent accidents are essential. Planning languages such as PDDL help build these rules in detail. The speed of a self-driving laboratory has meaning only on top of this safety device.

This approach also has limits. It is hard to write every condition in the world as a rule. Situations that were not written down may actually occur. If there are too many rules, the computational load for finding a plan increases. Unexpected failures cannot be prevented by rules alone. So a planning language alone is not enough. Other devices that monitor equipment status in real time must work together with it, and the next section discusses those devices.

3.3 Standard Communication Connecting Artificial Intelligence and Experimental Equipment (MCP)

This section explains how artificial intelligence and experimental equipment communicate. A language model handles sentences. A scale, an electric furnace, and a robot arm require precise commands. An intermediate layer is needed to connect the two. This section examines that intermediate layer. It looks at standard communication rules together with the safety devices placed on top of them.

A language model handles human writing well. It can understand a sentence telling it to raise the temperature slowly. But an electric furnace cannot understand sentences. An electric furnace must receive a precise signal telling it how many degrees per minute to increase. The same is true for a scale and a robot arm. So a translator is needed to turn sentences into the language of equipment.

In the past, this translation was built separately for each piece of equipment. A translator for a scale, a translator for an electric furnace, and a translator for a robot were each written separately. When one piece of equipment was changed, the translator also had to be rewritten. Many laboratories repeated the same work. This was a large waste. Without a standard, it was also hard to share code with one another.

Standard communication rules try to solve this problem. With standard rules, each piece of equipment is connected in the same way. New equipment can be attached right away if it follows the rules. It also becomes easier for laboratories to share code. Recent self-driving laboratory research looks for this standard in the Model Context Protocol.

There is an open rule called the Model Context Protocol. It is shortened to MCP. A protocol is an agreed communication method. MCP is a rule that connects artificial intelligence and software tools. It tells artificial intelligence, in a standard way, what tools are available. The artificial intelligence recognizes those tools by itself and calls them.

There is a point here that is easy to misunderstand. MCP is a high-level rule that connects software to software. It cannot directly move physical equipment such as scales and electric furnaces. The work of operating motors and valves is handled by drivers that exist separately for each piece of equipment. A driver is a lower-level program that directly handles equipment signals. MCP sits on top of this driver and connects it to artificial intelligence. So equipment is not connected directly by MCP alone. MCP is neither an industrial real-time control standard nor a safety standard. To connect equipment, a separate driver layer must be placed underneath it.

One case that brings this rule into a self-driving laboratory is the NIMO controller. It is an early study released in 2026. The researchers are Naruki Yoshikawa and Ryo Tamura. This study exposes several functions of a self-driving laboratory as MCP servers. Here, NIMO is a decision engine that selects the next experiment. The controller is an orchestration layer that connects that engine and several tools through MCP. In this case, artificial intelligence operated a self-driving laboratory that matched colors. It changed the components by itself and ran experiments until the target color appeared. What this study actually showed was this orchestration and color-matching experiment.

The purpose of NIMO stated in this study is clear. A self-driving laboratory has many optimization methods. It also has many types of tools. Connecting each method and tool one by one is a major bottleneck. The NIMO controller connects the two in a common way. People can build an experimental flow without writing code. They do this by choosing tools on a screen and connecting them. However, this study is still at an early stage, so it is too early to regard it as a widely used standard.

This communication layer matters in rocket materials laboratories. A materials laboratory contains equipment from different companies. The manufacturers of scales, electric furnaces, and diffraction equipment are all different. For one artificial intelligence system to handle them all, there must be a common language. Standard communication rules bind these many pieces of equipment into one flow. Even when new equipment is brought in, it can be attached without major construction. This makes it easier to expand or change a laboratory.

Standard communication has another advantage. It lets people design experiments even if they do not know code. Artificial intelligence recognizes by itself what tools are available. It can lay those tools out on a screen. A person only needs to choose the tools and connect them in order. The flow of weighing with a scale, mixing, heating, and measuring with diffraction equipment can be arranged like a diagram. When the barrier to designing experiments becomes lower, more researchers use self-driving laboratories.

This layer also has limits. The communication rules are still new and are still taking shape. Each equipment company follows the rules to a different degree. Old equipment is hard to adapt to the rules. If communication is cut off or signals are delayed, the experiment goes wrong. So the communication layer itself also needs a safety device. This device does not execute commands immediately but checks them first. That device is the state machine.

3.3.1 The Final Safety Line Protected by a State Machine

This subsection discusses the state machine that protects equipment safety. It looks at what a state machine is and how it blocks dangerous commands. It examines why this device is the final safety line in a self-driving laboratory. The state machine discussed here is not a function already fully built into a specific product. The NIMO controller discussed earlier is a study focused on orchestrating tools. A state machine and interlocks are a common safety design that must be placed separately on top of that orchestration. This book treats that safety design as a layer that a self-driving laboratory must have.

A state machine is a method that divides the current status of equipment into defined states. A finite state machine defines the possible states in advance. For example, there may be states such as standby, preparation, weighing, transfer, reaction, analysis, error, and emergency stop. Equipment is always in one of these states. The paths from one state to the next are also defined in advance.

The key is that only actions allowed in a given state are executed. In the reaction state, the electric furnace is hot. At this time, a command for the robot arm to open the furnace door is blocked. The door opens only after a signal confirms that the temperature has cooled below a safe value. A sensor confirms this signal. Before the confirmation signal arrives, the command is blocked at the state machine stage. This kind of blocking is called an interlock. An interlock is a safety lock that locks when conditions are not met.

A simple example shows why this method is safe. A microwave oven stops operating when the door opens. It runs only when the door is closed. If the door is open, it will not turn on no matter how many times the button is pressed. This is action restriction based on state. A state machine in a self-driving laboratory works on the same principle. In a dangerous state, dangerous commands are not executed at all.

Sentence commands from artificial intelligence must pass through these rules. No matter how confidently artificial intelligence gives a command, the rules come first. Only commands that pass the rules go to the equipment driver. The driver is the program that moves the actual motors and valves. Commands caught by the rules are not executed, and only a record is left. This record later becomes material for finding the cause.

In rocket materials experiments, this safety line is valuable. They handle ultra-high-temperature electric furnaces and reactive raw materials. Even a small mistake can lead to a major accident. It is even more dangerous at night when no one is present. The state machine is the last safety gate between artificial intelligence and equipment. Even if the AI makes a mistake, the equipment does not perform a dangerous action. The freedom of a self-driving laboratory is allowed only inside this gate.

This device also has limits. It is hard to handle states that were not defined in advance. Unexpected failures do not fit neatly into fixed boxes. If a sensor sends the wrong signal, the judgment is also wrong. So a state machine alone is not enough. It needs eyes that observe equipment in real time and detect abnormalities. Those eyes are the sensor fusion examined in the next section.

3.4 A Robot Execution Framework That Withstands Disorderly Laboratory Environments (PUDA)

This section deals with the difficulties robots face in real laboratory environments. A laboratory is different from a factory. Powders flow, clump, and carry static electricity. Crucibles become slightly misaligned. Sensor signals include noise. This section examines a robot execution framework that can withstand such disorderly environments. It discusses the PUDA study as a representative case.

Laboratory robots face harder tasks than factory robots. In a factory, the same part arrives at a fixed position. In a laboratory, the properties of powders differ each time. Some powders flow well. Some powders clump and block nozzles. Some powders carry static electricity and stick here and there. Crucibles also shift slightly each time they are placed.

This kind of disorder is called disturbance. A disturbance is an unplanned interference. When there is a disturbance, fixed motions alone cannot get the work done. The robot must have eyes and senses. The eyes are cameras, and the senses are force sensors. The robot reads this information in real time and corrects its motion.

PUDA is a case that deals with this kind of robot execution environment. The name stands for Physical Unified Device Architecture. It is also called an AI-friendly hardware harness for a self-driving laboratory. A harness is a framework that ties several pieces of equipment together and handles them as one. It is an early study released in 2026. The researchers are Zenkun Ren, Hongzhao Tan, Jiaen Yi, and Kedar Hippalgaonkar. What this study actually showed was a command-line execution environment that is easy for AI to handle, and a record system within it. The sensor fusion and error recovery discussed below were not demonstrated as performance results in this paper. They are the direction in which this kind of execution framework aims to move.

PUDA's method is somewhat different from earlier systems. It does not put a human-facing screen first. Instead, it creates a command-line environment that is easy for AI agents to handle. Each device appears as a tool that can be found from the command line. Commands and responses remain as standard records. A chain connects which command went to which sample and what result came out. This history is called provenance. Provenance is the history of where data came from and how it was produced.

Consider why this history record is valuable in rocket materials research. Aerospace parts require strict quality tracking. The record must show which raw materials were used and under what conditions they were made. If a problem occurs later, this record is used to find the cause. A system like PUDA leaves this history automatically. It has fewer gaps than records written by hand. Data produced by a self-driving laboratory can then be trusted and used.

There is a reason PUDA chose the command-line method. Screens for humans look good. But they are inconvenient for AI agents to handle. Buttons on a screen are hard for a machine to press. Command-line tools are easy for AI to call directly. AI observes the situation, makes judgments, and issues commands. The equipment executes those commands accurately and reversibly. It is as if the brain that plans the experiment and the hand that moves the equipment have been separated. With this division, the hand can remain the same even if the brain changes.

This method also has limits. A command-line environment is inconvenient for people to view directly. Command-line tools must be newly created for each device. Old equipment is hard to fit into this framework. Leaving standard records requires effort in storage and management. This research is still at an early stage. More time is needed before it can be widely used in many laboratories.

3.4.1 Combining Multiple Sensors to Judge and Recover from Failure

This subsection deals with sensor fusion, which combines information from several sensors to make judgments. It also examines the recovery process that reverses failures on its own. It looks at why these two functions protect experimental quality. Both functions are widely studied methods in robotics. An execution framework such as PUDA from the previous section is a vessel meant to contain these methods. Here, we examine the principles of the methods, not the achievement of any one system.

Sensor fusion is the work of combining information from several sensors to make a judgment. A camera sees position and shape. A force sensor detects the force applied to the robot wrist. Temperature sensors and pressure sensors check whether equipment is safe. If only one type of sensor is used, much can be missed. When several sensors are used together, judgment becomes more accurate.

Take the example of a robot gripping a crucible. If the camera is blocked, it cannot see the position. At that moment, the force sensor detects a collision. If the force is too small to sense clearly, the camera sees spilled powder. Different sensors fill one another's gaps. It is similar to how people use both hands and eyes. In a dark room, we feel with our hands, and in a bright place, we look with our eyes.

When a robot hits something, the force sensor value jumps sharply. If it exceeds a set limit, the robot stops moving at once. After stopping, it examines the situation again. It checks with a camera whether

powder has spilled or whether the crucible has slipped. This rechecking process is the beginning of error recovery.

Error recovery usually proceeds in three steps. First, it detects what went wrong. Next, it returns the robot to a safe state just before the failure. Finally, it slightly changes the gripping position or force and tries once more. Through these three steps, small mistakes can be handled without a person.

Error recovery is also important for experimental quality. If a wrongly weighed sample is carried forward as it is, bad data accumulates. Bad data leads an AI model in the wrong direction. So a self-driving laboratory records failures instead of hiding them. Recorded failures become data for the next round of learning. If the system knows where and why it failed, it avoids that mistake next time.

This function also has limits. Sensor values also include noise. If noise is misread as an abnormality, a normal experiment stops. If experiments stop too often, the process becomes slow. Conversely, if the limit is set too loosely, a real accident is missed. So the sensitivity of this device must be tuned well. This tuning still depends heavily on human experience. Autonomous recovery is not a device for removing people. It is a device for reducing dangerous repetition and calling a person when needed.

3.5 Berkeley A-Lab, a Self-Driving Laboratory That Actually Moved

This section examines a representative case of a self-driving laboratory that actually operated. It is A-Lab at Lawrence Berkeley National Laboratory in the United States. It looks at what this case achieved and what barriers it encountered. It also explains how to read its performance numbers.

A-Lab is an autonomous inorganic materials synthesis platform built by Lawrence Berkeley National Laboratory. Inorganic materials are substances made of elements such as metals and oxygen. Oxides and phosphates belong to this group. A-Lab selected promising candidates from computational data. Then it handled powders and heat-treated them with robots. It checked the results with X-ray diffraction equipment. It carried out this process with almost no human hands involved.

It is important to read the performance numbers accurately. A-Lab was reported in a 2023 Nature paper. The success number first reported was later reviewed again and lowered. Other materials chemists reanalyzed the diffraction patterns. In 2026, Nature published a correction. After the correction, the number was 36 out of 57 targets. This book uses the corrected number as its basis. It does not use the earlier, higher number.

This correction process itself offers a valuable lesson. The core of the correction has two parts. One is the misjudgment made by automated analysis. Success judgments were made by automated X-ray diffraction analysis. This analysis mistook signals from unreacted remaining raw materials for signals from the target new material. On the surface, it looked as if a new crystal had formed. This misjudgment became clear only after people examined the diffraction patterns again. The other issue is the meaning of the phrase new material. At first, it seemed to mean a material that had never existed in the world. On review, in many cases it meant a material that was not in this laboratory's data. That is different from a material that is new to the entire scientific community. The correction stated that the reported successes were rechecked by people, and some were left uncertain. Therefore, the judgment of a self-driving laboratory is not the final answer. It becomes a true success only after precise analysis and reproducibility testing. If verification is skipped because autonomous experimentation is fast, errors accumulate instead.

Consider what this case means for rocket materials research. A-Lab showed that autonomous synthesis is not an idea outside the laboratory. Oxides and ceramics are connected to thermal protection materials for rockets. A path has opened for exploring such materials faster than people can. But strict success

judgment must come with it. In aerospace parts, even a small impurity can become a major problem. So verification is as important as speed.

A-Lab revealed three clear barriers. The next subsection examines each barrier in detail. These barriers are also shared by other self-driving laboratories. Knowing the barriers accurately becomes the foundation for building the next system.

3.5.1 System Configuration and Operational Results

This subsection examines what parts made up A-Lab. It also looks at how those parts worked together. It draws a concrete picture of what a self-driving laboratory looks like in practice.

A-Lab's structure can be divided into three parts. The computational layer selects candidate materials and synthesis methods. In the execution layer, robots and equipment actually make the samples. In the measurement layer, X-ray diffraction equipment checks the results. These three layers run in a loop, from computation to execution and from execution to measurement.

The computational layer selected candidates from large databases. It used public data called the Materials Project. It also used computational data from Google. These data predict materials that are likely to be thermodynamically stable. Stable means that the material does not easily break apart and can be maintained. A-Lab used these predicted materials as its targets.

The execution layer had several robot arms. The robots moved samples among stations for weighing, mixing, pressing, and firing powders. They weighed powders accurately and mixed them with a ball mill. A ball mill is a device that grinds and mixes powder in a container filled with small balls. The mixed powder was pressed into small pellets. Those pellets were fired in a high-temperature electric furnace.

The key device in the measurement layer is an X-ray diffractometer. X-ray diffraction is a method for seeing crystal structure by shooting X-rays at a material. Each crystal has a different pattern of scattered X-rays. This pattern is like the fingerprint of a material. AI reads this pattern and judges whether the target material has been made. If the pattern matches the target, it is counted as a success.

A-Lab ran this process continuously for several days. Experiments continued even while people slept at night. This shows that the self-driving laboratory moved from an idea in a paper to actual equipment. It made several target materials without human intervention. This achievement remains a case that opened the door to autonomous materials synthesis.

But it was not fully unmanned operation. Human researchers prepared and supervised the system. People also examined failed experiments. Success judgments were later checked again by people. A self-driving laboratory reduces human work, but it does not erase people. People take charge of higher-level judgment and verification. Machines take charge of detailed repetition. This division of labor is the current reality of self-driving laboratories.

3.5.2 Three Barriers Revealed by A-Lab

This subsection examines the three barriers revealed by A-Lab one by one. These barriers are common tasks that self-driving laboratories must overcome. The problems must be understood accurately to build the next system well.

The first barrier is the difficulty of judging synthesis results. Some materials counted as successes were not single-phase. Single-phase means a state in which only the desired crystal exists in pure form. Some samples left unreacted raw materials. Some samples contained secondary phases. A secondary phase is another crystal that is not the target. Even if a diffraction pattern resembles the target, impurities may

be mixed in. So whether a material is single-phase must be checked strictly.

The second wall is powder handling. Each type of powder has different properties. Some powders carry static electricity. Titanium dioxide and nanopowders are examples. These powders stick inside the nozzle. Then the metering nozzle becomes clogged. Raw materials that absorb moisture easily are also hard to handle. Automatic dispensers often stop in front of these varied powders. A human can tap the container by hand and move on, but a machine gets stuck.

The third wall is environmental control. A-Lab operated in an environment open to the atmosphere. Air contains oxygen and moisture. Materials that are sensitive to oxygen and moisture cannot be handled in this environment. Their properties change or they decompose as soon as they are made. Sulfides and halide materials for all-solid-state batteries face this problem. These materials become unusable when they contact air. So A-Lab, open to the atmosphere, could not search this class of materials.

These walls clearly show why automation is difficult. Humans overcome these problems through the experience of their eyes and hands. If powder clumps, they tap the container. If a sample seems strange, they notice it by smell and color. A machine must be taught each of these experiences one by one. What humans do without thinking is hard for machines. The progress of the self-driving laboratory is the process of transferring these small senses to machines.

These three walls are intertwined. Poor powder handling leads to wrong measurements. Wrong measurements also shake the judgment of results. Without environmental control, some materials cannot be made at all. To expand self-driving laboratories, these walls must be crossed together. Fixing only one of them makes major progress difficult.

In rocket materials research, these walls feel even larger. New battery materials or special compounds for rockets are often highly reactive. Many materials are also sensitive to air. To handle these materials, environmental control must come first. A laboratory open to the atmosphere has difficulty touching this class of materials. That is why the third wall, environmental control, is especially large in rocket materials research.

The effort to cross this wall is the subject of the next section. A sealed environment is needed to handle materials that are sensitive to air. That environment is a glovebox. The next section looks at a self-driving laboratory that operates in a sealed environment. It is a system aimed directly at A-Lab's third wall.

3.6 Finding Air-Sensitive Materials in a Sealed Environment (A-Lab GPSS)

This section covers a self-driving laboratory that operates in a sealed environment. It is a system that tries to cross the third wall discussed in the previous section. It examines how to handle materials that are sensitive to air.

A glovebox is a sealed box with gloves attached. Thick rubber gloves are attached to the wall of the box. A person puts their hands into the gloves from outside and handles what is inside. The box is usually filled with an inert gas such as argon. An inert gas is a gas that does not react easily with other substances. So there is almost no oxygen or moisture inside the box.

This system is called A-Lab GPSS. GPSS means glovebox-protected synthesis system. It is the next generation, made to overcome the limits of A-Lab open to the atmosphere. The equipment was placed inside a large glovebox filled with ultrahigh-purity argon. Moisture and oxygen are kept below one part per million. Inside it, raw material storage, weighing, forming, sealing, and heat treatment are all done. The key is to remove every moment when the sample touches air.

This case is confirmed in a 2026 preprint. The researchers include Feiyu and others. It is an early study dealing with air-sensitive lithium halide spinel conductors. The paper states that 352 samples were made. There were 171 possible metal-pair combinations. Of these, 72 percent were handled by experiment. Here, 72 percent is not a success rate. It is the share of possible combinations covered by experiment. This number must not be misread as a success rate.

The paper states that results improved as experiments continued. In the first 75 samples, the rate of good results was low. In the last 75 samples, that rate increased. The artificial intelligence learned from earlier experiments and improved later experiments. This shows that closed-loop learning actually works. However, this study is still at an early stage, so it should not be expanded into broad conclusions.

Let us see why this sealed self-driving laboratory is valuable in the rocket industry. Spacecraft need high-performance batteries. All-solid-state batteries are safe and high-energy candidates. Their core is the solid electrolyte. But good solid electrolytes are often sensitive to air. A sealed self-driving laboratory can search these materials safely. The same environment is used for research on highly reactive rocket materials.

3.6.1 Air-Sensitive Lithium Halide Solid Electrolytes

This subsection looks closely at the materials covered here. It explains lithium halide solid electrolytes in simple terms. It examines why these materials are sensitive to air and why they are valuable.

First, let us unpack battery terms. A lithium ion is a small particle that carries charge inside a battery. This ion moves back and forth when the battery charges and discharges. An electrolyte is the path through which this ion travels. Common batteries use liquid electrolytes. Liquids can leak or catch fire.

A solid electrolyte makes this path out of a solid. A solid electrolyte must let lithium ions pass well. But it must block electrons well. A solid that passes ions and blocks electrons is a good electrolyte. A solid does not leak and resists fire. So all-solid-state batteries are considered safer batteries.

A lithium halide electrolyte is a solid electrolyte that contains halogen elements. Halogens are elements such as chlorine, bromine, and iodine. In chemical formulas, lithium, metals, and halogens are connected in fixed ratios. This material may work well with high-voltage electrodes. Higher voltage stores more energy in the same size. So it draws attention as a next-generation battery material.

The problem is that this material is very weak against water. This property is called deliquescence. Deliquescence is the property of absorbing moisture from the air and dissolving by itself. Even contact with trace moisture decomposes the phase. Harmful gases such as hydrogen chloride can be released during decomposition. So raw material storage, synthesis, and measurement must all be done in a sealed environment.

This condition makes the self-driving laboratory difficult. The inside of a sealed box is narrow and inconvenient to handle. Robots must move precisely in a confined space. Air exposure cannot be allowed even for a moment. Artificial intelligence proposes candidate compositions and synthesis conditions. Robots repeat the same procedure without air exposure. Analytical equipment quickly checks ionic conductivity and crystal phases. All these tasks must mesh inside the sealed space.

Measuring ionic conductivity is also important. Ionic conductivity is a value that shows how well lithium ions pass through. This value must be high for the battery to deliver power well. A self-driving laboratory measures the conductivity of a prepared sample right away. Then it uses that value as the target for the artificial intelligence. It chooses the next experiment toward compositions with higher conductivity. It narrows down good electrolytes by looking at crystal phases and conductivity together.

Limits still remain in searching these materials. In a sealed environment, putting equipment in and taking it out is difficult. The narrow space limits the equipment that can be handled. Deliquescent materials become unusable with even a small mistake. So every experiment must be careful. Even if automation increases speed, this need for care does not decrease. Sealed self-driving laboratories still handle only a narrow range of materials.

3.6.2 A Dual Loop Between Induction and Abduction

This subsection deals with two ways of thinking used by self-driving laboratories. They are induction and abduction. It looks at why search becomes faster when the two are used in turn. This method is the inference structure revealed by the A-Lab GPSS study in this section.

Induction is the way of thinking that finds common rules from many experimental results. Diffraction patterns and conductivity measurements are collected. Researchers examine which compositions were made well. They place many samples side by side and read the trends. They check how conductivity changes as the lattice expands. The rules built in this way become the basis for the next experiment.

Abduction is the way of thinking that searches for the cause when an unexpected result appears. Abduction is inference that chooses the cause that best explains a strange result. Suppose an unexpected impurity phase appears in one sample. Researchers ask whether the raw materials reacted incompletely or the temperature was too low. Among these, they choose the most plausible cause and change the next experiment. Abduction finds the next move directly from one surprising result. Its direction differs from induction, which extracts rules from many results.

A self-driving laboratory uses these two ways of thinking in turn. Density Functional Theory calculations first narrow down promising candidates. These calculations find the energy of a material at the atomic level. Experiments then test the selected candidates. When experimental results accumulate, induction finds rules. When unexpected results appear, abduction identifies causes and changes direction. In this way, calculation and experiment refine each other. Thinking does not flow in only one direction. It turns like a loop.

Let us look at this method through a simple example. Imagine finding your way in an unfamiliar city. Walking several times and finding the rules of shortcuts is induction. Guessing the cause and taking a detour when a road is suddenly blocked is abduction. If you trust only rules, you stop in unexpected situations. If you chase only guesses about causes, you lose the big picture. Using both helps you reach the destination faster. A self-driving laboratory also uses both ways of thinking to reach good materials faster.

Let us see why this dual loop is valuable in rocket materials research. New materials cannot be fully understood by calculation alone. Calculations assume ideal conditions. Real synthesis contains impurities and noise. Experiments alone also cannot reveal everything. Experiments take much time and cost. When the two methods are linked, calculation narrows the range and experiment confirms it. This combination is powerful in the search for expensive rocket materials.

This method also has limits. The languages of calculation and experiment are different. Matching calculation results with experimental results is difficult. If the calculation is wrong, experiments are wasted on the wrong candidates. If experimental noise is large, the calculation is corrected in the wrong way. So the bridge between the two loops must be built well. Building this bridge still depends heavily on human judgment.

3.6.3 Cooperation in Which Several Artificial Intelligences Divide Roles

This subsection deals with a method in which several artificial intelligences divide roles and cooperate. This is called a multi-agent system. It examines why several are better than one artificial intelligence, and what risks exist.

A multi-agent system is a structure in which several artificial intelligences divide roles. An agent is an artificial intelligence unit that judges and acts on its own. It is like a team dividing work instead of one person doing everything. In a self-driving laboratory, this division of roles is natural. Experiments require chemical knowledge, planning, safety, and analysis all together.

When roles are divided, each agent focuses on one task. The chemistry agent checks whether the reaction equation is correct. It checks whether mass is conserved and whether oxidation numbers match. It also examines whether raw materials react badly with the crucible. The planning agent organizes the order of equipment use. It divides bottleneck tasks such as long heat treatment in advance. It also decides which tasks can be done in parallel.

The safety agent is the eye that guards the laboratory. It watches the oxygen and moisture concentrations inside the glovebox. It also checks pressure changes and the flow of cooling water for the electric furnace. It repeatedly checks these values at very short intervals. If it sees an abnormal value, it raises an alert and leaves a record. However, this agent does not directly trigger an actual emergency stop. A generative agent may respond late or make a wrong judgment. If oxygen or pressure reaches a dangerous value, independent sensors and verified safety controllers stop the equipment immediately. This hardware interlock does not wait for the language model's judgment. The agent handles monitoring and records, while verified safety devices handle stopping. The analysis agent interprets measurement results. It matches diffraction patterns with databases. It assigns numbers to the amount of impurities. It puts those numbers into the learner's target values.

In 2026, research on these multi-agent laboratories increased greatly. One study organized several artificial intelligence agents under a coordinator. Each agent shared different tools. They were divided into groups that searched for knowledge, groups that converted designs into execution procedures, groups that handled robots, and groups that interpreted data. Grouped in this way, they achieved closed-loop materials discovery. Another study, ChatMat, automatically connected a materials property prediction workflow with multiple agents. Papers also appeared on the meaning of multi-agent systems in self-driving laboratory management.

The benefits of dividing roles are clear. It is easy to see what each agent is responsible for. When a problem occurs, it is easy to find the stage where it arose. One agent can be fixed without changing the whole system. New tools can also be attached easily. This is the same advantage that human team collaboration has.

Multi-agent systems also change the way science itself is done. In the past, human researchers recorded each experiment one by one. Now agents leave their decision process as data. The record shows why a condition was chosen and why a result was judged a success. This record is later used to trace the research back. But artificial intelligence decisions are not always transparent. It is hard for people to know every reason behind such decisions. This lack of transparency remains a problem to solve.

But the risks also grow. Large language models can invent facts that do not exist. One agent's error can spread to another agent. If several agents trust and pass work to one another, they may fail to catch the error. So final authority must remain with safety rules and human supervisors. Even when artificial intelligence speaks with confidence, the equipment may still be dangerous. As the number of agents grows, human cross-checking and hardware safety devices become more important.

Summary

This chapter examined how a self-driving laboratory operates. The key is a loop structure called a closed loop. Artificial intelligence chooses the next experiment. A robot makes it. Equipment measures it. The result goes back to the artificial intelligence. This loop runs day and night and quickly narrows a wide materials space.

The method artificial intelligence uses to choose the next experiment is Bayesian active learning. This method looks at expected performance and uncertainty together. It balances places that seem promising with places that are still unknown. In this way, it can reach good materials with fewer experiments. This is a major strength in the costly search for rocket materials.

Large language models extract synthesis methods from papers. Studies such as MSP-LLM perform this work. The extracted methods are safely converted into planning languages such as PDDL. A planning language becomes a fence that prevents robot accidents in advance. Artificial intelligence and equipment are connected through standards such as the Model Context Protocol. The NIMO controller is one example. A state machine is the final gate that blocks dangerous commands.

Robots must withstand disorderly experimental environments. Sensor fusion, which reads several sensors together, helps with this. Error recovery, which reverses failures on its own, works with it. Systems such as PUDA leave commands and results as records. These records are valuable for quality tracking of aerospace parts.

As real cases, we looked at A-Lab and A-Lab GPSS. Over seventeen days, A-Lab made 36 out of 57 targets. This figure reflects a 2026 correction. This case shows that judgments made by self-driving laboratories also need human review. A-Lab GPSS handled air-sensitive materials in a sealed environment. With 352 samples, it covered 72 percent of the possible metal-pair combinations. This number is combination coverage, not a success rate.

In 2026, this field expanded quickly. More multi-agent laboratories appeared, with several artificial intelligence systems dividing their roles. Attempts also continued to connect equipment through standard communication. Research was also active on adapting Bayesian optimization to experimental workflows. Review papers and guides that broadly examined self-driving laboratories also appeared. The larger direction is clear. It is to free human hands from detailed repetition. But judgment and verification still remain human responsibilities.

Most of the systems in this chapter are early-stage research. Many are preprints, and field standards are still limited. So this chapter looked at principles and limits together, rather than putting performance numbers first. A self-driving laboratory is not a tool that erases people. It assigns dangerous repetition to machines, while people take charge of judgment and verification. The true meaning of a self-driving laboratory is not fast experimentation, but trustworthy experimentation. Only on this principle can the automation of rocket materials research grow safely.

Related Supporting Materials

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., Sparkes, A., Whelan, K. E., & Clare, A. (2009). The automation of science. Science, 324(5923), 85-89. https://doi.org/10.1126/science.1165620

MacLeod, B. P., Parlane, F. G. L., Morrissey, T. D., Häse, F., Roch, L. M., Dettelbach, K. E., Moreira, R., Yunker, L. P. E., Rooney, M. B., Deeth, J. R., Lai, V., Ng, G. J., Situ, H., Zhang, R. H., Elliott, M. S., Haley, T. H., Dvorak, D. J., Aspuru-Guzik, A., Hein, J. E., & Berlinguette, C. P. (2020). Self-driving laboratory for accelerated discovery of thin-film materials. Science Advances, 6(20), eaaz8867. https://doi.org/10.1126/sciadv.aaz8867

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K. A., Zeng, Y., & Ceder, G. (2023). An autonomous laboratory for the accelerated synthesis of inorganic materials. Nature, 624(7990), 86-91. https://doi.org/10.1038/s41586-023-06734-w

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K. A., Zeng, Y., & Ceder, G. (2026). Author correction: An autonomous laboratory for the accelerated

synthesis of inorganic materials. Nature, 650, E1. https://doi.org/10.1038/s41586-025-09992-y

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. Nature, 624(7990), 80-85. https://doi.org/10.1038/s41586-023-06735-9

McDermott, M. J., Dwaraknath, S. S., & Persson, K. A. (2021). A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. Nature Communications, 12, 3097. https://doi.org/10.1038/s41467-021-23339-x

Tom, G., Schmid, S. P., Baird, S. G., Cao, Y., Darvish, K., Hao, H., Lo, S., Pablo-García, S., Rajaonson, E. M., Skreta, M., Yoshikawa, N., Corapi, S., Akkoc, G. D., Strieth-Kalthoff, F., Seifrid, M., & Aspuru-Guzik, A. (2024). Self-driving laboratories for chemistry and materials science. Chemical Reviews, 124(16), 9633-9732. https://doi.org/10.1021/acs.chemrev.4c00055

Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. Nature, 624, 570-578. https://doi.org/10.1038/s41586-023-06792-0

Noh, H., Na, G. S., Lee, N., & Park, C. (2026). MSP-LLM: A unified large language model framework for complete material synthesis planning. arXiv preprint, arXiv:2602.07543. https://arxiv.org/abs/2602.07543

Yoshikawa, N., & Tamura, R. (2026). NIMO Controller: a self-driving laboratory orchestrator based on the Model Context Protocol. arXiv preprint, arXiv:2605.15227. https://arxiv.org/abs/2605.15227

Tamura, R., & Yoshikawa, N. (2026). NIMO: A software platform for closed-loop materials exploration with diverse AI algorithms. arXiv preprint, arXiv:2606.15522. https://arxiv.org/abs/2606.15522

Ren, Z., Tan, H., Yee, J., & Hippalgaonkar, K. (2026). PUDA: An AI-native hardware harness for self-driving laboratories. arXiv preprint, arXiv:2607.26464. https://arxiv.org/abs/2607.26464

Fei, Y., Rendy, B., Yang, X., Woo, J., Huang, X., Li, C., Wang, S., Milsted, D., Zeng, Y., & Ceder, G. (2026). Agentic LLM reasoning in a self-driving laboratory for air-sensitive lithium halide spinel conductors. arXiv preprint, arXiv:2604.11957. https://arxiv.org/abs/2604.11957

Kusne, A. G., & McDannald, A. (2026). Managing autonomous materials labs with multi-agent AI and its implications for the science of science. Communications Materials, 7, 173. https://doi.org/10.1038/s43246-026-01219-5

Lv, S., Peng, L., Jiao, S., Yao, Y., Wu, W., & Hu, W. (2026). ChatMat: a multi-agent chemist for autonomous material prediction and exploration. Digital Discovery, 5(7), 2886-2898. https://doi.org/10.1039/d5dd00582e

Lee, H., Yoo, H. J., Jang, H. S., Park, B., Park, Y. J., & Han, S. S. (2026). Toward self-driving laboratory 2.0 for chemistry and materials discovery. Materials Horizons, 13(10), 4712-4739. https://doi.org/10.1039/d5mh01984b

Zhu, J., Zhang, L., Zhu, Y., Lin, X., Wu, Y., Di, S., Liu, B., Luo, Y., & Zhang, T. (2026). A survey of agentic materials science and engineering: where are we and where are we going? Journal of Materials Informatics, 6, 7. https://doi.org/10.20517/jmi.2026.07

Tobias, A. V., & Wahab, A. (2025). Autonomous 'self-driving' laboratories: a review of technology and policy implications. Royal Society Open Science, 12(7), 250646. https://doi.org/10.1098/rsos.250646

Wakabayashi, Y. K., & Otsuka, T. (2026). Bayesian optimization for self-driving materials laboratories: From algorithms to physics-informed workflows. arXiv preprint, arXiv:2608.26016. https://arxiv.org/abs/2608.26016

Ntagiantas, A., Tsilimidos, P., Giannakopoulos, G., Rekatsinas, C., & Krokidas, P. (2026). Active learning guided design space refinement for scalable multi-objective Bayesian optimization in materials discovery. arXiv preprint, arXiv:2608.04651. https://arxiv.org/abs/2608.04651

He, C., Singull, M., & Jacobsson, T. J. (2026). Bayesian optimisation for the experimental sciences: A practical guide to data-efficient optimisation of laboratory workflows. Advanced Intelligent Systems, 8(5), e202501149. https://doi.org/10.1002/aisy.202501149

Liang, H., Sun, Y., Paxson, R., Lee, C.-Y., Hall, A. T., Warecki, Z., Cumings, J., Koinuma, H., Kusne, A. G., Lippmaa, M., & Takeuchi, I. (2026). Autonomous epitaxial atomic-layer synthesis via real-time computer vision of electron diffraction. arXiv preprint, arXiv:2602.20432. https://arxiv.org/abs/2602.20432

Anthropic. (2025). Model Context Protocol specification (version 2025-06-18). https://modelcontextprotocol.io/specification/2025-06-18

Anthropic. (2024). Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol

Chapter 4 Physics-Consistent Prediction and Composition Optimization of High-Entropy Alloys (HEAs) and Superalloys

Introduction

When a pot is used in the kitchen for a long time, its bottom turns black. The part that often touches fire is damaged first. The hotter a metal gets, the softer and weaker it becomes. All metals share this property of weakening under heat.

Rocket engines face this problem more severely than almost any other machine. Inside the combustion chamber, burning fuel creates gas at thousands of degrees. That gas rushes out through a narrow passage called a nozzle. The nozzle wall must endure this hot gas right beside it. It is a far harsher place than the bottom of a pot.

Vehicles that reenter the atmosphere face the same concern. When they push into the air at high speed, the air is compressed and becomes hot. The front surface of the vehicle glows red. If the structure cannot withstand this heat, it melts or breaks.

You can feel this air heating with your palm. When you ride a bicycle fast, the wind hits your face hard. When you put your hand out the window, the wind pushes your hand. A flight vehicle meets that wind thousands of times faster. The pushed air is compressed and heated, and that heat is transferred to the front surface. The stagnation-point temperature on a reentry surface exceeds 1,600 degrees Celsius.

Hypersonic vehicles meet the same heat. Hypersonic speed means more than five times the speed of sound. At this speed, the leading edge of a wing becomes hotter than other parts. Heat gathers on narrow and sharp parts. So these parts need materials with excellent strength at high temperatures.

That is why people look for metals that do not lose strength even in very hot places. The metal that has long held this role is the nickel-based superalloy. A superalloy is an alloy specially designed to withstand high temperatures. But even this metal has a temperature limit.

This chapter deals with new metals that try to go beyond that limit. The main subject is the high-entropy alloy (HEA). It is an alloy made by mixing several metals in similar amounts. If only metals with high melting points are selected and mixed, the result becomes a new candidate for rocket materials. Such an alloy is called a refractory high-entropy alloy (RHEA).

This chapter explains a new way to find such metals. Instead of testing each one by hand, artificial intelligence (AI) quickly selects candidates. We will slowly explain how atomic arrangements, thermodynamic laws, and large-scale atomic calculations meet artificial intelligence. This is a story about metals, but it is explained through examples close to us, like seasons and school life.

Let us first picture why artificial intelligence is needed. To make one good alloy, we must choose the kinds and ratios of elements. Even with only five elements, the possible ratio combinations are hard to count. If we make and test them one by one, we cannot examine them all in a lifetime. Artificial intelligence narrows this vast field of candidates in advance.

The role of artificial intelligence is close to that of a guide. A guide who knows the mountain path shows the shortcut. But reaching the summit is still the person’s task. Artificial intelligence narrows the candidates, and people confirm them through experiments. This chapter shows through examples how this collaboration works.

4.1 Requirements of Extreme Aerospace Environments and Structural Limits of High-Temperature Alloys

What This Section Covers

This section first shows how harsh the thermal environment is for rockets and reentry vehicles. It then examines where long-used nickel-based superalloys meet their limits. Finally, it explains why refractory high-entropy alloys have emerged as a new alternative.

Put simply, this section defines the problem. We must understand the problem accurately to see the value of the artificial intelligence tools that appear later. So the focus is placed on understanding the situation before the numbers.

A rocket combustion chamber is a vessel with very high pressure. Fuel and oxidizer burn together inside it and create great pressure. This pressure reaches tens of megapascals. The pressure of a car tire is about 0.2 megapascals. A rocket combustion chamber withstands hundreds of times that pressure.

The gas made in the combustion chamber exits through the nozzle. Depending on the material, the temperature of this gas rises to thousands of degrees. But the temperature of the wall metal does not rise as high as the gas temperature. Cooling and the boundary layer block heat and keep the metal temperature lower. Even so, the amount of heat the wall receives is enormous. If the wall cannot withstand this heat, it melts, holes form, and the engine becomes unusable.

Heat does not come only once. An engine is turned on and off repeatedly. It heats up quickly when turned on and cools quickly when turned off. This rise and fall in temperature is called thermal shock. It is the same principle as a glass cup cracking when hot water is suddenly poured into it.

Force is added to this. When a metal receives force for a long time while hot, it slowly stretches. This phenomenon is called creep. It is similar to a stick of taffy sagging by itself when held for a long time. At high temperatures, creep and repeated loads work together and wear down the material.

Nickel-based superalloys have long endured these environments. The strength of these metals comes from their fine microstructure. Inside the gamma (γ) phase that forms the matrix, tiny particles called gamma prime (γ′) are densely embedded. These particles act as obstacles that stop the metal from deforming. It is like small nuts embedded in dough that keep the dough from stretching easily.

Each superalloy has slightly different kinds of obstacles. Inconel 718, which contains a large amount of niobium (Nb), is one example. In this alloy, a phase called gamma double prime (γ″, Ni3Nb) provides more strength than gamma prime. The names of the obstacles differ, but the principle is the same. Small particles block deformation and make the metal strong.

Rockets use different metals in different places. Copper alloys are often used in combustion chambers whose walls are cooled by fuel. This is because they are excellent at removing heat quickly. Nickel-based superalloys are used in parts such as turbines that receive high temperatures for a long time. For hotter nozzle throats or reentry surfaces, refractory metals or ceramics become candidates. This chapter deals with new metals that aim for the hottest places among them.

The problem appears when the temperature rises further. Above roughly 1,150 degrees, these small particles grow larger or collapse. This is called coarsening. When the obstacles disappear, the metal stretches easily and its strength drops quickly. Since the melting point of nickel-based superalloys is about 1,350 degrees, their effective operating limit is far below their melting point.

Harmful phases may also form. When a metal is kept at high temperature for a long time, atoms gather in certain patterns and form hard, brittle lumps. These phases are called harmful phases. When harmful

phases form, the metal breaks easily. Better superalloys are designed to suppress such phases.

Because of these problems, nickel-based superalloys need cooling. In real engines, small channels are made inside parts. Cold fuel flows through those channels and cools the wall. This cooling allows parts to be used at temperatures below their melting point. But cooling channels make the part structure more complex and increase weight.

The goal of refractory high-entropy alloys is to reduce this cooling. If the material itself can withstand hotter places, less cooling is needed. Less cooling makes the structure lighter and simpler. Researchers aim to develop new alloys that can endure places above 1,400 degrees without cooling.

So people looked for a completely different path. The idea was to build the framework from metals with high melting points. Metals such as niobium (Nb), molybdenum (Mo), tantalum (Ta), and tungsten (W) belong to this group. Their melting points are well above 2,000 degrees. Such metals are called refractory metals.

A refractory high-entropy alloy is an alloy made by mixing several of these refractory metals in similar amounts. Recent studies report that these alloys show higher strength than nickel-based superalloys even in compression tests at 1,600 degrees. That is why they are considered new candidates for rocket propulsion, hypersonic vehicle leading edges, and reentry heat shields.

But they do not have only advantages. These alloys tend to break easily at room temperature. This is called low ductility. And when they meet oxygen at high temperature, oxidation occurs and the surface crumbles. These two problems must be solved before the alloys can be used in real parts. Artificial intelligence is used to solve these problems quickly.

Let us look a little more closely at the oxidation problem. Alloys rich in tungsten or molybdenum react strongly with oxygen at high temperature. The surface oxide layer does not stick firmly and falls off like powder. This crumbly oxidation is called pest oxidation. The name comes from the idea that it eats away at the surface like pests eating grain.

The way to stop this problem is a protective oxide layer. When aluminum or chromium is added in the right amount, a dense oxide layer forms on the surface. This layer blocks oxygen from penetrating inward. It is similar to a layer stuck to the bottom of a pot that actually protects the metal inside. The key design question is how much to add and how to add it to make a good layer.

Compatibility with coatings is also important. An oxidation-resistant coating is usually added on top of the alloy. The problem is that the metal and the coating expand differently with temperature. The degree of this expansion is called the coefficient of thermal expansion. If the coefficients of thermal expansion of the two materials differ greatly, the boundary cracks when the temperature rises. So an alloy must be tuned not only for strength but also for compatibility with the coating.

In this way, a nozzle liner alloy must satisfy many conditions at once. It needs a high melting point, low density, high-temperature strength, oxidation resistance, and coating compatibility. These conditions often conflict with one another. Such a tangled problem is hard to solve by experiments alone. Artificial intelligence enters as a tool for quickly organizing this tangled problem.

4.2 Physical Modeling of the Four Core Effects and Phase-Equilibrium Thermodynamics

What This Section Covers

This section deals with the early explanatory framework for why high-entropy alloys are special. When this field first opened, researchers proposed four effects. These are called the four core effects. They are

configurational entropy, lattice distortion, sluggish diffusion, and the cocktail effect.

These four are not established physical laws. They are closer to early hypotheses. As we will see later, sluggish diffusion and single-phase stabilization from high entropy have not been proven as general laws. Still, we must understand this framework to see what artificial intelligence is trying to calculate. We will explain each one through examples from everyday life.

First, let us define what a high-entropy alloy is. Ordinary alloys add small amounts of other metals to one main metal. In steel, iron is the main actor, and a small amount of carbon is added. High-entropy alloys are different. They mix five or more metals in similar amounts. In other words, there are several main actors.

This concept was presented in 2004 by two research groups. They were Yeh's group in Taiwan and Cantor's group in the United Kingdom. They showed that stable metals could be made without relying on one main element. This discovery opened the door to a new way of designing alloys.

The first principle is configurational entropy. Entropy is a value related to the number of ways things can be mixed. The more ways there are to place items in a locker, the greater the disorder. When several metals are mixed in similar amounts, the number of ways to arrange the atoms becomes very large. This large disorder helps keep the metal in one mixed state.

This mixed state is called a solid solution. A solid solution is a state in which several elements are evenly dissolved in one crystal lattice. It resembles sugar dissolving in water to become one liquid. When entropy is large, it prevents elements from clustering separately and becoming hard compounds. That is why high-entropy alloys tend to form solid solutions rather than compounds.

The second principle is lattice distortion. Metal atoms differ in size. When large and small atoms mix in one lattice, the lattice becomes uneven and distorted. It is like a wall built from stones of different sizes rather than smooth bricks. This distorted lattice becomes a barrier that resists deformation.

Deformation occurs when small defects called dislocations slide through the lattice. When the lattice is uneven, dislocations have difficulty sliding. It is like trying to pull a cart over a gravel field. So alloys with large lattice distortion do not stretch easily and are strong.

The third principle is sluggish diffusion. Diffusion is the movement of atoms from one site to another. In high-entropy alloys, each atom has a different neighboring environment. Some sites are easy to move through, and others are difficult. This unevenness is thought to slow atomic movement on average.

When atoms move slowly, the microstructure does not change easily even at high temperature. As a result, crystal grains may grow slowly and resist creep. However, this effect is not always strong in every high-entropy alloy. The actual diffusion rate depends on the element combination, temperature, and crystal structure. For this reason, recent studies treat this effect not as a general law, but as a property that must be checked case by case.

The fourth principle is the cocktail effect. A cocktail is a drink that creates a new taste by mixing several ingredients. The taste after mixing is not the same as simply adding the taste of each ingredient. Metals are similar. When several elements are present together, properties appear that are hard to explain by simply adding the properties of each element.

These nonlinear properties appear in strength, corrosion resistance, and toughness. The problem is that people cannot easily know these combination effects in advance. There are too many possible combinations of elements. Artificial intelligence is used as a tool to find these hidden combination effects in data.

These four principles lead directly to rocket materials. Configurational entropy helps the microstructure avoid separation at high temperature. Lattice distortion increases strength. Slow diffusion can help resist creep. The cocktail effect can bring unexpected advantages such as oxidation resistance. The four principles must work together to withstand extreme environments.

However, these principles are not a master key. They appear strongly in some combinations and weakly in others. Therefore, even if we know the principles, each actual composition must be checked. The principles show the direction, and calculation and experiment do the checking. Artificial intelligence greatly speeds up this checking.

These four principles eventually come together in one question. Which composition will remain stable as a single phase? The field that answers this question is phase-equilibrium thermodynamics. In the next two subsections, we examine two indicators that help make this judgment.

4.2.1 The Omega Parameter for Measuring the Possibility of Remaining as One Phase

Thermodynamics is the study of energy and stability. Whether a state is stable is judged by free energy. A state with lower free energy is more stable. Just as water flows from a high place to a low place, matter tends to move toward a state with lower free energy.

Free energy is determined by two values together. One is enthalpy. Enthalpy is related to whether atoms favor or dislike one another. Atoms that go well together tend to mix, while atoms that dislike one another tend to separate. The other is entropy, which we saw earlier. Entropy represents the disorder of mixing.

The competition between these two changes with temperature. The higher the temperature, the stronger the effect of entropy becomes. It is like people spreading out more freely in a hot room. Therefore, at high temperature, a disordered solid solution is more likely to become stable. This is why high-entropy alloys often remain as one phase at high temperature.

Configurational entropy becomes large when elements are mixed in equimolar ratios. An equimolar ratio means adding each element in the same amount. If five elements are mixed in the same amount, configurational entropy reaches its maximum value. As the number of elements increases, this value becomes larger. The name high-entropy comes from this property of large configurational entropy.

The omega parameter is a simple value for estimating the possibility of a single-phase solid solution. It considers three things together. They are the average melting point, configurational entropy, and the magnitude of mixing enthalpy. If the favorable effect of entropy is larger than the unfavorable effect of enthalpy, the omega value becomes large. When omega is large, the alloy is considered more likely to remain as one phase.

The difference in atomic size is checked together with this. If atomic sizes differ too much, the lattice cannot withstand the strain and separates. Therefore, the size difference must be below a certain value for a single-phase solid solution to form easily. Researchers use an empirical rule that a single-phase solid solution forms when omega is large and the size difference is small.

This rule is convenient, but it is not a definitive tool. Real alloys are affected by local order, cooling rate, impurities, and heat treatment. Even with the same composition, the microstructure can differ depending on how the alloy is made. Therefore, the omega rule is used only for the first screening.

More accurate judgment is handled by methods that require more computation. A representative method is CALPHAD. CALPHAD is a method that calculates phase equilibrium by building data on the free energies of various phases. Density Functional Theory (DFT) calculations and experiments are added to this. Density functional theory is a quantum calculation method that computes energy from the

behavior of electrons.

Let us translate the competition between enthalpy and entropy into an everyday example. Suppose boys and girls sit mixed together in a classroom. The force that makes close friends gather together is similar to enthalpy. The force that makes seats mix freely is similar to entropy. During a quiet time, groups gather together. During an active time, people mix more freely. This is like the force for mixing becoming stronger as temperature rises.

Artificial intelligence makes this process faster. Simple indicators such as omega are used as input features for artificial intelligence when there is a very large amount of data. Artificial intelligence looks at these features and quickly screens which compositions will remain as one phase. Only the narrowed candidates are then checked with CALPHAD and experiments. This collaboration is the key to handling a wide composition space.

Recent studies have greatly improved this prediction accuracy. When several physical features are included together, phase prediction accuracy rises from about 70 percent to more than 90 percent. At this level, more than 1 million compositions can be screened quickly. Artificial intelligence greatly reduces the burden on human work. Time and resources are then focused on the small number of candidates selected in this way.

4.2.2 Estimating Crystal Structure with Valence Electron Concentration

This subsection deals with an indicator for estimating what crystal structure an alloy will have. That indicator is valence electron concentration (VEC). It means the average number of outer electrons. First, let us review what crystal structure is. Metal atoms are not placed randomly. They stack in a regular lattice. It is like stacking bricks according to a rule. Properties change depending on the stacking method.

Two structures are important for rocket materials. One is body-centered cubic (BCC). This is an arrangement with atoms at the eight corners of a cube and at its center. The other is face-centered cubic (FCC). This is an arrangement with atoms at the eight corners and at the centers of the six faces.

These two structures have different properties. Body-centered cubic structures are often favorable for high-temperature strength. However, they are likely to show brittleness and break easily at room temperature. Face-centered cubic structures often have good ductility and stretch well. If we know in advance which structure will form, design becomes easier.

Valence electron concentration is an indicator for estimating this structure. Valence electrons are the outer electrons of an atom. These electrons participate in bonding. Valence electron concentration is the average number of outer electrons contributed by each element, weighted by the composition ratio. This value indicates the likely direction of the crystal structure.

The empirical rule is as follows. If this value is roughly below 6.87, body-centered cubic is stable. If it is above 8, face-centered cubic is stable. Between these values, the two structures mix. Refractory high-entropy alloys often have body-centered cubic structures because this value is low. The famous Cantor alloy is face-centered cubic because this value is high.

The usefulness of this rule can also be seen in recent artificial intelligence research. One preprint was released in early 2026. A preprint is an early study that has not undergone peer review. This study predicted the melting points of refractory high-entropy alloys. It did not directly include crystal structure, but used only features obtained from elemental composition. Physically meaningful features such as valence electron concentration contributed greatly to the prediction. In this sense, physical knowledge built by humans and features selected by artificial intelligence meet each other.

This case shows two things. One is that simple indicators are still useful. The other is that artificial intelligence can learn in a way that does not conflict with physics. These two ideas run through this entire chapter.

Here we need to look a little more closely at why structure controls properties. Deformation occurs as dislocations glide through the lattice. Face-centered cubic structures have planes where atoms are closely packed, so dislocations pass through smoothly. As a result, they stretch well and do not break easily. Body-centered cubic structures do not have such closely packed planes. This is not because there are fewer paths for sliding, but because the energy barrier that dislocations must overcome is large. This barrier is sensitive to temperature. Therefore, body-centered cubic structures break easily at room temperature, but when temperature rises, dislocations move more easily and strength is maintained.

This difference leads directly to the selection of rocket materials. For a nozzle liner, high-temperature strength matters above all else. For this reason, refractory high-entropy alloys in the body-centered cubic family become candidates. However, they also carry the weakness of room-temperature brittleness. To reduce this weakness, the composition must be adjusted carefully. Valence electron concentration becomes the first compass for that adjustment.

However, the valence electron concentration rule also has limits. The boundaries of 6.87 and 8 are empirical values from a particular group of transition-metal alloys. They do not apply directly to other alloy groups. This value is an average, so it cannot capture the local environment. Even with the same average value, the actual structure can differ depending on the element combination. In addition, this rule cannot tell when hard and brittle intermetallic compounds will form. Laves phases and sigma phases are examples. When such phases are mixed in, the alloy becomes weaker. Therefore, this indicator is also used for the first screening and then checked by precise calculation. Artificial intelligence makes this screening much faster and broader.

4.3 Extracting Material Features from Research Papers

Artificial intelligence cannot see metals with its eyes. It understands materials only through features converted into numbers. These features are called descriptors. This section examines how to extract these features from research papers. First, we look at how wide the composition space is. When five or more elements are mixed and their ratios are changed, the number of cases explodes. If there are two ingredients to put in coffee, there are only a few combinations. If there are ten ingredients, the combinations become hard to count. Multicomponent alloys are exactly in this situation.

It is impossible to scan this wide space entirely by experiment. Making and testing one alloy takes time and money. CALPHAD calculations also take time for each composition. Therefore, a guide is needed to narrow the wide space quickly. Artificial intelligence is that guide.

To use artificial intelligence, materials must be converted into numbers. Descriptors are those numbers. Atomic radius, electronegativity, melting point, and the number of outer electrons are representative descriptors. Electronegativity is the force with which an atom attracts electrons. These values are calculated according to the composition and then entered into artificial intelligence.

The problem is that these numbers alone are not enough. Even with the same composition, properties can change greatly depending on how the alloy was made. Heat-treatment temperature, holding time, and cooling method change the microstructure. This process information is hard to capture in a single number. But this information is written in text in research papers.

Natural language-derived descriptors fill this gap. They are feature values made when artificial intelligence reads sentences in papers and converts them into numerical features. In effect, text read by

people is translated into numbers read by machines. Recent studies have used this method to include process descriptions in features.

A preprint released in 2026 is a good example. It is a result from researchers at Pennsylvania State University in the United States. They entered sentences describing heat-treatment processes into a large language model. From the numerical representation of those sentences, they could recover the original process values almost exactly. They then used this representation as a feature and improved alloy hardness prediction by about 20 percent. This study is still an early study that has not undergone peer review.

Literature mining is the work of collecting these features on a large scale. It automatically extracts phase transformation temperatures, solubility limits, strength, and ductility values from many metallurgy papers. In the past, people made tables by hand. Now large language models read sentences and fill tables. The data collected in this way becomes the foundation for artificial intelligence training.

Some studies have used language models directly for property prediction. One result was published in a journal in 2025. This study used a language model structure called a transformer. A transformer is an artificial intelligence system that reads relationships among words in a sentence. It treated alloy compositions and conditions like text and predicted properties. This is a case in which technology developed for language moved into materials research.

Recently, a study also used large language models to build a high-entropy alloy database. The result was published in a journal in 2026. This database automatically organized compositions and properties from research papers. It greatly accelerated work that would have taken a long time by hand. The data accumulated in this way becomes the raw material for the prediction models discussed in the next section.

This method also has limits. Measurement conditions and notation differ from paper to paper. If a machine reads a sentence incorrectly, wrong values enter the data. So people must review the collected data again. Collaboration in which artificial intelligence collects the data and people check it is the current standard.

The quality of the data determines the quality of the prediction. If a model learns from wrong data, it gives wrong answers. If the blueprint is wrong, even a skilled carpenter cannot build a sturdy house. So the stage of collecting data is as important as prediction. Good data is the basis of good artificial intelligence.

Recently, artificial intelligence has begun to expand rules on its own. One study published in a 2026 journal presented a knowledge-enhanced artificial intelligence framework. This study moved beyond the constraint that elements must be mixed in equal amounts. It freely changed element ratios for each purpose and designed single-phase high-entropy alloys. This was a result of adding the metallurgical knowledge built by humans to artificial intelligence.

This trend shows collaboration between artificial intelligence and human knowledge. If only artificial intelligence is used, predictions become unstable outside the data. If only human rules are used, it is hard to search a wide space. When the two are combined, they make up for each other's weaknesses. Design becomes stronger when literature mining and knowledge enhancement move together.

The connection with CALPHAD is also important. Artificial intelligence can quickly imitate CALPHAD calculations or choose the next calculation candidates. If data collected from the literature, CALPHAD calculations, density functional theory (DFT), and experiments are tied together, candidate design becomes faster. This integrated pipeline is the practical way to handle a wide composition space.

4.4 Graph Neural Networks That Represent Disordered Alloys

We now look at how artificial intelligence represents the disorder of high-entropy alloys. The core tool is the graph neural network (GNN). It is a method that learns by turning a material into a picture of points and lines. First, we need to understand what a graph is. A graph is a picture that shows relationships with points and lines. A subway map is a graph. Stations are points, and tracks are lines. A crystal can also be drawn as a graph. Atoms are points, and bonds between atoms are lines.

A graph neural network reads this picture and predicts the properties of a material. Each point contains information about an atom. Each line contains the distance and direction between atoms. The neural network understands the material as neighboring points exchange information. This exchange of information is called message passing. It is like understanding the mood of the whole class by passing notes with the friend sitting next to you.

This method works well for ideal single crystals. A single crystal is a crystal in which exactly one element fills each site. Its rules are clear, like a salt crystal. Existing crystal graph neural networks were made with this assumption. If one element is placed at each point, the calculation is clean.

The problem is that high-entropy alloys are not like this. In these alloys, several elements are mixed at one site by probability. One site may be niobium, molybdenum, or tantalum. This is called fractional occupancy. It is a state in which one site has several possibilities at the same time.

There is also local short-range order (SRO). Atoms are not mixed completely at random. Some elements like to be neighbors, while others avoid each other. This preference at short distances has a large effect on properties. Existing models cannot capture this subtle order.

Crystal fractional graph neural networks are an attempt to solve this problem. The idea in this direction is as follows. Instead of placing one element at each point, a composition vector is placed there. A composition vector is a list of numbers that records the probability that each element will occupy that site. For example, one site can be represented as 25 percent niobium, 25 percent molybdenum, and 50 percent tantalum. This naturally captures fractional occupancy.

What the earlier study that used this name actually showed is narrow and clear. This study predicted crystal energy by using both the local environment of about 16 atoms and the elemental fractions of the whole alloy. This was the result of training on about one thousand structures. It was not a study that matched short-range order, melting point, mixing enthalpy, and high-temperature strength all at once. Predicting these properties together is a future direction to explore, not an achievement already made. The method of adding element-pair information to edges to capture short-range order is also still at the proposal stage.

We now look at how message passing leads to learning. At first, each point knows only its own information. After one exchange of information, it knows about its neighbors. If this is repeated several times, it learns about places farther away. The model predicts the properties of the whole material using the information expanded in this way. If the prediction differs from the actual value, the values in the neural network are corrected. The model becomes smarter by repeating this correction countless times.

Another advantage is that one model can predict several properties together. This is called multi-objective prediction. Melting point and strength are related to each other. Information gained while learning one property helps predict another property. It is like deepening understanding by studying several subjects together.

However, this direction is still at an early stage. Crystal fractional graph neural networks have only recently been proposed as preliminary research. Large cells, complex short-range order, and real

heat-treatment histories are still difficult problems. So predictions from this model must be checked through experiments and atomistic simulations.

Meanwhile, other artificial intelligence studies on short-range order are increasing quickly. One study published in a 2025 journal compared methods for reproducing short-range order with machine learning potentials. A machine learning potential is a tool that calculates forces between atoms with artificial intelligence. This study showed that even high energy accuracy may fail to predict properties well. This means that how training data is chosen controls property prediction.

Another study in the same year accelerated Monte Carlo calculations with artificial intelligence. Monte Carlo is a calculation method that finds an equilibrium state through random trials. This study used an artificial intelligence surrogate model to reveal nanostructures at the scale of billions of atoms. It opened a path to observe subtle order inside high-entropy alloys at a large scale. Crystal fractional graph neural networks are also developing on top of this trend.

We now look at why this representation of disorder matters for rocket materials. The strength of a nozzle liner depends on how well dislocations are blocked. Short-range order changes the difficulty of the paths along which dislocations move. When certain elements cluster together, they block dislocations more strongly. So even with the same composition, strength changes depending on short-range order.

This subtle order changes depending on the manufacturing method. If the alloy is cooled quickly, disorder freezes in place. If it is cooled slowly, atoms settle into positions and order appears. Heat-treatment temperature and time also change order. So properties can be matched only by looking at both composition and process.

Artificial intelligence is used to learn these tangled relationships. It connects composition, process, order, and properties all at once. It is hard for people to handle so many variables at the same time. If there is enough data, artificial intelligence finds the connections. Crystal fractional graph neural networks are an attempt to capture these connections at the level of atomic arrangements.

This direction is still at an early stage. At present, both prediction accuracy and the amount of data are insufficient. As experimental data accumulates and calculations become faster, predictions will improve. Until then, artificial intelligence predictions must always be checked through experiments.

4.5 Large-Scale Molecular Dynamics That Sees Every Atom

The method of simulating the movement of each atom on a computer is called molecular dynamics (MD). It is a tool for watching, at the atomic level, how a metal breaks. The key to scaling this calculation to a very large size is an artificial intelligence method called the neuroevolution potential (NEP). First, we need to understand why a large scale is needed. Cracks in metals begin with a few atoms. That small gap grows, spreads to grain boundaries, and finally cuts the structure. To understand this process, we must look at the very small and the fairly large together. This kind of phenomenon is called a multi-scale phenomenon.

Traditional density functional theory (DFT) has difficulty doing this work. This calculation is accurate, but it requires a large amount of computation. It is hard to go beyond several hundred atoms. A real microstructure in which cracks spread contains more than tens of billions of atoms. So accurate quantum calculations alone cannot show the whole picture.

Empirical potentials have the opposite problem. A potential is a rule that calculates the forces acting between atoms. Empirical potentials are light and fast. But in complex alloys with more than five elements, their accuracy falls. Each element has a different neighboring environment, and they cannot

capture that subtlety.

So a tool is needed that combines only the strengths of the two methods. It must have both the accuracy of quantum calculations and the speed of empirical rules. Machine learning potentials fill this role. Artificial intelligence learns from quantum calculation data and predicts forces. The neuroevolution potential is a speed-focused tool in this family.

We should also look at what molecular dynamics is. This calculation applies Newton's laws of motion to every atom. It finds the force acting on each atom and moves it by a very short time step. This process is repeated countless times. Then we can watch how atoms move over time, like a movie.

4.5.1 Neuroevolution Potentials That Capture Both Accuracy and Speed

The neuroevolution potential is a tool that combines two ideas. One is the neural network. A neural network is artificial intelligence that learns rules from data. The other is the evolution strategy. An evolution strategy is a method that creates several candidates, chooses better ones, and refines them. It is a learning method that imitates natural selection.

This potential first turns the environment around an atom into numbers. It creates features from how far an atom is from its neighbors and at what angles they are arranged. These features are put into a neural network to predict the energy of one atom. When the energies of all atoms are added, the total energy is obtained. The forces acting on atoms are then obtained from changes in energy.

We now look at why both distances and angles are considered. If only distance is used, we know only how close atoms are. If angles are added, we know what shape the atoms form. Even at the same distance, a triangular arrangement and a straight-line arrangement have different properties. So both distances and angles must be included to capture the environment properly.

This method is connected to the nature of chemical bonding. Atoms bond differently depending on their surroundings. If neighbors are dense, bonding is strong. If they are sparse, bonding is weak. The neuroevolution potential learns this relationship from data. It uses the learned relationship to quickly estimate the energy of a new arrangement.

The advantage of this method lies in capturing both accuracy and speed. The neural network learns accuracy close to quantum calculations. The calculation itself is as light as empirical rules. So accurate forces can be obtained quickly. This is the key that opens large-scale calculation.

The secret of the speed lies in the graphics processing unit (GPU). This device was originally made to draw game screens. It is excellent at coloring many points on a screen at the same time. This parallel processing ability also fits atom calculations well. This is because the forces of many atoms can be calculated at once.

The neuroevolution potential was designed to run well on this device. It is used with a graphics processing unit program called GPUMD. This combination has been widely used in studies of thermal conduction, defects, and high-temperature deformation. Recently, follow-up studies have made the training method faster. These improvements make larger calculations possible.

We should also look at how this potential is made. First, density functional theory (DFT) is used to calculate the energies and forces of many atomic arrangements. These values become the textbook for artificial intelligence. Artificial intelligence refines its own predictions while studying this textbook. The evolution strategy selects values in the direction that makes the predictions match the data well. When this training is complete, a fast and accurate force calculator is finished.

Active learning is also used in training. Artificial intelligence finds arrangements about which it is uncertain on its own. Only those arrangements are selected and newly calculated with density functional theory (DFT). The data obtained in this way is put back into training. This is a smart way to cover a wide range with limited calculations.

However, a fast answer is not always a correct answer. Artificial intelligence potentials are accurate within the range they have learned. They can be wrong for compositions or temperatures that are not in the training data. In high-entropy alloys, the environment can change greatly even when the element combination changes only a little. So the representativeness of training data is most important. A separate process is needed to check prediction uncertainty.

4.5.2 Deformation and Cracking Seen Through Large-Scale Calculation

One trillion is a number that is hard to imagine. If one trillion atoms are arranged, the size becomes close to a real microstructure. This means that a specimen at the micrometer scale can be drawn down to every atom. In a small calculation, only one crack and a few boundaries are visible. In a large calculation, we can see many boundaries and defects affecting one another.

A preliminary study released in 2026 crossed this threshold. The study calculated 1.6 trillion atoms with a neuroevolution potential. It used about 45,000 computing devices together on a new supercomputer in China. What the study showed was scale and speed. It reported running the same calculation much faster than the previous record. It was not a study that actually analyzed dislocations and cracks in high-entropy alloys at this scale. Demonstrating access to a scale and revealing something at that scale are different matters. This study is still an early study that has not yet undergone peer review.

It is useful to compare what this number means. In the past, even a calculation with 1 million atoms was a major task. A few years ago, calculations passed 100 million atoms and soon reached 10 billion atoms. Now they have passed 1 trillion atoms. The size that can be calculated is growing quickly. Still, a 1-trillion-atom calculation is a symbolic demonstration that uses a very large supercomputer to its limit. The calculations used every day in actual alloy design range from thousands to hundreds of thousands of atoms. Artificial intelligence potentials and fast computers are driving this growth.

The benefit of large scale lies in statistics. The more atoms there are, the more likely rare events are to be captured. The seeds of cracks form very rarely. In a small calculation, those seeds may not be seen. In a large calculation, several seeds can be observed together. That makes the prediction closer to reality.

We first need to note why this calculation is useful for rocket materials. The microstructure of an actual nozzle liner contains an uncountable number of atoms. To know its life, we need to know what happens in this large structure. Large calculations give a picture close to this real structure. They can therefore be used to estimate which composition will last longer.

In principle, large-scale molecular dynamics of this kind can observe three events. The following are the goals of this method, not results already revealed by the demonstration above. The first is dislocation nucleation. A dislocation is a small defect where deformation begins. The calculation tracks in real time how dislocations form and glide in a body-centered cubic lattice. In high-entropy alloys, fluctuations in local composition pin dislocations. This pinning is the principle that raises strength.

The second is grain-boundary sliding. A grain boundary is the boundary between crystal grains. When a force is applied at high temperature, this boundary slides. Large calculations measure this sliding and the clustering of voids at the atomic level. This behavior determines creep life.

The third is crack growth. The crack tip is the very front of a crack. At this point, the atomic arrangement changes and deformation occurs. Large calculations observe how cracks grow and how

they stop. They provide clues for estimating which composition resists cracking well.

These three events are connected to one another. Dislocations pile up and move toward grain boundaries. Grain boundaries slide and voids form. Those voids connect and become cracks. In a small calculation, this connection is hard to see. In a large calculation, the process can be seen from start to finish at once.

We need to note why seeing this connection is useful. The life of a material is set by one weak link. A chain breaks at its weak link. If we know which stage fails first, we can reinforce that point. Large-scale calculations help find this weak link in advance.

This information is useful for research on rocket nozzle liners. A nozzle liner faces both hot gas and thermal shock. If we can estimate which composition will last longer before experiments, we can save time. Large-scale calculations act as a scale for narrowing the candidates.

However, these calculations also come with conditions. The accuracy of the results depends on the quality of the artificial intelligence potential. In situations the potential has not learned, the calculation will also be wrong. So results from large calculations must be verified by small-scale precise calculations and experiments. Being large and being correct are different matters.

4.6 Metal Design AI That Does Not Conflict with Physical Laws

Prediction based only on pure data learning can be risky. A method that blocks this risk with physical laws is called a physics-informed neural network (PINN). We will look at how this method is applied to metallurgy, the science of metals. First, we need to note the weakness of pure data learning. Artificial intelligence answers well within the data it has learned. But in areas without data, it gives strange answers. Such areas are called out-of-distribution. It is like a student facing a question outside the exam scope.

In metal design, these strange answers violate physical laws. For example, a stability value may move in the wrong direction as temperature rises. The model may also predict that lattice energy decreases as volume increases. Such answers cannot occur in nature. If these predictions are trusted as they are, the design will go wrong.

A physics-informed neural network prevents this problem during training. The method is to put physical laws into the loss function. The loss function is a yardstick that measures how wrong the artificial intelligence is. Usually, it measures only the difference between predictions and actual data. Here, the degree to which physical laws are violated is added.

Then the artificial intelligence tries to satisfy two things together. One is matching the data. The other is obeying physical laws. Even in areas without data, the laws guide the direction of the answer. It is like a student solving a question outside the exam scope by using basic principles.

What laws should be put into metal design? One example is how free energy should change with temperature. The relationship that strength is determined by the sum of several strengthening contributions can also be included. Conditions that phase equilibrium must satisfy can also be included. These laws reduce the freedom of the artificial intelligence. Reduced freedom instead produces more trustworthy answers.

Let us look a little more closely at how laws are put into the loss function. Physical laws are usually written as equations. When predicted values are inserted into those equations, both sides should match. If they do not match, a difference remains. This difference is called a residual. The method gives a penalty when the residual is large.

Then the artificial intelligence tries to avoid the penalty. It tries to match the data while also obeying the equation. It is like a student who gives an answer while also matching the calculation steps and units. Such a student is stronger on applied problems than a student who has only memorized answers. This is why physics-informed neural networks are strong outside the data.

The value of this method grows when data are scarce. Extreme materials for rockets have rare experimental data. One test requires great cost and time. When data are scarce, pure learning becomes unstable easily. Physical laws fill the gaps in insufficient data.

These two methods must be distinguished. The melting-point prediction study discussed in the previous section is not a physics-informed neural network. It is a regression model that selected physically meaningful features. It did not directly put physical equations into the loss function. Regression that uses good features and learning that puts physical laws into the loss function are different methods. Still, they move toward the same goal. The goal is to keep artificial intelligence predictions from conflicting with metallurgy. When good features and physical constraints go together, we get closer to that goal.

4.6.1 Bringing Tools Together for Ultra-High-Temperature Nozzle Liner Design

This subsection looks at how the tools discussed above are used together in actual design. The target is a refractory high-entropy alloy for an ultra-high-temperature nozzle liner. The focus is on the design and verification process rather than on a single completed result. It also notes how far recent research has come and what remains. A good nozzle liner alloy must satisfy several properties together. It must have a high melting point. If the density is too high, the rocket becomes heavy. Strength must be maintained at high temperature. The surface must withstand contact with oxygen. The coating and thermal expansion must match well. If only one property is good, it cannot be used as a part.

Handling these many goals together is difficult for people. Adding a large amount of tungsten raises the melting point. But it also increases density and may make the alloy brittle. Adding aluminum or chromium improves oxidation resistance. But other properties may become unstable. This kind of work, which handles intertwined goals, is called multi-objective optimization.

In rockets, density is a very sensitive issue. If a part is heavy, the whole rocket becomes heavy. A heavy rocket travels less far with the same fuel. So even if strength is good, the material cannot be used if it is too heavy. Melting point, strength, and weight must be placed on the same scale.

In this kind of weighing, there is no single correct answer. If strength is given up a little, weight can be reduced. If weight is given up a little, strength can be increased. A group of candidates with such trade-offs appears. A person then looks at this group and chooses one that fits the purpose. Artificial intelligence quickly draws this group.

Artificial intelligence handles these intertwined goals together. It turns compositions into numbers using the descriptors discussed above. Prediction models quickly screen melting point and strength. Physical constraints filter out candidates that make no sense. Large-scale atomic calculations examine deformation and cracks. This flow greatly narrows the candidates.

Recent studies show the usefulness of this flow. A study published in a journal in 2024 designed refractory high-entropy alloys for ultra-high-temperature use with machine learning. It used artificial intelligence to narrow a wide composition space and select promising candidates. It tried to improve both strength and ductility. Such studies show that artificial intelligence design works at the laboratory stage.

Artificial intelligence is also entering the oxidation problem. A preliminary study released in late 2025 explored refractory complex concentrated alloys resistant to high-temperature oxidation through active learning. Active learning is a method in which artificial intelligence chooses the next composition to test on its own. It is a path to finding good candidates with few experiments. This study is also an early study that has not yet undergone peer review.

However, it is too early to speak as if the results are settled. Any number stating exactly how much strength a certain composition has at a certain temperature requires checking the original source. This book does not use unverified performance values. Instead, this section emphasizes the verification process that design must pass through.

Verification has three stages. First, it checks whether the candidate composition is stable in calculation. CALPHAD and density functional theory (DFT) are used to confirm phase equilibrium. Next, it checks whether the alloy can actually be made. Even if it looks good in calculation, it is useless if it cannot be made. Finally, strength and oxidation resistance are measured at high temperature. A test close to the real environment is the final gate.

We must not forget that errors flow together through these stages. A small error in atomic calculation spreads into microstructure prediction and then grows again in part strength prediction. If oxidation removes the surface and the boundary moves, the prediction becomes even less stable. Variations from each manufacturing process are added to this. So a good design presents not one value but a range of life. In the end, it must pass a qualification test under the same conditions as the actual part.

These three stages do not move only in order. If a weakness appears in testing, the composition is adjusted again. The adjusted composition is calculated again and made again. This repeated cycle is called closed-loop design. A self-driving laboratory is an attempt to run this cycle automatically.

Artificial intelligence enters every stage of this cycle. It chooses the next composition to test, learns the results, and chooses again. The cycle runs even without a person watching overnight. The path to a good composition becomes that much shorter. This flow is changing the future of high-temperature alloy development.

Artificial intelligence design is also active in nickel-based superalloys. Studies in 2025 predicted the creep life of single-crystal superalloys with machine learning. They used composition and microstructure information to select compositions that would last longer. The volume fraction and shape of gamma prime particles were used as important features. Such studies show that artificial intelligence gives new uses even to old metals.

There is also research that combines phase-field calculations with artificial intelligence. A phase field is a calculation method that describes microstructure changes as smooth fields. A 2025 study connected this calculation with machine learning. It predicted the effects of composition and microstructure on creep deformation together. Methods that connect calculation and artificial intelligence are becoming more common.

These cases give a common lesson. Artificial intelligence becomes powerful when it has good features and enough data. When features are connected to physics, predictions become robust. If data are biased toward a certain area, predictions become unstable outside that area. So the first task is to collect broad data and choose features well.

The place of artificial intelligence in this process is clear. Artificial intelligence is a guide that reduces composition space. If artificial intelligence reduces 100 candidates to 3, people test only those 3 in depth. When roles are divided this way, development becomes faster and failures decrease. This is what metal design AI that does not conflict with physical laws looks like in practice.

Summary

This chapter dealt with metals that protect the hottest places in rockets. Those places are the combustion chamber nozzle and the reentry surface. There, heat, pressure, and force all attack the material together.

Long-used nickel-based superalloys reach their limit at about 1,150 degrees. A new candidate for temperatures above that is the refractory high-entropy alloy. This alloy mixes several metals with high melting points in roughly similar amounts.

Early research explained the properties of these alloys through four effects. They are configurational entropy, lattice distortion, sluggish diffusion, and the cocktail effect. Some of these later faced objections, so they are hard to treat as general laws. Even so, this framework is a starting point for thinking about which compositions will remain stable as a single phase. The omega parameter and valence electron concentration are simple indicators that help with this judgment.

Artificial intelligence is needed to handle the wide composition space. Artificial intelligence reads materials as numbers called descriptors. Large language models extract these features from papers. Graph neural networks learn materials as points and lines. Crystal fractional graph neural networks are a new attempt to capture the disorder of high-entropy alloys.

Atomic-level behavior is studied with molecular dynamics. An artificial intelligence method called a neuroevolution potential makes this calculation faster. Recent research has reached scales of more than 1 trillion atoms. This large calculation observes dislocations, grain boundary sliding, and cracks at the atomic level.

The idea that connects all these tools is physical consistency. Pure data learning can produce answers that violate physics. A physics-informed neural network (PINN) prevents this by putting laws into the loss function. Artificial intelligence is a guide that narrows the candidates, and the final judgment comes from experiments.

Most of the cases in this chapter are still early-stage studies. The numbers from prior studies are not fixed values. Company cases are industrial examples, not academic demonstrations. So good design always goes through both calculation and experimental confirmation. The current state of this field is collaboration in which artificial intelligence narrows the path and people choose the path.

This chapter can be reduced to one sentence. Artificial intelligence changes the speed of high-temperature metal design. In the past, testing one candidate took a long time. Now artificial intelligence screens many candidates in advance. Only the few that remain are confirmed through precise calculations and tests.

The future tasks are also clear. Experimental data must be collected more broadly and evenly. The uncertainty of artificial intelligence predictions must be reported together. Physical laws must be built more tightly into learning. As these three things come together, rocket material design will become faster.

Related Supporting Materials

Yeh, J. W., Chen, S. K., Lin, S. J., Gan, J. Y., Chin, T. S., Shun, T. T., Tsau, C. H., & Chang, S. Y. (2004). Nanostructured high-entropy alloys with multiple principal elements: Novel alloy design concepts and outcomes. Advanced Engineering Materials, 6(5), 299-303. https://doi.org/10.1002/adem.200300567

Cantor, B., Chang, I. T. H., Knight, P., & Vincent, A. J. B. (2004). Microstructural development in equiatomic multicomponent alloys. Materials Science and Engineering A, 375-377, 213-218. https://doi.org/10.1016/j.msea.2003.10.257

Senkov, O. N., Wilks, G. B., Miracle, D. B., Chuang, C. P., & Liaw, P. K. (2010). Refractory high-entropy alloys. Intermetallics, 18(9), 1758-1765. https://doi.org/10.1016/j.intermet.2010.05.014

Senkov, O. N., Wilks, G. B., Scott, J. M., & Miracle, D. B. (2011). Mechanical properties of Nb25Mo25Ta25W25 and V20Nb20Mo20Ta20W20 refractory high entropy alloys. Intermetallics, 19(5), 698-706. https://doi.org/10.1016/j.intermet.2011.01.004

Zhang, Y., Zuo, T. T., Tang, Z., Gao, M. C., Dahmen, K. A., Liaw, P. K., & Lu, Z. P. (2014). Microstructures and properties of high-entropy alloys. Progress in Materials Science, 61, 1-93. https://doi.org/10.1016/j.pmatsci.2013.10.001

Miracle, D. B., & Senkov, O. N. (2017). A critical review of high entropy alloys and related concepts. Acta Materialia, 122, 448-511. https://doi.org/10.1016/j.actamat.2016.08.081

George, E. P., Raabe, D., & Ritchie, R. O. (2019). High-entropy alloys. Nature Reviews Materials, 4(8), 515-534. https://doi.org/10.1038/s41578-019-0121-4

Rao, Z., Tung, P. Y., Xie, R., Wei, Y., Zhang, H., Ferrari, A., Klaver, T. P. C., Kormann, F., Sukumar, P. T., da Silva, A. K., Chen, Y., Li, Z., Ponge, D., Neugebauer, J., Gutfleisch, O., Bauer, S., & Raabe, D. (2022). Machine learning-enabled high-entropy alloy discovery. Science, 378(6615), 78-85. https://doi.org/10.1126/science.abo4940

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045

Fan, Z., Zeng, Z., Zhang, C., Wang, Y., Dong, H., Chen, Y., & Ala-Nissila, T. (2021). Neuroevolution machine learning potentials: Combining high accuracy and low cost in atomistic simulations and application to heat transport. Physical Review B, 104, 104309. https://doi.org/10.1103/PhysRevB.104.104309

Fan, Z., Wang, Y., Ying, P., Song, K., Wang, J., Wang, Y., Zeng, Z., Xu, K., Lindgren, E., Rahm, J. M., Gabourie, A., Liu, J., Dong, H., Wu, J., Chen, Y., Zhong, Z., Sun, J., Erhart, P., Su, Y., & Ala-Nissila, T. (2022). GPUMD: A package for constructing accurate machine-learned potentials and performing highly efficient atomistic simulations. Journal of Chemical Physics, 157, 114801. https://doi.org/10.1063/5.0106617

He, Q., Wu, D., Wang, H., et al. (2024). Machine learning assisted design of refractory high-entropy alloy for ultrahigh temperature applications. Materials & Design, 246, 113326. https://doi.org/10.1016/j.matdes.2024.113326

Divilov, S., et al. (2024). A comparative study of predicting high entropy alloy phase fractions with traditional machine learning and deep neural networks. npj Computational Materials, 10, 68. https://www.nature.com/articles/s41524-024-01335-1

Hasnain, M. (2026). Physically Consistent Machine Learning for Melting Temperature Prediction of Refractory High-Entropy Alloys. arXiv Preprint (not yet peer reviewed), arXiv:2601.03801. https://arxiv.org/abs/2601.03801

Suo, P., Cao, W., Wu, X., Zhang, W., Fan, Z., Xian, S., et al. (2026). Trillion-atom molecular dynamics simulations with ab initio accuracy. arXiv Preprint (not yet peer reviewed), arXiv:2604.24816. https://arxiv.org/abs/2604.24816

Hsiao, L. C., Liu, Z. K., & Reinhart, W. (2026). Modeling High Entropy Alloys' Mechanical Property through Natural Language-Derived Descriptors. arXiv Preprint (not yet peer reviewed), arXiv:2604.21807. https://arxiv.org/abs/2604.21807

Kotama, T., & Huang, Y. (2026). Crystal Fractional Graph Neural Network for Energy Prediction of High-Entropy Alloys. arXiv Preprint (not yet peer reviewed), arXiv:2605.08103. https://arxiv.org/abs/2605.08103

Bejjipurapu, A., Karumuri, S., Flanagan, J. C., Tucker, V., Bilionis, I., Strachan, A., Sandhage, K. H., & Titus, M. S. (2025). Active Learning Discovery of High Temperature Oxidation Resistant Refractory Complex Concentrated Alloys. arXiv Preprint (not yet peer reviewed), arXiv:2512.15958. https://arxiv.org/abs/2512.15958

Cao, Y., Sheriff, K., & Freitas, R. (2025). Capturing short-range order in high-entropy alloys with machine learning potentials. npj Computational Materials, 11, 268. https://doi.org/10.1038/s41524-025-01722-2

Liu, X., Yang, K., Liu, Y., Zhou, F., Fan, D., Pei, Z., Xu, P., & Tian, Y. (2025). Revealing nanostructures in high-entropy alloys via machine-learning accelerated scalable Monte Carlo simulation. npj Computational Materials, 11, 267. https://doi.org/10.1038/s41524-025-01762-8

Huang, H., Yu, Y., Deng, W., Qian, Q., & Zheng, H. (2026). Breaking equiatomic constraints: knowledge-enhanced AI framework for function-oriented single-phase high-entropy alloy design. npj Computational Materials, 12, 209. https://doi.org/10.1038/s41524-026-02071-4

Kamnis, S., & Delibasis, K. (2025). High entropy alloy property predictions using a transformer-based language model. Scientific Reports, 15, 11861. https://doi.org/10.1038/s41598-025-95170-z

Chizhevskiy, V., Marković, G., Benrazzouq, S., Cvelbar, U., Nominé, A., & Zavašnik, J. (2026). High Entropy Alloys Database generated with Large Language Model. Scientific Data, 13, 612. https://doi.org/10.1038/s41597-026-06930-z

Yao, J., Yu, W., Wang, J., Zhang, L., Liu, F., Li, W., Tan, L., Huang, L., & Liu, Y. (2025). Integrating Machine Learning and Thermodynamic Descriptors for Enhanced Ni-Based Single Crystal Superalloys Creep Life Prediction and Alloy Design. Metals and Materials International, 31, 2730-2748. https://doi.org/10.1007/s12540-025-01906-x

Song, J., Shan, Y., Zhang, Z., Wang, S., Zhang, H., Muhammad, S., Huang, H., & Li, Y. (2025). Phase-field-informed machine learning on creep behavior of Ni-based single-crystal superalloys. Journal of Materials Informatics, 5, 2. https://doi.org/10.20517/jmi.2025.02

Chapter 5 Microstructure Control and Heating-Life Evaluation of Aerospace Ceramic Matrix Composites (CMC) and Ultra-High-Temperature Ceramics (UHTC)

Introduction

A ceramic bowl does not melt even when placed in a hot oven. But if a hot bowl is put directly into cold water, it cracks. It is strong against heat but weak against sudden temperature changes and impact. Materials that protect the surfaces of rockets and hypersonic vehicles also seek an answer between these two properties. But the temperatures and forces there cannot be compared with those in a kitchen.

The inside of a rocket engine combustion chamber is filled with hot gas at thousands of degrees. The front part of a hypersonic vehicle heats itself as it collides with air. Surfaces in these places face heat, oxidation, wear, and large forces all at once. Oxidation is the process in which oxygen meets a material and eats away at it.

Metals are strong, but they soften at very high temperatures. It is similar to taffy bending in the hand. So researchers turn to ceramics. Like pottery, ceramics resist heat and hold up well even at high temperatures.

The two materials covered in this chapter are ceramic matrix composites (CMC) and ultra-high-temperature ceramics (UHTC). Ceramic matrix composites are made by placing thin fibers inside a ceramic matrix. Ultra-high-temperature ceramics are ceramics with very high melting points. Both aim to keep their shape in hot places.

The problem is that it is hard to know in advance when these materials will fail. It is difficult to reproduce real flight environments exactly in a laboratory. Each test is also expensive. So artificial intelligence (AI) appears as a tool to reduce this difficulty.

Artificial intelligence is used to read small structures inside materials and predict their lifetime. Small structures mean the shapes of grains, fibers, pores, and boundaries seen under a microscope. This chapter explains, step by step, how these small structures are turned into numbers and how the future is predicted. It explains the ideas in words, without equations.

This chapter has one main flow. It is the work of turning the small world of materials into numbers. Microscope images, crystal orientations, and atomic vibrations are all targets. This information is converted into a form that artificial intelligence can read. Then artificial intelligence predicts the future and shapes the structure.

The order of this chapter is as follows. First, we look at how materials degrade in extreme environments. Next, we look at a method for predicting lifetime by converting crystal grains into graphs. Then we look at a method for designing fiber architecture in reverse. Finally, we look at a method for connecting heat flow in a very small world with heat flow across an entire component.

One principle should be stated in advance. The artificial intelligence in this chapter does not replace testing. Performance numbers produced by artificial intelligence must be confirmed with real objects. Unconfirmed numbers remain only reference points. This book includes only verified data in the main text. This principle is the foundation for keeping safety in materials development.

5.1 Extreme Environments in Hypersonic Flight and Next-Generation Rocket Propulsion, and Degradation Mechanisms of Thermal Structures

What This Section Covers

This section first shows how harsh the environments are that materials must withstand. It uses hypersonic vehicles and the inside of high-pressure rocket engines as examples. It explains why ordinary metals are not enough and why ceramics are needed.

Next, it looks at how ceramic materials actually degrade. It examines, in order, oxidation of carbon fiber composites and melting erosion and phase changes in ultra-high-temperature ceramics. Melting erosion is the process in which a surface melts or is worn away by hot gas. We must understand this damage process to understand what the artificial intelligence predictions in the next section learn.

5.1.1 Nonequilibrium Aerodynamic Heating and High-Enthalpy Combustion Gas Environments

Aerodynamic heating is the process in which air heats a surface because of fast flight. Friction with the surface also adds heat. But in hypersonic flight, the larger share comes from air being strongly compressed and heated. The faster the speed, the more sharply this heat grows and heats the surface.

Hypersonic speed means more than five times the speed of sound. Mach 7 means a speed seven times faster than sound. At this speed, air does not flow in an ordinary way. A shock wave forms in front of the vehicle. A shock wave is a thin wall of pressure made when air is suddenly compressed.

After passing through the shock wave, the air slows and heats in an instant. When the temperature becomes very high, air molecules split apart. Oxygen molecules and nitrogen molecules separate into individual atoms. This separated atomic oxygen is more reactive than ordinary oxygen, so it attacks materials more easily.

The parts of a vehicle that receive heat more severely than anywhere else are the leading edge and the nose. The leading edge is the edge of a wing that meets the air first. The nose is the pointed front of the vehicle. The surface temperature there rises to thousands of degrees. On top of this, shear force that pushes the surface sideways acts at the same time.

The inside of a rocket engine is just as harsh. When methane and liquid oxygen burn, hot gas is produced. This gas contains water, carbon dioxide, and carbon monoxide. Some oxygen also remains. The combustion chamber pressure reaches several tens of atmospheres. This high-pressure gas strikes the combustion chamber wall without stopping.

This kind of environment is called a high-enthalpy environment. Enthalpy is the amount of thermal energy contained in a gas. High enthalpy means that the gas has a very large amount of heat. This gas pours heat into the wall. However, the wall temperature is lower than the gas temperature. That is because heat spreads into the material and cooling also takes place.

Time is another problem. A rocket burns for hundreds of seconds from launch to orbit. A hypersonic vehicle glides through the atmosphere for a long time. This means it receives heat for a long time, not just for a brief moment. So it is not enough for the material to survive only briefly.

Let us get a sense of how hot it is. The stagnation point at a hypersonic leading edge is close to the limit of the material. The stagnation point is the place where the flow stops and heat concentrates the most. The heat load at this place cannot be compared with a kitchen gas flame. The shear force that pushes the surface sideways is also applied at this place.

One material must withstand all these conditions. Heat, oxidation, shear force, and long exposure time overlap. A 2024 review of materials for hypersonic flight identifies this combined environment as a central challenge in materials selection. This challenge is the starting point of ceramic materials

research.

The role of artificial intelligence here is clear. Real tests are expensive and rare. So we must predict a wide range of conditions from data from only a few tests. Artificial intelligence is good at finding rules from small amounts of data. But without help from physical laws, it can produce wrong answers. This is why the methods in this chapter use physics and artificial intelligence together.

5.1.2 How Carbon Fiber Composites Are Damaged by Oxygen (C/SiC)

Carbon fiber reinforced silicon carbide (C/SiC) is a representative thermal structural material. Carbon fibers are thin and light, and they do not break easily. Silicon carbide (SiC) withstands high temperature and oxidation relatively well. Combining the two creates a lightweight and strong composite. This material is written below as C/SiC.

There are two main ways to make this material. One is chemical vapor infiltration. Gas is allowed to penetrate and fill the spaces between fibers with ceramic. The other is liquid resin infiltration and carbonization. Resin is inserted and fired to turn it into ceramic. Both methods fill the spaces between fiber bundles with ceramic.

The problem already begins during manufacturing. Carbon fibers and the silicon carbide matrix expand by different amounts when heated. This is called a difference in the coefficient of thermal expansion. When they cool from a hot state to room temperature, they shrink by different amounts. As a result, a fine network of cracks forms inside the matrix. These microcracks inevitably form during manufacturing.

Cracks become pathways for oxygen to enter. Even if the outside looks intact, oxygen penetrates inside. It is similar to water seeping through a small gap in wallpaper. Oxygen that enters attacks the carbon fibers and the interphase layer. The interphase layer is the thin boundary between the fibers and the matrix.

The silicon carbide matrix reacts with oxygen in two ways. When the temperature is not very high and there is enough oxygen, a glass film forms on the surface. This film is a thin protective layer made of silicon dioxide. This is called passive oxidation. This glass film blocks oxygen penetration and protects the inside of the material.

By contrast, when the temperature is very high or the oxygen pressure is low, the situation changes. A protective glass film cannot form. Instead, silicon carbide turns directly into gas and leaves. This is called active oxidation. In active oxidation, the surface keeps being worn away without a protective film. This risk increases in high-altitude hypersonic flight.

When the interphase layer disappears, the strength of the composite breaks down. Originally, fibers hold a crack like a bridge when the crack grows. This is called fiber bridging. When the interface disappears through oxidation, this holding force is lost. Then the material becomes likely to break suddenly. Delamination damage, in which layers separate from one another, also occurs.

This damage is not easily visible from the outside. Even if the surface looks intact, the interface inside is damaged first. So visual inspection alone is not enough. Oxygen penetration and the state of the interface inside the material must be examined. Artificial intelligence lifetime models read this hidden damage inside, not just the outside.

A study published in 2026 shows this damage in numbers. Wu and colleagues predicted oxidation damage in 2.5-dimensional woven C/SiC using a physical model. In this study, room-temperature tensile strength was about 264 megapascals. When the temperature rose to 1100 degrees, the strength fell to about 184 megapascals. The elastic modulus also decreased by about 23 percent over the same range.

This study uses a physics-based constitutive model, not an artificial intelligence model.

There is another way to block oxygen. It is to apply a protective coating to the surface. This is called an environmental barrier coating. This coating prevents oxygen and water vapor from entering. Hypersonic research aims for coatings that can withstand temperatures above 2000 degrees Celsius. Artificial intelligence is used to decide the coating thickness and composition.

What does this result suggest? If we look only at the average properties of the material, we miss interface damage. We must examine where oxygen entered and how much entered at each location. Temperature history and load history must also be considered. The artificial intelligence lifetime model discussed later takes on exactly this task.

A limitation should also be noted here. The model by Wu and colleagues is a physics-based calculation. It predicts damage even without artificial intelligence. Artificial intelligence is used to learn data from such physical models and increase speed. Physical models and artificial intelligence are not rivals. They are partners that fill each other's gaps.

5.1.3 How Ultra-High-Temperature Ceramics Melt and Spall (UHTC)

Ultra-high-temperature ceramics are ceramics with very high melting points. Representative materials are zirconium diboride and hafnium diboride. Zirconium diboride is a compound of zirconium and boron. The melting points of these materials exceed 3000 degrees Celsius. So they are studied as candidates for hypersonic leading edges and rocket components.

A high melting point does not mean we can relax. The real problem is oxidation. When zirconium diboride meets oxygen, two oxides form. One is a hard oxide called zirconia. The other is a boron oxide called diboron trioxide.

Diboron trioxide shows an interesting property at low temperature. It melts and becomes a liquid at about 450 degrees Celsius. This liquid fills small gaps on the surface. It works like ointment applied to a wound to fill gaps. This is called a self-healing film. In the early stage of oxidation, this film protects the material.

But when the temperature rises further, this defense breaks down. Above about 1100 degrees Celsius, diboron trioxide gradually evaporates. Around 1500 degrees Celsius, this film effectively disappears. The exact limit temperature changes depending on oxygen pressure, gas speed, and pressure. So it is hard to state as one fixed value.

Adding silicon carbide improves the situation. The liquid glass formed by oxidized silicon carbide mixes with diboron trioxide. The two components combine to form a glassy phase called borosilicate. This glass covers the surface up to a higher temperature.

When the temperature rises even higher, even this glass cannot hold up. Above about 1700 degrees Celsius, the viscosity of the glass decreases. The glass evaporates and is torn away by fast gas flow. When the protective film disappears in this way, the surface is rapidly eroded. In the end, only a porous zirconia skeleton remains on the surface. This porous skeleton lets oxygen pass farther inside more easily.

Zirconia has another trap. Its crystal structure changes with temperature. During cooling, it changes from the tetragonal phase to the monoclinic phase. These two names refer to differences in how atoms are stacked. When this change occurs, the volume increases by about 4.5 percent.

When volume suddenly increases, large forces develop inside the oxide film. It is similar to water freezing in winter and breaking a bottle. This force creates cracks in the oxide film. When the cracks

connect, the oxide film falls off in pieces. This is called spallation. A new surface is exposed where the pieces fell off, and oxidation begins again.

High-entropy borides are a new approach to this problem. A high-entropy ceramic is a ceramic made by mixing several metal elements in similar amounts. Five metals may be placed in one crystal. The range of possible compositions is wide, so it suits artificial intelligence search. A computer can first screen which combinations resist oxidation well.

A 2026 review of ultra-high-temperature ceramics shows this trend well. The review covers experiments, multiscale simulation, and data-driven design together. Multiscale means a method that connects several levels, from the atomic scale to the part scale. Data-driven design means a method in which artificial intelligence selects the composition and microstructure.

Artificial intelligence is also used to predict oxide layer thickness. One research team used machine learning to predict oxidation damage in ultra-high-temperature borides. They predicted oxide layer thickness with a method called random forest. Random forest is a method that combines many decision rules to produce an answer. However, the actual behavior of the oxide layer must be confirmed through wind tunnel and arc-jet tests. An arc jet is a device that blows ultra-high-temperature gas to imitate flight conditions.

Machine learning interatomic potential (MLIP) is also used to predict melting points. This is a method in which artificial intelligence quickly calculates the forces between atoms. In 2025, one research team used MLIP to predict the melting points of high-entropy ceramics. They first screened high-melting-point combinations by computer. This method narrows the candidates before experiments. It then reduces the number of costly synthesis and testing steps. When calculation screens candidates first in this way, the whole search for materials becomes faster.

Still, one question remains. Are the combinations chosen by the computer truly resistant to oxidation? Calculations assume ideal conditions. Real surfaces contain impurities and defects. Gas flow and pressure also differ from the calculation. So the final judgment comes from tests on real specimens. Artificial intelligence only helps choose better candidates for testing.

5.2 How to Predict Lifetime by Turning Crystal Grains into Graphs

What This Section Covers

This section explains how to turn the inside of a material into a form that artificial intelligence can read. The tool is a spatiotemporal graph neural network. A graph is a drawing that shows relationships with points and lines. A subway map is a good example. Stations are points, and route segments are lines.

Crystal grains inside a material are turned into points, and grain boundaries are turned into lines. Then the flow of time is added. This is because damage builds up when heat and force are applied repeatedly. This section explains how to build that graph, how damage grows, and how physical laws are put into learning. It is called a spatiotemporal model because it looks at time and space together.

5.2.1 Turning Microstructure into a Graph

A crystal grain is a small crystal particle inside a material. Metals and ceramics are not one large crystal. They are made from many small grains. In each grain, the atoms face in a slightly different direction. This direction is called crystal orientation.

A representative tool for reading crystal orientation is electron backscatter diffraction (EBSD). Inside an electron microscope, electrons are fired to read the direction of each grain. This equipment scans the surface and records the direction of each grain. The result appears as a colorful orientation map.

A grain boundary is the boundary where two grains meet. This boundary often becomes a weak link in a material. Oxygen seeps in along this path. Cracks also grow along this boundary. For this reason, knowing the properties of boundaries is central to lifetime prediction.

The method of turning this into a graph comes from a simple idea. One grain is treated as a point, or node. If two grains touch, the space between them is connected by a line, or edge. In this way, a complex microscope image becomes a clean relationship map. Artificial intelligence reads this relationship map better than a photograph.

The node contains information about the grain. It includes the grain's size, shape, and direction. It also includes the temperature and stress at that location. Stress is the strength of the force that presses or pulls an object. In this way, one point summarizes the state of that grain.

The edge contains information about the boundary. It includes the orientation difference between two grains. It also includes the area and direction of the boundary. It includes how much damage oxidation has caused. Boundaries with larger orientation differences tend to let oxygen penetrate more easily.

This method has already been tested in other materials. In 2023, one research team applied this graph method to steel. They classified locations with severe fatigue damage for each grain in steel. Fatigue is damage caused by repeated force. They converted EBSD maps into graphs and trained a model on them. In this study, the graph method proved useful for identifying damage locations by grain. However, this study dealt with fatigue classification in steel. It did not predict ceramic lifetime, creep rupture, or changes over time.

Applying this graph method to ceramic lifetime prediction is the direction proposed by this book. It has not yet been demonstrated for ceramics. The explanation below is not a verified fact. It is a sketch of where research can go.

New methods for handling boundary information are also emerging. In 2026, one research team filled missing areas in EBSD maps through graph learning. This is a technique that fills areas missing from measurements according to crystallographic rules. This method puts crystal rules into artificial intelligence so that it does not fill in unrealistic orientations. Such techniques are useful because materials data is not always complete.

The graph method has advantages over photographs. A photograph handles each pixel. So it does not know by itself where a grain starts and ends. A graph groups each grain as one unit. This is closer to how people view materials. For that reason, it is easier to learn relationships between grains.

Graphs are easy to handle even when their size changes. Photographs must have a fixed size. A graph can be handled in the same way even if the number of grains increases or decreases. This property is useful because each material has a different number of grains. Rules learned from a small specimen can be tested on a large part. However, when the size increases, the defect distribution and thermal gradient also change. So a step of verification in large parts is needed again.

A graph built in this way can also be applied to ceramics. This is because ceramics are also made of crystal grains and boundaries. Oxidation and creep in ultra-high-temperature ceramics also begin at boundaries. Creep is a phenomenon in which a material slowly stretches under force over a long time. For this reason, this graph method becomes a foundation for ceramic lifetime prediction.

5.2.2 Graph Learning That Looks at Neighbors and Time Together

Now the flow of time is added to the graph. Damage does not end in an instant. Heat and force repeat, and damage builds up little by little. So even the same material has a graph that changes over time. Tracking this change is the work of a spatiotemporal model.

First, consider space. The state of one grain is affected by its neighbors. If a neighboring grain receives a large force, force is also transferred to my grain. The calculation that gathers this neighboring information is graph convolution. Each node receives information from neighboring nodes and updates its own state.

Not all neighbors are treated in the same way. Some boundaries contribute greatly to damage, while others contribute less. So the model gives different weights to different neighbors. The method that decides these weights is attention. It gives greater attention to important neighbors.

The key is to put physical knowledge into attention. If the orientation difference at a boundary is large, the boundary energy is high. If the boundary energy is high, oxygen enters more easily. These physical rules are reflected in the weight calculation. Then the model avoids unreasonable conclusions even when data is limited.

Time is handled in a similar way. The past state creates the next state. Today's damage is added on top of the damage accumulated until yesterday. The model learns this chain of time. To do this, it uses a neural network structure that handles time order.

The important damage covered in this section has direction. In ceramic composites, fine cracks first form in the matrix. Along these cracks, the interface between the fiber and the matrix separates. Oxygen enters that space, and an oxidation layer grows. Damage usually spreads along the direction of force and gas flow.

This directional damage has a large effect on material lifetime. If cracks and interfacial debonding continue in one direction, that direction becomes weak. If pores connect to each other, the oxygen path becomes wider. A glassy phase trapped at boundaries may move and transfer damage. These changes alter the creep rate and bring rupture forward. Rupture means the moment when a material can no longer withstand load and breaks.

Grain-boundary sliding is also an important form of damage in this section. At high temperature, grains slide against each other along their boundaries. It is similar to wet bricks pushing past each other. This sliding increases creep deformation. The direction and properties of the boundary determine the amount of sliding. That is why edges containing boundary information are important.

Hot gas usually flows in one direction and pushes the surface. So damage in ceramics also grows with direction. A spatiotemporal graph model tries to include this directionality in learning. Then it can predict in which direction damage will grow. There is still not enough data to verify this kind of directional prediction.

There are two ways to handle time. One is to read the time sequence step by step. It is similar to reading a sentence from the beginning. The other is to handle the rate of change directly. It learns how fast damage grows. Both methods capture the process by which damage accumulates.

Attention is a central tool in today's artificial intelligence. Artificial intelligence that translates text also uses this method. The principle is to focus more on important words in a sentence. In a materials graph, it focuses on neighbors that matter for damage. The same principle of attention is used in both language translation and materials analysis.

This spatiotemporal graph method is still developing. There are not many large-scale demonstrations for ceramic composites. Good prediction requires a large amount of good data. Collecting ultra-high-temperature test data is itself difficult. So help from physical laws, discussed in the next subsection, is needed.

5.2.3 Directional Creep Rupture and Learning That Includes Physical Rules

Creep is a phenomenon in which a material slowly stretches under force over a long time. It is like a clothesline that sags little by little after hanging for a long time. At high temperature, this phenomenon becomes faster. Rocket parts receive force for a long time while they are hot. So creep becomes a major variable that determines the lifetime of rocket parts.

Creep damage builds up differently depending on direction. This is called anisotropy. Anisotropy is a property in which characteristics differ by direction. It is like wood splitting easily along the grain. A material receives forces from several directions at the same time. This is called multiaxial loading.

Researchers summarize this damage with one variable. This is called a damage variable. If the damage variable is 0, the material is intact. If it is 1, that location has broken. As damage grows, the remaining part receives more force. Then the damage grows faster. Because of this feedback, damage grows out of control and brings rupture forward.

When temperature rises, creep becomes faster. This is because atoms move more actively. This temperature dependence is captured with an exponential function. Even a small rise in temperature can greatly increase the damage rate. So it is important to lower the maximum surface temperature. The coating in the previous section and the heat flow in this section meet here.

Oxidation makes this damage faster. As the oxide layer thickens, boundaries become weaker. So oxide thickness is included in the damage law. This means creep and oxidation are handled together, not separately. In a rocket environment, the two always occur together.

There is one important physical rule here. Damage grows only in one direction. A crack that has formed does not heal on its own. Even as time passes, the damage variable does not decrease. This rule is called the irreversibility of damage. It is in the same context as the second law of thermodynamics.

If artificial intelligence violates this rule, its prediction is wrong. If it produces an answer in which damage decreases, it conflicts with reality. So a penalty is added during training. If an answer violates the rule, the loss is made large. Putting this penalty into the loss function is physics-informed learning.

Other rules are also included in physics-informed learning. Force balance is one of them. At one location, the sum of pushing and pulling forces must be consistent. A rule that mass does not disappear or appear is also included. These rules are added to the loss function. Then the model avoids answers that violate physics.

The rupture criterion also considers direction. Materials are weak against tensile force. So the direction with the greatest tensile force is examined separately. Internal swelling pressure is also considered. Forces in several directions are combined into one measure of risk. This risk determines the rate at which damage grows.

The roots of this method lie in physics-informed neural networks. In 2019, Raissi and his colleagues organized this concept. They put physical equations directly into neural network training. The idea is that physical rules fill the gaps even when data are scarce. This idea carries directly into predicting ceramic lifetime.

The real value of physical rules appears when predictions are extended. Test data are usually clustered at lower temperatures. What we really want to know is behavior near 2000 degrees Celsius. There is almost no data at this high temperature. With physical rules, we can make reasonable predictions even in regions without data.

This method still has limits. If the physical rules are put in incorrectly, prediction can become worse. Setting the weights among rules is also difficult. We cannot skip verification with real fracture data. For this reason, artificial intelligence prediction does not replace testing. It is used as a tool to reduce testing.

5.3 Generative Models for Designing Fiber Weaves and Coatings

The previous section predicted the lifetime of materials that had already been made. This section changes direction. First, we set the desired performance. Then we search backward for the structure that produces that performance. This is called inverse design. In cooking terms, it is like choosing the desired taste and then finding the mix of ingredients.

This section deals with three-dimensionally woven fiber structures and interface coatings. The question is what angle to weave the fibers at and how much coating to apply. For this purpose, we use a generative artificial intelligence method called a diffusion model. A diffusion model is a method that forms a clear structure from blurry noise. To this, we add a property that keeps the result stable even when the direction changes.

5.3.1 Three-Dimensionally Woven Fiber Structures and Direction-Stable Learning

A three-dimensional braided composite is a material made by weaving fiber bundles in three dimensions. It is similar to a basket woven tightly in three directions. An ordinary composite is made by stacking thin layers. Then the spaces between layers become weak and split easily. This is called delamination.

Three-dimensional braiding removes this weakness. Fibers are intertwined in several directions, so there is no layer boundary. As a result, damage from layer separation is less likely to occur. This is useful in places with large thermal shock, such as hypersonic leading edges. Thermal shock is damage that occurs when a material suddenly heats up or cools down.

There are several types of braiding. Fibers may be woven in four directions or in even more directions. The more directions there are, the more evenly the material resists different loads. But weaving becomes more difficult. So the weave must be chosen for the part's use. This choice of weave is also a subject of inverse design.

The problem is that the weave is not always perfect. In real manufacturing, fibers twist. They may also be pressed so that their cross sections become distorted. The weaving angle also shifts slightly. Coating thickness is not uniform from place to place. These variations make force concentrate in certain locations.

These variations are handled with probability. This is because not all products are identical. Each is slightly different. So we deal with a distribution of structures, not one correct structure. A diffusion model is well suited to handling this distribution. It can produce many plausible structures.

Here, a property called equivariance is important. Equivariance means that when an object is rotated, the result rotates in the same way. If a part is rotated to the left, the predicted force must also rotate to the left. Without this property, the same structure is judged differently depending on its orientation. That violates physics.

In three-dimensional space, an object can rotate and move. There is mathematics that handles these two motions together. This is called the special Euclidean group. The name is difficult, but the meaning is simple. It is a set of rules that treats rotation and translation together. Artificial intelligence is designed to always follow these rotation and translation rules.

Such equivariant diffusion models first developed in molecular design. In 2025, Chen and his colleagues presented an equivariant diffusion model that combined graphs at several scales. They used this model to design three-dimensional molecules. Molecules also rotate and translate in three-dimensional space. So the same mathematics can be brought into fiber structures. However, that study dealt with molecules, not ceramic fiber weaves. Fiber bundles, weave constraints, pores, and manufacturability

have not yet been verified with this model. Applying it to ceramic braiding is the direction this book sets out.

Let us look at why equivariance matters with an example. Place the same screw on a desk in different directions. The strength of the screw does not depend on its orientation. Artificial intelligence must also obey this fact. Without equivariance, it judges the same screw differently in each direction. Then data must be collected separately for each direction, which wastes a great deal of training.

Keeping equivariance saves data. Data from one direction can cover all directions. This saving is large when materials data are scarce. So it fits fields with little data, such as ultra-high-temperature ceramics. It can be seen as a way to build physical laws into the structure of the model.

In summary, three elements meet. They are three-dimensional structures, probabilistic variation, and a property that is stable under direction changes. These three must be included in one model. Only then can the diversity of real manufacturing be reflected as it is. The next subsection looks at how this model forms structures.

5.3.2 Diffusion Models That Recover Structure from Noise

The basic idea of a diffusion model is divided into two stages. The first stage is the process of blurring a structure into noise. The second stage is the process of recovering the structure from noise. Artificial intelligence learns the recovery process. It is like running backward the spread of a drop of ink in water.

Let us start with the first stage. Noise is added little by little to a clear fiber structure. As time passes, the structure becomes blurred. At the end, only complete noise remains. This process proceeds slowly according to rules. Those rules are called the noise schedule.

The second stage follows this process in reverse. It starts from noise and gradually recovers the structure. At each step, it removes a little noise. At this point, there is a value that tells which direction the noise should be removed in. This is called the score. The score points in the direction in which the structure becomes more plausible.

Artificial intelligence learns this score. It looks at a blurred structure and estimates the direction toward a clearer one. After training is complete, it can form a structure even from pure noise. This method is called score-based diffusion. Today's image-generating artificial intelligence draws pictures by the same principle.

This method suits materials design for a reason. A material structure does not have only one correct answer. Several structures can produce the same performance. A diffusion model creates these several candidates. Then a person chooses the one that is easier to make. It is useful because it gives several options, not just one answer.

Adding conditions is the core of inverse design. Creating any random structure is not useful. The model must create a structure with the desired performance. So the target performance is entered as a condition. The targets include thermal conductivity, tensile strength, fracture toughness, and ablation rate.

Fracture toughness is the ability to resist crack growth. Ablation rate is the speed at which the surface is worn away. These values are set at the desired levels. Then the model selects and forms structures that match those conditions. It is also possible to adjust whether the targets are reflected strongly or weakly.

This conditioning technique is already used in several materials. In 2025, a research group presented a three-dimensional diffusion model with physical constraints. They designed fiber-reinforced polymer composites to match target performance. They set a desired stress-strain curve and searched backward

for the structure. This method made the generative process directly follow physical rules.

When conditions are applied strongly, another problem arises. If the model pursues only target performance, it may produce a structure that cannot be made. It may look good on paper but cannot actually be woven. So the answer must be found within the range that can be made. This means adding manufacturing rules to the conditions. Finding this balance between performance and manufacturability is a difficult part of inverse design.

To use this method for ceramic composites, there are hurdles to overcome. A large amount of three-dimensional structural data for ceramics must be collected. This data is obtained by X-ray tomography. X-ray tomography works on the same principle as CT scanning in a hospital. It shows the inside of a part in three dimensions without cutting it open. The structures obtained this way become the foundation of the model.

Collecting data takes time. Scanning one specimen takes a long time. Dividing the scanned data into grains and fibers also requires much work. So data is always lacking. The equivariance discussed earlier helps reduce this shortage. It allows the model to learn more cases from less data.

5.3.3 Setting Fiber Angle and Coating Thickness Together

An inverse design model makes two decisions. One is the angle at which fibers are woven. The other is the thickness of the interface coating. These two strongly affect the material's strength and heat resistance.

Let us begin with the fiber angle. Force is usually applied strongly in one direction. A rocket nozzle receives force that pushes outward from the inside. A hypersonic leading edge receives shear force in the flow direction. When fibers are placed in the direction of the force, the material resists force well. So the model sets the angle with the force in mind.

Coating thickness is just as important. Here, two types of films must be distinguished. One is a thin interface layer placed between the fiber and the matrix. This interface layer is made of boron nitride or pyrolytic carbon. This film prevents a crack from cutting straight through the fiber. The crack slides along the film and spreads out the force. If the interface layer is too thin, this effect is weak. If it is too thick, other problems arise.

The other is an environmental barrier coating applied to the outside of the material. This coating blocks oxygen and water vapor from entering. Rare-earth disilicate is a candidate for this outer coating. This material blocks oxygen well at high temperatures. The interface layer and the outer coating differ in location and role. Several layers of the two films may be stacked so that each layer has a different role.

The inverse design model handles these two decisions at the same time. It tries to raise strength while lowering thermal stress. These two goals often conflict with each other. Gaining one means losing a little of the other. This kind of problem is called multi-objective optimization. The model finds a good balance between strength and thermal stress.

The concentration of force in one place is called stress concentration. It is like clothing tearing first at one thread. Cracks form first where stress concentration is large. So design tries to distribute force evenly. This is why fiber angle and coating thickness are refined.

When finding a balance, smoothness is also considered. If a fiber bends suddenly, force concentrates there. The same is true when coating thickness changes abruptly. So the model guides the angle and thickness to connect smoothly. Designs with sudden changes in angle or thickness receive a penalty.

This kind of inverse design is still close to the proof-of-concept stage. Claims about how much a specific performance value improves must be accepted with care. We must check whether the structure produced by the model actually delivers that performance. This verification again takes place through manufacturing and testing. Artificial intelligence takes the role of narrowing the candidates.

In 2026, generative research using text-based conditions also appeared. A research group created microstructures from sentences containing target properties. This is still early research that has not gone through peer review. Even so, the direction is clear. A person states the desired performance, and artificial intelligence forms the structure. This flow can lead to the design of ceramic composites.

5.4 Connecting the Very Small World to Heat Flow Across the Whole Part

This section deals with heat flow inside materials. Heat is the key that determines material lifetime. We must know where heat flows and how fast it flows to predict temperature. We must know the temperature to predict the later rates of oxidation and creep.

The problem is that heat flow is determined in a very small world. Atomic vibrations carry heat. But what we want to know is the temperature of the whole part. The very small world and the very large world must be connected. This connection is called multiscale matching. This section explains how to make that connection with a deep operator neural network.

5.4.1 Calculating Thermal Conduction from Atomic Vibrations (Green-Kubo Molecular Dynamics)

In ceramics, heat is transferred by atomic vibrations. The particles of these vibrations are called phonons. It is easy to think of phonons as particles of sound. Vibrations spread along the lattice and carry heat. How well these vibrations spread determines thermal conductivity.

At high temperatures, these vibrations collide with one another. When phonons collide, the flow of heat is disrupted. Grain boundaries, defects, and mixed atoms also scatter vibrations. The more scattering there is, the more slowly heat flows. To understand thermal conduction in ultra-high-temperature ceramics, we must understand this scattering.

High-entropy ceramics are unusual in this respect. Several elements are mixed, so the lattice is uneven. This disorder strongly scatters vibrations. As a result, heat does not spread well. If the goal is thermal shielding, this property helps. Less heat from the outer surface enters the inside.

The method for calculating this tiny world is molecular dynamics. It moves each atom inside a computer. It tracks how atoms vibrate and collide. This calculation covers very short times, on the order of nanoseconds. A nanosecond is one billionth of a second. It is a world that is that small and that fast.

The formula used to obtain thermal conductivity is the Green-Kubo formula. This formula watches fluctuations in heat flow over a long time. It is similar to learning the properties of water by watching ripples on calm water. It gathers the fluctuations in heat flow that atoms create on their own. Thermal conductivity is calculated from those fluctuations. This formula was established by Green and Kubo in the 1950s.

The accuracy of the calculation depends on how well we know the forces between atoms. The rules that handle these forces are called interatomic potentials. In the past, people used simple rules made by hand. These rules were fast, but their accuracy was low. Quantum calculations, by contrast, were accurate but too slow.

This is where the machine learning interatomic potential (MLIP) builds a bridge. Artificial intelligence learns from quantum calculation results and predicts forces. It produces accuracy close to quantum

calculations at high speed. This method is covered in detail in earlier chapters of this book. MLIP is also used in ultrahigh-temperature thermal conductivity calculations.

Recent studies show the power of this combination. In 2021, one research team used MLIP to handle zirconia. They calculated phase changes and thermal conduction with temperature at the same time. They extracted heat flow using the Green-Kubo method. This method captures the complex effects of atomic vibration without leaving them out.

There is a limitation in this calculation that should be noted. The Green-Kubo method obtains thermal conductivity when a material is in equilibrium. This value tells us the basic property of the inside of the material. But the surface where ablation occurs is not in equilibrium. There, radiation and surface chemical reactions strongly control heat. So the Green-Kubo value alone cannot explain all the heat at the surface. This value is best used as an input for heat conduction inside a part.

Why must we go all the way down to this tiny world? Could we not simply measure thermal conductivity directly? Measurement is difficult at ultrahigh temperatures. It is not easy to measure inside a ceramic at 2000 degrees Celsius. Oxidation proceeds, and the material itself changes. So we need a way to obtain values through calculation.

Calculation has another advantage. It can preview compositions that have not yet been made. High-entropy ceramics have countless possible combinations. It is impossible to make and measure all of them. Calculation tells us the thermal conductivity of candidates in advance. Then only promising combinations are selected and made in the real world.

This trend continues today. A 2026 study predicted very low thermal conductivity using Green-Kubo and MLIP. This combination is becoming established across several materials. Still, calculation requires large computers. Errors such as time truncation and statistical fluctuations must also be handled. So calculations are repeated several times and averaged.

The principle of validation is the same here. Calculated values must be compared with experimental values. The calculation is tested on several known materials. If the calculation agrees well with experiments, it gives confidence for new materials too. A value that has not gone through this comparison is only a reference. Calculation and experiment are two eyes that check each other.

5.4.2 Heat Flow Equation for the Whole Part

The previous subsection dealt with a very small world. Now we widen our view to the whole part. The temperature of a part differs from place to place and changes over time. The rule that handles this change is the heat conduction equation. The equation is written as a formula, but its meaning can be explained in words.

The meaning of the heat conduction equation is simple. The temperature change at a given location is determined by the difference between heat that enters and heat that leaves. Heat flows from the hot side to the cold side. The speed of this flow is thermal conductivity. The value calculated earlier enters here.

There are several reasons why this equation is difficult. First, thermal conductivity changes with temperature. When temperature rises, thermal conductivity changes. This is called nonlinear behavior. Nonlinear means a relationship in which cause and effect are not directly proportional. In addition, the material has different properties in different directions. Heat flows better along the direction in which fibers are laid.

This is joined by the problem that the material itself changes. When oxidation proceeds, thermal conductivity changes. When the surface is worn away, thickness decreases. So the values in the equation

change over time. The oxidation and creep from the previous section become intertwined with the heat flow in this section. Therefore, these many phenomena must be solved together, not separately.

What happens at the boundary is also complex. Several kinds of heat overlap on a hypersonic surface. Heat from heated air enters. There is also radiation, in which the surface emits heat as light. This is the phenomenon in which a hot object glows red and loses heat. Chemical heat release is added to this.

Chemical heat release occurs when atoms recombine at the surface. In hypersonic flight, air molecules split into atoms. When these atoms pair up again on the surface, they release heat. This is called catalytic recombination. The size of this heat changes depending on the surface material. So the choice of material strongly controls surface temperature.

All this heat is placed into the boundary condition of one equation. Incoming heat, outgoing radiation, and chemical heat release are placed together. Time flows with them. Heat keeps changing along the flight trajectory. This is called an unsteady state. It means dealing with a flowing process, not one frozen moment.

The traditional method for solving this equation is the finite element method. The part is divided into many small pieces. The equation is solved for each piece and then joined together. This method is accurate, but calculation takes a long time. Repeated calculation is needed at each time step. It is difficult to solve an entire Mach 7 trajectory in real time.

Rocket and hypersonic development needs fast answers. Each time the design changes, temperature must be calculated again. During testing, predictions close to real time are needed. Slow calculations cannot meet these demands. This is why the artificial intelligence operator in the next subsection appears.

5.4.3 A Neural Network That Converts Functions into Functions (DeepONet)

A deep operator neural network is artificial intelligence that learns a relationship that converts functions into functions. It is called DeepONet. An ordinary neural network takes numbers as input and produces numbers. DeepONet takes a function as input and produces a function. Here, a function means something like heat input that changes over time.

This idea was organized by Lu and colleagues in 2021. They established the theory and structure of neural networks that learn operators. This study is rooted in one mathematical theorem. It is the theorem that operators can be approximated by neural networks. This approximation theorem becomes the theoretical foundation that supports DeepONet.

DeepONet consists of two networks. One network reads the input function. The other network reads the location and time where the answer will be observed. The results of the two networks are multiplied to produce the answer. The first network is called the branch, and the second is called the trunk. The branch and trunk meet to produce temperature.

Why is it important to convert functions into functions? The input to a heat problem is not a single number. The input is aerodynamic heating that changes over time. The output is also a temperature field that changes with position and time. This type of problem is a relationship that connects one function to another. DeepONet learns this relationship as a whole.

The strength of DeepONet is speed. Once it has learned, it gives answers immediately for new inputs. It does not solve from the beginning each time like the finite element method. It can quickly scan many flight conditions. It is well suited to checking temperature while changing the design.

The problem is training data. To learn good answers, it needs a large amount of accurate data. But accurate data is expensive to make. This is because finite element calculations and molecular dynamics calculations take a long time. If there is too little data, artificial intelligence cannot learn well. The multifidelity method solves this problem.

Fidelity means how precise a calculation is. High fidelity is an accurate but expensive calculation. Low fidelity is a rough but cheap calculation. Multifidelity uses both together. It gathers a large amount of cheap data and only a small amount of expensive data.

The idea behind this method is close to a sketch and a finishing coat. Cheap calculations first draw the big picture. Expensive calculations add precise finishing on top of it. Artificial intelligence learns the difference between low-fidelity answers and high-fidelity answers. This difference is called the residual. By learning the residual, accuracy can be raised even with a small amount of expensive data.

In 2022, Lu and colleagues applied this multifidelity DeepONet to nanoscale heat transfer. They dealt with a heat flow problem in a very small world. The multifidelity method greatly reduced error with the same amount of expensive data. It also greatly reduced the amount of expensive data required. However, this study dealt with nanoscale heat transfer. It was not a demonstration linked to oxidation or ablation of ultrahigh-temperature ceramics, or to Mach 7 thermal loads.

The term homogenization should also be explained. Ceramic composites contain a mixture of fibers and pores. If this complex structure is handled one by one, calculation becomes overwhelming. So a small region is converted into average properties. This is called homogenization. Low-fidelity calculation uses these average properties to draw the big picture.

We should also note why linking the two worlds is difficult. The atomic world moves on the scale of nanoseconds. The part-level world deals with hundreds of seconds. The time scales differ by more than ten orders of magnitude. The size scales differ by a similar amount. It is impossible to calculate this gap all at once.

Connecting the two worlds in this way for ceramic thermal structures is the path proposed by this book. The low-fidelity side becomes a simple homogenization model. The high-fidelity side becomes a precise analysis linking molecular dynamics and finite elements. The picture is that artificial intelligence connects the two. Then thermal conduction in the very small world can lead to part temperature. This connection has not yet been widely validated in ceramic thermal structures.

5.4.4 Learning with Physical Rules and Increasing Speed

Artificial intelligence must not produce just any answer because it is fast. The answer must obey the laws of physics. So physical rules are included in learning. This method has the same roots as the life prediction in the previous section.

First, learning is performed to match the data. It matches both low-fidelity data and high-fidelity data. It guides the artificial intelligence answer to become close to the data. A penalty is given for the difference between the answer and the data. This penalty is called loss.

Learning that obeys the equation is then added. The answer from artificial intelligence is inserted into the heat conduction equation. If the answer satisfies the equation well, the penalty is small. If it deviates, the penalty is large. In this way, it gives physically consistent answers even at locations without data.

Adding physical rules helps the model hold up even with little data. Pure data learning needs a large amount of data. If data is scarce, it fills the gaps in strange ways. Physical rules fill these gaps with equations. This is well suited to ceramic problems, where ultrahigh-temperature data is rare. The ability to hold up with limited data is a major advantage of physics-informed learning.

Boundary conditions are added in the same way. The model is made to obey the heat balance at the surface. Incoming heat and outgoing radiation must match. If this rule is broken, a penalty is given. Surface temperature has a major effect on material life, so this condition is especially important.

These several penalties are gathered into one. The data penalty, equation penalty, and boundary penalty are added. Each penalty is given a weight. Artificial intelligence learns to make this total penalty as small as possible. Then it produces predictions that are fast and physically consistent.

To include physical rules, derivative calculations are needed. We must measure how much temperature changes with position and time. There is a technology that calculates these derivatives automatically. It is called automatic differentiation. However, this calculation is heavy and reduces speed.

So several methods are used to make computation lighter. The calculation is divided to reduce overlapping parts. Coordinates are changed with smooth functions to make differentiation easier. These optimizations provide inference much faster than traditional calculation. However, the specific acceleration factor or error value differs by problem. Those values should be cited only from validated cases.

The value of this method appears at the design stage. If the nozzle shape changes, the temperature distribution changes. With traditional calculation, each change requires a long wait. An artificial intelligence operator immediately shows the new temperature. Then dozens of shapes can be compared in one day. Designers reach better answers more quickly.

It is also useful at test sites. During an engine test, predictions close to real time are needed. We must know in advance which location will reach a dangerous temperature. An artificial intelligence operator produces this prediction quickly. However, it must be continually corrected with test data to be trusted. Work continues to reduce the difference between prediction and measurement.

The limits of this method are also clear. If physical rules are included incorrectly, the answer becomes worse. Setting the weights between rules also requires much work. Validation is difficult because real ultrahigh-temperature data is scarce. So this technology is still at a stage where research and testing proceed together. It is used as a tool to reduce the number of tests, rather than to replace tests.

Summary

This chapter dealt with the hot outer surfaces of rockets and hypersonic vehicles. Materials in these areas must withstand heat, oxidation, force, and long service times all at once. Metals are not enough, so ceramic materials are needed. Carbon fiber composites and ultra-high-temperature ceramics are the leading candidates. Predicting the lifetime of these materials is the key to development.

First, we looked at how materials degrade. Carbon fiber composites degrade when oxygen enters along cracks. Silicon carbide gains or loses a protective film depending on temperature and oxygen pressure. Ultra-high-temperature ceramics lose their surface as oxide films evaporate and phases change. This damage process is the target of lifetime prediction. Damage usually begins quietly inside the material, so we must read the inside, not just the surface.

Next, we looked at how artificial intelligence handles this damage. Crystal grains are turned into points, and boundaries are turned into lines to build a graph. Time is then added to track how damage accumulates. Physical laws are built into learning so that even ultra-high-temperature regions with little data can be predicted. A representative rule is that damage does not reverse.

We then looked at inverse design. The desired performance is set first, and the structure is found backward from it. A diffusion model shapes a structure from blurred noise. Direction-invariant

properties are added so the result fits physics. Fiber angle and interface coating thickness are determined together. A diffusion model gives not one answer but many candidates. A person then selects the structures that are easier to make from among those candidates.

Finally, we looked at multiscale matching of heat flow. The world of atomic vibration must be connected to the world of the whole component. Green-Kubo molecular dynamics calculates thermal conduction in the small world. A deep operator neural network transfers that value to the component temperature. A multifidelity method links cheap data with expensive data. This matching makes it possible to predict component temperature quickly. This fast prediction is useful both in the design stage and at the test site.

All these tools have one thing in common. Artificial intelligence does not replace testing. Artificial intelligence narrows the candidates and speeds up prediction. The final confirmation is done by wind tunnels, arc jets, and combustion tests. Development becomes faster only when artificial intelligence and real testing move together.

Another common point is the link with physics. The artificial intelligence in this chapter does not learn from data alone. It brings the laws of thermodynamics and force balance into learning. That is why it can handle ultra-high-temperature regions where data is scarce. This is a wall that pure data learning has difficulty crossing. The meeting of physics and artificial intelligence is the central key of this chapter.

The methods in this chapter are connected to one another. Heat flow prediction gives the surface temperature. Surface temperature determines the rates of oxidation and creep. Oxidation and creep lead to lifetime prediction. Lifetime prediction becomes the goal of inverse design. They are linked as one chain.

Many tasks remain. Real data at ultra-high temperatures is still scarce. Claims about the performance of inverse design models need physical validation. The way physical rules are built in must also be refined further. As these tasks are solved, the design of next-generation thermal protection systems will move forward. A thermal protection system means the whole structure that protects a vehicle from its hot outer surface.

Korea is also entering this field. Nuri and the next-generation launch vehicle deal with hot components. Reusable rockets must use components many times. This makes lifetime prediction more important. These methods that use artificial intelligence and physics together form the foundation. The ability to read the small world of materials protects the safety of large-scale flight.

Related Supporting Materials

Binner, J., Porter, M., Baker, B., Zou, J., Venkatachalam, V., Rubio Diaz, V., D'Angio, A., Ramanujam, P., Zhang, T., & Murthy, T. S. R. C. (2020). Selection, processing, properties and applications of ultra-high temperature ceramic matrix composites. International Materials Reviews, 65(7), 389-444. https://doi.org/10.1080/09506608.2019.1652006

Fahrenholtz, W. G., & Hilmas, G. E. (2017). Ultra-high temperature ceramics: Materials for extreme environments. Scripta Materialia, 129, 94-99. https://doi.org/10.1016/j.scriptamat.2016.10.018

DiCarlo, J. A., Yun, H. M., Morscher, G. N., & Bhatt, R. T. (2004). SiC/SiC composites for 1200°C and above. NASA Technical Reports Server, 20040191405. https://ntrs.nasa.gov/citations/20040191405

NASA. (2008). Ceramic Matrix Composite (CMC) thermal protection systems and hot structures for hypersonic vehicles. NASA Technical Reports Server, 20080017096. https://ntrs.nasa.gov/citations/20080017096

Wu, T., Wang, Y., Qi, W., Luo, X., Luo, P., Gao, X., & Song, Y. (2026). Predictive constitutive modelling of oxidation-induced degradation in 2.5D woven C/SiC composites. Materials, 19(2), 307. https://doi.org/10.3390/ma19020307

Qu, N., Zhou, W., Zhang, W., Liu, Y., Zheng, L., Cao, D., Tan, M., Zhu, J., & Zhang, X. (2026). Research progress and prospects of ultra-high-temperature ceramics: Experimentation, multiscale simulation and data-driven design. Nanomaterials, 16(11), 693. https://doi.org/10.3390/nano16110693

Baier, L., Frieß, M., Hensch, N., & Leisner, V. (2024). Development of ultra-high temperature ceramic matrix composites for hypersonic applications via reactive melt infiltration and mechanical testing under high temperature. CEAS Space Journal, 17(4), 635-644. https://doi.org/10.1007/s12567-024-00562-y

Peters, A. B., Zhang, D., Chen, S., Ott, C., Oses, C., Curtarolo, S., McCue, I., Pollock, T. M., & Eswarappa Prameela, S. (2024). Materials design for hypersonics. Nature Communications, 15, 3328. https://doi.org/10.1038/s41467-024-46753-3

Bianco, G., Nisar, A., Zhang, C., Boesl, B., & Agarwal, A. (2022). Predicting oxidation damage in ultra high-temperature borides: A machine learning approach. Ceramics International, 48(20), 29763-29769. https://doi.org/10.1016/j.ceramint.2022.06.236

Meng, H., Liu, Y., Yu, H., Zhuang, L., & Chu, Y. (2025). Machine-learning-potential-driven prediction of high-entropy ceramics with ultra-high melting points. Cell Reports Physical Science, 6, 102449. https://doi.org/10.1016/j.xcrp.2025.102449

Liu, Y., Meng, H., Zhu, Z., Yu, H., Zhuang, L., & Chu, Y. (2024). Exploring mechanical and thermal properties of high-entropy ceramics via general machine learning potentials. arXiv, 2406.08243. Preprint, not yet peer reviewed. https://arxiv.org/abs/2406.08243

Thomas, A., Durmaz, A. R., Alam, M., Gumbsch, P., Sack, H., & Eberl, C. (2023). Materials fatigue prediction using graph neural networks on microstructure representations. Scientific Reports, 13, 12562. https://doi.org/10.1038/s41598-023-39400-2

Zheng, B., Wen, J., Choy, Y.-S., & Fei, C. (2026). Physics-constrained EBSD image inpainting via adversarial graph learning: Bridging crystallographic rules and multimodal deep learning. Expert Systems with Applications, 298, 129667. https://doi.org/10.1016/j.eswa.2025.129667

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045

Chen, L., Li, Y., Ma, Y., Gao, L., & Yu, L. (2025). Multiscale graph equivariant diffusion model for 3D molecule design. Science Advances, 11(16), eadv0778. https://doi.org/10.1126/sciadv.adv0778

Xu, P., Wu, Y., Zarei, A., Ahmed, S., Pilla, S., Li, G., & Luo, F. (2025). Physically constrained 3D diffusion for inverse design of fiber-reinforced polymer composite materials. Composites Part B: Engineering, 303, 112515. https://doi.org/10.1016/j.compositesb.2025.112515

Dai, B., Wang, H., & Gui, J. (2026). Property-informed diffusion-based text-to-microstructure generation. arXiv, 2606.08150. Preprint, not yet peer reviewed. https://arxiv.org/abs/2606.08150

Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. E. (2021). Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. Nature Machine Intelligence, 3, 218-229. https://doi.org/10.1038/s42256-021-00302-5

Lu, L., Pestourie, R., Johnson, S. G., & Romano, G. (2022). Multifidelity deep neural operators for efficient learning of partial differential equations with application to fast inverse design of nanoscale heat transport. Physical Review Research, 4, 023210. https://doi.org/10.1103/PhysRevResearch.4.023210

Green, M. S. (1954). Markoff random processes and the statistical mechanics of time-dependent phenomena. Journal of Chemical Physics, 22, 398-413. https://doi.org/10.1063/1.1740082

Kubo, R. (1957). Statistical-mechanical theory of irreversible processes. Journal of the Physical Society of Japan, 12, 570-586. https://doi.org/10.1143/JPSJ.12.570

Verdi, C., Karsai, F., Liu, P., Jinnouchi, R., & Kresse, G. (2021). Thermal transport and phase transitions of zirconia by on-the-fly machine-learned interatomic potentials. npj Computational Materials, 7, 156. https://doi.org/10.1038/s41524-021-00630-5

Zhang, X., Liu, N., Guo, Q., Shi, G., & Wang, Y. (2026). Beyond Boltzmann transport: Green-Kubo prediction of lattice thermal conductivity with machine-learned potentials. Journal of Chemical Physics, 164(7), 074101. https://doi.org/10.1063/5.0297462

Mandal, S., Srivastava, A., Das, T., Singh, A. K., & Maiti, P. K. (2025). Probing lattice anharmonicity and thermal transport in ultralow-κ materials using machine learning interatomic potentials. Small, 22(8), 2513476. https://doi.org/10.1002/smll.202513476

Arabha, S., Shokri Aghbolagh, Z., Ghorbani, K., Hatam-Lee, S. M., & Rajabpour, A. (2021). Recent advances in lattice thermal conductivity calculation using machine-learning interatomic potentials. Journal of Applied Physics, 130(21), 210903. https://doi.org/10.1063/5.0069443

Chapter 6 Thermal Stress Modeling and In-situ Defect Detection in Metal Additive Manufacturing (LPBF/DED) Processes

Introduction

Imagine baking a cake in a kitchen. You pour batter into a pan and put it in the oven. One solid cake comes out. This old manufacturing method makes things by pouring material into a mold all at once. Rocket parts were made this way for a long time. Engineers cut away from large metal blocks or made separate pieces and joined them.

But hundreds of very narrow cooling channels run inside a rocket engine combustion chamber and nozzle. Cold fuel must flow through those channels to protect the wall from hot flames. This kind of complex internal structure is hard to make by cutting. If pieces are joined, the joints become weak spots.

Metal additive manufacturing opens another path. This method makes parts by stacking metal one layer at a time. Just as a 3D printer lays down plastic line by line, it melts and stacks metal powder or metal wire. It can make complex hollow channels in a single process. Because there are no joints, weak spots are reduced.

This method gives freedom, but it also brings new concerns. Tiny pores can form while metal is melted and stacked layer by layer. The part may warp as each layer heats and cools. The shape of the crystals may differ by direction, so strength may lean to one side. These defects are hard to see. But in rocket parts, one small defect can lead to a major accident.

This chapter looks at how artificial intelligence helps with this problem. Artificial intelligence does three things. It quickly finds defects during manufacturing. It predicts thermal stress and deformation quickly in place of slow calculations. It helps change process conditions on the fly. If we follow these three tasks in order, we can see the outline of a factory that makes nearly defect-free rocket parts by itself.

One idea underlies the technologies in this chapter. Artificial intelligence alone is not enough, and physics alone is not enough. Artificial intelligence trained only on data is shaken by noise. Physics calculations alone are too slow. When the two are combined, each fills the other's weakness. This combination is the core of physics-informed machine learning (PIML), and the technologies in this chapter are examples of it.

6.1 Two Ways to Build Rocket Parts by Stacking Metal

What This Section Covers

There are two main ways to build rocket engine parts from metal. One method fires a laser onto a thin layer of metal powder. The other melts and stacks metal powder or wire as it is fed into the laser focus. The two methods differ in the size and precision of the parts they make.

This section also introduces three metals used in rockets. The three metals have different properties. The points that require care during manufacturing also differ. We need to know what happens inside metal as it melts and solidifies. Then we can understand why the artificial intelligence technologies discussed later are needed. This section lays that foundation by introducing the properties of the three metals.

6.1.1 Next-Generation Rocket Propulsion Systems and Metal Additive Manufacturing

Rocket engines work under extreme conditions. The pressure inside the combustion chamber reaches hundreds of times the pressure of a car tire. The temperature where fuel burns exceeds 3,000 degrees. By contrast, liquid fuel close to minus 200 degrees flows through the cooling channels. One part must withstand this heat and cold at the same time.

Making such parts by old methods takes a lot of work. Cooling channels are cut one by one, and several pieces are joined by a method called brazing. Brazing joins metals by melting another metal between them. As the number of pieces grows, the number of joints increases, and manufacturing can take months. Joints can later become places where leaks or cracks occur.

Metal additive manufacturing greatly reduces this process. Laser Powder Bed Fusion (LPBF) lays down a thin layer of metal powder. A laser is fired onto it. Only the needed areas are melted and fused. Then another thin layer of powder is laid down and melted again. When this process is repeated thousands of times, the part is completed. It suits small, precise parts such as combustion chamber liners with dense cooling channels.

Directed Energy Deposition (DED) works a little differently. Metal powder or metal wire is fed into the laser focus, melted, and stacked. It is closer to a picture of adding layers of metal like painting over a surface with a brush. It builds quickly and is suitable for large parts. This method is used to make nozzle skirts with diameters greater than 1 meter.

The two methods have different strengths and weaknesses. Laser Powder Bed Fusion is very precise but slow. This is because thin layers must be laid down thousands of times. The size of the part that can be made is also limited by the size of the chamber. Directed Energy Deposition, by contrast, is fast and can make large parts, but the surface is rough. More machining and finishing are needed after manufacturing.

Powder quality also determines the result. Powder particles must be evenly round so they can spread well in thin layers. If particle sizes are uneven, gaps form. If the powder absorbs moisture, bubbles form during manufacturing. For this reason, powder is stored and managed in a dry state. Good parts begin with good powder.

Both methods can make a part as one solid piece. Combustion chamber liners, injector heads, turbopump impellers, and large nozzles can be formed without joints. NASA's Rapid Analysis and Manufacturing Propulsion Technology (RAMPT) project has used this method to make large nozzles. According to public data, it built a cooled-channel nozzle about 1.5 meters in diameter and about 1.8 meters long in 90 days. This nozzle is 65 percent the size of a large engine nozzle for a space launch vehicle.

The same project also reports test results. In development tests combining several injectors, combustion chambers, and nozzles, the parts withstood long periods of combustion. Total burn time far exceeded 10,000 seconds, and ignition was repeated hundreds of times. This means parts made by stacking material layer by layer can withstand real rocket combustion. These results show that additive manufacturing is moving beyond the laboratory and into flight parts.

The benefits of this method lie in time and part count. Old methods made and joined hundreds of separate pieces. Additive manufacturing reduces those hundreds of pieces to a few. Fewer pieces mean fewer joints and shorter manufacturing time. When engineers want to change a design, they can make and test it again quickly. This speed raises the pace of new engine development.

6.1.2 Metallurgical Characteristics and Phase Transformation Behavior by Alloy

Metals used in rocket parts each have different missions. This section looks at three representative metals. They are Inconel 718, GRCop-42, and C-103. Metallurgy is the study of the internal structure of metals and how that structure changes. Even the same metal can develop a different internal structure depending on how it cools.

The three metals each hold a different place. Inconel 718 is used for strong parts that bear loads. GRCop-42 is used for combustion chamber walls that remove heat quickly. C-103 is used for nozzle tips that directly face the flame. The three metals divide the work inside one engine. That is why metallurgy for making each metal well is needed.

Inconel 718 is a nickel-based superalloy. A superalloy is an alloy designed to keep its strength even at high temperatures. This metal is widely used in turbopumps, valves, and high-pressure pipes. It is strong, welds well, and withstands heat well. For this reason, it is suitable for engine parts that bear force.

The problem appears during rapid cooling. Metal melted by a laser solidifies in the blink of an eye. At this time, elements such as niobium and molybdenum gather at the boundaries between crystals. This concentration is called microsegregation. If segregation becomes severe, a weak compound called the Laves phase forms there. This compound takes away elements that strengthen the metal and lowers its strength.

That is why Inconel 718 needs heat treatment after manufacturing. The part is heated again to a high temperature to dissolve and remove the weak compound. Then the temperature is lowered to induce the formation of very small and uniform strengthening particles. Only after this process does the part become resistant to fatigue. Fatigue is the phenomenon in which a material slowly cracks after receiving force many times.

These strengthening particles have a name. They are a very small phase made of nickel and niobium, called gamma double-prime. In chemical formula terms, three nickel atoms are combined with one niobium atom. This phase creates most of the strength of Inconel 718. It differs from gamma prime, which is common in other nickel superalloys. Controlling niobium well is therefore the key to this metal.

GRCop-42 is a copper alloy developed by NASA for rocket combustion chambers. Copper conducts heat very well. This property is needed to remove hot heat quickly from the combustion chamber wall. But pure copper softens when it becomes hot. So small amounts of chromium and niobium are added to copper. These two elements form very small particles that hold the copper in place.

There are difficulties when making GRCop-42 with a laser. Copper reflects light well. Like a mirror, it bounces away a large portion of laser light. Therefore, a powerful laser or a green laser is needed to melt it enough. After manufacturing, the part goes through a process that removes internal pores by pressing it under high pressure and temperature. After this process, it conducts heat well and keeps its strength even in hot environments above 800 degrees.

Let us note why the combustion chamber wall must remove heat so well. Flames hotter than 3,000 degrees touch one side of the wall. Cold fuel flows on the other side. This thin wall separates the two temperatures. If heat is not removed quickly, the wall melts or cracks. That is why a copper alloy that conducts heat well is used as the wall material.

This alloy works together with cooling channels. Narrow channels are made to pass densely through the wall. Fuel flows through those channels and carries heat away from the wall. The more complex the channels are, the more evenly heat is removed. Additive manufacturing has the power to make these complex channels all at once. The material and the manufacturing method support each other.

C-103 is a high-temperature alloy based on niobium. Niobium has a melting point far above 2,000 degrees. That is why it is used for nozzle tips and attitude control thrusters that directly face flames. Small amounts of hafnium, titanium, and zirconium are added to adjust its properties. When large thin-wall nozzles are made, the DED method is used.

C-103 requires care with air during manufacturing. Molten niobium quickly absorbs oxygen and nitrogen in the air. If oxygen exceeds a certain amount, the metal becomes easy to break. It changes into a material that snaps like glass. For this reason, the manufacturing chamber is filled with a shielding gas such as argon, and oxygen is kept very low.

Oxidation is also a major problem during use. Oxygen in hot exhaust gas keeps attacking niobium. So a silicide protective coating containing silicon is applied to the surface. This coating prevents oxygen from reaching the metal. Atmosphere control during manufacturing and a protective coating during use are both needed. These two controls determine the success or failure of C-103 parts.

Recent research digs into the process for making this metal. One 2025 study made C-103 thin walls by laser wire deposition and examined their properties. Another study drew a defect map using a melt pool model that solves several kinds of physics together. This means mapping in advance the conditions under which defects form. With such a map, good conditions can be selected and used. It finds a path to avoid defects before manufacturing.

6.2 In-situ Defect Detection Mechanisms and Physics-Informed Vision Models

What This Section Covers

In-situ means observing something at the very place where it is being made. If defects are found after the part is finished, it is already too late. If defects are found during manufacturing, the process can be corrected or stopped on the spot. This section covers ways to see defects during manufacturing.

First, we look at why and how defects form. Next, we introduce devices that observe the small puddle made by the laser. Finally, we see how artificial intelligence reads the observed images. Here, we encounter a method that uses physical laws and artificial intelligence together. This method is one example of physics-informed machine learning (PIML), which runs through this whole book.

6.2.1 Thermofluid Mechanisms of Melt Pool Defect Formation

When a laser melts metal powder, a small puddle forms. This puddle is called a melt pool. Its diameter is only about the width of a few strands of hair. As the laser passes, this puddle moves with it. The quality of the part depends greatly on the state of this small puddle.

Inside this puddle, the metal keeps moving. Surface tension differs between hotter and less hot areas. That difference pushes the molten metal back and forth. This flow is called Marangoni convection. It is similar to the way the surface swirls in a pot of broth. This swirl can create defects, and it can also remove them.

One defect is keyhole porosity. It forms when the laser is too strong or moves too slowly. The intense light digs deep into the metal and makes a narrow, deep pit like a keyhole. This pit is unstable and suddenly collapses. When it collapses, gas inside it becomes trapped in the metal and forms a round bubble. This trapped bubble is keyhole porosity.

A keyhole is not always bad. A keyhole with the right depth melts the metal deeply and bonds it well. The problem arises when the keyhole is too deep and unstable. So the process tries to keep the keyhole depth within a proper range. It looks for a suitable window where the laser power is neither too high nor too low. Finding this suitable window is the core of process control.

Another defect is lack of fusion (Lack-of-Fusion; LoF). Unlike a keyhole, it forms when the laser is weak or the spacing between scan tracks is too wide. Neighboring laser tracks or the layer below do not melt and bond properly. A gap remains between them. This gap has a sharp shape and contains unmelted powder inside.

The two defects differ in the nature of their risk. Keyhole pores are round, so they are relatively less dangerous. Lack of fusion is sharp, so force concentrates at its tip. It is the same principle as the tip of a sharp glass shard cracking easily. So lack of fusion can easily become the starting point of a fatigue crack. Rocket parts undergo countless vibrations, so this defect becomes a major risk.

Recent research divides defects into finer categories. One 2026 study divided holes into three types. No hole, holes smaller than 15 micrometers, and holes larger than that. This 15 micrometers was the classification boundary set by that study. In that study, large keyhole pores were more harmful. Small gas pores were less harmful under those conditions. The study trained this distinction using both photodiode signals and precise X-ray images.

However, this boundary must not be used directly as an acceptance standard. The risk of a hole is not determined by size alone. Its location, shape, applied load, material, and required service life all matter. Even a hole of the same size is more dangerous if it is near the surface. So classification boundaries and structural safety standards must be considered separately.

There is a reason for dividing defects into finer categories this way. If all holes are treated the same, even sound parts may be rejected as defective. Conversely, if they are all ignored, dangerous holes may be missed. The type and size of a defect must be identified to respond properly. Artificial intelligence takes on this fine distinction.

Defects leave visible early signs. Just before a keyhole forms, the shape of the melt pool fluctuates. A pattern appears in which the pool becomes deeper and narrower. The human eye cannot keep up with these rapid changes. Cameras that capture tens of thousands of frames per second and artificial intelligence read these early signs. The goal is to detect a defect before it forms.

Powder spatter is also tied to defects. Spatter is the phenomenon in which molten metal splashes when the laser strikes strongly. The splashed droplets fall onto powder that has not yet melted. When the next laser pass reaches that spot, it does not melt properly. So frequent spatter increases lack of fusion. The degree of spatter also becomes a signal that warns of defects in advance.

6.2.2 Eyes That Measure Melt Pool Temperature and a Physics Baseline

Special eyes are needed to see the melt pool. The pool is very bright and changes very quickly. Pyrometry is a method of measuring temperature from the light emitted by a hot object. It uses the principle that hotter objects emit different light. This method measures temperature without touching the object.

Coaxial pyrometry observes light from the same direction as the laser path. In effect, the laser and the camera share the same lens. So wherever the laser moves, the pool can always be viewed head-on. A high-speed camera that captures tens of thousands of frames per second records changes in the pool's brightness and shape. Comparing light at two wavelengths gives a temperature that is less affected by surface conditions.

This two-wavelength method is not perfect either. Emissivity means the degree to which an object emits light. This method assumes that the ratio of emissivity at the two wavelengths is constant. If this assumption fails, errors appear in the measured temperature. Splashed metal vapor or a dirty lens can also disturb the values. So the measured values must be calibrated regularly.

A camera alone is not enough. There must be a reference for comparing what is normal and what is abnormal. One such reference is the Rosenthal heat conduction model. This model is an old physics theory from 1946. It is a formula solved by hand for the temperature distribution created when a moving heat source heats metal.

The Rosenthal model is only a rough sketch. It treats the metal as a semi-infinite body and places the heat source at a single point. It assumes that the properties of the metal remain constant regardless of temperature. It also ignores the hidden heat that enters and leaves during melting and boiling. It cannot capture the flow of molten metal or evaporation.

So this formula provides only an approximate reference temperature. It cannot explain the rapid motion of a collapsing keyhole. When a more precise reference is needed, finite element analysis or fluid calculation is used. Cases that use Rosenthal values directly as a baseline for real-time monitoring are still rare. Here, it is given as an example of the idea of setting a physics-based reference.

Artificial intelligence can take these deviations as input. It learns the difference between the actual temperature field and the physics baseline to predict defects. When a physics baseline exists, the burden that artificial intelligence must learn becomes lighter. This is because physics serves as a guide even when data are limited. The physics baseline reduces the learning burden on artificial intelligence.

Why is real time so important? The laser moves at several meters per second. The time needed to build one layer is short. A location where a defect forms is soon covered by the next layer. It must be detected before it is covered so that the spot can be repaired. So decisions must be fast, on the scale of milliseconds.

Observation devices continue to improve. Recent research has introduced a method that captures two wavelengths at the same time with one camera. This method records melt pool temperature changes at up to 100,000 frames per second. The precise temperature images obtained this way become the basis for identifying holes in several ways. Such precise eyes are needed to identify holes properly.

6.2.3 Design Principles of Physics-Informed Image Models

Now it is time to read the observed images. This section proposes design principles for an image model that includes physics. First, it looks at a widely used image neural network. A Vision Transformer is a neural network that divides an image into small patches and examines their relationships. It cuts a photo like a patchwork quilt and sees how each patch connects to the others. In melt pool images, the bright center, the trailing tail, and the light from scattered powder are all different. The model reads these relationships to identify defects.

The strength of this neural network lies in connecting distant parts. The temperature at the front of the melt pool and the tail behind it are connected. How heat flows cannot be known by looking at only one patch. A Vision Transformer compares all patches at once. So it reads the heat flow across the entire pool. This wide field of view helps capture early signs of defects.

This book proposes adding the law of heat conduction to this model. It makes artificial intelligence check for itself whether its answer matches physics. A loss function is a score that measures how wrong artificial intelligence is. The idea is to include defect classification error, temperature field error, and physics residuals in this score. A residual is a value that shows how much an equation is violated.

Including physics in this way has advantages. Camera images always contain noise. Scattered powder can also block the light. A model that learns only from data is easily disturbed by this noise. A model that learns physics as well has the law of heat conduction as a support. Even when the image is slightly blurred, it can preserve a judgment that fits physics.

Actual studies also support the power of physics-based features. One study extracted physics-based features from infrared camera signals. With these features, it detected holes in Inconel 718 parts. This is because the pattern of temperature change itself contains clues about defects. Features drawn from physics are resistant to noise and have clear meaning. So they also make it easier to explain the basis for a judgment.

Speed is also important. To make judgments during manufacturing, the answer must come in the blink of an eye. A 2026 study combined a pretrained neural network with a traditional classifier. This method classified melt pool images of a nickel alloy with an accuracy of about 0.95. The time needed to judge one frame was 1.15 milliseconds. This study is still an early study that has not yet undergone peer review. Still, it shows that judgment fast enough for factory use is possible.

The way physics is included requires weight control. If the physics residual is pressed too strongly, the model learns less from the data. If it follows only the data too much, it departs from physics. A good model requires the right balance between these two forces. Finding this balance point is the tuning task of physics-informed learning. The right weight differs by part and equipment.

Here, we should separate what has been verified from what is proposed. Finding holes with infrared features and a fast hybrid model are results shown in actual studies. The structure that adds heat conduction residuals to a Vision Transformer is the direction drawn by this book. It is not yet a method verified on actual parts. Judgments made by artificial intelligence must be checked again with cut cross sections and computed tomography.

6.3 Reducing Distortion in Large Nozzles

A nozzle skirt is a large trumpet-shaped part attached to the rear of an engine. It is made by building a thin plate in a rounded cone shape. Large and thin parts like this easily distort during manufacturing. This section deals with the causes of that distortion and ways to prevent it.

First, it examines why forces build up inside a part. Then it introduces a fast artificial intelligence model that replaces slow calculations. Finally, it looks at how to reduce distortion by using that model to control laser movement. The larger the part, the harder this distortion control becomes, and the more important it is.

6.3.1 Thermal Stress Accumulation and Buckling Distortion During Large Thin-Wall Additive Manufacturing

Residual stress is the tensile and compressive force left inside a part. It is invisible, but it always acts inside. Let us picture how this force forms. When a laser heats one line of metal, that area expands. This is because the heated metal swells.

After the laser passes, that line cools again. As it cools, it tries to shrink back. But the layer below has already solidified and holds this movement back. Stress remains inside as the force trying to shrink and the force holding it back oppose each other. Tensile force builds up in the upper part, and compressive force builds up in the lower part.

In a thin and large nozzle skirt, this force causes serious trouble. It is similar to paper wrinkling and bending when it is pulled from only one side. The part may distort in the radial direction, or the thin wall may suddenly bend. This sudden bending is called buckling. In severe cases, the part may even separate from the base plate.

The larger the part, the harder it is to manage. This is because heat continues to accumulate when the build reaches hundreds of layers. The temperature state during the first few layers differs from that

during the later layers. Even with the same laser conditions, the result changes. So for large parts, conditions must be changed while monitoring the thermal state.

This problem is more difficult for niobium alloys such as C-103. Because their melting point is high, large heat input is needed to melt them. Large heat leads to large stress. When making large nozzles from this metal, distortion control determines success or failure. Thermal analysis and path design must be planned together from the beginning.

Distortion is hard to fix after manufacturing. If a solidified part is forcibly straightened, new cracks form. So it is better to prevent stress from building up during manufacturing. It is like preventing an illness in advance rather than treating it after it appears. This prevention is the role of artificial intelligence prediction and active control. The surrogate model discussed later helps with that prevention.

6.3.2 Thermodynamics-Based Surrogate Model

Calculations are needed to predict stress inside a part. Finite element analysis is a method that divides a part into small pieces and calculates temperature and force. It is very accurate, but it takes a long time for large parts. This is because millions of small pieces must be solved one by one. A single analysis of a large nozzle can take several days.

A surrogate model replaces this slow calculation. A surrogate model is a fast model that learns the results of slow calculations in advance. A Multilayer Perceptron (MLP) is often used. It is a basic neural network that converts inputs into outputs through several layers. After learning many precise analysis results, it estimates the results for new conditions.

What does this model receive, and what does it produce? The inputs are laser power, travel speed, the path followed by the laser, and the waiting time between layers. The outputs are the residual stress distribution across the whole part and the final deformation shape. The model estimates in a short time what would take days with high-fidelity analysis. This makes it possible to compare many conditions quickly.

Speed opens a new path. If we must wait several days every time one condition changes, it is hard to find a good design. Even a rough estimate is useful if it is fast, because thousands of conditions can be screened in advance. Only good candidates are then checked again with high-fidelity analysis. A surrogate model acts like a net that narrows a wide field of candidates.

For a surrogate model to learn well, it needs a large amount of good data. High-fidelity analysis is run in advance under many conditions to collect data. Measurements from actual manufactured parts are added as well. The broader and more balanced the data, the more accurate the estimate becomes. Under unfamiliar conditions the model has never learned, the estimate becomes unstable. So a surrogate model should be trusted only within the range it has learned.

Recent studies have made progress in this direction. A 2025 study used machine learning to estimate residual stress quickly and reduce warping. The study reported that bending in a bridge-shaped test specimen was greatly reduced. The difference between the predicted stress and high-fidelity analysis was also small. This means the fast model came close to high-fidelity analysis.

Adding physical information improves it further. Another 2025 study put physical laws into a neural network and predicted warping in large parts almost in real time. It used a structure called a Neural Operator so that the model could learn across many conditions. This study is still an early study that has not yet gone through peer review. Even so, it shows a path to obtaining fast prediction and physical consistency together.

6.3.3 Active Optimization of Scan Paths and Interlayer Cooling Time

Now it is time to change the actual process based on prediction. A scan path is the direction and order in which the laser moves. If the laser keeps passing in only one direction, heat gathers on one side. Concentrated heat leaves large stress. So the direction is changed to spread heat evenly.

Many smart paths have been developed. One method divides a wide surface into small regions and fills them in alternating order. Another method slightly rotates the laser direction in each region. Curved paths that densely fill space are also used. The goal is one thing. It is to scatter heat flow evenly in all directions and prevent one-sided concentration.

The waiting time between layers is also important. If the next layer is built while the previous layer is still hot, heat piles up layer by layer. Stacked heat causes large warping. So the process waits until the previous layer cools to some extent. This waiting time is called interlayer cooling time.

Active optimization is a method that changes these conditions during the process. Pyrometry tells how hot the part is now. The surrogate model quickly predicts the stress risk of the next path. Based on these two inputs, laser power, speed, and waiting time are adjusted at each moment. It is a process of turning long human experience into numbers and rules.

The control methods are becoming smarter. Bayesian Optimization is a method for finding good conditions with few trials. It chooses the next condition intelligently based on results already tested. A Genetic Algorithm is a method that mixes good conditions to find better ones. These methods adjust waiting time to match the thermal state of the previous layer.

This control is done differently for each layer. Lower layers cool quickly because the base plate takes heat away. Heat may escape less easily in upper layers. In that case, the waiting time is increased in upper layers. But this is only one example. The degree of heat buildup depends on the part shape, the base plate, the layer area, and the build order. So sensors are used to observe the thermal state of each layer and set the waiting time.

If this method works well, warping is visibly reduced. Stress spreads more evenly, and peak stress falls. The part stays within the allowable tolerance. However, this control must always stay within a safe process range. Sensor delay, model error, and the response limits of equipment must be handled together. It can be applied to real parts only after it is first checked with test specimens.

The idea of a digital twin grows from this point. A digital twin means placing a twin of the real part inside a computer. Sensor values from the real process are continuously fed into the twin. The twin reflects what is happening inside the part now. It also shows in advance what may be risky in the next layer. The surrogate model is the brain that helps this twin think quickly.

6.4 Controlling the Shape of Crystals

When metal solidifies, crystals grow inside it. The shape of these crystals determines the strength of the part. If crystals grow long in only one direction, strength differs by direction. This section deals with ways to control crystal shape as desired.

First, we examine two crystal shapes and their differences. Next, we introduce a generative model that draws crystal shapes from process conditions. Finally, we look at reinforcement learning control that regulates crystal shape during manufacturing. The invisible world of crystals decides the fate of the part.

6.4.1 Columnar-to-Equiaxed Transition (CET) and Mechanical Anisotropy

When metal solidifies, crystals grow in two shapes. A columnar grain is a crystal that grows long in one direction like a pillar. It is close to the image of several wooden posts standing upright. An equiaxed grain is a crystal that grows to a similar size in all directions. It is close to the image of gravel spread evenly.

This difference divides the properties of a part. If there are many columnar grains, the part is strong along the column direction and weak sideways. Strength and fatigue life differ by direction. This property is called anisotropy. It is the same principle as wood that splits easily along the grain but not easily across it.

Anisotropy can be a concern in rocket parts. This is because the parts receive forces from many directions. If only one direction is strong, the weak direction cracks first. Equiaxed grains reduce this problem. They are strong evenly in all directions and resist cracking during solidification. So equiaxed grains are advantageous in parts that receive forces from many directions.

However, equiaxed grains are not always better. In parts where creep or thermal fatigue is a problem, columnar crystals can sometimes be better. Creep is a phenomenon in which a material slowly stretches under force over a long time. In parts where heat flows in only one direction, directional crystals can be advantageous. The desired crystal shape changes depending on the forces the part receives.

In additive manufacturing, columnar grains form easily. This is because heat escapes in the build direction of the laser. Crystals grow like pillars along the direction in which heat escapes. As a result, crystals line up in the build direction. A group of crystals whose directions are biased to one side in this way is called texture. The goal of this section is to change this bias as needed.

Crystal shape is strongly controlled by two values. The temperature gradient (G) means how steep the temperature is. It shows how quickly temperature changes with position. The solidification rate (R) is the speed at which the boundary between liquid and solid moves. The two values do different jobs.

The ratio G divided by R determines the shape of crystals. If this ratio is large, the shape tends toward columnar. If it is small, the shape tends toward equiaxed. The product G multiplied by R becomes the cooling rate. The larger this value, the finer the crystals become. So shape and size can be controlled separately. When laser power and speed change, these two values change together.

These two values are not the only factors that determine shape. Small particles that become seeds, the flow of molten metal, and the degree of remelting also have effects. The direction of crystals already grown below is also inherited. So equiaxed grains tend to form when the temperature gradient is low and solidification is fast, but this is not always the case. Many values must be considered together to control crystal shape properly.

There is a point where columnar grains change into equiaxed grains. This change is called the Columnar-to-Equiaxed Transition (CET). When the temperature gradient and solidification rate pass certain conditions, this change occurs. If we know these conditions, we can move closer to the desired crystal shape. Older theories have drawn this boundary with equations.

Recent studies refine this boundary more precisely. A 2025 study established a criterion for predicting CET from alloy composition. It used machine learning to reveal which elements strongly affect crystal shape. This means that crystal shape can be changed just by slightly changing composition. If process and composition are handled together, the range of crystal control becomes wider.

Machine learning that examines composition is also continuing. It uses data to identify which elements strongly affect crystal shape. Some results show that the amounts of elements such as silicon or carbon change crystal shape. These studies gradually fill in the map of crystal control.

6.4.2 Generative Models That Draw Crystal Shapes from Process Conditions

We want to see crystal shape in advance from process conditions. To do that, we need a tool that draws a microstructure image when conditions are entered. A Conditional GAN is one such tool. This neural network is a generative model that creates a desired image from conditions. Here, the conditions are values such as laser power, speed, temperature gradient, and solidification rate. Using this model for microstructure control is still at the proposal stage.

This model learns through competition between two parts. The Generator creates microstructure images. The Discriminator judges whether the image is real or fake. The generator tries to fool the discriminator, and the discriminator tries not to be fooled. If this competition is repeated, the generator gradually draws images that look real. It is close to the image of a counterfeiter and an appraiser improving each other's skills.

Here, the real images are electron backscatter diffraction (EBSD) images. This method reads crystal orientation with an electron microscope. It colors the orientation of each crystal and shows it like a map. The generator learns to imitate this map. When conditions are entered, it draws the crystal map that would appear under those conditions.

If this model learns well, it is useful. A designer can see in advance how crystals will change when process conditions change. The designer can estimate this without actually making and cutting a part. This saves a great deal of time and material. Many conditions can be compared on the screen, and the better option can be chosen.

Actually obtaining crystal maps takes a great deal of work. The part must be cut, and the surface must be polished smoothly. Crystal orientation must be read point by point with an electron microscope. Reading a wide surface takes a long time. A generative model becomes a shortcut that greatly shortens this slow process. Of course, a shortcut must sometimes be checked again against real measurements.

There are also cautions. Generative models are good at making plausible images. But plausibility is not the same as physical truth. Even if an image looks good, it may differ from reality. So physical indicators such as grain size, orientation distribution, and phase fraction must be checked together. The image can be trusted only when these indicators match reality.

Trust increases when physics is included in the loss function. If the crystal size in the generated image departs from the real distribution, a penalty is given. This guides the generator to draw images that fit physics. It narrows the gap between a pretty image and a correct image. A generative model can be used for rocket parts only when it is measured against the ruler of physics.

6.4.3 Reinforcement Learning That Controls Crystal Shape During Manufacturing

Now it is time to control crystal shape during manufacturing on its own. Reinforcement Learning is a method that learns good actions through trials and rewards. It is similar to a baby learning to walk while falling. A good action receives a reward, and a bad action receives a penalty. Actions that receive more rewards gradually increase.

In additive manufacturing, the action is laser control. The action is how much to raise or lower laser power and how much to change speed. However, crystal shape cannot be seen directly during manufacturing. So the temperature and shape of the melt pool are observed instead. Crystal shape is inferred from these values. Pyrometry provides this state in real time.

The reward is set by how close the result is to the target. A larger reward is given when the crystal shape is closer to the desired shape. The reward grows when the equiaxed grain fraction reaches the target. In contrast, a penalty is given if keyholing or lack of fusion occurs. Artificial intelligence learns laser

control by following this reward.

How the reward is designed determines success or failure. If the reward is too crude, the agent wanders down the wrong path. It may waste energy while chasing only the crystal shape. So the reward is designed to include crystal shape, defects, and energy together. This means placing several goals on the scale and balancing them. Designing this scale well is the core of reinforcement learning design.

Why is this real-time control needed? Because the shape of a part differs from place to place. Corners, thin walls, and protruding regions lose heat at different rates. Even with the same laser conditions, the crystal shape changes by location. Real-time control tries to keep the crystal shape uniform by changing conditions from place to place.

This direction is not yet an industrial standard. Many proposals remain at the simulation and test coupon stage. Cases where crystal shape is controlled in a closed loop in real time are rare. This is because crystal shape is hard to measure directly during the process. Sensor delay, model error, and equipment response limits must also be overcome.

Reinforcement learning is strong at experiencing risky attempts in advance. If bad conditions are tested on real equipment, the part is wasted. In a computer simulation, failure can happen freely. The system learns good control in advance from countless failures. After learning, it is carefully transferred to real equipment within a safe range. In effect, the computer pays the cost of failure instead.

That is why the order matters. First, control is learned in simulation. Then it is checked on limited test coupons. It is used as an auxiliary tool within a validated process window. Only by following these steps can reinforcement learning control be carefully introduced into real parts. After fabrication, the crystal shape is checked again with EBSD to see whether the control was correct.

6.5 Listening to the Inside Through Sound for Defect Monitoring

So far, defects have been found with light. But light sees only the surface. Cracks that form deep inside a part are hard to capture with a camera. This section discusses a method for listening to the inside through sound.

First, we examine what kind of sound defects make. Next, we look at how to turn that sound into an image. Finally, we look at how to judge defects better by using sound and light together. When different senses are combined, one sense fills in what another misses.

6.5.1 Physical Mechanisms of Broadband Acoustic Emission (AE) Signals

Acoustic emission (AE) is sound released when a sudden change occurs inside a material. More precisely, it is a very small elastic wave. It follows the same principle as the sharp sound made when a glass cup cracks. It is also like the cracking sound made when wood bends and breaks. Events inside a material leak out as sound.

In additive manufacturing, different events produce different sounds. Hot cracking releases a large amount of energy in an instant. So it produces a sharp, high-frequency sound. It is close to a short snapping burst. This kind of sound has large amplitude and rises suddenly.

Sound is also produced when a keyhole collapses. A hollow space inside the melt pool suddenly caves in and creates an impact. This sound oscillates repeatedly at a frequency somewhat lower than that of cracking. It is similar to the disturbance made when a bubble collapses in water. The texture of the sound differs from that of a crack.

Sound is also produced when powder spatters or equipment rubs against something. This sound lies at a lower frequency. Normal noise that is not a defect also mixes into the sound. So the sound must be separated carefully. The key is to distinguish the sound of a defect from ordinary noise.

There is a reason each sound has a different texture. When large energy is released suddenly, high frequencies are carried. When it is released slowly, low frequencies are carried. A crack opens in an instant, so it produces a high-pitched sound. Powder spatter is relatively slow, so it produces a low-pitched sound. However, the frequency actually captured depends on the sensor, the transmission path, and the filtering method. So this order should be viewed as a tendency observed in specific tests.

Listening through sound has a major strength. Light sees only the surface, but sound passes through the inside. Even cracks deep in an additively manufactured body can be captured through sound. If the first moment when a crack forms is detected, the process can respond quickly. The process can be stopped on the spot, and the region can be marked for reinspection. Repair by firing the laser again to remelt the area may also be considered. But remelting does not always remove cracks. It may also create new thermal stress and new cracks. A validated repair process and simply melting again must be treated differently.

There are two ways to listen to sound. One is to use a piezoelectric sensor that touches the part. A piezoelectric sensor is an element that produces electricity when it receives force. It is attached to the plate supporting the part and listens to elastic waves that have passed through the inside. The other is to listen to sound in the air with a microphone. The two methods differ in the paths the sound has traveled and in their frequencies.

6.5.2 Time-Frequency Wavelet Transform

It is hard to judge a sound signal by looking at it as it is. Vibration over time alone makes it hard to know what defect is present. This is because sounds of several pitches overlap at the same moment. We need to see both when a sound occurred and what pitch it had. Only then can we read the fingerprint of a defect.

The Wavelet Transform does this work. This method plots time and frequency together. It turns a sound signal into an image whose horizontal axis is time and whose vertical axis is pitch. This image is called a scalogram. It resembles a musical score, which shows time and pitch together.

This image draws a different pattern for each defect. The sharp sound of a crack appears as a bright spot at a high position. The fluctuation of a keyhole appears as a connected pattern in the middle. Noise lies faintly at low positions. Once sound becomes an image in this way, it becomes easier to compare by eye.

When reading patterns, the mother wavelet must be chosen well. A mother wavelet is a short waveform that serves as a yardstick for comparing sound. A yardstick similar to the texture of the sound makes the pattern clearer. If a yardstick with a different texture is used, the pattern becomes blurred. Choosing the right yardstick is also part of preprocessing. The right yardstick must be chosen to obtain an image with clear patterns.

Artificial intelligence learns patterns from this image. A neural network that reads photographs reads the scalogram. It learns from many examples which pattern corresponds to which defect. After learning, it looks at a new image and identifies the defect. By turning sound into an image, we can borrow the power of vision-based artificial intelligence.

Recent studies have achieved good results with this method. One study estimated keyhole pores using only airborne sound and laser speed. It learned by converting microphone-captured sound into scalograms. A value indicating the reliability of the result exceeded 0.8. This was a result published in an

academic journal. It showed that internal defects can be estimated on the millisecond scale using sound alone.

Listening to sound also has the advantage of low cost. High-speed cameras and precise X-ray systems are expensive and hard to handle. Piezoelectric sensors are relatively inexpensive and easy to attach. If several are attached, the source location of the sound can also be estimated. For this reason, sound-based monitoring is gaining attention as a practical path in the field.

The principle for locating a source with several sensors is simple. The same sound arrives at each sensor at a different time. It reaches a nearby sensor first and a distant sensor later. By comparing these differences in arrival time, the place where the sound occurred can be estimated. This is like finding direction from the time differences at several microphones. The location found in this way tells us where the defect is.

6.5.3 Defect Recognition Using Light and Sound Together

Multimodal means using several kinds of data together. It is like a person using eyes and ears together. Some things are missed when only sound is heard, and some things are missed when only light is seen. When the two are combined, they fill each other's gaps. This section discusses how to combine sound and light.

Light and sound see different things. Pyrometry shows the surface temperature well. Acoustic emission captures sudden changes that occur inside. A camera detects abnormal surface temperature first. Sound detects internal cracks first. When the two signals are combined, the surface and the inside are examined together.

The method for combining them is as follows. A sound-specific neural network reads the sound signal. An image-specific neural network reads the light image. The features extracted by the two neural networks are gathered together to make the final judgment. Each is assigned the task it does best, and they join forces at the conclusion.

Actual research has shown the power of this combination. One study collected pyrometry and sound together. By combining the two signals, it identified keyhole pores more effectively. Fewer pores were missed than when only one signal was used. Capturing cracks together in this way remains a future task.

Above all, time alignment is important when combining signals. The moment when the sound occurred and the moment when the light changed must be matched accurately. If the timing is shifted, different events may be seen as the same event. So the two sensors are synchronized very precisely. Because each sensor has different noise, preprocessing is also needed in advance.

There are several ways to combine the two signals. One method mixes the signals from the beginning and learns them together. Another method lets each learn separately and combines them only at the conclusion. Recent studies often use a method that combines separately learned features with a lightweight classifier. This method is computationally light and adapts well to new defects. Which method is best depends on the part and the equipment.

The trend toward combining several senses will expand further. In addition to light and sound, reflected laser signals are also used. Photographs of the fabricated layer and temperature maps are collected together. Many eyes and ears examine one part in overlapping layers. The more senses there are, the fewer defects are missed. It is the role of artificial intelligence to weave these many signals into one.

This method also has clear limits. It is safer to use artificial intelligence judgments for alarms during production and for selecting candidates for inspection. True pass-or-fail decisions are made by

nondestructive testing, cross-sectional observation, and computed tomography. Timing mismatch and noise between sensors can also cause wrong judgments. Artificial intelligence is a guide that narrows down where people should look.

Summary

This chapter examined technologies for building rocket parts with metal. We looked at methods that melt and stack powder with a laser and methods that melt and stack wire. We reviewed the properties of three metals, Inconel 718, GRCop-42, and C-103, and the points that require care. Each metal has a different mission, and each has different problems to control during fabrication.

Additive manufacturing gives freedom, but it also brings the risk of defects. Keyhole pores and lack of fusion become starting points for fatigue cracks. Residual stress distorts large parts. Directional bias in crystals makes strength lean to one side. These defects are directly tied to engine safety.

These defects are intertwined. One wrong laser condition can bring several defects at once. Trying to avoid keyholes may cause lack of fusion. Trying to reduce stress may worsen the crystal shape. So conditions must not be changed by looking at only one defect. Several defects must be placed on the scale together to find a balance.

Artificial intelligence helps this problem in three directions. The first is to find defects quickly during fabrication. A model that adds physical laws to pyrometry images identifies pores. A model that turns sound into images and reads them captures internal cracks. Combining light and sound examines the surface and the inside together.

The next task is to predict thermal stress and deformation quickly in place of slow calculations. A surrogate model shortens analysis that would take days. This fast prediction screens countless process conditions in advance. When physics is included, the prediction does not depart from physical behavior.

The last task is to help change process conditions as fabrication proceeds. Scan paths and cooling time are adjusted to reduce distortion. The laser is adjusted to keep the crystal shape uniform. A generative model previews the crystal shape from conditions. Reinforcement learning controls conditions on its own during fabrication.

One point must be made clear. Artificial intelligence cannot replace all inspections. The models discussed in this chapter are fast and smart, but they are not perfect. Much of the work is still early research, and little has hardened into industrial standards. It is safer to use artificial intelligence judgments for alarms and candidate selection. Final quality assurance is carried together by test coupons, nondestructive testing, and combustion tests.

There is also a new trend using language models. In additive manufacturing, papers and reports have piled up like mountains. A Large Language Model (LLM) reads these texts and organizes knowledge. We can ask under what conditions certain defects tend to form. One 2025 study attempted to predict defect conditions using a language model. Such tools become a bridge between human experience and the literature.

The big picture is this. Sensors become the eyes and ears, surrogate models become the brain, and controllers become the hands. When these three are connected as one, a factory can make parts while inspecting them on its own. NASA's success in building a large nozzle and having it withstand a long combustion test is the starting point. The next chapter examines how to finish parts made in this way and how to estimate whether those parts will last for a long time.

Related Supporting Materials

DebRoy, T., Wei, H. L., Zuback, J. S., Mukherjee, T., Elmer, J. W., Milewski, J. O., Beese, A. M., Wilson-Heid, A., De, A., & Zhang, W. (2018). Additive manufacturing of metallic components: Process, structure and properties. Progress in Materials Science, 92, 112-224. https://doi.org/10.1016/j.pmatsci.2017.10.001

King, W. E., Barth, H. D., Castillo, V. M., Gallegos, G. F., Gibbs, J. W., Hahn, D. E., Kamath, C., & Rubenchik, A. M. (2014). Observation of keyhole-mode laser melting in laser powder-bed fusion additive manufacturing. Journal of Materials Processing Technology, 214, 2915-2925. https://doi.org/10.1016/j.jmatprotec.2014.06.005

Rosenthal, D. (1946). The theory of moving sources of heat and its application to metal treatments. Transactions of the ASME, 68, 849-866.

Hunt, J. D. (1984). Steady state columnar and equiaxed growth of dendrites and eutectic. Materials Science and Engineering, 65, 75-83. https://doi.org/10.1016/0025-5416(84)90201-5

Mukherjee, T., Manvatkar, V., De, A., & DebRoy, T. (2017). Mitigation of thermal distortion during additive manufacturing. Scripta Materialia, 127, 79-83. https://doi.org/10.1016/j.scriptamat.2016.09.001

Gradl, P. R., Protz, C. S., Cooper, K., Ellis, D., Evans, L. J., & Garcia, C. (2019). GRCop-42 Development and Hot-fire Testing Using Additive Manufacturing Powder Bed Fusion for Channel-cooled Combustion Chambers. AIAA Propulsion and Energy 2019 Forum. https://doi.org/10.2514/6.2019-4228

Gradl, P., Mireles, O. R., Katsarelis, C., Smith, T. M., Sowards, J., Park, A., et al. (2023). Advancement of extreme environment additively manufactured alloys for next generation space propulsion applications. Acta Astronautica, 211, 483-497. https://doi.org/10.1016/j.actaastro.2023.06.035

NASA. (2023). Rapid Analysis and Manufacturing Propulsion Technology (RAMPT). https://www.nasa.gov/rapid-analysis-and-manufacturing-propulsion-technology/

NASA. (2023). GRCop-42 additive manufacturing for high heat flux thrust chamber liners (Document ID 20230007103). NASA Technical Reports Server. https://ntrs.nasa.gov/citations/20230007103

NIST. (2024). Additive Manufacturing Benchmark Test Series (AM Bench). National Institute of Standards and Technology. https://www.nist.gov/ambench

NIST. (2022). AM Bench 2022 measurement results data: in-situ thermography and scan strategy laser scans. National Institute of Standards and Technology. https://www.nist.gov/data-publications/am-bench-2022-measurement-results-data-situ-thermography-and-scan-strategy-laser

Levine, L. E., Lane, B. M., Heigel, J. C., et al. (2024). Outcomes and conclusions from the 2022 AM Bench measurements, challenge problems, and conference. Integrating Materials and Manufacturing Innovation, 13, 154-193. https://doi.org/10.1007/s40192-024-00372-4

Aydogan, B., & Chou, K. (2024). Review of in situ detection and ex situ characterization of porosity in laser powder bed fusion metal additive manufacturing. Metals, 14(6), 669. https://doi.org/10.3390/met14060669

Chen, L., Bi, G., Yao, X., Su, J., Tan, C., Feng, W., Benakis, M., Chew, Y., & Moon, S. K. (2024). In-situ process monitoring and adaptive quality enhancement in laser additive manufacturing: a critical review. arXiv:2404.13673. https://arxiv.org/abs/2404.13673

Emmanuel, I., Yang, Z., Yeung, H., & Zhang, X. (2026). Machine learning modeling for real-time melt pool monitoring in laser powder bed fusion additive manufacturing: a hybrid approach. arXiv:2606.23851. https://arxiv.org/abs/2606.23851

Liu, H., Guirguis, D., Zeng, X., Maurer, L., Rajan, V., Sanaei, N., Yang, C.-T., Beuth, J. L., Rollett, A. D., & Kara, L. B. (2026). Airborne acoustic emission enables sub-scanline keyhole porosity quantification and effective process characterization for metallic laser powder bed fusion. Additive Manufacturing, 116, 105062. https://doi.org/10.1016/j.addma.2025.105062

Tian, M., Mu, H., Ding, D., Li, M., Ding, Y., & Zhao, J. (2025). Real-time distortion prediction in metallic additive manufacturing via a physics-informed neural operator approach. arXiv:2511.13178. https://arxiv.org/abs/2511.13178

Zhang, R., Strickland, J., Hou, X., Yang, F., Li, X., de Oliveira, J. A., Li, J., & Zhang, S. (2025). Rapid residual stress simulation and distortion mitigation in laser additive manufacturing through machine learning. Additive Manufacturing, 102, 104721. https://doi.org/10.1016/j.addma.2025.104721

Bostan, B., Hinnebusch, S., Anderson, D., & To, A. C. (2025). Accurate detection of local porosity in laser powder bed fusion through deep learning of physics-based in-situ infrared camera signatures. Additive Manufacturing, 101, 104701. https://doi.org/10.1016/j.addma.2025.104701

Giri, S., Liu, S., Gorgannejad, S., Thampy, V., Quan, P., Nicolino, J. W., Strantza, M., Forien, J.-B., Martin, A. A., Calta, N. P., & Tassone, C. J. (2026). In-situ sensor monitoring of multi-class gas porosity formation in laser powder bed fusion using convolutional neural network. Journal of Intelligent Manufacturing. https://doi.org/10.1007/s10845-026-02861-z

Zhang, H., Caputo, A. N., Neu, R. W., Vallabh, C. K. P., To, A. C., Alam, M. J., & Zhao, X. (2026). Process-agnostic multi-metric nondestructive characterization of porosity in additive manufacturing via specialized single-camera two-wavelength pyrometry. Additive Manufacturing, 127, 105282. https://doi.org/10.1016/j.addma.2026.105282

Tempelman, J. R., Wachtor, A. J., Flynn, E. B., Depond, P. J., Forien, J.-B., Guss, G. M., Calta, N. P., & Matthews, M. J. (2022). Sensor fusion of pyrometry and acoustic measurements for localized keyhole pore identification in laser powder bed fusion. Journal of Materials Processing Technology, 308, 117656. https://doi.org/10.1016/j.jmatprotec.2022.117656

Li, Z., Zheng, H., Zhang, Z., Wang, J., Wen, G., Grasso, M., & Colosimo, B. M. (2026). On the use of acoustic emission methods for in-situ monitoring in metal additive manufacturing: a review study. Additive Manufacturing Frontiers, 5(2), 200347. https://doi.org/10.1016/j.amf.2026.200347

Brooke, R., Zhang, D., Qiu, D., Gibson, M. A., & Easton, M. (2025). Compositional criteria to predict columnar to equiaxed transitions in metal additive manufacturing. Nature Communications, 16, 5710. https://doi.org/10.1038/s41467-025-60162-0

Prabhune, B., Panicker, N., Jordan, B., Lim, B., Snyder, K., Delchini, M., See, N. D., Nag, S., Plotnikov, Y., & Lee, Y. (2025). Multi-physics melt pool modeling and process optimization for laser direct energy deposition of Nb-based refractory C103. Journal of Manufacturing Processes, 154, 444-461. https://doi.org/10.1016/j.jmapro.2025.09.082

Jeong, W., Cho, S., & Ryu, H. J. (2025). Optimization and characterization of C103 Nb alloy fabricated by laser wire directed energy deposition. International Journal of Refractory Metals and Hard Materials, 133, 107318. https://doi.org/10.1016/j.ijrmhm.2025.107318

Xu, Z. Y., Mao, J. Y., Ren, R. X., Wu, Z. Q., Lei, Y. J., Han, L. L., Chen, C., Fan, W., & Wang, Y. (2026). Research progress and challenges of intelligent in situ monitoring for laser additive manufacturing processes based on machine learning. Advances in Mechanical Engineering. https://doi.org/10.1177/16878132261442327

Pak, P., & Barati Farimani, A. (2025). AdditiveLLM: Large Language Models Predict Defects in Metals Additive Manufacturing. arXiv:2501.17784. https://arxiv.org/abs/2501.17784

Chapter 7 Post-Processing and High-Cycle Fatigue Life Assurance of 3D-Printed Aerospace Alloys

Introduction

Think of a 3D printer used at home. It melts plastic filament and builds an object one layer at a time. Rocket engine parts are made in a similar way. The material, however, is not plastic but metal powder. A laser melts the metal powder in an instant and solidifies it layer by layer. In this way, complex shapes can be printed as one piece.

This method has greatly changed the making of rocket parts. In the past, several pieces were made separately and joined by welding. Now even parts with cooling channels inside can be made at once. The number of parts falls, and the weight becomes lighter. As the process of joining many pieces disappears, the production time also becomes shorter.

But there is one problem. Metal fresh from the printer cannot be mounted on a rocket as it is. Its surface is rough, and small pores remain inside. Invisible stress is also trapped inside. If the part spins fast in this state, it may break.

That is why finishing after printing matters. This finishing is called post-processing. The surface is cut and smoothed. Internal pores are removed with heat and pressure. Heat treatment makes the structure uniform. This chapter covers the post-processing steps in order, from surface finishing to removing internal defects.

Think of cutting tough meat in a kitchen. A knife passes easily through soft tofu. It does not enter tough meat easily. Superalloys for rockets are like tough meat. They are very hard to cut. This difficult-to-cut property is called poor machinability.

There is one more major concern. It is fatigue. Fatigue is a phenomenon in which a material breaks after a small force is repeated many times. If a paper clip is bent and straightened several times, it snaps. It snaps even though no large force was applied at once. The rotating blades of a rocket turbopump face this risk.

This chapter looks at how artificial intelligence helps with two tasks. The first is finding conditions for cutting hard-to-machine metals well. The second is knowing in advance how many repeated loads a part can withstand. Both tasks take a long time if they rely only on human experience. Artificial intelligence learns from data and narrows the answer quickly.

However, artificial intelligence does not decide everything. Rocket parts involve human lives and large costs. Predictions by artificial intelligence must be checked again through experiments and inspections. This chapter explains how artificial intelligence and experiments work together while checking each other.

After reading this chapter, two points become clear. One is why 3D-printed metal needs finishing. The other is how artificial intelligence helps with that finishing and inspection. Difficult equations are not used. Each idea is explained with examples from everyday life. If you follow slowly, the overall picture of this chapter will not be hard to grasp.

7.1 Rocket Parts Moving to 3D Printing and New Challenges

First, let us unpack what this section covers. The heart of a rocket engine is an extremely harsh place. Cryogenic fuel below minus 200 degrees Celsius flows on one side. Hot gas above several hundred

degrees passes on the other side. The pressure is high, and the vibration is strong. This section looks at why metals used in such places are made by 3D printing, and what weaknesses appear in the process.

Rocket engine parts must be light and strong. If they are heavy, the payload that can be carried is reduced. If they are weak, they cannot withstand high pressure. These two requirements conflict with each other. This is because making a part lighter often makes it weaker. Additive Manufacturing (AM) opened a path to solving this problem.

A regenerative cooling combustion chamber is a good example. The combustion chamber wall touches hot flame. If left as it is, the wall melts. So thin cooling channels are made inside the wall. Cold fuel passes through those channels and cools the wall. A structure with channels hidden inside the wall is hard to make by older methods. Additive manufacturing makes such complex internal structures at once.

A representative technology for making parts is Laser Powder Bed Fusion (LPBF). A laser passes over a thin layer of metal powder. The spot touched by the laser melts instantly and then soon solidifies. Powder is spread again on top, and the laser is fired again. This process is repeated thousands of times to build up the part.

A metal widely used in rocket parts is nickel-based superalloy. A superalloy is a special metal that does not lose strength even at high temperature. A representative material is Inconel 718. Haynes-series alloys are also used. These metals endure for a long time even in hot gas.

The problem lies inside the metal melted by the laser. The place where the laser passes cools very quickly. It cools so fast that the temperature can fall by hundreds of thousands of degrees in one second. When it cools so abruptly, the structure solidifies unevenly. Crystals grow long in the build direction. These long crystals are called columnar grains.

A phenomenon also appears in which properties change with direction. This is called anisotropy. Wood makes this easy to understand. Wood splits easily along the grain. Cutting across the grain requires more force. Additively manufactured metal also has different strength in different directions. This property makes part design and machining difficult.

Rapid cooling leaves other traces as well. If hot water is poured into a frozen glass cup in winter, the cup cracks. This happens because the inside and outside expand at different rates. A similar thing happens inside additively manufactured metal. The solidification rate differs from place to place. So force becomes trapped inside the part. This trapped force is called residual stress.

Element segregation is added to this. When molten metal solidifies, certain elements gather to one side. In Inconel 718, the element niobium (Nb) gathers. Hard particles form in those places. These particles are called the Laves phase. The Laves phase is much harder than the matrix metal and breaks easily.

These internal conditions create two challenges. The first is that the material is hard to cut. Hard particles damage cutting tools. The second is that internal pores become seeds of fatigue cracks. In fast-spinning parts, these seeds are dangerous. So the following sections deal with these two challenges in order.

There is a reason additive manufacturing has quickly taken hold in the rocket industry. It greatly reduces the time needed to make one part. In the past, making a casting mold alone took several months. Additive manufacturing builds the part directly without a mold. If the design is changed, a new part can be printed the next day. Because design and production can be repeated in short cycles, development becomes faster.

Its ability to reduce weight is also powerful. When several pieces are combined into one, joints disappear. Joints add weight and can easily become weak points. Removing joints makes the part lighter and stronger. That is why rocket companies invest heavily in this technology. The U.S. National Aeronautics and Space Administration has also tested additively manufactured combustion chambers for a long time.

Post-treatment must include heat treatment. Heat treatment means heating and cooling metal to change its structure. First, the metal is heated to a high temperature to dissolve hard particles back into the matrix. This step is called solution treatment. Next, it is held for a long time at a suitable temperature to grow strengthening particles evenly. This step is called aging. A superalloy can show its full strength only after these two steps.

In short, additive manufacturing has brought major advantages and new challenges together. The advantages are complex shapes and fast production. The challenges are rough surfaces and internal defects. The keys to solving these challenges are post-processing and inspection. Artificial intelligence assists the whole process of post-processing and inspection.

7.2 Why Additively Manufactured Superalloys Are Hard to Cut

First, let us unpack what this section covers. Poor machinability means a property that makes a material hard to cut. Additively manufactured superalloys make metals that are already hard to cut even harder to cut. This section examines three reasons. They are the directionality of the crystal structure, trapped stress and work hardening, and the process by which tools wear.

7.2.1 Crystal Structure That Changes with Direction

This subsection deals with how the directionality of the crystal structure affects cutting. Earlier, we said that columnar grains grow long in the build direction. We now look closely at why this structure makes cutting difficult. We also examine the impact that the hard Laves phase has on tools.

Metal is made of many small crystals. In each crystal, atoms are arranged in a regular order. The direction in which these crystals stand determines the properties. In additively manufactured metal, crystals are biased in one direction. Because crystals grow long along the build direction, properties differ by direction.

When a cutting tool passes through this structure, the resistance changes. The angle between the crystal direction and the cutting edge direction differs from place to place. Some places are cut smoothly. Some places resist strongly. So cutting force rises and falls regularly. Because the force is uneven, machining becomes unstable, and the surface also becomes uneven.

This rise and fall gives the tool repeated impacts. As impacts accumulate, the cutting edge is damaged little by little. Vibration also grows. This vibration is called chatter. If chatter becomes severe, wave patterns remain on the surface. In precision parts, these patterns become serious defects.

Hard particles add to the problem. When Inconel 718 solidifies, niobium gathers at crystal boundaries. Plate-shaped or lump-shaped Laves phases form there. These particles are much harder than the matrix metal, while also being brittle.

What happens when a cutting tool hits these hard particles? It is like passing a knife through gravel mixed into soft soil. Small chips break off the cutting edge. When this microscopic damage keeps accumulating, the tool edge eventually becomes blunt and worn.

To solve this problem, the structure must first be made uniform. Heat treatment, discussed later, plays that role. Heat treatment dissolves hard particles and returns them to the matrix. It also changes long

columnar grains into slightly rounder shapes. Then the degree to which properties differ by direction decreases.

This structural difference is directly connected to part performance. If strength differs by direction, design becomes difficult. The part must be placed to avoid the weak direction. Making the structure uniform reduces this concern. That is why post-treatment improves not only machining but also performance.

Artificial intelligence includes this structural information in prediction. The direction of crystals and the amount of hard particles become inputs. It learns how cutting resistance changes according to these values. It is hard for humans to inspect each structure image by eye. Artificial intelligence reads many images quickly and finds rules.

Crystal direction is measured with special equipment. When an electron beam is fired at the surface of a specimen, each crystal produces a different pattern. These patterns show which direction each crystal faces. This direction information becomes a good input for artificial intelligence. Humans have difficulty counting directions by eye. Machines read the directions of countless crystals at once.

The build angle also changes properties. The structure changes depending on whether the part is built upright or lying down. Even with the same drawing, strength differs if the orientation differs. So the build direction is set from the design stage. This angle also affects machinability and fatigue. Artificial intelligence includes the angle as a variable and learns its effect.

However, the structure differs slightly from part to part. Even when parts are made with the same settings, there are variations from place to place. Small laser fluctuations and differences in powder size create variation. So variation remains in model predictions as well. Later sections examine in detail how to handle this variation.

7.2.2 Trapped Stress and the Property of Becoming Harder as It Is Cut

This subsection deals with trapped stress and work hardening. Earlier, we said that force becomes trapped inside additively manufactured metal. We examine what problems this trapped force causes during cutting. We also look at the property of superalloys becoming harder while they are being cut.

First, let us revisit residual stress. Forces are trapped inside a part in several directions. This bundle of forces is represented by a mathematical box called a tensor. A tensor can be seen as a table that holds values in several directions together. It is invisible, but inside the part these forces push and pull against one another.

This trapped force reveals its problem at the moment of cutting. When the surface is cut away, the force at that place is released. Then the remaining part loses balance. Thin blades or ring-shaped parts bend. While trying to meet precise dimensions, the shape instead becomes distorted. So even the order in which surfaces are cut must be decided in advance.

Next is work hardening. Work hardening is the property by which metal becomes harder as it is deformed. If a wire is bent several times, that spot becomes stiff. Superalloys have this property strongly. The thin layer in front of the cutting edge becomes harder in an instant.

Then the tool must cut again through the layer that has just hardened. This takes more force. It also generates more heat. Inconel 718 is a metal that does not release heat well. Its thermal conductivity is only about one-third that of ordinary steel. So heat that cannot escape collects between the cutting edge and the part, raising the temperature further.

The strengthening phases also make work hard for the tool. Very small strengthening particles are spread through Inconel 718. The main strengthening particle in this alloy is niobium-containing gamma double prime (γ″, Ni3Nb). These particles hold the metal firmly. They do not dissolve easily even at high temperatures. So the metal does not soften even in the hot cutting zone, and the tool keeps receiving large forces.

All these factors combine to cause surface damage. A thin layer below the cut surface changes. The microstructure breaks into fine pieces, and its color turns white. This layer is called the white layer. Tensile residual stress and microcracks are likely to remain together in the white layer. This layer becomes the starting point of cracks and greatly reduces the fatigue life of the part.

Why is the white layer dangerous? This layer contains invisible tensile residual stress. Tensile stress is a force that opens cracks. When microcracks are added to it, they become seeds. Under repeated loading, cracks grow from these seeds. So the final finishing step must not leave a white layer.

There are ways to reduce this problem. After cutting, a post-treatment is added that strikes the surface with small particles. This treatment leaves compressive stress on the surface. Compressive stress is a force that presses cracks closed. It acts in the opposite direction to tensile stress. When compressive stress is planted on the surface in this way, the fatigue life of the part increases.

Artificial intelligence also learns the effects of these different post-treatments together. Cutting conditions, peening treatment, and heat treatment are all entered as inputs. Then it predicts which combinations will produce good results. It greatly reduces the work that people used to test one by one. Even so, the final check is done through actual testing.

Artificial intelligence is used to detect this risk in advance. When cutting conditions and residual stress values are entered, it predicts the surface condition. It learns under which conditions a white layer forms. Then dangerous conditions can be avoided. For titanium additively manufactured parts, a deep learning model has been developed that predicts surface roughness and hardness from process conditions. This book proposes applying the same approach to cut surfaces of superalloys.

However, predictions can be trusted only within the range learned by the model. Errors become larger under conditions it has never seen before. So actual parts must always be checked by surface inspection. The white layer is thin, so it is easy to miss with the eye. Artificial intelligence filters out dangerous conditions in advance and narrows the candidates.

7.2.3 How Tools Wear and How Surfaces Are Damaged

This subsection covers the different ways tools wear. Tool wear is directly linked to production quality. When a tool wears, the surface becomes rough. A rough surface becomes the starting point of fatigue cracks. So knowing in advance how much a tool has worn is important in quality control.

There are several types of tool wear. The first is notch wear. A deep groove forms in the tool at the boundary of the cutting depth. This is where oxygen in the air and the hardened metal layer rub together. Oxygen and the hard metal layer rub together, and the groove in this area gradually becomes deeper.

The second is peeling caused by adhesion. Nickel sticks easily to tool materials. Cut metal chips become welded onto the cutting edge. This lump is called a built-up edge. When this lump breaks away, it tears off the coating on the tool surface with it. As the coating is ripped away, the tool surface is quickly damaged.

The third is diffusion wear. The cutting zone has a very high temperature. At high temperatures, atoms diffuse into each other. Atoms inside the tool move toward the metal chips. Then a hollow mark remains

on the tool surface, and this mark is called a crater.

The fourth is the surface white layer and microcracks. The white layer discussed earlier appears here as well. Rapid deformation and rapid cooling combine to make a thin white layer. Cracks are likely to be hidden in this layer. Even if the outside looks smooth, flaws such as cracks may be hidden inside.

All these types of wear are linked to each other. When one becomes larger, others also tend to grow. So the cause must not be viewed as only one thing. Several signals must be examined together. Cutting force, vibration, sound, and temperature are all signals of wear.

Artificial intelligence reads these signals together. Sensors collect signals during cutting in real time. The model uses these signals to estimate how much the tool has worn at that moment. Recent research used deep learning to read photos of worn tools and directly predict the degree of wear. A machine now does work that once required a person to stop the tool and inspect it with a magnifier.

The same approach works for other metals. One study predicted tool wear during cutting of a nickel alloy called Hastelloy C276 using a data-based model. This is because several superalloys show similar wear patterns. These results are also useful for rocket superalloys.

If tool wear can be predicted, unmanned machining becomes possible. The machine checks the tool condition by itself. If there is danger, it lowers the speed or changes the tool. A person does not have to watch it all the time. In factories that machine parts overnight without people, predictive models support that safety.

These four types of wear depend greatly on temperature. The hotter the cutting zone becomes, the faster wear proceeds. So cooling is important. Cutting fluid is sprayed to remove heat. Sometimes cutting fluid is shot directly at the edge under high pressure. This cooling method also becomes an input variable for artificial intelligence.

In summary, tool wear prediction becomes the eye of quality control. The tool can be changed before wear becomes severe. Conditions can be corrected before defective parts are produced. However, sensor signals contain noise. Vibration also includes shaking from the machine itself. So the model must be trained well to withstand noise. Using several sensors together reduces the effect of noise.

7.3 Calculating Forces for Better Cutting and Optimizing Post-Treatment

The previous section examined why cutting is difficult. This section looks at how to overcome that difficulty. It covers three topics in order. First, it looks at how to calculate the forces involved in cutting. Next, it examines how machinability changes depending on post-treatment. Finally, it looks at how artificial intelligence finds optimal conditions.

7.3.1 Cutting Force and Models That Handle It

This subsection looks at cutting force and the models that handle it. Cutting force is the force with which the tool pushes the metal. If this force is known in advance, tools and conditions can be chosen well. We examine the basic idea of calculating force together with the problem of vibration.

Cutting force does not act in only one direction. There is a force pushing forward, a force pushing sideways, and a force pressing downward. So cutting force is expressed as a vector with components in three directions. A vector is a quantity that has both magnitude and direction. By contrast, stress inside a material is expressed as a tensor. A tensor is a quantity that contains values in several directions together. The cutting force vector and the stress tensor are different quantities. The size of this force depends on the cutting thickness and the strength of the material.

During cutting, three things happen at the same time. The metal deforms, heat is generated, and friction occurs. A good model must handle all three together. The strength of a material when it is compressed quickly changes with temperature. When it becomes hot, the metal softens a little. This relationship is put into the model as a table, not as an equation.

This model uses different constants for each material. The constants are determined by experiment. Even for the same Inconel 718, the constants differ between additively manufactured material and wrought material. This is because additively manufactured material has a different microstructure. So experiments dedicated to additively manufactured material are needed. The model becomes accurate only when enough experimental data for additively manufactured material is accumulated.

Artificial intelligence also helps determine the constants. There is a method that works backward from cutting signals to find the constants. Force and temperature are measured, and constants that match those values are selected. A machine quickly performs work that people used to fit by hand. Then the model can be adapted quickly even to a new material.

The vibration problem is also important. Chatter was discussed earlier. Chatter is vibration in which the tool and the part shake each other. Once vibration occurs, the tool cuts the mark again in the next rotation. Then the vibration grows larger. A small vibration grows larger by itself and interferes with machining.

To prevent this vibration, the safe cutting depth must be known. If the cut is too deep, vibration grows. If the cut is shallow, vibration dies down. This boundary changes with rotational speed. A diagram showing this boundary is called a stability lobe diagram. Conditions without vibration are chosen by looking at this diagram.

Artificial intelligence helps speed up this calculation. Calculation using only a physics model takes a long time. Every time conditions change slightly, the problem must be solved again from the beginning. Artificial intelligence learns the results of the physics model in advance. After that, it immediately gives the answer for new conditions. A model that replaces heavy calculation in this way is called a surrogate. Surrogate means substitute. Accurate calculations are performed only a few times in advance. The substitute model that has learned those results handles the rest.

This method greatly reduces time in the field. In the past, many trial cuts were made while changing conditions. Now the model filters out dangerous conditions first. Only the few remaining conditions are actually tested. However, if the physics model is wrong, the artificial intelligence will also be wrong. So it is safer to use physics and data together.

The method of using physics and data together is called physics-informed machine learning (PIML). This method puts physical laws into artificial intelligence as well. Even when data is scarce, physical laws help hold the answer in place. They become a safety device that prevents absurd predictions. This idea appears repeatedly in later chapters of this book. It is highly useful in fields where data is scarce, such as rocket engineering.

7.3.2 Comparing the Machinability of Wrought Material, As-Built Material, and HIP-Treated Material

This subsection compares metals in three states. Even the same Inconel 718 has different properties depending on how it was made and post-treated. We examine wrought material, the as-built state, and the state after hot isostatic pressing side by side.

First, let us look at wrought material. Wrought material is made by pressing and working metal. It has the same roots as the way iron is made by hammering it in a blacksmith's shop. Its microstructure is fine

and uniform. The crystals are small and round. So variation in cutting force is relatively small among the three materials. Tool wear is also smaller, so prediction is much easier.

Next is the as-built state. This state is also called as-built. It is the state directly from the printer without post-treatment. Residual stress is large, and hard Laves phases are concentrated. Columnar grains are also thick and long. So cutting resistance rises much higher than in wrought material. The risk of tool breakage is also relatively high among the three materials.

The third is the state after Hot Isostatic Pressing (HIP). HIP is a post-treatment that applies high temperature and high pressure together. It is usually performed at temperatures above 1000 degrees Celsius. The pressure also exceeds 100 megapascals. Under these conditions, pores inside the metal are pressed closed.

It is easy to understand by thinking of bread dough. Small air bubbles are inside the dough. If the dough is pressed and kneaded for a long time, these bubbles escape and the dough becomes dense. HIP is similar. High pressure presses pores inside the metal shut. High pressure and heat move atoms so that the walls of the pores bond to each other.

HIP also changes the microstructure. However, the degree of change depends on temperature and holding time. If the temperature is high enough, the hard Laves phase dissolves into the matrix. If the temperature is low or the time is short, some Laves phase remains. Long columnar grains may also remain, depending on the conditions. When the microstructure becomes more uniform, directional differences and variation in cutting force decrease. So HIP is used together with solution treatment, aging, and homogenization heat treatment under matched conditions.

However, HIP also has limits. It cannot remove every pore. Pores connected to the surface and pores trapped inside behave differently. Argon bubbles trapped during additive manufacturing cause problems again even if they are closed by HIP. This is because argon is a gas that does not dissolve in metal. If it later receives high temperature again, the closed bubbles swell and reopen. This phenomenon is called thermally induced porosity. So it is important to reduce gas bubbles from the start.

New problems also appear after HIP. When the microstructure becomes softer, metal chips do not break easily. Chips continue as long strands and wrap around the tool. If this wrapping becomes severe, it scratches the surface. So the tool shape must be chosen so that chips break well.

Artificial intelligence organizes this comparison as data. It gathers cutting data for each condition and learns the differences. It can then judge which condition a new part is closest to. It also suggests which post-processing steps are needed. Recent studies broadly examine ways to optimize HIP and heat treatment together.

HIP is a post-processing step that costs time and money. The part must be heated and pressed for a long time inside large equipment. So it is hard to use HIP for every part. Engineers must decide which parts truly need HIP. Artificial intelligence helps with this decision as well. It looks at defect information and estimates the effect of HIP in advance.

Public research also provides useful guidance. The U.S. National Aeronautics and Space Administration examined through simulation how HIP reduces defects in additively manufactured Inconel 718. One study measured fatigue strength after treating parts with many initial pores by HIP. Most pores were closed, but argon bubbles remained. Results like these build the foundation for post-processing design.

There is a reason to use wrought material as the standard. Wrought material has been verified for a long time. Its properties are well known, and there is a large amount of data. The goal is to raise

additively manufactured material to the level of wrought material. Post-processing is a path toward that goal. Artificial intelligence helps find that path faster.

7.3.3 Finding Good Cutting Conditions with Few Experiments

This subsection explains how to find good conditions with few experiments. There are many cutting conditions. Speed, feed rate, depth, and cooling method are all variables. If all these variables are combined, the number of experiments explodes. The Taguchi method and artificial neural networks are combined to solve this problem.

The Taguchi method is a way to compare many conditions with few experiments. It does not test every possible combination. It selects only representative combinations and tests them. With a few well-chosen combinations, it reads the overall trend. Since the number of experiments decreases, time and cost are saved as well.

This method greatly reduces the number of experiments. For example, hundreds of combinations can be reduced to eighteen experiments. In each experiment, surface roughness and tool wear are measured. The energy used for cutting is also measured. Values collected from these combinations become training data for artificial intelligence.

An artificial neural network (ANN) learns rules from this data. An artificial neural network is a computing structure modeled on neural connections in the human brain. The inputs include cutting conditions and heat-treatment state. The outputs include surface roughness and tool wear. The model finds the relationship between inputs and outputs by itself and learns it as rules.

After training, it predicts conditions that were not tested. If a new speed and a new feed rate are entered, it tells the result in advance. Then a search for optimal conditions is added. It selects conditions that make the surface smooth and tool wear low. When the two goals conflict, it finds a balance point.

Actual cases are also accumulating. One study examined micro-turning of additively manufactured Inconel 718. It looked at how machinability changes in heat-treated material. It compared cutting speeds between 30 meters and 60 meters per minute. Data like this becomes training material for the model.

This combination is highly useful in the field. First, engineers make a small experimental table. Then they collect data through sensors and measurements. Finally, artificial intelligence learns the relationship between conditions and quality. This makes it possible to set conditions for a new part quickly.

When there are several goals, a balance point must be found. Engineers want a smooth surface, low tool wear, and a short machining time. These goals conflict with each other. If cutting is fast, the surface is likely to become rough. So if one goal is improved alone, another often becomes worse.

At this point, a method that handles several goals together is used. One goal is yielded a little to gain another. Good combinations obtained in this way are collected. Among them, the combination that fits the part is selected. Artificial intelligence finds these combinations quickly.

The actual goal is to reduce surface roughness. Rocket parts must have smooth surfaces to resist fatigue. Well-chosen cutting conditions and good tools together move toward this goal. A smooth surface reduces crack initiation sites. So surface finishing is the first step in fatigue assurance.

The limits are also clear. A model can be trusted only within the range it has learned. Its prediction becomes unstable at speeds or feed rates not included in training. If data is scarce, error grows. So for critical parts, engineers confirm the conditions through actual cutting tests. They verify whether the prediction was correct by measuring surface roughness and tool wear.

Tool selection is also an important variable. Superalloys are cut with tools that have special coatings. Coatings help tools withstand heat and wear. Tools made from ceramic materials are also used. Artificial intelligence also learns which tool fits which condition. This makes it possible to balance tool cost and machining quality.

The future task is to broaden the data. If experimental data from many institutions is collected, the model becomes more robust. However, each institution uses slightly different measurement methods. The way cutting conditions are recorded also differs. So rules are needed to align the data into one form. Once such standards are in place, prediction reliability increases.

7.4 Very High Cycle Fatigue Failure of Turbopump Blades

So far, we have discussed how to cut well. From this section, we discuss how long a part can endure. We examine why the rotating blades of a rocket turbopump are especially dangerous and what traces they leave when they break.

7.4.1 A Harsh Environment with Both Cryogenic Temperatures and High Temperatures

This subsection covers the harsh environment around turbopump blades. A turbopump is a pump that pushes fuel and oxidizer at high pressure. The blades that rotate rapidly inside it are called an impeller. We look at why the loads on this part are special.

The impeller rotates extremely fast. It rotates 30,000 to 80,000 times per minute. This is more than ten times faster than an automobile engine. During rotation, centrifugal force pulls the blades outward. If the rotation is steady, this centrifugal force is a steady load whose magnitude changes very little.

The temperature environment is also harsh. The impeller itself contacts cryogenic propellant. Liquid hydrogen is colder than minus 250 degrees Celsius. Liquid oxygen is also near minus 180 degrees Celsius. Hot working gas is not at the impeller but at the turbine attached to the same shaft. This heat gradually moves toward the pump through the shaft, bearings, and seals. So the two ends of one shaft face a very cold region and a hot region at the same time.

A large temperature difference creates contraction and expansion. The cold side contracts, and the hot side expands. This difference creates force inside the part. Steady centrifugal force lies underneath it. Repeated vibration loads are then added on top and shake the part.

Repeated loads come from many places. Each time a blade passes a stationary part, force is applied. Tiny imbalance in the rotating shaft and pressure pulsation also add force. Startup and shutdown each add a large load. Blade-passing forces repeat thousands of times per second. So the number of cycles accumulates very quickly.

Fatigue with such a large number of cycles is called very high cycle fatigue (VHCF). Ordinary design uses ten million cycles as a standard. The actual number of repeated loads a part receives is obtained by multiplying the excitation frequency by the accumulated operating time. Specimen studies cover the range of 100 million and 1 billion cycles. In this range, fatigue rules change, so it must be examined separately.

Ordinary fatigue tests have difficulty covering this range. If 1 billion cycles are tested at a normal speed, it takes too long. So special high-frequency testing machines are used. These machines vibrate a specimen about 20,000 times per second. Even with such fast vibration, one test takes several days to finish.

Because testing takes a long time, data is valuable. If data is scarce, prediction becomes unstable. So artificial intelligence needs methods that use data sparingly. It must produce reliable predictions even

with little data. This problem of using data efficiently is discussed again in detail in a later section.

Material selection is also tied to this environment. Some metals become soft or brittle in very cold places. Inconel 718 has relatively stable properties even at cryogenic temperatures. So it is widely used for cryogenic parts. Haynes alloys perform well at higher temperatures. Materials must be chosen according to where the part is located.

Even within one part, conditions differ from place to place. The blade root receives large forces. The blade tip receives much vibration. So engineers must first find the weak locations that cannot easily withstand force. These locations are called critical sections. Fatigue assurance is based on these critical sections.

Calculation is also needed to find critical sections. The forces acting on the part are solved by computer. This calculation is called finite element analysis. The part is divided into small pieces, and the force on each piece is calculated. The place where force concentrates is the critical section. Artificial intelligence sometimes performs this calculation quickly in its place.

Impeller vibration has special rules. The number of times a blade passes a stationary part is fixed. This number is multiplied by the rotational speed to become the vibration period. At certain rotational speeds, this vibration overlaps with the part's natural vibration. Then the vibration grows greatly. This overlap is called resonance. Resonance quickly damages parts. So dangerous rotational speeds are avoided in design.

Fatigue assurance of the impeller is central to rocket safety. If one blade breaks, the whole pump is destroyed. Then the engine stops or explodes. So this part receives stricter inspection than almost any other part. Both surface quality and internal defects must be controlled.

7.4.2 Failure That Starts Inside and the Fish-Eye Pattern

This subsection explains how very high cycle fatigue breaks a part. Earlier, we said that the rules in this range are different. Here we look at how they differ in detail. We also examine the unusual pattern left on the fracture surface.

Ordinary fatigue starts at the surface. The surface is where scratches easily form. Under repeated loads, slip marks appear on the surface. Cracks start from these marks. So if the surface is polished smooth, life increases.

Very high cycle fatigue is different. Cracks start inside, not at the surface. When the number of cycles is extremely large, internal defects cannot endure even weak force. Small holes or hard particles inside become seeds. A small argon bubble that HIP cannot remove is a dangerous example.

A distinctive region forms around the crack seed. Hundreds of millions of tiny deformations accumulate around the seed. A crack that starts inside grows where air cannot reach. The two faces of a narrow gap rub against each other and crush the microstructure into very fine pieces. This finely crushed region is called a fine granular area (FGA). Under a microscope, this region looks darker than its surroundings. This dark region sits at the center of the fish-eye.

It takes a long time for this dark region to form. Most of the total fatigue life is spent in this stage. The period before the crack grows large enough to be seen is long. Even when the outside looks intact, damage is already progressing inside.

This point matters for setting inspection timing. Damage appears late on the outside. So engineers must not feel safe by looking only at the surface. They must inspect the inside at fixed intervals. In reusable rockets, managing this interval is even more important. Before a used part is used again, its inside must

be checked.

When a crack grows from the seed, it leaves a round pattern. This pattern looks like a fish eye. So it is called a fish-eye. A fish-eye is evidence that the crack started inside. By looking at the fracture surface, engineers identify the cause from this pattern.

At the end, the crack reaches the surface. At that moment, the part suddenly breaks. A crack that had been growing slowly bursts all at once. This sudden failure, which is hard to detect in advance, is the most dangerous.

This failure mode has major meaning for inspection. If only the surface is examined, the danger is missed. The inside must be inspected. So X-ray CT and ultrasonic inspection are used together. These inspections look inside a part without breaking it. Such inspection is called non-destructive testing.

Artificial intelligence works together with these inspections. It uses the size and location of internal defects found by CT as inputs. With these values, it predicts fatigue life. It ranks which defects are dangerous. Recent research proposed a method to find critical defects in additively manufactured parts non-destructively. Machines first identify internal defects that humans can easily miss.

It is difficult to eliminate internal defects completely. So parts are designed to remain safe even when defects exist. This idea is called damage-tolerant design. Engineers decide in advance how large a defect the part can withstand. Parts with defects larger than that are discarded. Artificial intelligence creates the data used to set this standard.

This method also connects to inspection standards. Inspection equipment cannot find every very small defect. There is a size limit to the defects it can detect. Design makes defects above this limit safe. Then defects missed during inspection are not dangerous. In this way, inspection and design fill each other's gaps and protect safety.

7.5 Fatigue Reliability Assurance That Predicts Scatter Together

Fatigue life does not come out as one exact number. Even parts under the same conditions have different lives. So prediction must include the range of scatter. This section explains artificial intelligence that handles that scatter. It looks at how to turn defects into numbers, how to build a model that contains scatter, and how to connect it to safety standards.

7.5.1 Turning Material State into Numbers

This subsection deals with ways to turn material state into numbers. Artificial intelligence learns by taking in numbers. So defects and surfaces must be converted into numbers. These numbers are called descriptors. Here, we examine one by one which numbers matter for fatigue life.

First, the shape of a defect is captured as numbers. When a part is scanned with CT, the three-dimensional shape of a void appears. This shape is organized as a tensor. As noted earlier, a tensor is a box that holds values in several directions. It contains information on which direction the void is elongated in and how flat it is.

X-ray CT works on the same principle as CT in a hospital. X-rays are sent from several directions to see inside. That information is combined to make a three-dimensional image. Internal voids can be seen without cutting the part. The size and shape of each void are measured. The size and shape information collected in this way becomes basic data for fatigue prediction.

The size of a defect is also important. A good way to measure size is the Murakami square root of area method. This method treats a defect like a small crack. The defect is projected onto a plane

perpendicular to the loading direction, and that area is measured. The square root of that area is used as the defect size. This is because the square root is closely connected to the strength of the force that opens a crack. So a complex three-dimensional void is converted into this one value and handled. The larger this value is, the worse its effect on fatigue becomes.

The location of the defect is also considered. Even at the same size, a defect near the surface is more dangerous. A surface defect has less surrounding material to hold it, so force concentrates more. Environmental effects from contact with air also make the crack grow slightly faster. A defect deep inside is surrounded by material on all sides and is less dangerous. So different correction values are given depending on location.

Surface roughness is also converted into numbers. A machined surface has tiny valleys and peaks. The deeper and sharper the valleys are, the more force concentrates there. The degree of this concentration is called the stress concentration factor. It is like paper tearing easily if you make a small notch before cutting it. A small notch gathers dispersed force into one point and weakens that spot.

These numbers are connected to fatigue life. There is a graph that shows the relationship between force level and the number of repeated cycles. This graph is called an S-N curve. Information on defects and roughness moves this curve up or down. If defects are large and the surface is rough, the curve moves downward. When the curve moves downward, it means the part breaks sooner under the same force.

Recent studies have put these descriptors into artificial intelligence and achieved results. One study predicted the fatigue life of additively manufactured metal using defect information measured by CT. More studies are also applying the Murakami method to various additive manufacturing cases. They repeatedly confirm that defect size, location, and shape have a large effect on life.

Material hardness is also included among the descriptors. Hardness shows how hard a material is. In the Murakami equation, higher hardness is treated as improving resistance to defects. But hardness alone cannot explain all resistance to defects. A very hard material can instead become more sensitive to defects. So hardness, defect size, location, and load ratio must be considered together.

Loading conditions are also entered as numbers. How large the force is and in which direction it acts are important. It also matters whether the force comes from one direction only or is mixed from several directions. Temperature is included as well. This is because material properties differ in hot and cold places.

Recent research has achieved good results with these descriptors. One study predicted the high-cycle fatigue life of additively manufactured Inconel 718. It used defect size and load level as inputs. It solved the problem of limited data by increasing the data with artificial intelligence. As a result, the coefficient of determination of the prediction rose to 0.975. The average error was only 1.13 percent. The coefficient of determination shows how well a prediction matches reality. The closer this value is to 1, the better the prediction matches reality. However, this number came from limited data for one material. Because it was obtained using expanded data, it is hard to apply directly to other processes and materials.

However, choosing descriptors is not easy. Judgment is needed to decide which numbers to include and which to remove. If an important number is left out, the prediction goes wrong. If a useless number is added, learning becomes blurred. So people who know materials well and artificial intelligence must choose them together.

7.5.2 Bayesian Neural Network (BNN) That Produces a Range of Answers

This subsection deals with artificial intelligence that contains scatter. An ordinary neural network gives only one answer. This is not enough for fatigue prediction. Life spreads widely even under the same

conditions. So a model that gives the range of answers together is needed.

This kind of model is called a Bayesian Neural Network (BNN). In an ordinary neural network, each internal value is fixed as one value. A Bayesian Neural Network treats internal values as distributions. It has a range of values, not a single value. So the prediction output also comes out as a range, not as one value.

The word Gaussian often appears here. Gaussian means a bell-shaped distribution. It is high in the middle and becomes lower on both sides. It is easy to think of the distribution of test scores. Many people are near the average, and fewer people are at the ends. But fatigue life itself is often not bell-shaped. It scatters in other shapes, such as lognormal or Weibull distributions. So a bell-shaped assumption is placed on the logarithm of life. This bell-shaped assumption is also a convenience to make calculation easier.

Two values are important in this distribution. One is the mean. The mean is the representative value. The other is the width of spread. A wide width means large scatter. A Bayesian Neural Network gives these two values together.

Let us compare why a prediction has a range to a weather forecast. A forecast says there is a 70 percent chance of rain tomorrow. It does not say with certainty that it is 100 percent. This is because the future has uncertainty. A Bayesian Neural Network also reports a probability range instead of a single answer.

The way this model is trained is also special. Because the internal values are distributions, calculation is complex. Exact calculation is practically impossible. So a method is used that replaces them with a nearby distribution. The core of training is this process of reducing the amount of calculation by changing them into a nearby distribution.

After training is complete, predictions are drawn several times. Because the internal values are distributions, a slightly different answer comes out each time. These answers are collected to find the mean and the width. If the answers are similar to each other, the width is narrow. If the answers are scattered, the width is wide. The size of this width shows how much the prediction can be trusted.

This method resembles human judgment. Experienced engineers also distinguish certainty from guesswork. They speak with confidence about parts they know well. They speak cautiously about parts they see for the first time. A Bayesian Neural Network also communicates its own confidence in this way. So it is easier for people to look at that level of confidence and use the result with trust.

This approach has one more advantage. When the model meets conditions it has never seen before, it gives a wide range. It sends a signal by itself that it does not know well. People see this signal and proceed carefully. A prediction with a wide range is used only after more checking. A model that reveals its own limits is a great help to safety.

This approach fits fatigue prediction well. Fatigue is a phenomenon with large scatter. A range is more honest than a single number. Several recent studies are following this direction. Studies have predicted the fatigue life of aerospace metals together with scatter. Other studies have raised trust by adding physical laws.

Why is it useful to add physical laws? Fatigue has long-established theories. They are equations that contain the relationship between force and life. These equations are added to artificial intelligence. Then the answer does not stray far even when data are limited. Studies have also handled loads from several directions together. This trend makes life assurance for rocket parts more robust.

This method also has a cost. Predictions must be drawn several times, so computation increases. Training is also harder than with an ordinary neural network. Still, this cost is worthwhile for rocket

parts. Knowing the scatter protects safety. If an accident occurs after trusting only one number, the loss is far greater.

7.5.3 Separating Scatter and Connecting It to Safety Standards

This subsection deals with how to separate scatter. Earlier, we said that prediction has a range. This range is not one kind. Two kinds of uncertainty are mixed in it. When the two are separated, it becomes clear what should be done.

The first is epistemic uncertainty. This is scatter that occurs because the model does not know enough. It appears strongly under conditions with limited data. If the situation has not been learned before, the model lacks confidence. This uncertainty gradually decreases when the missing data are added.

The second is aleatoric uncertainty. This occurs because the material and load themselves vary. Even parts under the same conditions have different internal defects. Noise is also mixed into measurement. This uncertainty is built into the phenomenon itself, so it does not decrease much even when more data are added.

Why is it important to separate these two? The prescriptions are different. If epistemic uncertainty is large, more experiments are needed. If aleatoric uncertainty is large, more safety margin is needed. If the two are lumped together, it is easy to prescribe the wrong response. When the two kinds of uncertainty are separated, the next step becomes clear.

Now prediction is connected to safety standards. Rocket parts must exceed defined safety standards. These standards include two things together. One is what percentage of parts survive. The other is how much we can trust that value. For example, a lower bound may be set where 90 percent survive with 95 percent confidence. A value that sets survival probability and confidence level together in this way is called a material allowable. The A-basis and B-basis values used in aerospace materials are examples. This value cannot be guaranteed simply by subtracting scatter from the mean. It is calculated statistically by including both the number of test data points and their dispersion.

This lower bound is called the design allowable. A part must last longer than this allowable. A safety factor may be multiplied once more. A safety factor is a margin against unexpected risk. Accidents are prevented by placing several layers of margin in this way. If the margin is excessive, the part becomes heavy, so a balance must be found.

This leads judgment toward safety. Under conditions where the model lacks confidence, it sets life lower. Under conditions where it is confident, it gives a little more margin. This attitude fits rocket parts. Even if the estimate is somewhat conservative, preventing an accident is better.

Certification is not completed by artificial intelligence alone. Fatigue assurance uses several tools together. Artificial intelligence prediction, standard fatigue testing, nondestructive inspection, and safety factors are all needed. Standard test methods are also defined. There are international standards for how force is applied and how data are analyzed.

The role of artificial intelligence is clear. It quickly finds risky candidates. It shows where more testing is needed. It helps obtain much information from a small number of experiments. People use that information to design test plans well and greatly reduce time and cost.

The final judgment is made by people and experiments. Artificial intelligence prediction is the beginning, not the end. Actual test data must support the prediction. Quality documents must record that process. Only when all of this is in place can a part go onto a rocket.

Tasks remain for the future. Fatigue data are still limited. Data in the very high-cycle region are even scarcer. This is because testing takes a long time. So methods for learning from small amounts of data continue to be studied. There are also methods for transferring data from other materials. As these methods develop, prediction becomes more robust.

Data quality is also important. Each dataset measures defects differently. Test methods also differ from one dataset to another. If records are poor, the model learns incorrectly. So a culture of careful data recording is needed. Good models require good, well-recorded data.

Standards and certification must advance as well. Current certification rules were made for older materials. Additive materials have different properties. New rules are therefore needed. Experts are also discussing how to include artificial intelligence predictions in certification. Once this discussion is settled, additive parts will be used more widely.

Summary

This chapter covered two major problems. One is how to machine metals that are hard to cut. The other is how to know how long a part will last. Both problems arise in rocket parts made by 3D printing. Artificial intelligence gives major help with both problems.

The two problems are in fact connected. Good machining makes the surface smooth. A smooth surface is strong against fatigue. Removing internal pores reduces the seeds of cracks. In the end, good processing leads to long life. That is why machining and fatigue must always be considered together. This chapter placed the two problems, machining and fatigue, side by side and examined them together.

The first part looked at why additive superalloys are hard to machine. Structures that cool rapidly have different properties depending on direction. Stress is trapped inside, and hard particles are concentrated in some areas. These properties make tools wear quickly. They also leave white layers and microcracks on the surface.

We also looked at ways to overcome these difficulties. Cutting force and vibration are calculated in advance to choose safe conditions. Hot isostatic pressing reduces internal pores and makes the structure more uniform. The Taguchi method and artificial neural networks find good conditions from a small number of experiments. Artificial intelligence quickly narrows the candidates in this process.

The later part dealt with fatigue. Rocket turbopump blades face very-high-cycle fatigue. In this range, cracks begin inside the material, not at the surface. A round pattern called a fish-eye remains on the fracture surface. For this reason, nondestructive testing is needed to look inside a part without breaking it.

Fatigue prediction must include variability. Even parts made under the same conditions have different lifetimes. A Bayesian neural network gives a range of answers instead of one answer. If this range is divided into two types, the next task becomes clear. We can decide whether to increase the data or increase the safety margin.

The key is cooperation between artificial intelligence and people. Artificial intelligence reads data quickly and points out risks. It reduces the experimental plan intelligently. But final certification depends on experiments and inspections. Rocket parts are protected by these two forces working together. This cooperation between artificial intelligence and experiment becomes the foundation that supports safe flight.

Let us reduce this chapter to one line. We must machine well, finish well, and inspect well. Then we must predict the results together with their variability. Artificial intelligence helps every one of these

steps move faster. People use that help to make safer parts. The future of 3D-printed rocket parts ultimately depends on this cooperation.

Related Supporting Materials

Ojo, O. O., & Taban, E. (2023). Post-processing treatments-microstructure-performance interrelationship of metal additive manufactured aerospace alloys: A review. Materials Science and Technology, 39(1), 1-41. https://doi.org/10.1080/02670836.2022.2130530

Song, Z., Peng, J., Zhu, L., Deng, C., Zhao, Y., Guo, Q., & Zhu, A. (2025). High-Cycle Fatigue Life Prediction of Additive Manufacturing Inconel 718 Alloy via Machine Learning. Materials, 18(11), 2604. https://doi.org/10.3390/ma18112604

Yu, C., Huang, Z., Zhang, Z., Shen, J., Wang, J., & Xu, Z. (2021). Influence of post-processing on very high cycle fatigue resistance of Inconel 718 obtained with laser powder bed fusion. International Journal of Fatigue, 153, 106510. https://doi.org/10.1016/j.ijfatigue.2021.106510

ASTM International. (2021). ASTM E466-21: Standard Practice for Conducting Force Controlled Constant Amplitude Axial Fatigue Tests of Metallic Materials. ASTM International, West Conshohocken, PA. https://www.astm.org/e0466-21.html

ASTM International. (2015). ASTM E739-10(2015): Standard Practice for Statistical Analysis of Linear or Linearized Stress-Life and Strain-Life Fatigue Data. ASTM International, West Conshohocken, PA. https://www.astm.org/e0739-10r15.html

Yi, M., Tang, W., Zhu, Y., Liang, C., Tang, Z., & Yin, Y. (2023). A holistic review on fatigue properties of additively manufactured metals. arXiv preprint. (Preprint, not yet peer reviewed) https://arxiv.org/abs/2311.07046

Altıparmak, S. C., Lombardi, M., Bondioli, F., Fino, P., Biamino, S., & Ugues, D. (2025). Hot isostatic pressing for powder-based additive manufacturing of metals: State-of-the-art review and future perspectives. Journal of Materials Research and Technology, 39, 4794-4822. https://doi.org/10.1016/j.jmrt.2025.10.135

Murakami, Y. (2002). Metal Fatigue: Effects of Small Defects and Nonmetallic Inclusions. Elsevier Science, Oxford, UK.

Basquin, O. H. (1910). The exponential law of endurance tests. Proceedings of the American Society for Testing and Materials, 10(2), 625-630.

Coffin, L. F. (1954). A study of the effects of cyclic thermal stresses on a ductile metal. Transactions of the American Society of Mechanical Engineers, 76(6), 931-950.

Manson, S. S. (1953). Behavior of materials under conditions of thermal stress. National Advisory Committee for Aeronautics (NACA) Technical Note, NACA-TN-2933.

Arola, D., & Ramulu, M. (1999). An examination of the effects of surface texture on the strength, fatigue life, and stress concentration of machined components. Journal of Engineering Materials and Technology, 121(3), 376-381.

Blavette, D., Cadel, E., Fraczkiewicz, A., & Menand, A. (1999). Three-dimensional atomic-scale imaging of segregation and precipitation in superalloys. Science, 286(5448), 2317-2319. https://doi.org/10.1126/science.286.5448.2317

Wang, Q., Yao, G., & Hou, M. (2026). Fatigue life prediction and uncertainty quantification of aerospace metals: A Bayesian physics-informed neural network model. Reliability Engineering & System Safety, 266, 111724. https://doi.org/10.1016/j.ress.2025.111724

Zhang, D., Dai, W., & Ding, Y. (2025). Bayesian optimization-based physics-informed neural network for adaptive prediction of multiaxial fatigue life in metallic materials. Fatigue & Fracture of Engineering Materials & Structures, 49(1), 28-51. https://doi.org/10.1111/ffe.70101

Chang, J., Basvoju, D., Vakanski, A., Charit, I., & Xian, M. (2025). Predictive modeling and uncertainty quantification of fatigue life in metal alloys using machine learning. arXiv preprint. (Preprint, not yet peer reviewed) https://arxiv.org/abs/2501.15057

Li, L., Chang, J., Vakanski, A., Wang, Y., Yao, T., & Xian, M. (2024). Uncertainty quantification in multivariable regression for material property prediction with Bayesian neural networks. Scientific Reports, 14, 10543. https://doi.org/10.1038/s41598-024-61189-x

Vishnu, S., Kuriachen, B., & Mathew, J. (2026). Microturning response of LPBF Inconel 718: Effect of homogenization annealing. Materials and Manufacturing Processes, 41(7), 964-973. https://doi.org/10.1080/10426914.2026.2660062

Truong, T. T., Airao, J., Fattahi, S., Azarhoushang, B., Karras, P., & Aghababaei, R. (2025). Image-based machine learning model for tool wear estimation in milling Inconel 718. Wear, 571, 205865. https://doi.org/10.1016/j.wear.2025.205865

Abdullah, M., Rao, A. C. U., Ramachandran, T., Biswal, S., Chaudhary, S. P., Singh, R., Mishra, A., & Mann, V. S. (2026). A data-driven approach for high-accuracy tool wear prediction in machining Hastelloy C276. Scientific Reports, 16, 19442. https://doi.org/10.1038/s41598-026-50824-4

Medibew, T. M., Zieliński, D., Agebo, S. W., & Deja, M. (2025). Recent research progress in the abrasive machining and finishing of additively manufactured metal parts. Materials, 18(6), 1249. https://doi.org/10.3390/ma18061249

Liu, Y., Zhao, L., Huang, S., Meng, X., Zhang, Z., Lu, H., Zhou, S., & Dai, G. (2026). Recent progress in surface post-treatment techniques for additive manufactured alloy metal parts: A comprehensive review. Additive Manufacturing Frontiers, 200339. https://doi.org/10.1016/j.amf.2026.200339

Hassan, M., Hussain, M. A., & Jahan, M. P. (2026). Post-processing of additively manufactured metallic parts - A review. The International Journal of Advanced Manufacturing Technology. https://doi.org/10.1007/s00170-026-18615-3

Sommer, D., Truetsch, B., Esen, C., & Hellmann, R. (2025). Effect of hot isostatic pressing and sequenced heat treatment on the mechanical properties of hybrid additive manufactured Inconel 718 components. Journal of Manufacturing and Materials Processing, 9(11), 378. https://doi.org/10.3390/jmmp9110378

Kaletsch, A., Qin, S., Herzog, S., & Broeckmann, C. (2021). Influence of high initial porosity introduced by laser powder bed fusion on the fatigue strength of Inconel 718 after post-processing with hot isostatic pressing. Additive Manufacturing, 47, 102331. https://doi.org/10.1016/j.addma.2021.102331

Fody, J. M., & Lang, C. G. (2021). Hot isostatic pressing for reduction of defects in additively manufactured Inconel-718: Preliminary simulation approach and results. NASA Technical Reports Server, Document ID 20210015451. https://ntrs.nasa.gov/citations/20210015451

Baig, S., Jam, A., Beretta, S., Shao, S., & Shamsaei, N. (2025). Non-destructive detection of critical defects in additive manufacturing. Scientific Reports, 15, 6740. https://doi.org/10.1038/s41598-025-91608-6

Novelino, A. L. B., de Oliveira, D., Araújo, J. A., & Ziberov, M. (2026). Defects characterization of wall components fabricated via WAAM and fatigue strength estimation: A theoretical prediction approach based on Murakami's square root of area method. Progress in Engineering Science, 3(1), 100238. https://doi.org/10.1016/j.pes.2026.100238

He, P., Huang, Y., Wang, W., Guo, Y., Yu, Z., Chen, W., Li, B., & Wang, X. (2025). Deep learning driven prediction method for roughness and hardness of laser in-situ forging additive manufacturing Ti-6Al-4V. Virtual and Physical Prototyping, 20(1), e2569541. https://doi.org/10.1080/17452759.2025.2569541

Chapter 8 Modeling Gas-Phase Reactive Flow and Combustion Chemistry Through Physics-Informed Machine Learning (PIML/PINNs) and Operator Learning (FNO)

Introduction

When you turn on a kitchen gas stove, a blue flame appears. The fuel gas meets oxygen in the air and produces heat. This process is called combustion. Combustion is everywhere in daily life. Car engines and boilers also run by combustion.

A rocket engine produces thrust by the same principle. But the scale and conditions are very different. Inside a rocket combustion chamber, the temperature is much higher and the pressure is high. Fuel and oxygen mix and react extremely fast. Inside it, gas flows, swirls, and makes sound.

It is hard to calculate all of this at once. Chemical reactions happen in the blink of an eye. By contrast, the flow moving through the whole combustion chamber is slower. Fast changes and slow changes are mixed in one space. If a computer must track both at the same time, the amount of computation grows sharply.

Until now, engineers have solved the equations by dividing them into very small pieces. This method is called computational fluid dynamics (CFD). It is accurate, but the computational cost is high. To solve a large rocket combustion chamber in detail, a supercomputer may need to run for days. Designers have long wanted a faster tool.

Artificial intelligence helps at this point. A well-trained neural network can greatly reduce computation time. It can also fill in missing information when there are few sensors. But in combustion problems, it is dangerous if artificial intelligence gives answers as it pleases. In rockets, even a small error can lead to a major accident.

So a method emerged that puts physical laws directly into learning. This method is called physics-informed machine learning (PIML). The neural network does not merely imitate data. It is also made to obey laws such as conservation of mass and conservation of energy. A representative method is the physics-informed neural network (PINN).

This chapter covers several methods of physics-informed learning. First, it explains in words the equations that govern combustion. Next, it examines the computational difficulties created by fast chemical reactions. It then looks at a new architecture called the Kolmogorov-Arnold Network (KAN). It also introduces the Fourier neural operator (FNO), which transforms functions into functions. Finally, it concludes with a case that reconstructs the inside of a rotating detonation engine.

Here, let us use one analogy. Think of two ways to draw a map. One method is to walk across the entire land step by step and measure it. It is accurate, but it takes a long time. The other method is to measure only a few places and fill in the rest. It is fast, but if the filling follows no rules, the map becomes wrong. Physics-informed learning means putting the rules of nature into that filling process.

After reading this chapter, you will have the big picture. You will see how artificial intelligence is changing combustion calculations. You will also be able to tell what already works and what remains difficult. New tools are powerful, but they are not. They become trustworthy only when physics and verification are present together. Keeping this balance is the idea that runs through this chapter.

This book handles difficult names and numbers with care. Some recent studies have names that sound impressive but little more. Some performance numbers are not verified. So this chapter selects and explains only confirmed research. Numbers that cannot be verified are not used in the text. Early-stage research is clearly identified as early research.

8.1 The Paradigm Shift in Computational Combustion Engineering and Artificial Intelligence

What This Section Covers

This section first explains why combustion calculations are difficult. It looks at how complex the events inside a rocket combustion chamber are. It then examines the limits of traditional computational methods. It explains why artificial intelligence has appeared as a way to move beyond those limits.

This section does not use difficult equations. Instead, it translates the concepts into scenes from daily life. If you understand this section, all later sections become easier. That is because this whole chapter is ultimately about several ways to solve this problem.

The inside of a rocket combustion chamber is an extreme environment. The temperature rises to several thousand degrees. The pressure reaches hundreds of times atmospheric pressure. Fuel and oxidizer react violently in a short time. Engines that use methane and liquid oxygen are a representative example.

These reactions are far more complex than they appear. Methane does not have only one path to meet oxygen and become water and carbon dioxide. Dozens of intermediate substances appear along the way. These substances, called radicals, are highly reactive. So hundreds of elementary reactions become intertwined and proceed at the same time.

Think of radicals as runners in a relay race. The baton passes quickly from hand to hand. No runner holds the baton for long. Radicals also appear in an instant and then pass on to the next substance. This busy exchange drives combustion. If we miss that speed, we cannot draw the flame correctly.

Turbulence is added on top of this. Turbulence is a flow in which gas swirls and mixes chaotically. It is like a river turning white as it breaks against rocks. Gas inside a rocket combustion chamber mixes in this way. Turbulence makes fuel and oxygen meet quickly. But its irregular motion makes the calculation harder.

To solve this reaction precisely, space and time must be divided very finely. The thin layer where a flame stands can be thinner than a human hair. Inside it, temperature and concentration change sharply. If this thin layer is missed, the calculation goes wrong. So the grid must be very dense to capture this layer.

Let us feel this multiscale problem with numbers. The thin layer where the flame stands is tens of micrometers thick. By contrast, the total length of the combustion chamber is tens of centimeters. The size difference reaches thousands of times. The same gap appears in time. To capture this wide range all at once, the number of grid cells grows explosively.

Traditional computational methods include direct numerical simulation (DNS) and large eddy simulation (LES). Direct numerical simulation solves every vortex, even the very small ones, without leaving any out. Large eddy simulation solves only large vortices and approximates the small ones. Both are accurate, but they require large amounts of computation. If a large combustion chamber is solved in detail, it takes a long time.

Let us translate high computational cost into everyday language. It is like making one high-definition movie. If every scene is drawn in detail, it can take months. Combustion calculation is the same. Every time the conditions change, the picture must be drawn again from the beginning. Designers want to test dozens of conditions. But in practice, it is difficult to wait for days every time a condition changes.

Artificial intelligence opens a new path here. Once a neural network learns well, it can answer similar problems quickly. It becomes a lightweight surrogate model that replaces heavy computation. A surrogate model is a fast approximate model that imitates the original calculation. Designers can scan many conditions in a short time.

Think of a surrogate model as a map app. It tells you how long a route will take without actually walking it. It compares many routes in an instant. A surrogate model does the same. It estimates the result without running the full calculation. Thanks to this, many design options can be examined in a short time.

But a map app can be wrong on unfamiliar roads. A surrogate model also goes wrong when it leaves the range it has learned. So we do not simply trust its estimated answer. Before important decisions, we check again with precise calculations. A surrogate model is a compass that points the direction, not the final judgment. This attitude runs consistently through the whole chapter.

Let us explain the supercritical state a little more. Water makes it easy to understand. Usually, water boils and changes into steam. The boundary between liquid and gas is clear. But when pressure and temperature become very high, that boundary becomes blurred. Fuel inside the high-pressure combustion chamber of a rocket is in this kind of state. Accurately calculating this blurred boundary between liquid and gas is difficult.

There are real cases as well. DeepFlame is software that uses neural networks to calculate supercritical combustion quickly. Supercritical means a high-pressure state in which the boundary between liquid and gas becomes blurred. Researchers optimized this code and solved turbulent combustion of methane and oxygen at a large scale. According to public materials, they increased the computational scale to the level of tens of billions of cells. This is a much wider range than before. This result was presented as early research at a high-performance computing conference in 2025.

The role of artificial intelligence in this case is clear. Chemical reaction calculation is a time-consuming part. A neural network quickly replaces this calculation. At the same time, it preserves the properties of the real fluid. It is an attempt to gain both speed and accuracy. This direction advances large-scale rocket combustion calculations.

Let us consider why large-scale calculation matters. A real rocket engine has hundreds of injectors. An injector is a part that sprays and mixes fuel and oxygen. The flames from these many injectors affect one another. Small-scale calculations cannot easily capture this interaction. Large-scale calculation must become possible to get closer to a real engine.

Artificial intelligence enters combustion in three main ways. One is a fast surrogate model that replaces time-consuming calculations. Another is reconstruction, which restores the whole field from sparse measurements. The third is real-time diagnosis and control of the state. This chapter focuses on the first two methods, and the third control method is covered in a later chapter.

Let us take an example of why it is dangerous when artificial intelligence violates physics. Suppose a model gives a temperature lower than the real value. The designer judges that the wall can withstand it. But if the real temperature is higher, the wall melts. A small error leads to a major accident. That is why obeying physics is directly tied to safety in rockets.

Of course, the limits are also clear. Artificial intelligence often fails outside the range it has learned. Rockets constantly encounter new conditions. If physical laws are not obeyed, predictions become dangerous. So this whole chapter deals with ways to combine physics and learning. In 2026, a review also appeared on physics-informed learning for aerospace propulsion. This trend shows that the field is growing quickly.

8.2 Governing Equations for Multicomponent Compressible Reactive Flow and Physics-Informed Learning

What This Section Covers

This section explains in words the basic equations that govern combustion. The word equation may sound difficult. But the meaning is simple. It is the rule that nothing simply disappears or appears from nowhere.

Next, it looks at how to put this rule into a neural network. It examines how a physics-informed neural network obeys laws. These two concepts are the backbone of this chapter. All the methods that appear later branch out from these two concepts.

8.2.1 Governing Equations of the Reactive Navier-Stokes System

The Navier-Stokes equations are rules that describe the motion of fluids. They are used to solve phenomena such as flowing water and blowing air. In rocket combustion, chemical reactions and heat are added to them. So the number of equations that must be solved together increases.

For an ordinary fluid without reactions, three balances are enough: mass, momentum, and energy. But a reacting fluid such as combustion needs one more balance, species conservation. So these equations contain four conservation laws. The first is conservation of mass. The amount of gas entering and leaving must match. Gas does not appear or disappear by itself. This balance, in which mass adds up correctly, becomes the basis of every calculation.

Next is conservation of momentum. Momentum is the quantity obtained by multiplying mass by velocity. When a force is applied, momentum changes. Pressure differences and viscosity create this force. That is why gas is pushed out rapidly inside a rocket.

Think of conservation of momentum with a bicycle. When you press the pedals, the bicycle moves forward. The pedaling force creates a change in speed. Inside a rocket, the pressure difference becomes that force. Viscosity is friction that holds flows together and slows them down. These two forces, pressure difference and viscosity, change the velocity of the gas.

Another is species conservation. A species is a substance that participates in a reaction, such as oxygen, methane, or water. When a reaction occurs, some substances decrease and others increase. The reaction rate determines how much they increase or decrease. A reacting fluid sets up this conservation equation separately for each substance. So equations are added according to the number of species. The last is conservation of energy. The heat produced by reactions warms the gas and pushes the flow outward.

Conservation of energy can be compared to a piggy bank. If you match the money put in with the money spent, the remaining money is known. In combustion, chemical reactions do not create new energy. The chemical energy trapped inside the fuel simply changes into heat. That heat warms the gas and does work as it leaves. Some heat also escapes as radiation. Conservation of energy is the rule that balances all these inflows and outflows.

The word compressible is also important. Compressible means that the density of a gas changes greatly with pressure. A rocket combustion chamber has high pressure and large changes. So density cannot be

fixed. This continual change in density makes the calculation harder.

Let us look at why these four balances matter so much. A calculation that breaks balance soon becomes false. If mass increases by itself, it is as if thrust has appeared from nowhere. If energy leaks away, the temperature prediction goes wrong. In rocket design, this kind of error can lead straight to an accident. So artificial intelligence must also keep this balance.

The chemical reaction rate is set by the Arrhenius law. The Arrhenius law is the rule that a reaction becomes sharply faster as temperature rises. It resembles the way a match flame spreads in an instant once it catches. This sharp behavior becomes the root of the calculation problem we will see later. Rocket designers must predict this reaction precisely to keep systems safe.

Diffusion and heat transfer must also be solved together. Diffusion is the phenomenon in which a substance spreads from a place of high concentration to a place of low concentration. It is like a smell spreading from the kitchen into a room. Heat transfer is the phenomenon in which heat moves from a hot place to a cold place. In a combustion chamber, these two occur at the same time as chemical reactions. So the equations are intertwined and solved as one system.

8.2.2 Physics-Informed Composite Loss Functions

The core of a physics-informed neural network lies in the loss function. A loss function is a measure of how wrong the neural network is. An ordinary neural network usually measures only the difference from data. A physics-informed neural network adds physical error to that.

This is how it works. The neural network receives position and time as inputs. It then outputs the pressure, temperature, velocity, and concentration at that point. These values are correct only if they satisfy the conservation equations. If they do not satisfy the equations, they receive a penalty by that amount.

To measure this physical error, derivatives are needed. This is where automatic differentiation is used. Automatic differentiation is a technique that obtains derivatives by following the calculation process of the neural network. It differs from finite differences, which approximate derivatives from grid differences. So it has no truncation error of the kind that occurs in finite differences. However, approximation error in the neural network itself and error from choosing sample points remain. What automatic differentiation removes is the error that arises when derivatives are approximated.

Let us look at why derivatives are needed. Conservation equations deal with rates of change. They examine how quickly pressure changes and how velocity changes. This rate of change is the derivative. The change in the values produced by the neural network is obtained through differentiation. Then we check whether the resulting rate of change satisfies the equations.

Traditional computation approximated derivatives on a grid. It estimated change from differences between neighboring grid points. In this method, the error grows when the grid is coarse. Automatic differentiation does not use this approximation. It differentiates the neural network's formula directly. So there is no need to lay down a separate fine grid for differentiation. However, the points where physical error is measured still must be chosen. These points are called collocation points. If the collocation points are too sparse, the solution becomes unstable. Being free from a grid does not mean being accurate everywhere.

A loss function contains several terms together. The difference from data is one term. The errors in the mass, momentum, energy, and species equations are each separate terms. Boundary conditions and the initial state are also included as terms. Training weighs these several terms and reduces them together.

This creates a major advantage. Even when data is limited, the laws of physics fill the gaps. In a laboratory, only wall pressure sensors or optical probes may be used. Such sparse measurements alone cannot reveal the entire flow. If physical laws are set as rules, the spaces between measurements can be filled.

This property shines in inverse problems. Ordinary computation takes causes as inputs and obtains results. An inverse problem looks at results and finds causes. One example is finding the internal heat source from wall temperature alone. Physics-informed neural networks are strong at such inverse problems. This is because they handle measurements and physics together in one loss function.

This ability is valuable in rocket tests. It is difficult to place many sensors inside a combustion chamber. The chamber is hot and narrow, so sensors cannot withstand it. For that reason, only a few points are measured from the outside. The key task is to recover the internal state from these few points. A physics-informed neural network builds a bridge between the measurements and the internal state.

This loss function contains the difficulty of weighting. The problem is how to set the importance of each term. If too much weight is given to one term, the other terms are ignored. If the model follows only physical error, it misses the data. If it follows only the data, it violates physics. How to strike this balance is a major research topic.

We can compare this weighting to seasoning food. If the salt is too strong, the food is salty. If the sugar is too strong, it is sweet. If any one element goes too far, the taste breaks down. The several terms in a loss function are the same. If one term is too strong, training leans to one side. So methods that adjust the weights by themselves are also being studied.

The study that first made this concept widely known was the 2019 paper by Raissi and others. Since then, it has spread widely into the combustion field. In 2025, a comprehensive review also appeared that organized physics-informed learning for combustion. This review divides the field into chemical reactions, reacting flows, and other combustion problems. It shows that physics-informed learning has become a major current.

We must also look honestly at the limits. A physics-informed neural network is good at finding a solution tailored to one condition. But when the condition changes, it often must be trained again. It is still difficult to cover a wide range of conditions at once. Training can also become unstable near strong shock waves or sharp flame fronts. Research continues to fill these gaps.

8.3 Stiff Reaction Kinetics and Conservation-Constrained Physics Learning

8.3.1 Numerical Stiffness in Chemical Reactions and the Collapse of Standard Learning

Stiffness occurs when some changes are very fast and others are slow. This situation is common in combustion chemistry. Intermediate substances such as radicals appear and disappear in an instant. By contrast, temperature and main components change relatively slowly.

Let us look at how large this time difference is. Fast reactions proceed on the scale of one billionth of a second. Slow flows move on the scale of one thousandth of a second. The difference between the two scales exceeds a million times. This large gap becomes the root of stiffness that destabilizes computation.

Problems arise when fast and slow changes are solved with the same interval. If the interval is matched to the fast reaction, the calculation becomes endlessly long. If it is matched to the slow flow, the fast reaction is missed and the calculation diverges. Traditional numerical analysis has struggled with this point for a long time. So traditional numerical analysis has handled it with special time integration methods.

This time gap resembles photography. Suppose we try to capture a hummingbird's wingbeats and a turtle's steps in one frame. If we match the shutter to the hummingbird, we must press it countless times. If we match it to the turtle, the hummingbird becomes blurred. It is hard to capture both in one frame. The fast reactions and slow flows of combustion are the same.

A standard physics-informed neural network often collapses here. This neural network tries to reduce all errors at once. The large error created by fast reactions dominates training. Then it cannot properly learn the balance of slow variables. This phenomenon, in which training is pulled to one side, is called gradient pathology.

Let us picture this pathology. Suppose one loud voice takes over a meeting. The words of quieter people are drowned out. The meeting leans in only one direction. The large error from fast reactions is that loud voice. So the balance of slow variables is pushed to the background in training.

Two further causes are hidden here. One is the tendency of neural networks to learn abrupt changes late. Neural networks first learn smooth flows and put off sharp changes until later. This property is called spectral bias. The steep change at the moment of ignition is exactly this sharp part. The other is the failure to preserve temporal causality. In combustion, an earlier moment causes a later moment. But standard training tries to fit all times at once. Then the order of cause and effect becomes blurred. In severe cases, the solution collapses into a flat result in which no reaction occurs. For this reason, methods have emerged that divide time and learn it step by step, or assign different weights to different species.

Another problem is the collapse of conservation laws. When several terms are weighted at the same time, balance can break. The sum of species concentrations must always remain constant. The number of atoms must also be the same before and after a reaction. If these rules are broken, nonsensical values such as negative concentrations appear.

Let us look at why negative concentration is a serious problem. In reality, the amount of a substance cannot be negative. Water cannot be minus one cup. If a calculation produces such a value, the next step breaks down. The reaction rate calculation either runs out of control or stops. So concentration must be prevented from falling below zero.

Several early studies dealt with this problem. Stiff-PINN by Ji and others in 2021 is a representative example. This study added special treatment to handle stiff chemistry. Multiscale physics-informed neural networks appeared around the same time. These studies confirmed how large a barrier stiffness is for training.

There is a reason this barrier becomes higher in rockets. The combustion of methane and oxygen has many reaction pathways. Dozens of substances must be handled. As the number of substances increases, stiffness grows with it. So rocket combustion sits at the difficult end of the stiffness problem. Methods that impose physics strongly are therefore all the more urgent.

8.3.2 Combining Conservation Constraints and Time Integration

The key to solving stiffness problems is to impose physics more firmly. There are two main ways to impose physics. One is a soft constraint, which gives a penalty when a rule is broken. Adding physical error terms to the loss function belongs to this approach. The other is a hard constraint, which builds the rule into the structure so it cannot be broken at all. The more severe the stiffness, the more these two methods are used together.

The first tool is to connect a verified time integrator to the neural network. The Runge-Kutta method is a standard technique that solves differential equations by dividing time into small steps. But the type

must be chosen carefully. Explicit methods, which push values forward directly from the previous step, are weak for stiff problems. To solve them stably, the time interval must be split excessively finely. So stable integrators in the implicit family are used for stiff problems. This stable integration structure, whose stability has been verified, is placed inside the neural network. Then deep learning inherits the stability built up over many years of numerical analysis.

The second tool is projection that enforces conservation. The concentrations produced by the neural network are not used as they are. First, the values are converted so they are always positive. Next, they are adjusted again so the sum remains constant. Finally, the values are adjusted so the number of atoms is conserved. After this process, values that violate physics cannot appear at all.

We can compare projection to a gatekeeper. Whatever value the neural network produces is checked at the gate. If it does not meet the rules, it is corrected to a nearby valid value. Negative values are turned into positive values. If the sum is off, it is adjusted again. So the values that go out always follow the rules.

This approach differs from the soft approach. The soft approach only gives a penalty when a rule is broken. A penalty guides training, but it does not prevent violations. The hard approach blocks the violation itself through the structure. Even if training is imperfect, physics is preserved. So it suits rockets, where safety matters above all.

A real example is CRK-PINN. This study solved combustion reaction rate equations with a physics-informed neural network. The authors are Zhang Shihong, Zhang Qi, and Wang Bosen. The paper passed peer review and was published in the journal Combustion and Flame. This method uses soft constraints, which put physical laws into loss function terms. It included the law of mass action and the Arrhenius law in the loss. Enthalpy conservation, element conservation, and concentration-sum conservation were imposed in the same way. Unlike the hard projection described above, it guides physics through penalties.

The performance is also public. The calculation became faster in a hydrogen-air reaction system. It handled ten species and twenty-one reactions. It was about six to fourteen times faster than standard computation. When connected to actual reacting flow computation, it was still about two to five times faster. This is a case that gained computation speed while preserving physical laws.

Let us look at what it means to become several times faster. Chemical reaction calculation is a large part of combustion analysis that takes a long time. A large share of the total computation time is spent here. If this part becomes ten times faster, the whole calculation is greatly reduced. A calculation that took several days can shrink to one day. The cycle of revising the design and testing it again becomes that much faster.

It is worth clarifying here so that soft constraints and hard constraints are not mixed up. CRK-PINN is a soft method that guides physics through the loss. It is different in nature from hard projection, which forcibly corrects values. The two methods can be used together or separately. Soft constraints are flexible, but they do not completely prevent violations of physics. Hard constraints prevent violations, but the structure becomes more complex. How much of each to use is a design judgment.

The lesson from this method is simple and clear. It does not discard verified knowledge from numerical analysis. Methods such as Runge-Kutta have been refined for decades. Artificial intelligence grows stronger when it is placed on top of this knowledge. Old physics knowledge and new learning do not oppose each other. They help each other.

The limits are also clear. These methods have been verified within the reaction systems they handled. Methane and oxygen combustion in rockets is more complex. The number of substances and reactions is

much larger. To move them to a new reaction system, they must be trained and verified again. There is also a trade-off. The stronger the physics constraints are, the heavier the computation becomes.

8.4 Reaction Flow Modeling Based on Kolmogorov-Arnold Networks

8.4.1 Basics of Placing Functions on Edges and Adaptive Activation Functions

First, think about the structure of an ordinary neural network. A neural network is made of nodes and the edges that connect them. A fixed activation function is placed at each node. Only a number called a weight is attached to each edge. Learning means adjusting these weight numbers.

Let us first explain the term activation function. A neural network receives an input and sends it out through a certain curve. This curve is the activation function. An ordinary neural network fixes the shape of this curve. It overlaps many fixed curves to make complex relationships. Learning keeps the shape of the curves unchanged and changes only the strength of the overlap.

A Kolmogorov-Arnold network reverses this arrangement. It places functions on the edges, not on the nodes. It changes the functions themselves through learning. It does not use fixed shapes. It shapes the curves flexibly. This idea comes from a mathematical theorem proved long ago.

That theorem is the Kolmogorov-Arnold representation theorem. It says that a complex function can be expressed as a sum of simple one-dimensional functions. This structure represents the function on each edge as a flexible curve. This curve can change shape locally. So it can refine only the places where change is sharp, such as a flame.

Let us translate this theorem into block building. No matter how complex a shape is, it is a combination of a few basic blocks. A large structure is divided into small pieces and connected again. A complex function can also be broken into a sum of simple one-dimensional pieces. A Kolmogorov-Arnold network uses this decomposition as its structure. Old mathematics has become the foundation for a new neural network.

The material used to make these curves is the B-spline. A B-spline is a method that smoothly connects several curve pieces. It belongs to the same family as the tools car designers use to draw smooth curves. Each piece affects only its own interval. So changing one place does not disturb other places.

This locality gives learning an advantage. Only the narrow interval where the flame stands can be divided finely. The remaining wide intervals can be left coarse. Resources are concentrated where they are needed. This is called grid refinement. It saves computation while obtaining the needed precision.

We can see why this property is useful. It is easier to see which input affects which output. Only the needed interval can be expressed precisely. It is also easier to inspect and examine the function visually. It suits problems such as combustion, where local changes are severe.

Let us turn the difference between nodes and edges into an everyday example. An ordinary neural network is like stacking several lenses with fixed shapes. The lens shapes are fixed, and only the strength of overlap is adjusted. A Kolmogorov-Arnold network carves the lens shapes themselves. So it can create the desired refraction more freely. But carving the lenses takes more work.

Of course, this structure is not. Learning can be more difficult. The more finely a curve is divided, the more values must be adjusted. Depending on the problem, an ordinary neural network may be better. It is right to treat the new structure as one option.

8.4.2 Flame Surface Expressive Power and Parameter Reduction

A flame surface is a thin region where reactions are concentrated. In this region, temperature rises sharply over a short distance. Gas at room temperature rises to thousands of degrees in an instant. Concentrations also change abruptly. It is difficult to express the steep changes in this narrow interval through computation.

Ordinary neural networks are weak at this kind of rapid change. With fixed activation functions, it is hard to draw a sharp layer smoothly. If the model is forced to fit it, ripple-like errors appear around it. They also tend to learn high-frequency information slowly. Because of these properties, flame surface calculation results become blurred.

Let us compare these ripples to a flag drawing. Suppose we draw a straight pole only with a soft brush. Wavy patterns appear next to the sharp edge of the pole. In computation, false patterns also appear next to a sharp flame surface. These patterns are fake. They do not exist in reality. Such fake patterns blur predictions as signals that do not exist in reality.

A Kolmogorov-Arnold network can show strength at this point. It draws steep layers better with flexible curves. It can choose only the narrow interval where the flame stands and divide it finely. It leaves unnecessary places coarse and reduces waste. This adaptive ability to shift resources fits combustion calculation well.

A real example is ChemKANs. The authors are König, Kim, and Deng. This study was published in 2025 in the journal Physical Chemistry Chemical Physics. It embedded reaction rates and the flow of thermodynamic laws in a Kolmogorov-Arnold network. The goal was to learn combustion chemistry models and speed up computation. However, what this study verified was not a flame surface spread through space. It learned a chemical reactor problem in which substances react over time. Restoring spatial flame surfaces or shock waves in a rocket combustion chamber remains a future task.

The published results are practical. This model worked with a very small number of adjustable values. It learned well without overfitting, even from noisy data. It withstood noise of up to fifteen percent. It also showed results about twice as fast as standard computation. The same research team also released a follow-up early study on learning reaction rates according to pressure.

Let us explain what it means to be robust to noise. Real measurements always contain noise. This is because sensors are not perfect. A model that is weak against noise memorizes even this noise. This is called overfitting. A model that is robust to noise learns only the real rules.

Let us compare overfitting to studying for an exam. A student who memorizes only the answers gets the question wrong when it changes. A student who understands the principle can solve new questions. A model that is robust to noise is like a student who has learned the principle. Rocket measurement data contains a lot of noise. So this robustness to noise is valuable in rocket measurement.

The strengths of this model must be stated within the verified range. Three things have been verified. It has few adjustable values. It is robust to noise. It is about twice as fast as standard computation. Numbers that claim greater performance than this still need verification. So this book states only this verified range.

What does a Kolmogorov-Arnold network mean for rockets? Flame surfaces also stand as thin layers inside rocket combustion chambers. This thin layer must be drawn well to predict heat and pressure correctly. Fewer parameters make computation lighter. A light model is good for real-time computation. So this structure is attracting interest as a candidate for real-time diagnosis.

Let us consider the weight of saying that there are few parameters. Parameters are the number of values a model adjusts. If this number is large, learning requires a lot of data. Computation is also heavy,

and memory use is large. Doing the same work with fewer parameters is beneficial in many ways. It can also run on a small computer carried on a rocket.

The limits must also be seen. These cases were verified in relatively simple reaction systems. If the number of chemical species grows to dozens, the number of edges increases explosively. Then memory and training time can grow greatly. For this reason, research continues to make the structure lighter. High-pressure turbulent combustion in rockets is far more complex. There is still a long way to go before this structure can be used in actual engine design. Public code, comparative experiments, and independent verification are still needed. For now, it is right to treat it as a promising candidate structure.

8.5 Fourier Neural Operators That Transform Functions into Functions

8.5.1 Principles of Operator Learning and Grid Independence

An operator is a rule that goes from a function to a function. The wording is difficult, so let us change it into an example. Suppose we input the flow-rate distribution at the combustion chamber inlet. Then the temperature distribution of the entire combustion chamber comes out. The rule that changes this inlet picture into a temperature picture is an operator.

Let us compare the mapping from function to function to cooking. Suppose we input the entire list of ingredients, and a photo of the finished dish comes out. If the ingredients change a little, the dish changes with them. Operator learning learns this whole transformation. It learns the entire relationship, not the value of each ingredient. So it gives answers quickly even when conditions change.

An ordinary neural network changes a few numbers into a few numbers. Operator learning changes one picture into another picture. So it handles the entire flow field at once. DeepONet is one study that first broadened this idea. The Fourier neural operator is also a method that belongs to this operator learning.

Let us look at why this difference is large. It is slow to compute again for every condition. Once an operator is learned, it answers many conditions quickly. Even if the inlet condition changes, it does not need to learn again. In design, conditions are changed countless times. So operator learning is drawing attention as a design tool.

The core of the Fourier neural operator lies in the Fourier transform. The Fourier transform is a technique that changes a complex shape into a sum of many waves. Low-frequency waves contain the large structures of the flow. High-frequency waves contain details such as fine patterns. This neural network mainly learns low-frequency components and quickly draws the large flow.

This method has its own advantages. The neural network treats the flow as a sum of waves. It quickly captures the large flow by summarizing it with a few low frequencies. It is useful for seeing relationships across a wide region at a glance. However, this method is not a numerical method that directly solves the derivatives in the equations. Keeping only low frequencies and cutting high frequencies makes computation lighter. But that cut can instead blur sharp parts such as shock waves or flame surfaces. So care is needed where details are important.

Let us compare this blurring to a photograph. A photo that is out of focus has blurry boundaries. Sharp lines look smeared. The more high frequencies are cut, the more this kind of smearing appears. The sharp boundary of a flame can be crushed. So these places need separate processing that preserves high frequencies.

A notable property of this method is grid independence. This model is defined regardless of how fine the computational grid is. It can learn from low-resolution data and then be applied to a finer grid. This is called zero-shot super-resolution. It resembles restoring a low-resolution photo into high resolution. However, this is a theoretical property and a result shown in specific experiments. Accuracy at a new resolution is not automatically guaranteed.

Ordinary neural networks cannot do this kind of work. The sizes of the input and output are fixed from the beginning. If the grid changes, the model must be made again. The Fourier neural operator is different. It treats the flow as a function, not as a grid. So the same model is used even when the resolution changes.

Let us compare the Fourier transform to radio. Radio receives many broadcasts by separating them by frequency. Low frequencies and high frequencies contain different information. The Fourier neural operator also divides the flow by frequency. The large flow is contained in low frequencies, and fine patterns are contained in high frequencies. This neural network selects the low-frequency side and quickly captures the large flow.

Grid independence gives rocket computation a large advantage. In the early stage of development, many options are quickly reviewed with a coarse grid. After a good option is chosen, it is checked precisely with a fine grid. At this time, the model does not need to learn again. The same model is used as it is at both resolutions, so time and cost are reduced together.

However, this does not happen automatically for every problem. Places where values break abruptly, such as shock waves, are difficult. Thin flame surfaces and turbulence near walls also require care. In these places, high-frequency information is important. Separate processing and verification must follow before the results can be trusted.

8.5.2 Neural Operators for Combustion Flow

Combustion is a difficult problem in which heat, flow, and chemistry are intertwined. Research continues to adapt neural operators to this problem. Several teams are trying different methods. Here, we look at verified recent cases mainly through concepts.

One direction is to learn stiff chemistry as an operator. The extended Fourier neural operator (EFNO) is an example. The authors are Wong, Li, Zhang, Chen, and Zhou. This study tried to solve stiff chemistry even under unseen conditions. It was tested on zero-dimensional hydrogen autoignition and a three-dimensional hydrogen-ammonia jet flame. It showed the ability to generalize well even in untrained states. This extended operator study was published in the journal Combustion and Flame.

Let us unpack the term autoignition. When fuel and oxygen are kept hot, they ignite on their own. They burn without a match. The chemistry at this moment is very fast and highly stiff. The extension operator tried to learn this difficult moment. This problem is often used as a test bed for stiff chemistry.

Let us look at why the extension operator matters. Ordinary neural networks work well only under the conditions they learned. They can go wrong when temperature or pressure shifts only slightly. The extension operator tried to lower this wall. In tests, it generalized to some degree even to conditions it had not learned. This showed room for use under new rocket conditions.

Another direction is to learn three-dimensional reacting turbulence as a whole. One study handled this with a Fourier neural operator that added a U-Net structure. A U-Net is a neural network structure that handles the big picture and details together. It reproduced the frequency characteristics of velocity, temperature, and density in compressible reacting turbulence well. It predicted the flow much faster than conventional computation. These results show the potential of surrogate models. Many studies in

this line are still at an early stage before peer review.

Let us also unpack what it means to match frequency characteristics well. A flow contains both large vortices and small vortices. Vortices of each size contain energy. A key measure is whether this energy distribution is drawn correctly. If only the large flow is right and the small vortices are missed, the result is incomplete. A good operator broadly reproduces this distribution across both large flows and small vortices.

A third direction is to handle radiative heat transfer with an operator. A hot flame also loses heat in the form of light. This is called radiative heat transfer. Zhao and others proposed a neural operator for solving radiative heat transfer in fires. This was an early study on fires, not rockets. So it is used here only as a neighboring example, not as rocket validation.

Let us understand radiative heat transfer through a campfire. When you stand in front of a campfire, your face feels hot. Light carries heat before the air warms up. Flames inside a rocket combustion chamber also release heat as light in this way. This radiation heats the wall and cools the flame. So accurate prediction must include this radiative heat as well.

It is difficult to include radiation in computation. Light spreads in all directions and mixes with itself. Its intensity differs by direction. An attempt to learn this complex flow with an operator appeared in fire research. To transfer it to rockets, the conditions must be adjusted and verified again. Even so, it is a meaningful early case that shows a future direction.

These cases show a common benefit. When a model learns high-fidelity computation results, it becomes a fast surrogate model. It can restore low-resolution computation into a high-resolution form. It becomes an auxiliary tool for estimating the whole field when there are few sensors. It can be used to scan the design space in a short time.

There is also a point where operator learning meets physics-informed learning. A neural operator is fast, but it does not always obey physics. Conservation constraints can be added to its loss. This is an attempt to obtain both the speed of an operator and the reliability of physics. This combination appears again later in the rotating detonation case. So these two tools do not compete with each other. They work as a pair.

Here, let us clarify how to read these cases. This book does not provide a performance comparison table for specific models. Numbers from early studies can vary greatly depending on conditions. Instead, it explains the role of neural operators through verified cases. The focus is on how the method is used, not on the numbers.

There is also a case of large-scale computation. The DeepFlame research team solved supercritical combustion at large scale by including a neural network. The authors are Guo, Mao, Liu, Tan, Zha, and Chen. According to public materials, they computed turbulent combustion of methane and oxygen at the scale of tens of billions of cells. This range greatly exceeded previous capabilities. It is early research toward rocket combustion with more than one hundred injectors.

Many checks are needed before use in actual design. We must see whether mass and energy conservation are maintained. We must confirm whether shock wave positions are correct. We must check whether chemical species concentrations fall below zero. Only after passing this verification does the model become a tool. A surrogate model is a partner that helps high-fidelity computation, not a replacement.

8.6 Reconstruction of Shock Wave-Combustion Interfaces in Rotating Detonation Engines

8.6.1 Analysis Challenges for Rotating Detonation Engines

A rotating detonation engine (RDRE) uses an annular combustion chamber. Inside it, a detonation wave rapidly circles around. Detonation is explosive combustion in which a shock wave and combustion are attached and move together. This wave runs at thousands of meters per second, several times faster than sound.

This method has the advantage that combustion itself raises pressure. Ordinary engines raise pressure with a compressor or pump and then burn the mixture. In detonation, a shock wave instantly compresses the gas and burns it. So the pressure rises by itself during burning. This is called pressure-gain combustion. However, injection pressure is still needed to push in fuel and oxygen. Even if pressure rises inside the combustor, losses occur in injection, mixing, and exhaust. So the overall engine benefit depends on the conditions. If used well, it may raise efficiency, and many organizations around the world have entered its development.

Let us look at the difference between detonation and ordinary combustion. Ordinary combustion spreads slowly, like a flame moving outward. Detonation spreads like an explosion led by a shock wave. The shock wave compresses and heats the air in an instant. Fuel burns immediately behind it. This structure, in which the shock wave and flame are attached, creates high pressure.

The problem is that the interior is extremely complex. The leading shock wave and the heat-release reaction zone are attached to each other. Oblique shock waves, contact surfaces, and unburned mixing layers become tangled. These structures interact on the scale of millionths of a second. That is why this is among the hardest classes of problems in fluid dynamics.

Let us compare this interaction to an accident on a road. Several cars become tangled at a narrow intersection at the same time. A change in one place spreads through the whole scene in an instant. Inside detonation, shock waves and flames become tangled in this way. A small disturbance changes the whole wave. That is why it is difficult to predict and control the inside of this engine.

Measurement is even more frustrating. What can be seen in a ground test is limited. The pressure signal on the outside wall of the combustion chamber is almost all there is. Sometimes the light from the flame is filmed with a high-speed camera. This light is called self-luminescence. It is light emitted by certain molecules during reaction. But these measurements cannot directly reveal the internal three-dimensional state.

We should also know why rotating detonation engines are being developed. Ordinary rocket engines burn fuel at a nearly constant pressure. Detonation combustion burns while raising pressure by itself. It may produce greater force from the same fuel. It may also make the engine shorter and lighter. That is why this method is considered a candidate for next-generation propulsion technology.

Let us compare the difficulty of this engine to a camera. Suppose you take a picture of a fast-spinning top. If the shutter is slow, the top blurs. A detonation wave is much faster than a spinning top. Even the fastest camera captures only the outside appearance. To look inside, the empty parts must be filled in by computation.

Let us feel the speed of a detonation wave in numbers. This wave runs at thousands of meters per second. That is faster than a bullet. It takes less than a millisecond to circle an annular combustion chamber once. In that short time, pressure and temperature rise and fall sharply. The change is so fast that the human eye cannot follow it.

That is why reconstruction models are urgently needed. A tool is needed to recover the whole flow field from sparse measurements. Pressure, temperature, velocity, and chemical species distributions must be

estimated in a short time. This is where physics-informed learning and operator learning meet. The methods learned in this chapter join forces here.

8.6.2 Reconstruction and Digital Twins Combining Data and Physics

The first tool for reconstruction is super-resolution. It restores sparse or coarse measurements into a detailed field. A 2026 study by Yao and Zhang attempted this with a cycle-consistent generative adversarial network. A generative adversarial network is a structure in which a generator and a discriminator compete and learn. It resembles a counterfeiter and an inspector training each other. This method learned from engine computation data for hydrogen and air. It showed that a coarse flow field could be reconstructed at eight times finer detail. This result is a study published in an academic journal.

Let us see why super-resolution is needed through an image. When an old photograph is enlarged, it becomes rough like stairs. Super-resolution technology restores this rough photograph smoothly. Combustion computation also produces coarse grid results. If this coarse field is restored into a detailed field, information increases. It is a way to save computation cost with coarse computation while recovering details.

The next tool is multi-fidelity fusion. It connects cheap low-fidelity computation with expensive high-fidelity computation. The idea is to obtain accurate estimates even with a small amount of high-fidelity data. A study called Cheap2Rich applied this to rotating detonation engines. The authors are Bao, Jayak, Powers, Raman, and Kurtz. This multi-fidelity study is still early research before peer review.

Let us compare the advantage of multi-fidelity to household budgeting. If only expensive materials are used, the household budget becomes tight. Cheap materials set the base, and good materials are used only where they matter. Computation is the same. Low-fidelity computation sets the broad frame. Only important places are refined with high-fidelity computation. This is a way to approach high accuracy at low cost.

This fusion also includes the concept of data assimilation. Data assimilation is a method that keeps aligning computation with real measurements. It is like a weather forecast correcting its prediction with observations. Sensor signals also enter a rotating detonation engine. These signals continuously correct the computation. Then the model prediction moves closer to the actual engine state.

Another tool is a generative model that reflects observations. A generative model is a neural network that creates plausible scenes by itself. If sparse observations are given as conditions, it draws the whole field. Even if observation locations change, there is room to use it without retraining. Wang and others proposed such an attempt for reconstruction in a rotating detonation combustor. So it is useful when there are few sensors and observations are insufficient.

The last tool is the combination of neural operators and physics-informed neural networks. A neural operator quickly approximates the whole field. A physics-informed neural network holds it to conservation laws. When the two methods are combined, the result is an estimator that is both fast and consistent with physics. When sensor signals come in, it reconstructs the interior in a short time.

These reconstruction results lead to a digital twin. A digital twin is a technology in which a computer model follows the state of a real device. From the reconstructed flow field, wall heat load and combustion efficiency can be calculated. The risk of an injector being eroded by heat can be detected in advance. It also connects to control that manages combustion instability in a short time.

Let us compare combustion instability to a swing. If a swing is pushed in rhythm, it moves higher and higher. Pressure oscillations inside a combustion chamber also grow when they match the rhythm of

the flame. If this oscillation grows, the engine can break. So the oscillation must be detected quickly and its rhythm must be broken. Real-time reconstruction and control take on the work of detecting and managing this oscillation.

However, there are many points to be careful about here as well. Most of these reconstruction studies are at an early stage. Numbers that emphasize large speedups or high accuracy are tied to conditions. So this book does not use such numbers in the main text. This part is explained only through design concepts and verified cases.

Let us examine why a digital twin is useful. The inside of a real engine cannot be seen directly during a test. A digital twin estimates the internal state from sensor signals. It detects danger signs faster than a person can. It can send a signal before the problem grows. In reusable rockets, this early warning is valuable.

Self-updating also has hidden risks. If a model corrects itself with its own predictions, errors can accumulate. It can push itself in the wrong direction. One study in 2026 pointed out this problem. It proposed a self-updating digital twin stabilized by physics. This study showed the concept with a general thermal problem, not a rocket. It is a method in which physical rules hold the reins of updating.

Let us look at why limiters are needed. Artificial intelligence sometimes produces strange values. If those values enter control as they are, it is dangerous. A limiter prevents values from exceeding the safe range. It is a fence that keeps the physical limits set by people. Such fences are necessary for autonomous control.

More preparation is needed before use in real engine control. Enough high-speed test data must be accumulated. Independent verification must follow. Limiters that protect safety must also be in place. This flow shows that the technology is still maturing. Still, the direction is clear. Reconstruction that uses data and physics together is taking hold.

Summary

This chapter covered several ways to bring artificial intelligence into combustion computation. The starting point was one problem. Rocket combustion requires heavy computation because fast chemistry and slow flow are intertwined. Traditional computation is accurate but slow. Artificial intelligence is fast, but it becomes risky if it violates physics.

The answer is physics-informed machine learning. The neural network does not merely imitate data. It also follows laws such as conservation of mass and energy. The physics-informed neural network is the leading method. Even when data is scarce, physics fills in the blanks. This strength is especially clear in inverse problems.

At this point, let us sort out the relationship between artificial intelligence and physics. Artificial intelligence is strong at finding rules in data. Physics contains laws that have long been tested. Either one alone is not enough. If we use only data, physics is violated. If we use only physics, computation becomes heavy. Using the two together is therefore the core of this chapter.

In stiff problems, physics must be imposed more strongly. Penalties alone are often not enough. A time integrator is built into the neural network to gain stability. Projection that enforces conservation prevents impossible values. CRK-PINN is an example of a soft constraint that imposes physics through the loss.

Let us also review the equations of multicomponent reacting flow. The four balances of mass, momentum, species, and energy form the framework. These balances state that nothing simply appears

or disappears. Physics-informed learning puts this rule into the loss function. Automatic differentiation measures the equation residuals accurately. As a result, it can reconstruct the whole even from sparse data.

Kolmogorov-Arnold Networks present a new structure. They place flexible functions on the connections. They can describe steep flame fronts better. ChemKANs showed this possibility in combustion chemistry. They produced noise-robust results even with a small number of values.

Fourier neural operators turn functions into functions. They can learn at low resolution and be used at high resolution. They become fast surrogate models and reconstruction tools. Studies combining extended operators and U-Nets showed results in reacting flow. Operator studies dealing with radiative heat transfer have also appeared in nearby fields. However, shock waves and thin flame fronts still require caution.

The rotating detonation engine is a stage where these methods come together. The goal is reconstruction that recovers the interior from sparse measurements. Super-resolution, multifidelity methods, and combinations of operators and physics neural networks are used. The results lead to digital twins and real-time control. The field is still at an early stage, and more validation is needed.

Let us put the methods in this chapter into one sentence. They are all paths that use physics and data together. Physics provides trust, and data provides speed. It is difficult to handle rocket combustion with only one of them. The answer lies where the two forces meet. This book identified that place through validated examples.

Let us also summarize the nature of the examples used in this chapter. Raissi's physics-informed neural network and Li's Fourier neural operator are widely cited foundational studies. CRK-PINN and ChemKANs are peer-reviewed journal papers. DeepFlame, extended operators, and multifidelity reconstruction are recent studies. Many of them are still early-stage preprints. Early-stage research means that it may change in the future.

The attitude running through this chapter is also worth recalling. This book chose examples by putting evidence first. Models with only plausible names were lowered to the level of concepts. Unverified performance figures were not included in the main text. Early-stage research was clearly identified as early-stage research. This was done to help readers distinguish facts from expectations.

The tasks ahead are also clear. We need models that fit well across broad conditions in a single framework. They must handle shock waves and thin flame fronts stably. More cases must be validated with real engine data. Safety-preserving limiters must develop alongside them. As these tasks are solved, the tools will move closer to the field.

The lesson of this chapter is clear. Artificial intelligence can greatly change the speed of combustion computation. However, it can be trusted only when physics and validation come with it. Verified evidence matters more than flashy model names. On this balance, next-generation propulsion technology grows one step at a time. When artificial intelligence and physics work together, those steps become firm.

Related Supporting Materials

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3, 422-440. https://doi.org/10.1038/s42254-021-00314-5

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2021). Fourier Neural Operator for Parametric Partial Differential Equations. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.08895

Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2023). Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs. Journal of Machine Learning Research, 24, 1-97. https://www.jmlr.org/papers/v24/21-1524.html

Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. E. (2021). Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. Nature Machine Intelligence, 3, 218-229. https://doi.org/10.1038/s42256-021-00302-5

Lu, L., Meng, X., Mao, Z., & Karniadakis, G. E. (2021). DeepXDE: A deep learning library for solving differential equations. SIAM Review, 63(1), 208-228. https://doi.org/10.1137/19M1274067

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljacic, M., Hou, T. Y., & Tegmark, M. (2025). KAN: Kolmogorov-Arnold Networks. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2404.19756

Ji, W., Qiu, W., Shi, Z., Pan, S., & Deng, S. (2021). Stiff-PINN: Physics-Informed Neural Network for Stiff Chemical Kinetics. The Journal of Physical Chemistry A, 125(36), 8098-8106. https://doi.org/10.1021/acs.jpca.1c05102

Zhang, S., Zhang, C., & Wang, B. (2024). CRK-PINN: A physics-informed neural network for solving combustion reaction kinetics ordinary differential equations. Combustion and Flame, 269, 113647. https://doi.org/10.1016/j.combustflame.2024.113647

Koenig, B. C., Kim, S., & Deng, S. (2025). ChemKANs for combustion chemistry modeling and acceleration. Physical Chemistry Chemical Physics, 27(33), 17313-17330. https://doi.org/10.1039/D5CP02009C

Koenig, B. C., & Deng, S. (2025). Kolmogorov-Arnold Chemical Reaction Neural Networks for learning pressure-dependent kinetic rate laws. arXiv preprint(Preprint, not yet peer reviewed). arXiv:2511.07686. https://arxiv.org/abs/2511.07686

Weng, Y., Li, H., Zhang, H., Chen, Z. X., & Zhou, D. (2025). Extended Fourier Neural Operators to learn stiff chemical kinetics under unseen conditions. Combustion and Flame, 272, 113847. https://doi.org/10.1016/j.combustflame.2024.113847

Wu, J., Wang, X., Zhang, G., Liu, J., Li, X., Zhang, Y., Zhang, H., Lyu, J., Wang, B., & Wu, Y. (2025). Physics-informed machine learning for combustion: A review. arXiv preprint(Preprint, not yet peer reviewed). arXiv:2509.03347. https://arxiv.org/abs/2509.03347

Zhang, Z., Li, Z., Wang, Y., Yang, H., Peng, W., Teng, J., & Wang, J. (2026). Prediction of three-dimensional chemically reacting compressible turbulence based on implicit U-Net enhanced Fourier neural operator. Acta Mechanica Sinica, 42(7), 725274. https://doi.org/10.1007/s10409-025-25274-x

Guo, Z., Mao, R., Liu, L., Tan, G., Jia, W., & Chen, Z. X. (2025). Deep Learning-Enabled Supercritical Flame Simulation at Detailed Chemistry and Real-Fluid Accuracy Towards Trillion-Cell Scale. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC25). arXiv:2508.18969. https://doi.org/10.1145/3712285.3759850

DeepModeling. (2026). DeepFlame development repository. https://github.com/deepmodeling/deepflame-dev

Jiao, A., Jiang, W., Lu, X., Wang, Y., & Lu, L. (2026). Nested Fourier-enhanced neural operator for efficient modeling of radiation transfer in fires. arXiv preprint(Preprint, not yet peer reviewed). arXiv:2604.13919. https://arxiv.org/abs/2604.13919

Bao, Y., Zajac, J., Powers, M., Raman, V., & Kutz, J. N. (2026). Cheap2Rich: A Multi-Fidelity Framework for Data Assimilation and System Identification of Multiscale Physics -- Rotating Detonation Engines. arXiv preprint(Preprint, not yet peer reviewed). arXiv:2601.20295. https://arxiv.org/abs/2601.20295

Yao, S., & Zhang, W. (2026). Super-resolution reconstruction of high-fidelity rotating detonation flow fields based on cycle-consistent generative adversarial networks. Physics of Fluids, 38(4), 041704. https://doi.org/10.1063/5.0333031

Wang, X., Li, Z., Zhang, Y., Wen, H., & Wang, B. (2025). Observation-guided generative flow field reconstruction of rotating detonation combustor. Physics of Fluids, 37(11), 116110. https://doi.org/10.1063/5.0300394

Karkadakattil, A. (2026). A Physics-Stabilized Self-Updating Digital Twin Framework Using Physics-Informed Neural Networks for Thermal Field Prediction. Applied Computer Systems, 31(1), 30-40. https://doi.org/10.2478/acss-2026-0003

Shateri, A., Yang, Z., Yan, Y., Paul, M. C., & Xie, J. (2026). AI-Powered Surrogate Modelling for Multiscale Combustion: A Critical Review and Opportunities. arXiv preprint(Preprint, not yet peer reviewed). arXiv:2604.25617. https://arxiv.org/abs/2604.25617

Mousavi, M., Caldwell, C., Baltes, J., Aljasem, M., & Lee, B. J. (2025). Physics-Informed Neural Networks in Clean Combustion: A Pathway to Sustainable Aerospace Propulsion. arXiv preprint(Preprint, not yet peer reviewed). arXiv:2509.08094. https://arxiv.org/abs/2509.08094

Poinsot, T., & Veynante, D. (2012). Theoretical and Numerical Combustion (3rd ed.). R.T. Edwards, Inc., Philadelphia, PA.

Chapter 9 Topology Inverse Design Optimization of Regenerative Cooling Channels, Film Cooling Holes, and Inducers

Introduction

If you grab a hot pot by hand in summer, you get burned. So we make pot handles from wood or plastic. This keeps heat from reaching the hand. Rocket engines face the same problem. Inside the engine, fuel burns and makes very hot gas. This gas is hot enough to melt metal. Yet the wall that holds this hot gas is made of metal. How to protect this wall is a major challenge in rocket engine design.

Rockets solve this problem with ingenuity. They do not burn the fuel entering the engine right away. First, they send it through thin passages made inside the hot wall. As the fuel passes through the wall, it takes heat away from the wall. The wall cools, and the fuel warms up. The warmed fuel then enters the combustion chamber and burns. This method both protects the wall and preheats the fuel. It is called regenerative cooling. The word regenerative is used because the heated fuel is not thrown away but used again.

Regenerative cooling is an old method. It has been used since early rockets. But it was hard to design cooling passages freely. The machining method used to cut metal fixed the passage shapes. Recently, two things loosened this constraint. One is metal 3D printing. The other is artificial intelligence that quickly finds good designs. Together, they have brought cooling design into a new phase.

This chapter deals with thermal management structures in rocket engines. It examines cooling passages inside the wall and film cooling holes that coat the wall with a thin cooling film. It also covers the inducer, the front blade of the pump that pushes fuel upward. All three structures are problems of balance between heat and flow. If we try to cool the wall more strongly, the resistance that blocks the flow grows. If we try to reduce resistance, cooling becomes weaker. It is hard to find this balance point by human effort alone.

This is where artificial intelligence enters. The artificial intelligence discussed in this chapter is not the kind that draws pictures or writes text. It is a tool that quickly finds good answers in a complex design space. Some AI replaces slow flow calculations. Some AI chooses the next experiment on its own. Some AI reduces complex cooling passage shapes into small groups of numbers. This chapter explains these tools one by one in simple terms.

Let us note why this chapter matters. In rocket engines, thermal management and pumps are frequent trouble spots. If the wall melts or the pump vibrates, the engine stops. So these parts take much time and money when a new engine is built. Even making design a little faster and a little safer is a big help. Artificial intelligence helps at exactly this point. It scans a wide design space faster than people and selects good candidates.

One point should be made in advance. Rocket engine cooling design is a field with large safety margins. If the wall melts once, the whole engine breaks apart. So engineers do not simply trust and use the answers produced by AI. AI is used to find good candidates quickly. The final judgment is always checked with high-fidelity calculations and real hot-fire tests. The cases in this chapter should be read on that basis.

9.1 How Rockets Protect Hot Combustion Chamber Walls

This section explains how rocket engines protect hot walls. First, it looks at why rocket combustion chambers are so hot. Then it explains what regenerative cooling is and why it was difficult to make. Finally, it examines how metal 3D printing is changing this design. This section is the foundation for everything that follows.

In the main combustion chamber of a rocket engine, fuel and oxygen meet and burn. The temperature of the gas produced at this time easily exceeds 3,000 degrees. The melting point of steel is about 1,500 degrees. Combustion gas is far hotter than the melting point of steel. The pressure inside the combustion chamber is also high. In high-performance engines, the internal pressure exceeds one hundred times atmospheric pressure. Hot, high-pressure gas pushes and heats a thin metal wall from the inside.

High pressure is also a heavy burden. The high pressure inside the combustion chamber pushes the thin wall outward. It is like the skin of a balloon becoming tight when the balloon is strongly inflated. On top of this comes a large temperature difference between the hot inside and the cold outside. The hot side tries to expand, while the cold side tends to remain as it is. This conflict leaves large stress in the wall. Pressure and heat are tormenting the wall together. So the wall must be thin, well cooled, and strong.

The place where heat pouring into the wall reaches its maximum is the nozzle throat. A nozzle is a trumpet-shaped part that creates thrust by expelling gas backward. The narrowed waist of that trumpet is called the throat. At this point, the gas accelerates to the speed of sound. The temperature of the gas itself actually drops a little as it passes through the nozzle. Even so, the heat entering the wall is greatest at the throat. This is because fast gas scrapes fiercely along the narrow wall. The amount of heat entering the wall per unit area is called heat flux. The heat flux at the nozzle throat is very large. It is as if the heat from dozens of powerful heaters were concentrated into an area about the size of a palm.

Let us estimate how large the heat flux is. The heat from one burner on a household gas stove spreads across an area about the size of a palm. At the nozzle throat, the heat from dozens of such burners is concentrated into that same area. That heat pours into a single thin metal wall. The temperature difference between the inner and outer surfaces of the wall reaches hundreds of degrees. This large difference bends the metal and causes cracks. If the wall cannot be cooled, it collapses within seconds.

If the wall receives this heat as it is, it melts or weakens. So regenerative cooling is used. Many thin passages are made inside the wall. Fuel is first sent through these passages. As the fuel passes through them, it takes heat away from the wall. The wall cools, and the fuel warms up. The warmed fuel enters the combustion chamber and burns. Because the method protects the wall and preheats the fuel, it does not waste heat. Film cooling may also be added. Film cooling is a method that lays a thin film of cooling fluid along the wall. Using the two methods together can protect the wall more safely.

The problem is making these cooling passages. In the past, passages were cut into a block of metal. A cutting process forced the passage shapes to remain simple. Usually, straight rectangular grooves were cut side by side. With this kind of passage, it is hard to change the width or direction greatly from place to place. It is difficult to cool only the nozzle throat, where heat is concentrated, more strongly. If the passages are forced to become narrower, fuel has a harder time passing through them and pressure loss grows. If pressure loss is large, the pump that pushes in the fuel must work harder.

Metal 3D printing is removing this limit. Metal 3D printing spreads metal powder in a thin layer and melts it with a laser to bind it. Parts are made by stacking layers one by one. This method can create complex internal passages that cannot be made by cutting. It is possible to design passages that split like tree branches and then merge again. Passages can also curve along the curved surface of the wall. Passages can be placed densely where heat is concentrated and sparsely where it is not. Design freedom has expanded greatly.

Let us look at how much 3D printing changes cooling passages. The old method made passages by cutting grooves into the inner wall and putting a cover over the outside. Several parts had to be made and joined together. If there are many joints, those places can become weak. 3D printing produces a wall with passages inside it as one piece. With fewer joints, the wall becomes stronger. The manufacturing steps are also reduced, saving time and material. This makes it easier to revise and test the design many times.

Materials for cooling passages have also advanced. The National Aeronautics and Space Administration (NASA) developed a copper alloy called GRCop-42. Copper is a metal that transfers heat well. GRCop-42 is an alloy made by mixing small amounts of chromium and niobium into copper. It transfers heat well and also withstands high temperatures. With 3D printing, this alloy can be used to print an entire combustion chamber with passages inside it. The National Aeronautics and Space Administration has conducted actual hot-fire tests of combustion chambers made from this alloy. The development and test results are organized in public reports. The National Aeronautics and Space Administration advanced this material and manufacturing technology within its next-generation propulsion technology program. It is a case where materials, manufacturing, and design progressed as one package.

Film cooling can also be used to protect the wall more safely. Many small holes are drilled in the wall. A small amount of cold fluid is sent through those holes. The fluid that comes out forms a thin film along the wall. This film gets between the hot gas and the wall. If regenerative cooling cools the wall from the inside, film cooling covers the wall surface with a protective film. The two methods help each other. Near the nozzle throat, where heat is concentrated, the two methods are sometimes used together.

The shape of the cooling passages strongly affects thermal management performance. If the passages are narrow and dense, the area in contact with the wall becomes larger. If the contact area is larger, more heat is removed. But if the passages are too narrow, fuel has difficulty passing through. More force is needed to push the fuel in. This loss of force is called pressure loss. If pressure loss is large, the pump must spin harder. A good cooling passage cools the wall well while keeping pressure loss low. These two goals pull against each other.

The new method is not easy. A study published in 2025 analyzed a failure case of a 3D-printed GRCop-42 combustion chamber. In a hot-fire test in February 2021, the combustion chamber broke after 9 seconds. When the cause was examined, the problem was a place where the printing process had stopped once during production. Far more tiny pores than usual formed there. If there are many pores, that part becomes weak. This case shows how sensitive the new manufacturing method is to quality control. In exchange for free design, the manufacturing process must be watched more carefully.

The expanded design freedom also brings new concerns. Once passage shapes can be changed almost without limit, choosing which shape is good becomes harder. The method of drawing a few shapes by hand and calculating them cannot scan the whole wide design space. Even calculating the flow and heat of one passage precisely takes a long time on a computer. Artificial intelligence enters the task of exploring this wide space quickly and intelligently. The following sections examine those methods one by one.

9.2 Surrogate Models That Predict Rapid Changes in Fuel Inside Cooling Passages

This section explains why fuel inside cooling passages is hard to predict. Methane or kerosene used as coolant changes its properties sharply at high pressure and temperature. High-fidelity analysis that calculates this changeable flow is slow. So AI surrogate models are used to replace the calculations. This section explains the principles and limits of those surrogate models.

Methane or RP-1 is used as a rocket coolant. RP-1 is kerosene refined for rockets. It is a close relative of the kerosene we use. These fuels are under very high pressure inside cooling passages. When the pressure exceeds a certain value, the boundary between liquid and gas becomes blurred. With water, it is like a state in which the clear distinction of boiling water forming bubbles in a pot disappears. This state is called the supercritical state. Fuel in the supercritical state has intermediate properties that are neither liquid nor gas.

The word supercritical may sound difficult. Let us explain it simply. When water is boiled in a pot, the water boils as it turns into bubbles. Liquid and gas are clearly separated. But if the pressure is raised very high, this boiling disappears. The liquid gradually connects to gas-like properties. The clear boundary where bubbles form disappears. This boundary-free state is the supercritical state. Fuel inside cooling passages flows in this state because it is under high pressure.

The properties of the supercritical state change rapidly with temperature. Across a certain temperature range, density drops sharply. In the same range, specific heat suddenly shoots up. Specific heat is the amount of heat needed to raise the temperature by 1 degree. The point where density and specific heat fluctuate like this is called the pseudo-critical line. Near this region, the fuel changes properties rapidly, as if it were boiling. It does not actually boil, but it resembles boiling, so it is called pseudo-boiling.

A surge in specific heat can be good for cooling. If specific heat is large, the fuel can hold a lot of heat. The same amount of fuel carries away more heat. But if density drops sharply, the fuel expands and speeds up. If the faster flow fluctuates near the wall, the way heat is carried becomes disturbed. The good side and the bad side appear together at the same point. This is why predicting cooling in this region is difficult.

Pseudo-boiling causes problems in cooling. The wall cools only when the fuel draws heat well from the wall. But near the pseudo-critical line, the fuel's ability to draw heat can suddenly fall. The heated fuel layer near the wall blocks heat like a thin blanket. This is called heat transfer deterioration. When heat transfer deteriorates, the wall temperature at that spot suddenly rises. A 2020 study summarized that pseudo-boiling and heat transfer deterioration appear together when rocket fuel is heated. A 2024 study examined why this deterioration occurs using a microscopic two-state flow model. This heat transfer deterioration is also connected to the coking problem discussed later. When the wall suddenly becomes hot, the fuel breaks down faster at that spot.

There is a precise method for calculating this kind of flow. It is computational fluid dynamics (CFD). The flow is divided into small grid cells. The velocity, pressure, and temperature in each cell are solved according to physical laws. The underlying laws are three conservation laws. The first is conservation of mass. The amount of fluid flowing in must match the amount flowing out. The next is conservation of momentum. The forces that push or pull the fluid must match the change in velocity. The last is conservation of energy. The heat entering and leaving must match the change in temperature. Solving these three laws together in every cell is CFD. CFD is accurate but slow. A computer may take nearly an hour just to solve one short section of a cooling channel in detail. In cooling channel design, the channel shape and flow rate must be checked across thousands of variations. If each calculation takes an hour, the design work will never finish.

That is why a surrogate model is used. A surrogate model is a fast approximate model that replaces a slow calculation. First, many cases are calculated in advance with CFD, and data are collected. An artificial neural network is trained with that data. The neural network takes the channel shape, flow rate, pressure, and inlet temperature as inputs. It then immediately outputs the wall temperature, heat flux, and pressure loss. Researchers at the German Aerospace Center (DLR) built this kind of neural network surrogate model. What they actually used was an ordinary neural network with several hidden

layers. A calculation that took CFD an hour was finished by this model in less than 1 second. At first, it handled only straight channels. Later, it was expanded to handle curved channels as well.

The process for making a surrogate model is as follows. First, the ranges of the design variables are set. These are the ranges for channel width, channel height, flow rate, and inlet temperature. Within those ranges, several cases are selected and calculated with CFD. The calculation results are collected as input-output pairs. These pairs become the textbook for the neural network. The neural network studies the pairs and learns the rule that maps inputs to outputs. After training is complete, it can answer quickly for new inputs as well. This is similar to how a person learns problem types by solving many examples.

There are other ways to improve the accuracy of a surrogate model. These methods were not used directly in the DLR study above. They are widely used techniques for making surrogate models more robust. One method is to include physical knowledge. If a neural network learns only from data, it may produce answers that violate physical laws. So the laws that mass, momentum, and energy are conserved are inserted into the training rules. If the neural network violates those laws, it receives a penalty. A neural network built this way is called a physics-informed neural network (PINN). A structure called a residual connection is sometimes added to this. A residual connection is a path that sends information from an earlier layer directly to a later layer by skipping intermediate layers. This structure helps train deep neural networks stably. Applying these techniques to cooling channel surrogate models is a direction proposed by this book.

The property values of supercritical fluids must also be entered accurately. The National Institute of Standards and Technology (NIST) publishes values such as density and specific heat for many fluids by temperature and pressure. For a pure fluid with one component, such as methane, these standard values can be connected and used directly. This is needed so the sharp changes near the pseudo-critical line are not missed. But RP-1, a kerosene, is different. RP-1 is a mixture of several hydrocarbons, so its composition changes slightly with each refining process. For that reason, pure-fluid tables cannot be used as they are. A separate surrogate mixture model must be made to match its properties. If the property values are inaccurate, the surrogate model will also produce the wrong wall temperature. Supercritical fluids are sensitive in their properties, so this base data is important.

Some studies focus on choosing the coolant itself well. A 2026 study designed a surrogate fuel that imitates rocket-grade kerosene. It used a genetic algorithm to find a combination that matches properties well in the supercritical state. A genetic algorithm is a method that improves the answer by selecting and mixing good combinations. The more accurate the fuel model used in cooling calculations is, the more accurate the prediction becomes. Artificial intelligence also helps in choosing fuel combinations.

There is a point to be careful about when using surrogate models. One should not rely on model names used like trademarks or on unverified performance numbers. For that reason, this book explains the idea using the general concept of a CFD surrogate model instead of a specific model name. If the principle is the same, the lesson is the same no matter what the model is called.

This surrogate model has also been used in real engines. The German Aerospace Center used this method to examine the cooling channel performance of an engine it developed. It improved wall temperature prediction and reflected that in the design. Because the calculation became faster, several channel designs could be compared in a short time. The designers explored the design space broadly without being tied down by slow, precise calculations.

A surrogate model has clear limits. A surrogate model can be trusted only within the range it has learned. It can be badly wrong for unfamiliar channel shapes or operating conditions that it did not see

during training. This risk is greater for supercritical fluids because their properties are sensitive. So the final design must not be decided by the surrogate model alone. The surrogate model is used to quickly select good candidates. Only those candidates are then checked again with precise CFD and actual combustion tests. A surrogate model is not a tool that replaces the answer. It is a tool that quickly filters candidates.

9.3 Active Learning for Finding Film Cooling Holes with Less Computation

Let us first look at why film cooling is needed. Regenerative cooling cools the wall from the inside. But where heat is too concentrated, cooling only from inside the wall may not be enough. This is because the wall surface itself touches hot gas directly and is heated. So a thin protective film is added over the surface. This is film cooling. Regenerative cooling and film cooling protect the wall together from the inside and the outside.

Film cooling is a method that makes a thin film of cold fluid directly in front of a hot wall. Many small holes are drilled in the wall. Cold fluid is blown through those holes at an angle. The injected fluid spreads thinly along the wall and covers it like a blanket. This film separates the hot gas from the wall. The wall is heated less because it does not directly touch the hot gas. It is similar to air for preventing frost flowing upward from the bottom of a car windshield.

The performance of film cooling is determined by several values. One is the strength of the injected fluid. This is called the blowing ratio. It is the flow strength of the cold fluid divided by the flow strength of the hot gas. The inclination angle of the hole, the spacing between holes, and the hole shape also determine performance. Even small changes in these values can greatly change the film's performance. Whether the film stays attached to the wall or lifts off and peels away depends on these values.

Problems also occur when the blowing ratio is too high. If the fluid is blown too strongly, the film lifts off from the wall. The injected jet fails to cover the wall and rises upward. Then the wall touches the hot gas again. If the blowing ratio is too low, the film is thin and is quickly swept away. So there is a suitable value for the blowing strength as well. The angle, spacing, and strength must be adjusted together to make a good film.

One reason the film peels away is vortices. When fluid is injected from a hole, a pair of vortices rotating left and right forms. These vortices pull hot gas toward the wall. Then the cooling film that was made with effort is lifted and stripped away. The wall is exposed to hot gas again. If the hole is designed well, these vortices can be weakened. That is why fan-shaped holes, whose exits widen like a fan, are often used today. The injected fluid spreads widely to the sides and covers the wall evenly.

Fan-shaped holes spread the cooling film wider than round holes. A round hole injects fluid as one jet. This jet can easily lift off from the wall. A fan-shaped hole widens the exit to the sides and injects the fluid broadly. The widely spread fluid attaches better to the wall. When the attached area becomes larger, it covers the wall evenly. That is why gas turbines and aircraft engines often use fan-shaped holes. Performance depends on how much the hole is widened and at what angle it is drilled.

To find a good hole shape, the flow must be calculated. A precise analysis is needed to see even the vortices in detail. This analysis is called large eddy simulation (LES). LES is a method that directly calculates large vortices in the flow. It is precise, but computationally very expensive. Even solving one condition takes a long time. It is difficult to calculate hundreds of cases while changing the hole angle, spacing, and strength. Since the computation budget is limited, not every condition can be calculated at random.

This is where the Gaussian process helps. A Gaussian process is a method for making predictions with a small amount of data. Its feature is that it does not output only the predicted value. It also outputs how uncertain the prediction is. Where there is much data, it predicts with confidence. Where there is no data, it marks the prediction as uncertain. It is like drawing a map in detail for roads already traveled, and drawing untraveled roads faintly. This method connects values using a smooth rule called Matern.

Let us also look at how a Gaussian process connects values. Conditions that are close to each other are likely to have similar performance. If the hole angle changes only slightly, the cooling performance will also change only slightly. A kernel expresses this idea mathematically. A kernel is a yardstick for measuring how close two conditions are. If they are close, their values are treated as similar. If they are far apart, their values may differ. This yardstick smoothly connects the conditions that have already been calculated. It also estimates values for conditions that have not been calculated.

Active learning adds smart selection to this process. It chooses by itself which condition to calculate next with LES. When choosing, it looks at two things together. One is where performance appears good. The other is where uncertainty is high. Exploring where performance looks good moves the search closer to a better answer. Exploring uncertain areas makes the unclear parts of the map sharper. The rule that balances these two is called an acquisition function. A commonly used rule measures the expected degree of improvement over the current result.

The advantage of this method is that it chooses the order of calculations intelligently. It does not calculate random conditions without direction. It selects and calculates the conditions from which it can learn the most. So it can approach a good hole shape with only a small number of LES calculations. It saves expensive computation while narrowing the design quickly. This approach first became established in the design of cooling holes for gas turbine blades. Studies verified so far mainly focus on gas turbines. Bayesian optimization that considers several objectives together improved both cooling performance and uniformity. One study reported higher cooling performance than an earlier design under the same pressure-loss condition.

Artificial intelligence is also used for the flow calculation itself. Studies around 2024 built neural networks that predicted film cooling effectiveness. One study predicted cooling effectiveness even when wall curvature and hole angle changed. These prediction models produce values similar to precise calculations much faster. They help compare many hole shapes quickly in the early stage of design. A Gaussian process may choose the next calculation, while a neural network quickly estimates that calculation. The two can also be combined in this way.

Film cooling in rocket combustion chambers borrows the same method. But rocket conditions are much harsher. The gas is hotter and faster. So optimal hole values obtained from aircraft engines must not be transferred directly to rockets. When the hole size, fluid velocity, wall curvature, and fuel type change, the result changes. A good design method can be transferred, but the specific optimal numbers must be found again each time. Artificial intelligence is a tool that makes that repeated search faster. The final shape must be verified again under rocket conditions.

9.4 Designing Cooling Passage Shapes with Artificial Intelligence

Cooling passage design is a problem where solids and fluids are intertwined. Heat from hot gas first enters the metal wall. That heat travels through the metal toward the cooling passage. In the passage, the flowing fuel takes that heat away. Heat conduction through the metal and convection by the fluid meet in one place. A calculation that solves both together is called conjugate heat transfer (CHT). In rocket cooling passages, the solid and the fluid are strongly coupled, so this calculation is important.

A winter window makes this easier to understand. The room is warm, and the outside is cold. The inner and outer surfaces of the glass have different temperatures. Heat flows through the glass, and cold air outside takes heat away. Heat conduction in the glass and convection in the outside air meet at the window surface. A rocket wall is the same. Heat enters from the hot gas side, and fuel takes heat away on the cooling passage side. The two flows meet at the wall and reach a balance.

The key to conjugate heat transfer is the condition at the boundary. Consider the surface where the metal and the fuel touch. The heat that leaves the metal at that surface must equal the heat that the fuel receives at that surface. Heat does not disappear or appear from nowhere. Wall temperature must be calculated while keeping this condition. This calculation shows where the wall reaches its highest temperature and how that place should be cooled.

Conjugate heat transfer calculations require a large amount of computation. This is because heat conduction in the solid and convection in the fluid must be solved at the same time. In addition, the fluid is turbulent. Turbulence is a flow where large and small vortices are mixed together. The more finely this flow is resolved, the longer the calculation takes. Even solving one cooling passage in detail takes a computer a long time. If there are thousands of design candidates, that calculation must be repeated thousands of times. So a way to reduce the computation is needed.

Now it is time to design the passage shape. Topology optimization is used here. Topology optimization is a method for finding whether a fixed space should be filled with material or left empty. For a cooling passage, the problem becomes deciding where to leave metal and where to leave empty passages for fuel flow. A person does not set the passage shape in advance. If only the goal is given, the optimization draws the passage shape by itself. The goal is to lower the maximum wall temperature. At the same time, the pressure loss must not become too large. Studies that design cooling passages with topology optimization already exist. A representative method is a density-based method that changes the material layout little by little while considering flow and heat together.

This problem has too many variables. The space is divided into hundreds of thousands of small cells. For each cell, one must decide whether it is metal or empty space. The number of possible cases is countless. Searching this wide space as it is would take endless time. So a way to reduce the problem is needed. This is where a convolutional autoencoder (CAE) comes in.

An autoencoder is a neural network that reduces complex data to a smaller form and then restores it. The front part compresses the passage shape into a small group of numbers. This group of numbers is called a latent vector. The back part restores the passage shape from that latent vector. If it is trained well, it can contain the key shape of the passage with only a few numbers. It is similar to compressing a photo into a small file and then opening it again. Convolution is a calculation method that looks at neighboring parts of an image together. It is well suited to handling spatial data such as passage shapes.

There are still few validation cases that directly apply this method of reducing the design space with an autoencoder to rocket cooling passages. This book proposes that direction. Topology optimization itself is a verified method, and the idea is to add an autoencoder to it to speed up the search.

This compression makes design easier. The optimizer does not directly handle hundreds of thousands of cells. Instead, it changes only a small latent vector in different ways. If the latent vector is changed slightly, the back part generates a plausible passage shape. In effect, a wide design space is reduced to a small space and searched. The optimizer finds a good latent vector in this small space. The passage made by that vector becomes a good cooling passage.

Using a latent vector has another advantage. The passage shapes that are generated are generally plausible. This is because the autoencoder learned from actual passage data. It creates new shapes

within the form it learned through training. As a result, physically unreasonable strange shapes appear less often. It reduces wasted effort that occurs when a wide space is searched at random. The search is fast because it looks for answers only in a small and useful space. However, similarity to the training data alone does not guarantee a good passage. It must be checked separately whether the passage is properly connected from start to finish, whether the wall is not too thin, and whether flow resistance and structural strength are acceptable. These conditions must be added separately as rules inside the optimization.

The condition that the shape can be manufactured must also be included. 3D printing builds a part upward layer by layer. A section that extends too far sideways without support below can sag during fabrication. So a constraint is imposed so that the passage wall does not lean below a certain angle. If this constraint is met, the passage can stand by itself without support. If these conditions are included in the optimization, the resulting design can be printed directly. This prevents a design that is good in calculation from becoming a design that cannot actually be made.

There is also research on replacing conjugate heat transfer calculations with artificial intelligence. One 2024 study built a neural network by stacking encoders and decoders in layers. This neural network received the part shape as input and directly drew the temperature distribution. It replaced repeated high-fidelity conjugate heat transfer calculations with fast prediction. Another study in 2025 treated the solid and the fluid with separate neural networks. The two neural networks were trained to match values at the boundary. This makes it easier to capture sharp temperature changes at the surface where the wall and the fuel touch. These studies dealt with general conjugate heat transfer problems. If such fast prediction models are used inside topology optimization of rocket cooling passages, design exploration becomes faster. That combination is a direction to refine in the future.

The passages drawn by topology optimization have shapes that are difficult for people to draw by hand. Branches split and merge, and they bend along curved wall surfaces. Slanted fins may also line up densely like fish bones. These shapes distribute heat evenly and guide the flow smoothly. Even if they do not look good to the human eye, they may be good shapes by calculation. Artificial intelligence helps find these unexpected shapes.

There is an industrial example of this method. One company applied this topology optimization technology to rocket cooling passages. It announced that the result greatly lowered the maximum wall temperature. Recently, research has also appeared on using regular curved surfaces in cooling passages. This surface is called a triply periodic minimal surface (TPMS). A TPMS is a structure in which smoothly curved surfaces, like soap bubble films, repeat regularly in space. This structure folds a large surface into a narrow space. A larger surface exchanges more heat. One 2025 study made TPMS passages that wrapped along the curved wall surface. It reported that cooling performance and temperature uniformity around the wall both improved. In 2026, a study also appeared that increased heat transfer with cooling passages filled inside a lattice. These curved passages and lattice passages can be made only with 3D printing. However, company cases differ from academic papers in how they are verified. This book treats them only as reference cases that show an industrial trend. Final performance should be used as a design basis only when independent test data exists.

9.5 Multi-Objective Optimization to Reduce Coking, the Buildup of Soot

RP-1, a kerosene-based fuel, has one weakness when it is used as a coolant. When the fuel passes near a hot wall, it breaks down because of heat. This is called pyrolysis. Among the broken fragments, carbon sticks and burns onto the wall. It is like oil turning black and sticking to a frying pan when left there for a long time. This carbon deposit on the wall is called coking. Coking gradually builds up inside the

cooling passage.

The breakdown of fuel by heat is called pyrolysis. It is a process in which large molecules receive heat and split into smaller pieces. The higher the temperature, the faster the splitting occurs. Among the broken pieces, those rich in carbon clump together. The clumped carbon sticks to the wall and forms a deposit. So the hotter the wall is, the more easily coking forms. Coking depends on wall temperature, fuel type, and elapsed time together.

Coking often occurs at the nozzle throat. The nozzle throat is where the heat flux is highest. Because the wall is that hot, fuel decomposition is active. Coking causes two problems. One is that carbon deposits do not transfer heat well. The deposit blocks the space between the wall and the fuel, making the wall hotter. The other is that the deposit narrows the passage. When the passage becomes narrower, fuel has more difficulty passing through, and pressure loss increases. When the two problems overlap, the wall becomes dangerous. In reusable engines, coking becomes a greater problem because it builds up over repeated tests.

There is a rule in the relationship between temperature and coking. Even a small rise in temperature greatly speeds up chemical reactions. When wall temperature rises, fuel decomposition becomes noticeably faster. Coking then builds up faster as well. So even a small decrease in wall temperature can greatly reduce coking. Better cooling cools the wall, and a cooler wall reduces coking. Cooling and coking suppression are connected in this way.

The heat transfer deterioration discussed in the previous section is linked to this coking. When supercritical fuel crosses a pseudo-critical line, heat transfer near the wall suddenly worsens. Then the wall temperature at that location rises sharply. When the wall becomes hotter, the fuel breaks down faster, and coking builds up faster. The accumulated coking blocks more heat and makes the wall even hotter. If this feedback continues, the wall can fail suddenly. Heat transfer deterioration and coking amplify each other.

To reduce coking, wall temperature must be lowered. If the wall is less hot, the fuel breaks down less. The flow near the wall must also be considered. If fuel passes quickly along the wall, there is less chance for deposits to stick. The force with which a fluid brushes along a wall near the surface is called wall shear stress. This force must be strong enough to reduce deposit buildup. Coking suppression design is a problem of controlling wall temperature and wall shear together.

The problem is that there are several goals, and they conflict with each other. There is a goal of lowering wall temperature. There is a goal of reducing pressure loss. There is a goal of suppressing coking. These goals do not all improve together. If the passage is narrowed to make the flow faster, the wall cools, but pressure loss increases. If the passage is widened to slow the flow, pressure loss decreases, but cooling becomes weaker. Gaining one side means losing another. A problem where several goals conflict in this way is called multi-objective optimization.

This kind of trade-off is familiar in daily life. Think about choosing a bag. It is difficult to have a bag that is light, sturdy, and cheap all at once. If it is light, it is likely to be weak, and if it is sturdy, it is heavy or expensive. So we choose a balance point that fits the purpose. Cooling passage design is the same. Cooling, pressure loss, and coking suppression cannot all be made best at the same time. If one side is gained, another side must be given up a little.

Multi-objective optimization does not have a single answer. Instead, it produces a set of good answers that trade off against each other. This set is called the Pareto front. Each answer on the Pareto front has a different balance point. Some answers put cooling first, and others put pressure loss first. The designer looks at this set and chooses the balance point that fits the engine. Instead of deciding on one answer in

advance, the designer lays out good candidates and chooses among them.

Bayesian optimization is used here. It is a method that extends the Gaussian process and active learning from the previous section to multiple goals. First, it draws a rough map of the goals with a small number of calculations. It then uses that map to choose which conditions to calculate next. It first calculates the conditions that will greatly help expand the Pareto front. It finds good balance points while using fewer expensive combustion analyses. If coking is included as a goal, the result is a design that weighs wall temperature, shear, and pressure loss together. However, verified cases of optimization that use coking as a goal in rocket cooling passages are still difficult to find. The physics of pyrolysis and deposition are well known, so this book proposes a direction in which that physics is included as a design goal.

Coking becomes a bigger challenge in reusable engines. But that does not mean expendable engines are free from coking. Even during one burn, deposits can narrow a passage or worsen heat transfer, causing wall failure. What matters is how fast deposits build up within the mission time, whether the engine is reused or not. Engines used many times face an added problem of accumulation. With each firing test, deposits build up little by little. The accumulated deposits make cooling worse over time. So reusable engines must be designed from the start to suppress coking. Artificial intelligence optimization helps create designs that look ahead to these long-term changes.

This method can also be connected to rocket test data. The Lampoldshausen test site of the German Aerospace Center obtains data by actually firing rocket engines. With this kind of test data, the base model for optimization becomes more accurate. Data fills in real coking behavior that calculations alone can miss. This is a process of refining the design by moving back and forth between tests and calculations.

Channels with a high aspect ratio are sometimes used to suppress coking. The aspect ratio is the channel height divided by its width. A narrow and deep channel has a high aspect ratio. This kind of channel has a large area in contact with the wall, so it removes heat well. As the wall becomes cooler, fuel decomposition decreases. Spiral channels are also used. When the flow is made to swirl, it brushes more strongly near the wall. But these channels can increase pressure loss, so that must be weighed together.

There is a caution in this section too. Figures for coking reduction can be trusted only when they are backed by actual tests. So this book focuses on the principle of the method instead of specific numbers. In an actual engine, fuel composition, surface roughness, catalytic action of the wall, and operating time all affect coking. There are many variables, so prediction is difficult. Coking predictions not supported by tests should be treated conservatively. Artificial intelligence is used to narrow down good candidates, and the final decision is made through actual firing tests.

9.6 Physics-Enhanced Machine Learning for Designing a Pump Front-Blade Inducer

A rocket engine must push fuel and oxygen strongly into the combustion chamber. The turbopump does this work. A turbopump is a pump that spins very fast. But the main blades of the pump cannot work properly if the inlet pressure is low. So an auxiliary blade is placed in front of the main blades. This auxiliary blade is the inducer. The inducer first draws in the liquid smoothly and raises the pressure slightly. Thanks to this, the main blades behind it rotate stably. It is a part like the gatekeeper of the pump.

Let us first look at why a turbopump is needed. The pressure inside the combustion chamber is very high. To push fuel into it, an even higher pressure is needed. Low-pressure fuel cannot enter a high-pressure combustion chamber. It is like water not flowing by itself from a low place to a high place. So the pump greatly raises the pressure of the fuel. A rocket turbopump does this through very fast

rotation. The turbopump of a large engine rotates tens of thousands of times per minute.

An inducer is more difficult to handle when it deals with liquid hydrogen. Liquid hydrogen is an extremely cold liquid at about minus 253 degrees Celsius. Its density is low, so it is light. When the inducer spins fast, a region where pressure drops sharply forms near the front of the blade. A liquid boils when the pressure becomes low enough. Bubbles can form even when it is not hot, if only the pressure is low. This is the same principle as water boiling more easily at a low temperature in the mountains. The formation of bubbles in a flowing liquid is called cavitation.

Cavitation troubles the inducer. The bubbles that form collapse in an instant when they flow back into a higher-pressure region. When a bubble collapses, it creates a very strong pressure wave. This wave gradually erodes the blade surface. If this continues for a long time, the blade becomes pitted. Cavitation also lowers pump performance. When bubbles increase, the pressure raised by the pump suddenly collapses. If this collapse point is passed, the whole engine becomes dangerous. So the inducer must be designed to delay cavitation.

There is one helpful property. Cryogenic liquids have a thermodynamic suppression effect. For a bubble to form, it must take latent heat of vaporization from the surrounding liquid. Then the temperature around the bubble drops slightly. When the temperature drops, the boiling pressure at that spot also drops. This makes it harder for the bubble to grow further. In effect, the bubble suppresses its own growth. Liquid hydrogen has a large suppression effect, so cavitation is delayed a little. Inducer design must include this effect in the calculation. This effect must also be included in the physical rules of physics-enhanced learning. Otherwise, cavitation performance is wrongly judged to be worse than it really is.

Inducer design is the task of deciding the blade shape. Blade angle, curvature, and forward sweep determine performance. There are three goals. It must raise pressure well. It must have good efficiency. It must delay cavitation. These goals also conflict with one another. If the blade is made more upright to raise pressure strongly, cavitation forms more easily. If the blade is laid down to avoid cavitation, the pressure rises less. It is the same kind of balancing problem as in the earlier sections.

There is a limit to the inducer's ability to raise pressure. Suppose we run the pump while lowering the inlet pressure little by little. Up to a certain point, the pump holds up well. But when the inlet pressure falls below a certain value, cavitation becomes severe. At that moment, the pressure raised by the pump suddenly collapses. This collapse point is called the head breakdown point. Head is the size of the force with which a pump pushes up a fluid. In an actual pump, the point where bubbles first form is distinguished from the point where head completely collapses. The point where head drops by 3 percent between them is sometimes used as a standard. This value is called NPSH3. Incipient cavitation, NPSH3, and complete head breakdown are different points. A good inducer delays this breakdown point to a lower inlet pressure. That means it can rotate stably even at lower pressure. This property is also connected to rocket weight. If the inducer can tolerate low pressure well, the pressure of the fuel tank can be kept low. When tank pressure is low, the tank wall can be made thinner. When the tank becomes lighter, the whole rocket becomes lighter. The design of one small blade can change the weight of the rocket.

Physics-enhanced machine learning is used here. Machine learning that learns only from data can violate physical laws. So the conservation laws of fluids and cavitation physics are included in learning. The flow field is quickly approximated with a Fourier neural operator (FNO). This operator is a neural network that quickly predicts the whole picture of a flow at once. If an ordinary neural network predicts the value at one point, this operator predicts the entire flow field as a whole. Because it handles the flow by dividing it into waves of several sizes, it captures both large flows and small ripples. So it is well

suited to quickly approximating values spread widely through space, such as a flow field. By changing the blade shape, engineers can quickly examine the flow and cavitation risk. This is much faster than running a high-precision calculation every time. The Fourier neural operator itself has been validated in several flow problems. Using it for inverse design of a liquid hydrogen inducer is the direction proposed in this book.

Artificial intelligence examines several kinds of information in inducer design. It looks at the pressure distribution near the front of the blade. It checks where pressure drops sharply and how much cavitation will grow there. It also checks how much pressure recovers behind the blade. It combines these pieces of information to find a good blade shape. A forward-swept blade shape helps delay cavitation. Artificial intelligence creates several such shapes and compares them quickly.

Artificial intelligence is also used to diagnose cavitation. One study from 2025 filmed inducer cavitation with a high-speed camera and analyzed the video with a neural network. It used a neural network that reads spatial patterns together with a neural network that reads changes over time. This makes it possible to detect the moment when cavitation becomes unstable. Artificial intelligence captures rapid changes that are hard to see with the eye. This study concerns diagnosis that recognizes cavitation. Its use differs from a surrogate model that computes the flow in place of full simulation or inverse design that finds the blade shape. The three should be viewed separately, not lumped together.

Inducer design does not end with flow calculation. The blades rotate very fast, so they receive large forces. The waves created when cavitation bubbles collapse also strike the blades. So engineers must examine whether the blades can withstand these forces. If cooling channel design is a balance of heat and flow, inducer design is a balance of flow, force, and cavitation. Artificial intelligence is used to weigh these multiple goals together.

Here too, the limits are clear. Cryogenic hydrogen testing is difficult and expensive. There are not many test facilities that can handle liquids colder than minus 250 degrees Celsius. So there is not enough validation data for inducer design. Figures for cavitation reduction can be trusted only when they are backed by actual pump tests. So this book does not use unverified numbers. Instead, it focuses on what information physics-enhanced models examine to help design. Even if artificial intelligence narrows down good blade candidates, the final judgment must be made through actual pump tests.

Summary

This chapter examined how to design rocket engine thermal management structures with artificial intelligence. It covered three structures. They are cooling channels inside the wall, film cooling holes that cover the wall with a cooling film, and the pump front-blade inducer that pushes up the fuel. All three structures have the same kind of problem. It is a balance between heat and flow. If one side is gained strongly, the other side is lost. Artificial intelligence has entered this process to find the balance point quickly.

Two major changes lay behind this. One is metal 3D printing. It became possible to make complex internal channels that could not be made by cutting. Design freedom expanded greatly. The other is new materials such as the copper alloy GRCop-42, which transfers heat well. Free design and good materials came together to make new cooling structures possible. But new manufacturing methods are sensitive to quality control, so the manufacturing process must be watched carefully.

Artificial intelligence was used in several ways. Surrogate models quickly approximated the behavior of supercritical fuel in place of slow fluid calculations. Gaussian process active learning found good film cooling holes while saving expensive calculations. Convolutional autoencoders reduced complex channel shapes into a small numerical space, making topology design easier. Multi-objective Bayesian

optimization found balance points among conflicting goals and suppressed coking. Physics-enhanced machine learning and Fourier neural operators helped design inducers that control cavitation.

The methods in this chapter are connected to one another. A surrogate model quickly approximates the flow. Active learning chooses the next calculation based on that approximation. An autoencoder reduces the design space and makes exploration easier. Bayesian optimization finds balance points among several goals. Physics-enhanced learning makes the model obey physical laws. The point is not to use just one tool, but to connect several tools. It is the designer's job to choose and combine them to fit the problem.

There is one principle running through this chapter. Artificial intelligence is not a tool that decides the answer for us. It is a tool that quickly narrows a wide design space to good candidates. A surrogate model can be trusted only within the range it has learned. Performance figures from company cases are different from academic validation. Figures whose sources cannot be verified are not used. Rocket engine cooling is a field where the engine breaks if the wall melts even once. So the final judgment is always made with high-precision calculations and actual firing tests.

The methods in this chapter have something in common. They are all techniques for saving expensive calculations. High-precision calculations are accurate but slow. If those slow calculations are run every time, the design will never finish. So artificial intelligence takes over fast approximation. Surrogate models replace calculations, active learning chooses the order of calculations, and autoencoders reduce the problem. The saved time is used to examine more candidates. The more candidates are examined, the higher the chance of getting closer to a good answer.

The future direction also rests on this principle. Machine learning that embeds deeper physical knowledge is growing. The flow that connects design and diagnosis is also expanding. Artificial intelligence proposes design candidates, tests verify them, and the test data then trains the artificial intelligence again. This cycle is gradually pushing rocket engine thermal management design forward. This chapter can be reduced to one sentence. It helps artificial intelligence quickly solve the difficult balance between heat and flow, within a wider space of design freedom. And the answer is always confirmed by testing. The tools have become faster, but the principle of verification remains the same. The next chapter covers how artificial intelligence diagnoses and controls the engine when an engine made in this way is actually operated.

Related Supporting Materials

Pizzarelli, M. (2017). Regenerative cooling of liquid rocket engine thrust chambers: Modeling and simulation. Journal of Propulsion and Power, 33(4), 802-821. https://doi.org/10.2514/1.B36261

Waxenegger-Wilfing, G., Dresia, K., Deeken, J., & Oschwald, M. (2020). Heat transfer prediction for methane in regenerative cooling channels with neural networks. Journal of Thermophysics and Heat Transfer, 34(2), 347-357. https://doi.org/10.2514/1.T5865

Dresia, K., Waxenegger-Wilfing, G., Riccius, J., Deeken, J., & Oschwald, M. (2023). Improved wall temperature prediction for the LUMEN rocket combustion chamber with neural networks. Aerospace, 10(5), 450. https://doi.org/10.3390/aerospace10050450

Waxenegger-Wilfing, G., Dresia, K., Deeken, J., & Oschwald, M. (2021). Machine learning methods for the design and operation of liquid rocket engines: Research activities at the DLR Institute of Space Propulsion. arXiv. https://arxiv.org/abs/2102.07109

Gradl, P. R., Protz, C. S., Cooper, K., Ellis, D., & Evans, L. J. (2019). GRCop-42 development and hot-fire testing using additive manufacturing powder bed fusion for channel-cooled combustion chambers. AIAA Propulsion and Energy 2019 Forum. https://doi.org/10.2514/6.2019-4228

NASA NTRS. (2019). GRCop-42 development and hot-fire testing using additive manufacturing powder bed fusion for channel-cooled combustion chambers. Document ID 20190030433. https://ntrs.nasa.gov/citations/20190030433

Gradl, P., Mireles, O. R., Katsarelis, C., Smith, T. M., Sowards, J., Park, A., et al. (2023). Advancement of extreme environment additively manufactured alloys for next generation space propulsion applications. Acta Astronautica, 211, 483-497. https://doi.org/10.1016/j.actaastro.2023.06.035

NASA. (2023). GRCop-42 additive manufacturing for high heat flux thrust chamber liners. NASA Technical Reports Server. https://ntrs.nasa.gov/citations/20230007103

NASA. (2023). Rapid Analysis and Manufacturing Propulsion Technology (RAMPT) project. NASA. https://www.nasa.gov/rapid-analysis-and-manufacturing-propulsion-technology/

Demeneghi, G., Williams, B. B., Katsarelis, C., Tilson, W., & Gradl, P. (2025). Additively manufactured GRCop-42 copper-alloy combustion chamber failure analysis: The role of build interruptions. Engineering Failure Analysis, 174, 109509. https://doi.org/10.1016/j.engfailanal.2025.109509

Ferster, K. K., et al. (2025). An open tool for the design of cooling channels within regeneratively cooled liquid rocket engines. AIAA SciTech 2025 Forum. https://doi.org/10.2514/6.2025-2127

NIST. (2026). Thermophysical properties of fluid systems. NIST Chemistry WebBook. https://webbook.nist.gov/chemistry/fluid/

Banuti, D. T. (2020). Pseudo-boiling and heat transfer deterioration while heating supercritical liquid rocket engine propellants. Journal of Supercritical Fluids, 165, 104895. https://doi.org/10.1016/j.supflu.2020.104895

Zhu, B., Zhang, Q., Ma, L., & Shi, H. (2025). Analysis of heat transfer deterioration mechanism of supercritical fluid based on VOF two-phase method and pseudo-boiling theory. International Journal of Heat and Mass Transfer, 240, 126643. https://doi.org/10.1016/j.ijheatmasstransfer.2024.126643

Xiao, R., Zhang, P., Chen, L., & Hou, Y. (2023). Machine learning based prediction of heat transfer deterioration of supercritical fluid in upward vertical tubes. Applied Thermal Engineering, 228, 120477. https://doi.org/10.1016/j.applthermaleng.2023.120477

Gong, K., Zhang, Y., Cao, Y., Feng, Y., & Qin, J. (2024). Deep learning approach for predicting the flow field and heat transfer of supercritical hydrocarbon fuels. International Journal of Heat and Mass Transfer, 219, 124869. https://doi.org/10.1016/j.ijheatmasstransfer.2023.124869

Ko, S., Gil, Y., & Park, S. (2026). Development of RP-3 surrogate fuels via multi-objective genetic algorithm for regenerative cooling CFD with supercritical property fidelity. Aerospace, 13(4), 307. https://doi.org/10.3390/aerospace13040307

Zhang, H., Gou, J., Yin, P., Su, X., & Yuan, X. (2023). Film-cooling hole optimization and experimental validation considering the lateral pressure gradient. Frontiers in Mechanical Engineering, 8, 973293. https://doi.org/10.3389/fmech.2022.973293

Wang, Y., Wang, Z., Qian, S., Qiu, X., Shen, W., Zhang, X., Lyu, B., & Cui, J. (2024). Data-driven framework for prediction and optimization of gas turbine blade film cooling. Physics of Fluids, 36(3), 035160. https://doi.org/10.1063/5.0186087

Yao, G., Li, D., Zhu, J., Tao, Z., & Qiu, L. (2024). Hybrid AI framework for the predictions of film cooling effectiveness distribution with various surface curvatures and compound angles. Applied Thermal Engineering, 257, 124147. https://doi.org/10.1016/j.applthermaleng.2024.124147

Navah, F., Lamarche-Gagnon, M.-E., & Ilinca, F. (2023). Thermofluid topology optimization for cooling channel design. arXiv:2302.04745. https://arxiv.org/abs/2302.04745

Gao, C., Xu, W., Zhu, X., Cui, J., Luo, T., Wang, D., Sun, L., Ling, W., Li, X., & Zhou, W. (2025). Enhanced regenerative cooling performance with conformal TPMS channels. Energy, 322, 135530. https://doi.org/10.1016/j.energy.2025.135530

Fu, R., Miao, X., Gao, L., Duan, L., Xu, Y., Guo, M., Luo, P., & Hu, J. (2026). Numerical study of flow and heat-transfer enhancement in lattice-core regenerative cooling channels for liquid rocket engines. International Communications in Heat and Mass Transfer, 175, 111069. https://doi.org/10.1016/j.icheatmasstransfer.2026.111069

Ebbs-Picken, T., Romero, D. A., Da Silva, C. M., & Amon, C. H. (2024). Deep encoder-decoder hierarchical convolutional neural networks for conjugate heat transfer surrogate modeling. Applied Energy, 372, 123723. https://doi.org/10.1016/j.apenergy.2024.123723

Lee, J., Shin, S., Choi, H., Lee, A., Park, B., & Lee, S. (2025). Extended multiphysics-informed neural network for conjugate heat transfer problems. International Journal of Heat and Mass Transfer, 246, 127098. https://doi.org/10.1016/j.ijheatmasstransfer.2025.127098

Lee, S., Yoon, Y., & Song, S. J. (2025). Deep learning-based identification of turbopump inducer cavitation instability using high-speed camera images. AIAA SciTech 2025 Forum. https://doi.org/10.2514/6.2025-0760

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2021). Fourier neural operator for parametric partial differential equations. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.08895

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045

DLR Institute of Space Propulsion. (2026). Test facilities at Lampoldshausen and rocket propulsion research. DLR. https://www.dlr.de/en/ra

Chapter 10 Real-Time Fault Diagnosis of Liquid Rocket Engines (LREs), Digital Twins, and Reinforcement Learning-Based Real-Time Adaptive Propulsion Control

Introduction

When people drive a car for many years, they learn its condition by sound. If the engine makes an unfamiliar noise, they go to a repair shop. If a warning light turns on, they slow down. People judge in this way by looking at several signals together. A rocket engine is the same. But a rocket engine is much faster, hotter, and more dangerous than a car.

A liquid rocket engine burns liquid fuel and a liquid oxidizer to create great force. This engine handles enormous heat and pressure in a short time. If even one part is slightly out of place, the whole system can fail. So it is important to keep watching the engine while it is running. Watching the condition in this way throughout engine operation is called real-time condition monitoring.

This chapter explains how artificial intelligence watches and controls a rocket engine. Artificial intelligence is written in English as Artificial Intelligence, or AI for short. The chapter covers four main topics. The first is diagnosis, which quickly finds abnormalities in sensor signals. The next is a digital twin (Digital Twin), which uses a computer to visualize the inside of an engine that cannot be seen. Another is a way to measure uncertainty even when an engine has aged or has manufacturing errors. The last is smart control that adjusts several valves together.

When these technologies are combined, the engine manages its own health. This trend is essential in the age of reusable rockets. When one engine is used many times, its condition changes slightly with each use. It is difficult for people to adjust it by hand every time. Artificial intelligence is a tool that reads those changes quickly and responds to them. From here, we will unpack one by one how these technologies work.

Before reading this chapter, keep one attitude in mind. Artificial intelligence is not magic. It is a tool that looks at data and finds rules. If the data is poor, the answers are poor too. So this chapter explains both the uses and the limits of each technology. It does not use unverified performance figures. Instead, it focuses on explaining the principles clearly. The goal is for readers to understand the framework of the technology.

10.1 Challenges in Real-Time Condition Monitoring of LREs

What This Section Covers

This section explains why it is so difficult to monitor a rocket engine in real time. It first shows how harsh the environment of a liquid rocket engine is. Then it points out the limits of monitoring methods used so far. Finally, it explains why artificial intelligence monitoring is needed.

First, we need to define one term. LRE, which appears often in this chapter, is short for liquid rocket engine (Liquid Rocket Engine). From here on, it will be called a liquid rocket engine or simply an engine. This engine is widely used to lift heavy objects into space.

A liquid rocket engine is a complex machine in which many parts move together. The combustion chamber is the room where fuel and oxidizer are burned. The injector is the part that sprays fuel and oxidizer into fine streams and mixes them. The turbopump is a pump that pushes propellant in at high pressure. Valves open and close the flow of propellant. Cooling channels remove heat so the walls do not

melt. These parts must move like one body for the engine to run properly.

The problem is that this environment is extremely harsh. The pressure inside the combustion chamber reaches several dozen atmospheres. The flame temperature exceeds 3,000 degrees. The turbopump rotates tens of thousands of times per minute. These values are on a completely different scale from a kitchen pressure cooker or a car engine. Even a small vibration can spread into a large force. So the engine always runs with the risk of damage.

This harsh environment is severe for sensors too. Sensors are the eyes that read the engine's condition. But if a sensor is close to a hot area, it can fail by itself. If vibration is strong, wires can break. It is like having to look at the engine with blurred eyes. So it is hard to trust just one sensor value. Several sensors must be viewed and filtered together. Even if one sensor is abnormal, other sensors can be used to check it.

The amount of signals is also serious. One sensor that reads densely can produce tens of thousands of values per second. An engine has tens to hundreds of such sensors attached. Then millions of values pour out every second. A person cannot look at all these values with the naked eye. On top of that, judgment must be completed within milliseconds. A millisecond is one-thousandth of a second. Hundreds of judgments must be made in the time it takes to blink once. This speed and volume cannot be handled without machine help.

Here is an example of how quickly an engine abnormality can turn into an accident. Pressure inside the combustion chamber begins to oscillate regularly and strongly. This oscillation can grow several times larger in just a few milliseconds. The larger oscillation strikes the injector and the wall. When this striking repeats, the metal cracks. Hot gas leaks through the crack. This process happens in the blink of an eye. That is why the first signs of abnormality must be caught as early as possible.

Another difficulty is that ground tests and flight have different conditions. In a ground test, the engine is firmly fixed. Many sensors can be attached. If a problem occurs, the engine can be shut down immediately. During flight, the situation is different. Many sensors cannot be attached because of weight. People cannot watch from nearby. So a flight engine must judge for itself with only a small number of sensors, and artificial intelligence helps with this judgment.

Rocket engine diagnosis has the same goal as diagnosis for other machines. But the conditions are much harsher. Time is short, data is scarce, and the risk is high. So methods from other fields are hard to transfer as they are. They must be refined for rockets. The technologies introduced in this chapter are all results of that refinement, and that work is still continuing.

Ground testing is also dangerous. A test in which an engine is fixed to the ground and fired is called a hot-fire test. In this test, the engine produces the same force as in real operation. In space, the situation is more difficult. For a rocket to land again, the engine is turned off and then turned on again in the air. At this time, the propellant sloshes inside the tank. This sloshing is called sloshing. The timing of valve opening also shifts little by little. These sudden fluctuations appear and disappear in an instant, so they are easy to miss.

It is easier to understand if we picture a rocket landing again. During launch, the engine produces maximum force. During landing, it must greatly reduce its force and descend gently. In this process, the engine's force rises and falls over a wide range. The liquid inside the propellant tank sloshes too. Valves open and close quickly. Temperature and pressure change rapidly. Abnormalities are likely to occur in this short and harsh moment. So this moment must be watched without missing it.

Restarting an engine in the air is also difficult. Even in orbit, Earth's gravity is still large. But because the rocket and propellant fall freely together, weight is hardly felt. This state is called microgravity. In

microgravity, liquid does not stay in place inside the tank. So small acceleration or surface tension must be used to gather the propellant to one side. If the engine is started without this process, the flow is uneven. Pressure and flow rate jump for a moment. Diagnosis can become unstable in this transient state. A normal condition may be mistaken for an abnormal one. So diagnosis must learn to withstand this transient state.

Until now, engine monitoring has often used threshold methods. If a value crosses a line set in advance, the engine is shut down immediately. This line is called a redline. For example, if pressure exceeds a set value, the engine is stopped. This method is easy to understand and fast, so it has long been used as a reliable method.

The redline method resembles a home circuit breaker. If too much electricity is used, the breaker trips. When the limit is exceeded, it cuts off without exception. This method is good for preventing major accidents. The judgment is simple, so it is fast and certain. Rocket engines also use this kind of safety shutdown as a basic measure. If a value crosses a dangerous line, the engine is turned off at once.

However, the redline method can easily miss weak abnormal signals. Sensor signals always contain noise. If the noise is large, an alarm may sound even when nothing is wrong. This is called a false alarm. If false alarms occur often, people stop trusting alarms. Conversely, tiny injector vibrations or slight seal wear remain hidden within the threshold. These weak signals are hard to see until they grow, and once they grow, it is already too late to act.

Artificial intelligence diagnosis tries to fill this gap. A person can look at only one or two instruments. Artificial intelligence looks at hundreds of sensor signals at once. It examines the shapes and relationships of many signals together. So it can read signs of abnormality before a value crosses the limit line. A review paper published in an aerospace journal in 2026 summarized this trend. The paper explains that data-driven methods have greatly increased in liquid rocket engine diagnosis. At the same time, it points out the difficulty that real fault data is scarce.

In summary, rocket engine monitoring faces three barriers. The signals are fast and noisy. Weak abnormalities appear late. A balance must also be struck between false alarms and missed detections. Artificial intelligence is an attempt to overcome these three barriers together. From the next section, we will examine the methods one by one in detail.

Artificial intelligence monitoring does not eliminate the redline method. The two methods are used together. The redline is the last line of defense against major accidents. Artificial intelligence reads weak signs in advance before that point. It is a layered safety net. The inner net catches what the outer net misses. Rocket safety is protected by overlapping several methods in this way, without relying on only one.

10.2 Finding Abnormalities Early by Looking at Multiple Sensors Together

What This Section Covers

This section deals with artificial intelligence structures that quickly detect abnormalities in engine sensor signals. It first explains telemetry and ultra-high-frequency signals. Then it introduces a neural network that looks at both time flow and channel relationships. This kind of neural network is called a temporal convolutional attention network (TCAN). This structure is not a model verified specifically for rockets, but a method used in many fields. Finally, it also points out the limits that this kind of diagnosis has not yet overcome.

First, let us define the terms. Telemetry is the sending and receiving of signals about the condition of a device that is far away. Sensors that measure pressure, vibration, temperature, and flow rate are

attached to the engine. These sensor values enter a computer in real time. Ultra-high-frequency means reading these signals at very dense intervals. Some sensors read tens of thousands to hundreds of thousands of times per second. They must read this densely to catch fast oscillations.

Piezoelectric elements are used in sensors that measure pressure. A piezoelectric element is a material that produces a small electric signal when pressed. When pressure changes quickly, the electrical signal also changes quickly. So this sensor is good at catching fast pressure oscillations. An accelerometer measures how much a part vibrates. When these two sensors are used together, pressure and vibration are viewed at the same time. In effect, the engine's condition is read from several angles.

Slow monitoring misses fast abnormalities. Imagine taking a picture of a hummingbird's wingbeat with an old camera. If the shutter is slow, the wings blur together. Only a fast shutter can capture the instant of the wing. Sensors are the same. If the reading interval is sparse, fast oscillations disappear. So dangerous high-frequency oscillations must be read densely.

Let us see why such fast monitoring is needed. Abnormalities in a rocket engine grow in the blink of an eye. There is a phenomenon in which pressure near the injector oscillates regularly. This is called chugging. It is a low-frequency lurching of the engine. If it becomes severe, valves and pipes can shake and break. This kind of oscillation grows in an instant. It is too fast for a person to see and respond. So a machine must judge in milliseconds instead.

An abnormality in one part may first appear as a signal in another part. For example, the shaft of a turbopump may vibrate first. A short time later, that vibration may lead to a change in combustion chamber pressure. It is difficult for a person to see this time delay and connection at a glance. Because a neural network looks at several sensors together, it can learn these time delays and relationships between channels.

The structure introduced here combines two methods. One is temporal convolution. Temporal convolution is a method that extracts patterns from the time flow of a signal. It widens the range from short windows to long windows and observes longer flows. The other is attention. Attention is a method that decides where to focus more among several signals. It gives more weight to the channel that matters now. When these two methods are combined, short oscillations and channel relationships are examined together. This combined structure is called a temporal convolutional attention network (TCAN).

Let us look more closely at how temporal convolution sees a long flow. Convolution is a method that scans a signal with a small window. If the window is narrow, it sees only a short moment. Rocket anomalies can also appear over a long time. So the model must look across a wide span of time. The method used here is to widen the spacing of the window. This is called dilated convolution. It skips values one by one and scans a wider span of time. It is like climbing stairs two steps at a time and reaching higher with the same number of steps. In this way, it sees a long flow with less computation.

Attention also provides an example of how focus is divided. Think of listening for a friend's voice in a noisy classroom. The ear focuses on the important sound among many sounds. Attention does the same work. It gives weight to the channel that most clearly shows the current anomaly. When the situation changes, the weight changes too. So it selects the important signals among many sensors at each moment. Temporal convolution sees the flow of time, and attention selects the channels.

This combined approach is widely used in recent research. A 2025 study proposes a temporal convolutional neural network with added attention. The study explains that it captures both short dependencies and long dependencies. Another study in the same year combines temporal convolution with graph attention. It views multiple sensors as a graph and learns their relationships together. These

studies are not specific to rockets. But they fit well with the problem of finding anomalies across multiple sensors. Rocket engine diagnosis is the same kind of problem, because it must find anomalies across multiple sensors. What has been verified is that this structure found anomalies well in general time-series data. There is still no empirical proof that it works directly for rocket sensors and failure types. This book proposes this structure as a direction to apply to rocket diagnosis.

Other research in the rocket field points in a similar direction. Deep learning diagnosis that reads multiple sensor signals together is increasing. A 2022 study uses an interpretable long short-term memory network. This neural network classifies engine anomalies using data from multiple sensors. A 2024 study deals with fault detection based on feature learning. All of these efforts try to combine multiple signals and find anomalies quickly.

Where this diagnosis is computed is also important. If signals are sent to a distant computer for judgment, time is lost. To produce an answer within milliseconds, computation must happen next to the engine. This method of computing right next to the device is called edge computing. A neural network is placed on a small dedicated chip and run there. As soon as it receives a signal, it judges the situation on the spot. Only then can it detect danger within milliseconds. So a diagnostic neural network must be small and fast. Even if a large neural network is accurate, its value falls if it is slow.

A diagnostic model usually goes through three stages. First, it smooths and aligns the sensor signals. This is called standardization. Each sensor has a different scale and unit. Next, it cuts out a short time window and feeds it into the model. Finally, it divides the state into several types. It classifies whether the state is normal, a sensor anomaly, a pump anomaly, or combustion oscillation. The diagnostic model repeats these three stages continuously at very high speed.

There are many types of anomalies that this diagnosis tries to catch. There is chugging, which sloshes with low-frequency vibration near the injector. There is high-frequency oscillation, which rings with faster vibration. There is leakage caused by wear in the turbopump seal. There is also a problem in which worn bearings increase vibration. Each anomaly leaves a different signal shape. The diagnostic model learns these differences in shape. It looks at which shape appears and identifies the type of anomaly.

One point requires care here. Some writings on this structure cite numbers such as accuracy or latency. But those numbers cannot be confirmed in published papers. If unconfirmed numbers are written as facts, they mislead readers. So this book does not use such numbers. Instead, it explains this structure only as a principle of early fault diagnosis. The key point is that it looks at multiple sensors together and catches anomalies early.

In a real engine, diagnostic results are handled by classifying the risk level of a failure. A critical signal, such as pressure far beyond the danger line, shuts down the engine immediately by itself. If multiple confirmations are forced for such a signal, shutdown is delayed and the accident grows. By contrast, weak signs are acted on when several signals point to the same anomaly. This prevents a healthy test from being ruined by one mistaken judgment from a model. So diagnostic outputs are used together with safety logic. Human supervision and rule-based safety devices must also be present.

To trust a diagnostic model, we must know why it made its judgment. A neural network is like a box whose inside is hard to see. It gives an answer, but it does not explain the reason. In a dangerous place such as a rocket, the reason matters. So research is increasing on showing which sensor led the judgment. This is called interpretable diagnosis. People look at the reason and decide whether to trust the diagnosis. Only when the reason can be known can artificial intelligence be used for safety judgment.

Another difficulty is the lack of data. Real failure data from rocket engines is very rare. A failure usually means an expensive accident. So there is a lot of normal data and almost no failure data. Artificial intelligence has difficulty learning well when data is scarce. Researchers use several methods to solve this problem. One is to create synthetic data that imitates failure conditions with a physics model. Another is to deliberately insert predefined failures into a test model. This procedure of deliberately inserting a failure is called fault injection. Synthetic data generation and fault injection do not mean the same thing. The lack of data remains a major challenge in this field.

There is also a method for transferring what was learned from one engine to another engine. This is called transfer learning. It is similar to learning a motorcycle faster if you already know how to ride a bicycle. Knowledge already learned is carried over to a new situation. Then the model can learn quickly even with little data from the new engine. A 2026 review paper in an aerospace journal also points out this difficulty. The issue is that real failure data is scarce and failure types are numerous. The paper organizes diagnostic research around these three practical problems.

The goal of diagnosis is not only to prevent accidents. It is also used to perform maintenance properly. If there is no anomaly, parts do not need to be replaced. If signs of an anomaly appear, repairs are made in advance. This method of maintenance based on condition is called condition-based maintenance. Reusable engines reduce costs with this method. Diagnosis is a tool that protects both safety and cost.

10.3 A Digital Twin That Draws the Inside of an Invisible Combustion Chamber

This section deals with dangerous oscillations inside the combustion chamber. These oscillations are called combustion instability. First, it explains how sound and flame strengthen each other. Then it looks at how a computer reconstructs the invisible flow of pressure. This method uses a neural network that follows physical equations. Finally, it reviews the usefulness and limits of this method.

Let us begin with the term. Combustion instability is a phenomenon in which pressure inside the combustion chamber oscillates strongly. When flame and sound move in step, this oscillation grows. A small pressure oscillation shakes the heat release of the flame. The shaken flame then increases the pressure oscillation again. If this feedback repeats, the oscillation grows like a snowball. In severe cases, the combustion chamber wall and injector break. In rocket history, this phenomenon has long been a difficult problem.

This oscillation has been a major problem since the early days of rocket development. Even the large engines that sent people to the Moon struggled with this problem. Pressure rang loudly inside the combustion chamber and destroyed engines. Researchers controlled it by changing the injector shape many times. That shows how difficult combustion oscillation is to handle. Even today, this problem is checked whenever a new engine is built.

Oscillation has several shapes. Think of the combustion chamber as a cylinder. Sound can resonate along the length of the cylinder. It can also resonate while rotating around the circumference. It can also resonate across the diameter. Each shape has a different frequency. If we know which shape is growing, response becomes easier. That is why it is important to know the shape and strength of the oscillation.

There is an old principle that explains this feedback. If more heat is released at the moment when pressure rises, the oscillation grows. This principle is called the Rayleigh criterion. It is easy to understand by thinking of pushing a swing. If you push the swing when it moves forward, it rises higher and higher. If you push at the opposite moment, the swing stops. Flame and sound also strengthen or calm each other by the same principle. The timing of exchanged force separates stability from instability.

This feedback also resembles the relationship between a tuning fork and a microphone. In an auditorium, if a microphone is brought close to a speaker, a sharp sound occurs. This happens because sound from the speaker enters the microphone and is amplified again. This feedback makes the sound grow like a snowball. In a combustion chamber, the flame is like the speaker and the sound is like the microphone. If their phases line up, the sound grows out of control. So a design is needed to break this feedback.

The problem is that it is hard to look directly inside the combustion chamber. The inside is too hot and the pressure is high. Sensors cannot be placed just anywhere. Pressure sensors can barely be attached to a few points on the wall. But dangerous oscillation occurs throughout the combustion chamber. It is hard to know the full picture from only a few points on the wall. So a method is needed to draw the whole picture from limited observations.

This is where the digital twin appears. A digital twin is a computer model that follows a real device exactly. While the real engine runs, the model runs with it. It receives sensor values and continually adjusts the model to match reality. This makes it possible to estimate the state at places without sensors. The name comes from a model that runs beside the real system like a twin.

A digital twin is somewhat like car navigation. Navigation receives real road conditions and redraws the route. A digital twin also receives signals from the real engine and redraws its state. The difference is speed and precision. A rocket twin must match the state every millisecond. Only then can it follow rapid oscillations. To reach this speed, the model must be made light.

Finding the inside from a few wall sensors is a problem solved backward. Usually, we know the cause and find the result. Here, we know only the wall pressure, which is the result. From that result, we trace back the internal pressure field, which is the cause. A problem that finds the cause from the result in this way is called an inverse problem. An inverse problem is difficult because it can have many answers. Physical equations leave only the answers that match physics among these many answers.

The method used to fill the inside of this digital twin is a physics-based neural network. This neural network is called a physics-informed neural network (PINN). From here on, it will be called a physics neural network. An ordinary neural network learns only from data. A physics neural network learns from data while also following physical equations. During training, it is penalized when it violates the equations. So even with little data, it produces answers that match physics.

Let us look a little more at how a physics neural network learns. An ordinary neural network reduces only the difference from the correct answer. A physics neural network reduces two things together. One is the difference from sensor values. The other is the degree to which it violates physical equations. It learns to lower both penalties together. Then the answer follows physics even in places without sensors. Physical rules fill the empty spaces in the data. It is similar to how outline lines guide the colors in a coloring book.

This method can become a low-cost monitoring tool. To calculate the inside of a combustion chamber precisely, a large computer must run for a long time. During a test, there is no time for that. Once a physics neural network has been trained, it produces answers quickly. So it can draw the internal state even during a test. This fast picture is used to watch for signs of dangerous oscillation. It is like looking inside without stopping the test.

The equation that describes the sound shapes of a combustion chamber is the Helmholtz equation. This equation tells us what shape sound takes as it resonates inside a closed space. It is similar to the way certain notes resonate well inside a wind instrument. A physics neural network learns while following this equation and the boundary conditions. It receives values from several wall sensors as input. Then it

reconstructs the pressure flow across the entire combustion chamber. It shows, like a picture, where the oscillation is strong.

The sound in a combustion chamber changes depending on how it responds at the boundaries. The injector face and nozzle throat reflect some sound and absorb some sound. The degree of this reflection and absorption is called an acoustic boundary condition. A physics neural network learns while following these boundary conditions as well. Only then does it draw the sound of a real combustion chamber, not the sound of an empty room. The Rayleigh criterion discussed above and these boundary conditions determine rocket combustion instability.

This direction is active in recent research. Physics-informed neural networks are strong at reconstructing acoustic pressure fields from limited observations. This is because pressure is measured at only a few points, and velocity is rarely measured. One 2024 study uses a physics-informed neural network to learn thermoacoustic interactions in a combustor. Thermoacoustics means a phenomenon in which heat and sound are coupled. The study identifies values such as pressure damping and heat-release coupling from data. One 2026 study estimates the acoustic absorption properties of a combustion chamber wall using a domain-decomposed physics-informed neural network. These studies deal with thermoacoustic learning in general combustors and estimation of boundary properties. There is still no demonstration that reconstructs the full pressure field in real time from a small number of sensors in an actual liquid rocket combustion chamber. This book proposes that kind of real-time reconstruction as a direction for future progress.

Physics-informed neural networks are also used in clean combustion research for aerospace propulsion. One 2026 paper sees this trend as a path toward sustainable propulsion. It connects physical equations with data to describe combustion more accurately. This research extends not only to rockets but also to aircraft engines. The idea of a neural network that respects physics is spreading across many fields.

Digital twins are already used in aircraft engine maintenance. One large aircraft engine reportedly used a digital twin to detect several abnormalities in advance. It is said to have predicted turbine blade wear and combustion problems. This case shows that monitoring models can be useful in practice. Rocket engines operate in harsher conditions than aircraft engines and have less data. So moving this direction to rockets requires more verification.

Physics-informed neural networks also have weaknesses. Adding equations to learning makes training difficult. The weights of the physics penalty and the data penalty must be balanced well. If the weights are off, the answer leans to one side. This tuning becomes harder as combustion becomes more complex. So physics-informed neural networks are not a cure-all. The equations and structure must be chosen for the problem. Human knowledge is still needed for this choice.

More broadly, physics-based machine learning is growing quickly across combustion research. One review paper published in 2025 summarizes this trend. It explains the field by dividing it into combustion chemistry, reacting flows, and several other problems. It argues that adding physical knowledge to learning can keep answers consistent with physics even when data is scarce. Another review paper from 2026 focuses on turbulent combustion in aerospace propulsion. It explains that the method connects the rigor of physics with the power of data.

The strength of this method is that it uses data and physics together. When there are few sensors, data alone cannot easily fill the gaps. Physical equations narrow those gaps. So the full state can be described even with a few inexpensive sensors. Its ability to run quickly for monitoring is also highly useful.

The limits are clear too. Flame reactions and turbulence are highly complex. The Helmholtz equation is a reduced model that treats sound in a simple way. It cannot include every detail of real combustion. So

this digital twin is used for fast monitoring. Precise judgment must be supplemented with high-fidelity analysis and real tests. Detailed numbers such as specific frequencies or pressure amplitudes have not been confirmed in published papers. So this book does not use such numbers. Instead, it explains only the principle of how this method works and where it is useful.

For a digital twin to run well, it must be kept aligned with the real system. If the model gradually drifts from reality, the picture becomes distorted. So the model is continuously corrected with sensor values. If this correction is slow, the twin loses track of the real system. If correction is too frequent, the twin follows even noise and becomes unstable. Finding the balance between the two is the key technical issue. A good twin does not lose the real system, but it is not swayed by noise.

Monitoring is not the only way to control combustion oscillation. Structures that absorb sound are sometimes placed in the wall. These structures are called acoustic absorbers. The injector arrangement may also be changed to break feedback. Artificial intelligence monitoring is a tool that helps this design. It tells in advance under which conditions oscillation will grow. Oscillation can be controlled only when design and monitoring work together.

10.4 How to Measure Engine Aging and Unit-to-Unit Differences

This section deals with problems caused by engine aging and differences between individual units. It first explains why reusable engines change little by little. It then looks at a way to measure the uncertainty of that change. This method is a neural network that gives both an answer and the width of that answer. Finally, it explains why this uncertainty matters for safety. One point should be made clear in advance. The studies cited here predict the life of general machines. They have not been verified with rocket engine data, and rocket application remains a future possibility.

A reusable engine uses the same engine many times. It is like moving from using a paper cup once and throwing it away to using a ceramic cup many times. This approach greatly lowers the cost of space transportation. But each use wears parts down little by little. Injector surfaces wear. Turbine blades erode. Cooling passages change because of thermal fatigue. This kind of performance change over time is called aging.

Another variable is manufacturing variation. Even when parts are made from the same drawing, their actual dimensions differ slightly. Each part has tiny errors. Because of these errors, each engine responds differently. When aging and manufacturing variation overlap, each engine develops its own character. Even with the same valve command, each engine produces a slightly different result. So the controller must know this difference.

This change directly affects engine performance. Wear and erosion gradually reduce thrust. When thrust falls, the rocket trajectory changes. The completeness of combustion also changes. Characteristic exhaust velocity is a value that represents this completeness. When this value falls, the same propellant produces less force. So it is important to measure aging and variation in advance.

The term characteristic exhaust velocity needs a short explanation. This value is an indicator of how completely the combustion chamber burns propellant. It is determined by combustion chamber pressure, nozzle throat area, and propellant flow rate. It is different from the actual exhaust velocity from the nozzle or nozzle efficiency. It is a measure only of how well combustion occurs. If this value is high, combustion is complete. If it is low, the propellant burns less completely. When an injector wears, mixing gets worse and this value falls. So changes in this value can indicate the health of an engine.

This change can be compared to old sneakers. New sneakers have firm soles. After long use, the soles wear down and become slippery. Even with the same effort, a runner moves less well. An engine also

loses performance little by little as it is used. The problem is that this decline differs from engine to engine. Some engines wear quickly, and others wear slowly. So the condition of each engine must be examined separately.

Knowing aging is the key to safe reuse. If the condition is unknown, two mistakes can happen. An engine that is still usable may be discarded too early. That is waste. Conversely, an engine that is already fatigued may be used too hard. That can lead to an accident. If the condition is known accurately, the engine can be used as much as possible without crossing the danger line, avoiding both mistakes.

This is where a variational Bayesian neural network is used. In English, it is called a Variational Bayesian Neural Network. From here on, this book calls it a Bayesian neural network. An ordinary neural network gives only one answer. For example, it answers that thrust loss is exactly 3 percent. A Bayesian neural network gives the range of the answer as well. It reports both the answer and the degree of confidence, such as saying that thrust loss is likely between 2 and 4 percent.

The way a Bayesian neural network gives a width can be pictured simply. Suppose we ask many people instead of only one person. If many answers are similar, confidence is high. If the answers differ widely, confidence is low. A Bayesian neural network contains these many opinions inside one model. So it gives both the central value of the answer and the spread around it. A narrow width means the model is confident. A wide width means it still does not know well.

There are two kinds of uncertainty. One comes from variation in the data itself. This is unavoidable variation, like sensor noise. The other comes from the model having learned too little. It grows in situations the model has not experienced. When the model first encounters a new kind of aging, this uncertainty becomes large. Separating the two kinds makes it easier to respond. It helps decide whether to collect more data or run more tests.

A Bayesian neural network is harder to train than an ordinary neural network. This is because it must learn not one answer, but a distribution of answers. Finding this distribution exactly requires a large amount of computation. So researchers use approximation methods. They learn the distribution by changing it into a shape that is easier to handle. This method is called variational inference. The word variational comes from this approximation. Because it uses approximation, the results must be checked against real data.

Aging prediction is not only a rocket problem. Aircraft engines, wind turbines, and factory machines face the same concern. They are all machines used for a long time. So research in this field moves back and forth. Methods that worked well for other machines are moved to rockets. The demanding conditions of rockets can also produce new methods. When fields exchange methods in this way, technology grows faster.

This degree of confidence is called uncertainty. Knowing uncertainty is important for safety. A model should be handled differently when it answers with confidence and when it does not know well. When the model says it does not know, people act more cautiously. They leave a larger safety margin. They confirm the result with additional tests. It is the same principle as giving the chance of rain in a weather forecast. If the probability is half and half, people take an umbrella and watch the situation.

This direction is growing quickly in several fields. Research that predicts the remaining life of parts is a representative example. The remaining life is called remaining useful life. One 2024 study predicts this life using a Bayesian neural network. It gives the answer as a probability distribution, not as a single value. One 2025 study uses a Bayesian recurrent neural network to handle several uncertainties together. These studies are not rocket-specific. But the problem of dealing with the condition of aging machines is the same. The same idea can be used to predict aging in rocket engines.

Aircraft engine data is one widely tested target for this method. Researchers compete using public engine aging data. They compare the predictions of several methods on the same data. This makes it possible to judge which method is better. Rocket data is scarce, but public data like this can build the framework. The established method can later be moved to rocket data and checked.

The controller receives this uncertainty information and makes decisions. If uncertainty is large, it operates conservatively. It may set thrust slightly lower or leave more margin. If uncertainty is small, it draws out more performance. This makes it possible to operate according to the condition of the engine. It is safer than handling all engines with one fixed rule.

Uncertainty information is also used in maintenance decisions. If the model predicts aging with confidence, planning becomes easier. If the model says it does not know well, the part is opened and inspected directly. This reduces blind maintenance. Because maintenance is based on condition, parts are saved. At the same time, dangerous parts are not missed. Knowing uncertainty protects both economy and safety.

There are limits too. Just because artificial intelligence gives a probability does not mean that the probability is automatically correct. If the model learns incorrectly, it gives false confidence. So it must be calibrated and verified with real test data. This means checking whether the uncertainty given by the model matches the actual error. A probability that has not gone through this check is hard to trust. In dangerous systems such as rockets, this verification must be stricter. Numbers such as detailed confidence intervals have not been confirmed in published papers. So this book does not use such numbers. Instead, it explains only the framework of the idea of measuring uncertainty.

10.5 Reinforcement Learning Control That Adjusts Multiple Valves Together (LUMEN, DLR P8.3)

This section deals with intelligent control that adjusts several engine valves together. It first explains what reinforcement learning is. It then looks at why rocket engine control is difficult. Next, it introduces a real experimental case from Germany. Finally, it examines the remaining tasks before this technology can be used in flight.

Reinforcement learning is a method that learns through trial and error. It is similar to how a baby learns to walk. Think of receiving a penalty for falling and a reward for walking well. Good actions receive rewards, and bad actions receive penalties. By trying many times, the system finds a way to receive more rewards. A method that adds a neural network to this is called Deep Reinforcement Learning (DRL). From here on, this book calls it deep reinforcement learning.

In reinforcement learning, two impulses collide. One is the impulse to use a good method already known. The other is the impulse to try a new method. If it always uses only what it knows, it cannot find a better path. If it always tries only new things, it takes risks and wanders. So a balance is needed between the two. This balance can be seen as a tug-of-war between the wish to find something new and the wish to use what is already known. Early in learning, the system tries many things. Later, it uses what it knows.

In rockets, this trial and error cannot be done with a real engine. We cannot learn by breaking engines. So the system first learns in a model inside a computer. This model is called a simulator. In a simulator, it is fine to fall thousands of times. The system tries freely and finds good control. It is like learning the knack by playing many rounds of a game.

Let us see why rocket engine control is difficult. The values inside an engine are linked to one another. If the oxidizer valve opens, pressure rises. If pressure rises, turbopump rotation changes. The ratio of fuel

to oxidizer also moves with it. This ratio is called the mixture ratio. If one value is touched, other values move one after another. So one valve cannot be controlled in isolation. Control that handles many values at the same time is needed.

In a landing rocket, this control is even more difficult. For the rocket to come back down, thrust must be greatly reduced. It may be reduced to below one third of maximum thrust. Adjusting thrust across such a wide range is called deep throttling. It means throttling deeply. Even then, the mixture ratio must be kept proper. If the mixture ratio is wrong, combustion worsens and parts are damaged. Many values must be matched at the same time, quickly, and safely.

The German Aerospace Center studies this problem with a real engine. The German Aerospace Center is abbreviated as DLR. DLR built an experimental engine called LUMEN. LUMEN is a pump-fed engine that uses liquid oxygen and methane. Its thrust class is 25 kilonewtons. It is the size of a small rocket engine. This engine first ran on the P8.3 test stand in March 2024. P8.3 is a test facility at DLR Lampoldshausen. LUMEN is a proving ground for reusable engine technology. It studies condition monitoring and smart control together.

LUMEN uses a specific cycle. This cycle is called expander bleed. It drives the turbine with part of the fuel heated in the cooling channels. This cycle has a relatively simple and safe structure. So it is well suited to testing many technologies. LUMEN is an open test stand used by industry and research institutes together. New control methods can be applied to a real engine.

The P8.3 test stand began operation in 2021. This test stand was built by expanding an earlier facility. In the past, only combustion chambers could be tested. P8.3 tests the whole engine and the turbopump. So control methods can be checked under real conditions. LUMEN and P8.3 form the basis for research on artificial intelligence control.

DLR attaches artificial intelligence control to this engine. DLR official materials explain that neural networks model heat transfer in the cooling channels. They state that artificial intelligence control matches thrust and monitors combustion instability. Kai Dresia's 2025 doctoral dissertation covers this research in detail. The dissertation controls a rocket engine with deep reinforcement learning. It compares the results with conventional model predictive control. It reports an average error of about 1.3 percent while controlling several values at the same time. This value is the result presented in Dresia's dissertation. It comes from simulation and test stand research. It is not a value demonstrated in flight across the full operating range of a complete engine.

In earlier research, DLR applied neural network control to real tests. In 2023, it verified neural network control in a LUMEN turbopump test. It also conducted research that models heat transfer in cooling channels with a neural network. In 2024, it announced the results and progress of LUMEN component tests. These studies build up and lead to real engine control. This is not a single success. It is an accumulation over many years.

Let us look a little more at the framework in which a controller handles states and actions. At every moment, the controller reads the state and produces an action. That action creates the next state. In the next state, it produces another action. This process continues without stopping. A problem in which states and actions follow one another in this way is called a sequential decision problem. Reinforcement learning is well suited to handling this sequential decision-making. It learns not one judgment, but a continuing flow of judgments.

Let us see the principle by which a reinforcement learning controller makes judgments. The controller first reads the state of the engine. These are values such as combustion chamber pressure, valve position, pump rotation, and temperature. These values are called the state. The controller looks at the

state and decides how much to move the valves. This movement is called an action. If the result of the action is good, it receives a reward. If it is bad, it receives a penalty. If the engine is close to the target thrust, it receives a reward. If the mixture ratio is wrong or vibration grows, it receives a penalty. The rule that contains these rewards and penalties is called the reward function. The controller refines valve operation in the direction that earns more rewards.

Designing the reward function well decides the success or failure of control. The rules for rewards and penalties must contain the goal as it is. If only thrust is rewarded, the controller may ignore vibration. So thrust, mixture ratio, and vibration are all included in the rewards and penalties. Moving a valve too suddenly is also penalized. Sudden operation puts strain on parts. The weights of several goals must be balanced well. This weighting is the core task in control design.

In industry, automated operation of reusable engines is also a major topic. The latest engine from one U.S. company aims at repeated reuse. This engine is known to have reduced the number of parts and changed them to make production easier. It is said to have removed heat shields to reduce weight and complexity. These changes reflect a trend toward using engines often and servicing them quickly. But this is a company development case. It is not a performance value verified by scholarship. This book cites such cases only as references for the trend.

The strength of this method lies in handling entangled problems as a whole. It is difficult for people to write rules by hand for each valve. Because values are linked to one another, the rules become complex. Reinforcement learning learns these linked relationships on its own from data. It looks at many values together and finds a point of balance. But a controller that has learned once does not automatically adapt to aging it has not experienced. There are several ways to respond to changes in an engine. One way is to read the state first and reflect it in control. There is robust control, which makes the system stay steady even when changes occur. Another way is to make the controller learn again from new data. Online adaptation during operation creates a new risk: unverified changes in control. So the path to use is decided by considering safety.

Reinforcement learning control can be compared with existing methods. One long-used method is model predictive control. This method looks ahead with a mathematical model of the engine and decides the valves. If there is an accurate model, it works well. But if the engine ages or the model becomes inaccurate, performance declines. Reinforcement learning learns directly from data, so it is flexible with such changes. The two methods each have strengths and weaknesses. So researchers compare them and combine them.

Let us see why keeping the mixture ratio proper is important. The ratio of fuel to oxidizer is the mixture ratio. Flame temperature is usually highest near the ratio at which the fuel and oxidizer match exactly. If the ratio moves away from this point, the temperature falls whether fuel remains or oxidizer remains. Rockets often operate with a little more fuel on purpose. This is to lower the flame temperature touching the wall and protect the wall. But if too much fuel is added, thrust falls. Kerosene-based fuels create soot at this point. Engines that use hydrogen produce almost no soot. In this way, the proper mixture ratio differs for each propellant combination. So the mixture ratio must be maintained even while thrust is changed. This is why control that matches several values at the same time is needed.

Here, numbers are handled carefully. Detailed performance values such as fast deep throttling time, mixture ratio error range, or vibration reduction rate have not been confirmed in published papers. Using unconfirmed performance values would exaggerate the facts. So this book does not use such numbers. The confirmed value is an average error of about 1.3 percent in multivariable control. Only this value is used as evidence.

It is worth noting what this average error of 1.3 percent means. This value is the result of matching several variables at the same time. It means that several goals, not just one, were maintained together. It is difficult to achieve this balance with hand-written rules. Artificial intelligence learned the linked relationships and found a balance. But this value comes from the limited conditions of a specific study. It is not a value that can be transferred as it is to every engine. If conditions change, it must be checked again.

There is still a long way to go before this technology can be used directly in flight engines. A reinforcement learning controller first learns in simulation. Inside a computer, mistakes are not dangerous. Then hardware is connected and tested. This test is called hardware-in-the-loop testing. Real parts and the neural network are run together. Next, safety is checked through ground firing tests. Only after all these stages are passed can flight be discussed. Rocket certification is very strict.

One major difficulty here is the gap between the simulator and reality. A computer model is not exactly the same as a real engine. Control that learned well in the model may be off in reality. This difference is called the simulator-to-reality gap. To reduce this gap, the model must be made more accurate. The model must be continuously adjusted with real test data. Test stands such as LUMEN are used for this adjustment. Artificial intelligence can be trusted only when real data exists.

Another principle is to place safety devices in layers. A reinforcement learning controller may issue a strange command. Rule-based safety devices are kept together to prepare for that case. If the controller's command crosses a danger line, the safety device blocks it. Artificial intelligence raises performance, and rules protect the floor. These two layers are needed to apply artificial intelligence to a real engine.

Valves also have physical limits. A valve cannot open or close faster than a set speed. This maximum speed is called the valve actuator rate limit. If artificial intelligence gives a sudden command beyond this limit, shock reaches the piping. Water hammer, in which liquid suddenly stops and strikes the pipe, is one such example. So the controller's commands are limited to speeds the valve can produce. In addition, if the artificial intelligence becomes unstable, a safety transfer device immediately hands control to a classical controller. The classical controller protects the engine safely with rules that have long been verified.

The value of reinforcement learning control is clear in reuse. When one engine is used many times, its condition differs each time. Fixed control rules cannot keep up with this change. A reinforcement learning controller reads the state and adjusts the valves accordingly. It handles an old engine as old and a new engine as new. This flexibility lowers the cost of reusable rockets. That is why many organizations around the world put effort into this research.

However, artificial intelligence control is still at the ground test stage. It is not widely used in actual flight engines. Flight certification requires long testing and records. It must be proved that new control is safe in any situation. This proof takes much data and time. So artificial intelligence control is now growing in laboratories and on test stands. When the results accumulate, they will someday lead to flight.

10.6 Conclusion and Future Outlook

This section looks at how the four technologies above are tied together as one. It then summarizes why this bundle is needed in the age of reusable rockets. Finally, it reviews the safety principles needed for artificial intelligence engine management.

Intelligent operation of a liquid rocket engine connects four technologies. Sensor diagnosis looks at many signals together, reads weak signs early, and finds early abnormalities. A digital twin restores

invisible internal pressure flows with a neural network that follows physical equations. A Bayesian model measures the uncertainty of aging and variation, and gives both an answer and the range of the answer. A reinforcement learning controller adjusts several valves together and finds a balance point among linked values.

These four technologies exchange data with one another. When diagnosis finds an abnormality, the controller responds. When the digital twin reports the internal state, diagnosis becomes more accurate. When the Bayesian model reports uncertainty, the controller becomes more cautious. In this way, the four technologies turn as one loop. It is a structure in which the engine checks its own condition and adjusts itself. This structure is called autonomous health management.

These four technologies make up for one another's weaknesses. Diagnosis alone makes it hard to know the exact cause of an anomaly. When the digital twin shows the inside, the cause becomes clearer. A controller alone does not know how old the engine has become. When the Bayesian model reports uncertainty, control becomes safer. One technology fills the gap left by another. That is why they are tied together, not treated separately. Only when they are tied together does autonomous management become complete.

This flow is not yet a finished product. Most of it is still in the research and testing stage. Many results are also early studies that have not gone through peer review. So it is too early to state the achievements as settled facts. This book makes that point clear. For now, the task is to understand the framework and direction of the technology. More time is needed before it is widely used in real flight engines.

This flow is essential in the age of reusable rockets. When one engine is used many times, its condition changes each time. Fixed rules alone make it hard to handle every change. The values change too quickly and are too interconnected for people to adjust them by hand each time. Artificial intelligence is a tool for reading and responding to these changes quickly. Autonomous management is needed to lower the cost of space transportation.

But artificial intelligence control must have safeguards. One is explainability. People must be able to know why the model made that judgment. Another is independent verification. The model's judgment must be checked again by another method. The last is transition to a safe state when failure occurs. Even if the model becomes unstable, the engine must be able to stop safely. Without these three safeguards, it is hard to entrust an engine to artificial intelligence.

Artificial intelligence is not a tool for trusting the engine in place of people. Artificial intelligence is a tool for looking at the engine in greater detail. People borrow the eyes of artificial intelligence to understand the engine more deeply. The final judgment and responsibility still belong to people. On this principle, artificial intelligence makes rockets safer and cheaper.

Summary

Let us review the main points of this chapter. A liquid rocket engine is a very fast, hot, and dangerous machine. To use this engine safely, real-time monitoring and control are needed. The traditional threshold method can easily miss weak anomalies. Artificial intelligence fills this gap by looking at many sensors together. It tries to read weak signs before values cross the danger line.

Early diagnosis looks at both time flow and channel relationships. A structure that combines temporal convolution and attention is a typical example. This structure finds anomalies early across many sensors. The judgment must be completed within milliseconds on a small chip beside the engine. However, rocket-specific performance figures are still hard to confirm. So this book explains only the principle. Diagnosis is used together with human supervision and rule-based safety devices.

A digital twin draws the inside of the combustion chamber that cannot be seen. It uses a physics-informed neural network that follows physical equations. With a few sensors on the wall, it restores the full pressure flow. It is very useful for monitoring combustion instability. But it cannot capture every detail of a complex flame. It is used for fast monitoring and supplemented with detailed analysis. A digital twin must be constantly corrected with real signals to do its job.

A Bayesian neural network measures the uncertainty of aging and variation. It gives both an answer and the range of that answer. A narrow range means it is confident, and a wide range means it does not know well. This information is used by the controller to decide whether to act cautiously. It is also used to decide when maintenance should be done. But the model's probabilities must be verified with real data.

A reinforcement learning controller adjusts several valves together. Germany's DLR is studying this with the LUMEN engine and the P8.3 test bench. Public materials report an average error of about 1.3 percent for controlling several variables. Some detailed performance figures are not used because they lack confirmation. Several stages of verification remain before this technology can be used in flight.

Let us return to one principle that runs through this chapter. Artificial intelligence is a tool for improving performance. But safety is protected by rules and people. Both layers must be present before artificial intelligence is connected to a real engine. Unverified numbers should not be trusted. Early research and verified achievements should be read separately. This attitude is the right way to handle a dangerous technology like rockets.

When these four technologies are tied together, the engine takes care of its own health. The four technologies exchange data with one another and move as one loop. This autonomous management is the foundation of the reusable rocket age. That is because when one engine is used many times, its condition changes each time. But for safety, explainability, independent verification, and a safe transition device must be present together.

The technologies discussed in this chapter answer different questions. Diagnosis asks what is wrong now. The digital twin asks what the unseen inside is like. The Bayesian model asks how much this judgment should be trusted. The controller asks what should be done as a result. When these four questions come together, the engine's condition and response are connected as one. Good artificial intelligence engine management handles all four questions together.

Finally, this chapter states how its materials should be treated. Much of the latest research cited in this chapter is at an early stage. Many papers are preprints that have not gone through peer review. Company engine cases only show the direction of industry. They do not carry the same weight as academically verified results. So this chapter gave more weight to confirmed values and principles. Figures that could not be confirmed were left out. I hope readers understand the direction of the technology correctly. Artificial intelligence is a tool in the long journey toward making rockets safer and cheaper.

Related Supporting Materials

Yang, S., Wu, J., Xie, C., Zhong, Z., Cheng, Y., Wang, B., Liu, Y., & Lu, J. (2026). Data-Driven Intelligent Fault Diagnosis: A Review from the Perspectives of Three Practical Issues in Liquid Rocket Engines. International Journal of Aeronautical and Space Sciences, 27(4), 3438-3463. https://doi.org/10.1007/s42405-026-01131-9

Wang, T., Ding, L., & Yu, H. (2022). Research and Development of Fault Diagnosis Methods for Liquid Rocket Engines. Aerospace, 9(9), 481. https://doi.org/10.3390/aerospace9090481

Deng, L., Cheng, Y., & Shi, Y. (2022). Fault Detection and Diagnosis for Liquid Rocket Engines Based on Long Short-Term Memory and Generative Adversarial Networks. Aerospace, 9(8), 399. https://doi.org/10.3390/aerospace9080399

Urgolo, A., Pill, I., Waxenegger-Wilfing, G., & Freiberger, M. (2024). Property Learning-Based Fault Detection for Liquid Propellant Rocket Engine Control Systems. OASIcs, 125, 15:1-15:20. https://doi.org/10.4230/OASIcs.DX.2024.15

Waxenegger-Wilfing, G., Dresia, K., Deeken, J., & Oschwald, M. (2021). Machine Learning Methods for the Design and Operation of Liquid Rocket Engines: Research Activities at the DLR Institute of Space Propulsion. arXiv. https://arxiv.org/abs/2102.07109

NASA. (2009). Distributed Health Monitoring System for Reusable Liquid Rocket Engines. NASA Technical Reports Server. https://ntrs.nasa.gov/api/citations/20090040299/downloads/20090040299.pdf

Wang, C., Liu, Q., Qu, L., Chang, W., Wang, T., Duan, P., & Cao, Y. (2025). Self-attention-enhanced Temporal Convolutional Networks for Anomaly Detection in Time Series. Proceedings of the 2025 4th International Conference on Cryptography, Network Security and Communication Technology. https://doi.org/10.1145/3723890.3723911

Wang, D., Zhang, H., & Yu, Z. (2026). MST-MFGAT: An adaptive temporal convolution with feature graph attention network for time series anomaly detection. Neurocomputing, 664, 132134. https://doi.org/10.1016/j.neucom.2025.132134

Harrje, D. T., & Reardon, F. H. (1972). Liquid Propellant Rocket Combustion Instability. NASA SP-194. https://ntrs.nasa.gov/citations/19720026079

Culick, F. E. C. (2006). Unsteady Combustion in Liquid Rocket Engines: Dynamics, Instabilities, and Control. NATO RTO-EN-AVT-143.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045

Mariappan, S., Nath, K., & Karniadakis, G. E. (2024). Learning thermoacoustic interactions in combustors using a physics-informed neural network. Engineering Applications of Artificial Intelligence, 138, 109388. https://doi.org/10.1016/j.engappai.2024.109388

Li, A., Du, J., & Liu, Y. (2026). Domain-decomposition physics-informed neural network approach for the acoustic boundary impedance estimation of a combustion chamber model. Journal of Vibration and Control. https://doi.org/10.1177/10775463261433496

Wu, J., Wang, X., Zhang, G., Liu, J., Li, X., Zhang, Y., Zhang, H., Lyu, J., Wang, B., & Wu, Y. (2025). Physics-informed machine learning for combustion: A review. arXiv. https://arxiv.org/abs/2509.03347

Mousavi, M., Caldwell, C., Baltes, J., Aljasem, M., & Lee, B. J. (2025). Physics-Informed Neural Networks in Clean Combustion: A Pathway to Sustainable Aerospace Propulsion. arXiv:2509.08094. https://arxiv.org/abs/2509.08094

Selvam D., C., Devarajan, Y., Raja, T., Tiwari, A., Sunil Kumar, M., Sharma, K., Agrawal, A. P., Khandual, A., & Mehar, K. (2026). Physics-Informed machine learning for turbulent combustion in aerospace propulsion: bridging physical rigour and data intelligence. Engineering Applications of Computational Fluid Mechanics, 20, 2678066. https://doi.org/10.1080/19942060.2026.2678066

Ochella, S., Dinmohammadi, F., & Shafiee, M. (2024). Bayesian neural networks for uncertainty quantification in remaining useful life prediction of systems with sensor monitoring. Journal of Mechanical Engineering Science. https://doi.org/10.1177/16878132241239802

Chen, C., Wang, C., Guo, J., Cui, P., Zheng, J., & Liu, Z. (2025). Remaining useful life prediction considering multiple uncertainty information via Bayesian BiGRU-based method. Reliability Engineering & System Safety, 264, 111431. https://doi.org/10.1016/j.ress.2025.111431

Dresia, K. (2025). Rocket Engine Control with Deep Reinforcement Learning. DLR-Forschungsbericht DLR-FB-2025-16. https://doi.org/10.57676/bwbm-p528

Dresia, K., Adler, A., Saraf, A., Hahn, R., Kurudzija, E., Traudt, T., Deeken, J., & Waxenegger-Wilfing, G. (2023). Rocket Engine Control with Neural Networks: Experimental Results of the LUMEN Turbopump Test Campaign. AIAA SciTech 2023 Forum, AIAA 2023-0137. https://doi.org/10.2514/6.2023-0137

Börner, M., Traudt, T., Dresia, K., Armbruster, W., Suslov, D., Groll, C., Hahn, R. D., Saraf, A. M., Deeken, J., Hardi, J., Oschwald, M., & Schlechtriem, S. (2024). Component Test Results and Project Progress of the DLR Liquid Upper Stage Demonstrator Engine (LUMEN). AIAA SciTech 2024 Forum, AIAA 2024-1398. https://doi.org/10.2514/6.2024-1398

DLR. (2026). LUMEN: AI-assisted reusable rocket engines. German Aerospace Center. https://www.dlr.de/en/research-and-transfer/featured-topics/reusable-space-transportation/lumen

DLR. (2026). LUMEN, Liquid Upper Stage Demonstrator Engine. German Aerospace Center. https://www.dlr.de/en/research-and-transfer/research-infrastructure/large-scale-research-facilities/lumen-en

DLR Institute of Space Propulsion. (2026). Test cell P8.3. German Aerospace Center. https://www.dlr.de/en/ra/research-transfer/research-and-test-infrastructure/test-benches-for-space-propulsion/research-and-technology-test-bench-p8-1/test-cell-p8.3

Support This AI Library Book

If this book was useful, you can support future translations, public editions, and open AI knowledge projects through PayPal.

PayPal

https://paypal.me/kimkjkj

Scan the QR code or open the link above.