Deliverable D3.8

Title: Final report on integrative management platform

Lead Beneficiary	IDE R&D
Delivery Date	30th November 2025

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101000794

Deliverable No.	D3.8
Dissemination Level	Public
Work Package	3
Task	3.2
Lead Beneficiary	IDE R&D
Contributing beneficiary(ies)	IDE R&D
Reviewers	Maria Lopez
Due date of deliverable	30/11/2025
Actual submission date	30/11/2025
Main Author/Contributors	Manuel Salvador and Alex Hall (IDE R&D)

Project Acronym	SECRETed
Project Title	Sustainable Exploitation of bio-based Compounds Revealed and Engineered from naTural sources
Activity	RIA Research and Innovation action
Call	H2020-FNR-2020-2
Funding Scheme	Grant Agreement No: 101000794

Document History				
Version	Date	Beneficiary	Author/Reviewer	Notes
1.0	26/11/2025	IDE R&D	Manuel Salvador	First draft of deliverable
2.0	27/11/2025	IDE R&D	Alex Hall	Updated figures
3.0	29/11/2025	IDE R&D	Manuel Salvador, Alex Hall	Final text revision.

All the contributors to this deliverable declare that they:

- *Are aware that plagiarism and/or literal utilization (copy) of materials and texts from other Projects, works, and deliverables must be avoided and may be subject to disciplinary actions against the related partners and/or the Project consortium by the EU.*
- *Confirm that all their individual contributions to this deliverable are genuine and their own work or the work of their teams working in the Project, except where is explicitly indicated otherwise.*

Have followed the required conventions in referencing the thoughts, ideas, and texts made outside the Project.

the SECRETed Coordinator, IDENER RESEARCH & DEVELOPMENT AGRUPACION DE INTERES ECONOMICO (www.idener.com). The information reported in this document reflects only the author's view and that the Agency is not responsible for any use that may be made of the information it contains.

Executive Summary

This deliverable describes how the SECRETed Integrative Management Platform (IMP) provides the first unified, prediction-enabled environment dedicated to biosurfactants and siderophores. It consolidates all project evidence streams—curated chemical structures, biosynthetic gene clusters, producer organisms, GEASS-derived fragments, and external predictive models—into a harmonised, interoperable system.

The platform integrates descriptor-based and fragment-based representations, PanBGC-supported biosynthetic inference, and a broad suite of predictive outputs including QSAR/QSPR, GCN-based surfactant property prediction, and physics-based solubility modelling. All these computations run externally and are incorporated as versioned datasets with full traceability.

A coordinated set of analytical and visual modules enables users to navigate chemical space, inspect biosynthetic logic, analyse producer distributions and compare molecular families. Cross-linked interactions allow holistic interpretation across chemical, genetic and ecological dimensions.

Overall, the IMP establishes a unique, industry-oriented decision-support system that connects natural chemical diversity to biosynthetic feasibility and industrial performance indicators, positioning SECRETed as a major contributor to enabling biobased surfactant substitution

Table of contents

1. Introduction

One of the central scientific challenges addressed by SECRETed is how to systematically enable the substitution of conventional synthetic surfactants with biologically derived amphiphilic molecules, leveraging the remarkable yet underexploited diversity of microbial biosurfactants and siderophores. Although these natural molecules possess physicochemical properties with clear industrial potential, the field has historically lacked an integrated, prediction-oriented framework capable of reconciling chemical, genetic and process-relevant information at the scale required for rational industrial uptake. No prior initiative has provided a unified, data-driven environment specifically designed to identify, compare, prioritise and ultimately enable the engineering of natural or nature-inspired amphiphiles suitable for biotechnology.

The Integrative Management Platform (IMP) developed under SECRETed constitutes a unique effort in this regard. It performs large-scale data reconciliation, computational prediction and biosynthetic inference within a single decision-support architecture. Rather than operating as a static database, the IMP functions as a computationally enriched analytical ecosystem where chemical structures, predicted physicochemical behaviours, producer metadata, biosynthetic gene cluster (BGC) information and metabolic logic are consistently interconnected.

A major pillar of this platform is the expansion, standardisation and computational characterisation of the SECRETed chemical space. Across Deliverables D3.2–D3.4, biosurfactants, siderophores and structurally related natural products were collected, harmonised and embedded into a descriptor-based space from which surfactant-relevant and metal-binding properties could be inferred. Classical molecular descriptors (Mordred) were complemented by fragment-based ISIDA representations, enabling predictive modelling centred on functional groups known to drive amphiphilicity and metal chelation. This combined descriptor framework allowed, for the first time, the derivation of quantitative property profiles across the chemical space, capturing:

- amphiphilicity gradients and surface-activity indicators,
- structural motifs associated with metal chelation,
- molecular family relationships and substructural organisation,
- and multidimensional similarity between natural compounds and synthetic surfactants.

This analysis also revealed the presence of “dual molecules”—natural products simultaneously displaying biosurfactant-like and siderophore-like properties—a previously under-recognised category of potential relevance for applications requiring both emulsification and metal capture.

The platform incorporates the producer dimension as well, linking molecules to their reported microbial origins whenever data are available. This adds ecological and taxonomic context, supports assessments of culturing feasibility, and underpins genotype–phenotype interpretation. Crucially, thanks to extensive contributions from partners specializing in biosynthetic analysis, producer information is complemented by a deeply refined layer of BGC annotation. Through the expansion and curation of MIBiG entries, the construction of pan-genome BGC datasets, and the delineation of

gene cluster families, the IMP associates a growing proportion of the chemical space with putative or experimentally supported biosynthetic architectures. Beyond full clusters, **PanBGC analyses enable the inference of additional BGC candidates whenever only partial biosynthetic signatures are detected in genome sequences**, thus extending genetic coverage to cryptic or incompletely annotated regions. This integrative approach establishes a robust foundation for linking chemical diversity with biosynthetic potential—an essential prerequisite for rational metabolic engineering.

A further cornerstone of the IMP is the inclusion of GEASS, an enzymatic-rule-based fragmentation and annotation engine that decomposes molecules into biochemically interpretable building units. GEASS reveals how complex amphiphilic structures can be reconstructed from coherent metabolic fragments, and how these fragments map to KEGG or LipidMaps building units. This fragmentation logic not only enhances interpretability but also supports forward-looking design, enabling the construction of chimeric or modified biosurfactants. This design dimension links directly to work later extended in D4.3, where IDE R&D generated new-to-nature amphiphiles subsequently explored experimentally by partners.

On the predictive front, the IMP consolidates a broad suite of machine-learning and physics-informed models that collectively characterise biosurfactant and siderophore behaviour. These externally executed models include:

- QSPR and QSAR models predicting surfactant-like and chelating-like behaviour,
- classifiers capable of detecting dual-function molecules,
- graph convolutional neural networks estimating CMC, Krafft temperature, HLB and surface tension,
- descriptor- and fragment-based models quantifying similarity to commercial surfactants,
- and solubility/partitioning predictions (Joback, UNIFAC, Rackett, empirical correlations) across 26 solvents.

Together, these predictive layers allow high-resolution comparison between natural amphiphiles and industrial benchmarks, thereby identifying biological molecules that already fall within commercially relevant physicochemical ranges. This capability effectively creates an “application network” linking natural diversity to industrial specifications.

In sum, the IMP represents not merely a project-specific tool but a scientific contribution in its own right. It aggregates data types never before combined at this breadth—chemical structures, biosynthetic inference, fragment-level decomposition, machine learning profiles, solubility landscapes and producer information—and reconciles them within a unified multidimensional environment. The result is a comprehensive, model-driven gateway that fills a long-standing gap in the field: a systematic, prediction-enabled path to support the transition from synthetic to biologically derived surfactants.

This deliverable (D3.8) presents the IMP as such an integrative achievement. It documents how the constituent data layers and predictive models converge within a coherent platform, and outlines how this infrastructure advances the overarching

objective of enabling robust, sustainable biological alternatives to synthetic surfactants. The following sections describe the platform's objectives, architecture, computation pipelines, analytical modules and predictive capabilities, forming the foundation for its forthcoming scientific dissemination and industrial impact.

2. Objectives and Scope of the Integrative Management Platform (IMP)

The overarching objective of the Integrative Management Platform (IMP) is to provide a unified, scientifically robust and industry-ready computational environment that consolidates all evidence streams generated across the SECRETed project into a coherent system supporting analysis, comparison, prioritisation and future engineering of biosurfactants and siderophores. The platform responds to the original challenge set out in the project vision: the field lacks a harmonised, data-driven resource capable of connecting chemical structures, biosynthetic logic, producer biology and physicochemical behaviour at the scale and depth necessary to support industrial decision-making.
To address this gap, the IMP operationalises five major objectives:

2.1. Objective 1 — Integrate heterogeneous data sources into a unified analytical environment.

The IMP brings together multiple categories of information—chemical structures, biosynthetic gene clusters, producer organisms, predicted physicochemical properties, solubility profiles and biosynthetic fragmentation outputs—into a single platform governed by a consistent data schema.
This integration ensures that:

- structurally diverse natural products are comparable under a shared descriptor space;
- biosynthetic information is standardised across sources of varying annotation quality;
- external predictive models can be ingested as harmonised, versioned datasets;
- cross-links between chemical, genetic and biological layers are explicitly traceable.

The IMP therefore acts as a data reconciliation hub (Figure 1), resolving discrepancies across databases and ensuring interoperability for downstream analysis.

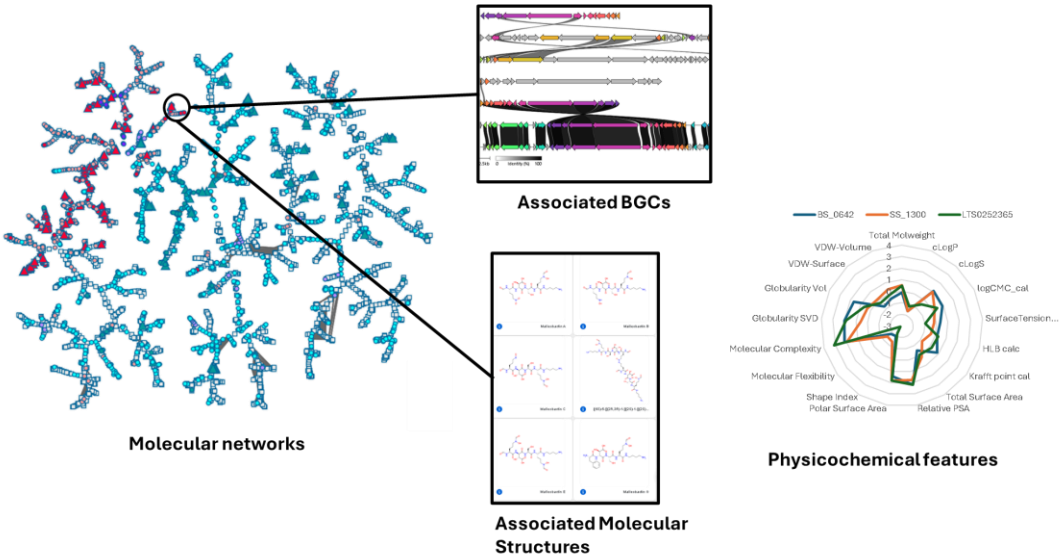

Figure 1. Overview of SECRETed chemical, genetic, and physicochemical space. Left: Molecular network of biosurfactants and siderophores clustered by structural similarity. Top-center: Representative BGCs showing gene organisation and synteny. Bottom-center: Example molecular structures from selected clusters. Right: Radar plots comparing key physicochemical properties across natural molecules, synthetic surfactants, and predicted chimeras.

2.2. Objective 2 — Enable multi-scale interpretation of biosurfactants and siderophores.

The IMP is designed to capture amphiphilic behaviour across levels of biological and chemical organisation, allowing users to move fluidly between:

- **substructural logic** (via GEASS fragmentation),
- **molecular behaviour** (via QSAR/QSPR, ISIDA-derived functional fragments and GCN-based predictions),
- **biosynthetic architecture** (via curated and PanBGC-inferred BGCs),
- **producer ecology** (via harmonised origin metadata),
- **solubility and extractability** (via UNIFAC/Joback/Rackett computations), and
- **industrial comparators** (through supervised similarity mapping to synthetic surfactants).

This multi-layered interpretability is essential for assessing which biological molecules are viable substitutes for synthetic surfactants, and under what biochemical or process constraints.

2.3. Objective 3 — Support predictive and design-oriented workflows.

The IMP is not limited to retrospective analysis. It incorporates externally computed predictive layers that enable anticipatory assessment of molecular candidates. These include:

- surfactant-like activity predictions,
- chelation potential and metal-binding fragments,
- dual-function behaviour classifiers,
- CMC, Krafft temperature and HLB estimates,
- surface-tension predictions,
- solubility and partitioning profiles across 26 solvents,
- biosynthetic feasibility estimates from BGC inference and PanBGC cluster reconstruction.

Through these capabilities, the platform supports *forward-looking design*, allowing researchers and industrial users to identify promising natural molecules, assess experimental tractability, and prioritise targets for engineering.

2.4. Objective 4 —Facilitate genotype-to-phenotype inference and biosynthetic navigation

A key requirement for enabling industrial uptake is understanding the genetic basis and biosynthetic plausibility of target molecules. The IMP integrates:

- curated BGCs from MiBIG and literature;
- BGC expansions inferred from structural anchoring and BLAST homology;
- PanBGC-driven cluster boundaries when only partial genomic signatures are available;
- gene cluster family organisation and accessory gene profiling.

This enables users to:

- predict which organisms are likely to produce a given amphiphile,
- assess cluster completeness and potential for expression,
- evaluate engineering feasibility,
- and identify biosynthetic modules suitable for recombination or modification.

2.5. Objective 5 — Provide an industry-relevant, user-oriented decision-support system

Beyond scientific integration, the IMP has been conceived with **industrial applicability** in mind. It supports:

- rapid identification of molecules falling within industrially relevant CMC, HLB or surface-tension ranges,
- comparison of natural molecules with commercial surfactant analogues,
- prioritisation of candidates combining favourable physicochemical behaviour with tractable biosynthesis,
- early-stage evaluation of downstream process compatibility based on predicted solubility and extractability,

- and transparent traceability for regulatory and techno-economic workflows.

The IMP thus acts as a **strategic interface** between biosynthetic diversity and industrial requirements, enabling the translation of SECRETed's scientific findings into actionable innovation pathways.

3. Architectural Principles and Data Model of the IMP

The Integrative Management Platform (IMP) has been conceived as a modular, scalable and methodologically transparent environment capable of reconciling multiple evidence layers generated within SECRETed. Its architecture reflects the need to integrate heterogeneous datasets—chemical, genetic, biosynthetic, taxonomic and predictive—while preserving data provenance, harmonisation consistency and reproducibility.

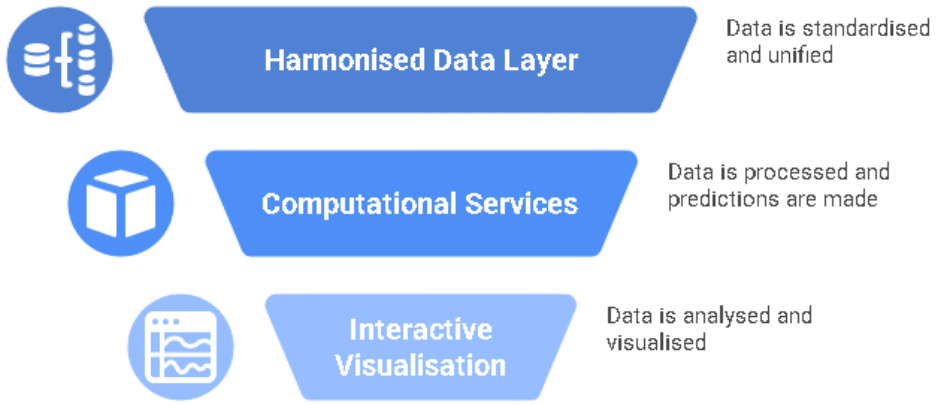

Figure 2. Data processing flow in the Integrative Management Platform.

The IMP therefore adopts a **three-tier organisational model** comprising:

1. **A harmonised data layer**,
2. **A computational services and prediction integration layer**, and
3. **An interactive analytical and visualisation layer**.

These layers are supported by a unified data model designed to ensure that every entity—compound, BGC, producer, descriptor, predicted property or fragmentation unit—is addressable, version-controlled and interoperable.

3.1. Architectural Principles

Modularity and separation of concerns

The IMP is designed around clear boundaries between data, computational logic and user-facing visualisation. This separation ensures that:

- new predictive models can be incorporated without altering the platform's user interface;
- updates to chemical or genetic datasets propagate through harmonised schemas;

- and external simulations (e.g., solubility, QSPR, GCN outputs) can be ingested as versioned datasets without modifying the core architecture.

This modularity is essential for sustainability, given the fast-evolving nature of biosynthetic databases and predictive chemistry tools.

Data provenance, traceability and versioning

Every dataset integrated into the IMP—whether curated structures from T3.x, PanBGC-inferred clusters, GEASS fragmentation outputs, or external predictions—is tagged with:

- **a source identifier** (e.g., LOTUS, NPAtlas, MiBIG, literature, SECRETed-generated),
- **a version**,
- **a timestamp**,
- **a processing pipeline identifier**, and
- **a reproducible checksum**, when applicable.

This enables full backward traceability and ensures that scientific conclusions derived from the platform can be transparently reproduced, audited and published.

Standardisation and harmonisation as a prerequisite for integration

Given the heterogeneity of upstream sources, the platform enforces a rigorous standardisation procedure before any element enters the harmonised data layer. This includes:

- molecule canonicalisation (RDKit),
- normalisation of charges, stereochemistry and tautomeric forms,
- removal of duplicates and salts,
- hierarchical taxonomic reconciliation and downgrading of inconsistent annotations,
- structural alignment of BGCs (MiBIG-compatible annotation),
- harmonisation of physicochemical descriptors,
- and normalisation of predictive modules (feature scaling, descriptor alignment, ID mapping).

This ensures coherent cross-layer comparisons and removes artefacts that would otherwise compromise predictive interpretation.

Extensibility through external prediction modules

A central architectural principle is that **the IMP does not compute machine-learning or physics-based predictions internally**. Instead:

- all predictive models operate as **external pipelines**;
- their outputs are **imported** into the IMP as **harmonised property tables**, cluster-level summaries or entity-level metadata;
- and each imported model retains a **version identifier** and **methodological descriptor**, ensuring interpretability and traceability.

This approach minimises computational load within the platform, maximises methodological flexibility, and ensures compliance with scientific reproducibility standards.

Interoperability and cross-linking between data types

The architecture is designed so that **any entity**—a molecule, a BGC, a producer organism, a GEASS fragment, a predicted physicochemical property—can be:

- queried,
- linked to its counterparts,
- filtered using user-defined criteria,
- and visualised in multiple coordinated views.

This multi-entity interoperability is fundamental to the IMP's scientific purpose: enabling users to relate chemical structure to biosynthetic architecture, biological origin and application-level physicochemical behaviour

3.2. Harmonised Data Layer

The harmonised data layer consolidates all information flowing into the IMP and defines how each entity category is represented. It consists of **five core entity classes**:

Compound entities (chemical space)

Each compound is associated with:

- canonical SMILES/InChI,
- compound ID,
- source provenance (LOTUS, NPAtlas, NPASS, literature, GEASS-derived fragment assembly),
- descriptor vectors (Mordred, ISIDA fragments),
- predicted physicochemical properties (CMC, HLB, Krafft, surface tension),
- solubility/partitioning profiles across solvents,
- GEASS fragment graphs,
- cluster assignments (MAP4-based molecular families),
- similarity mappings to synthetic surfactants.

This uniform representation enables cross-model comparison and multi-criteria filtering.

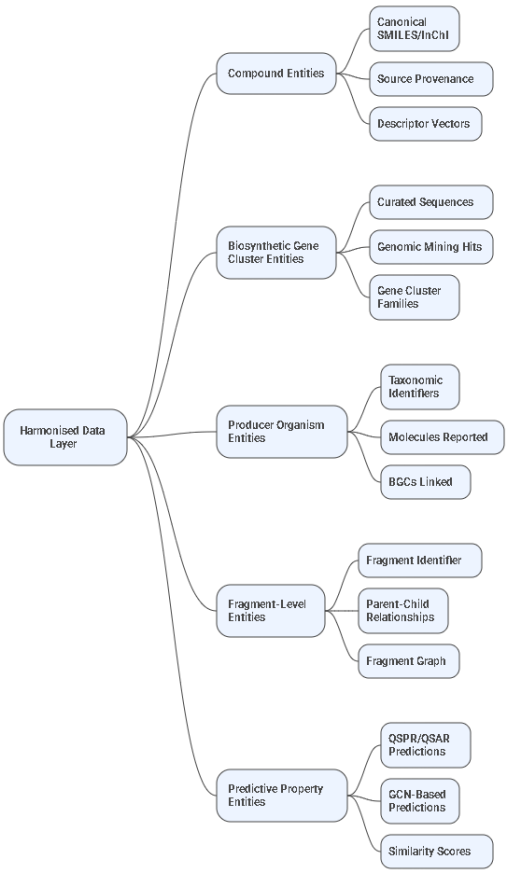

Figure 3. Harmonised Data Layer of the IMP.

Biosynthetic gene cluster (BGC) entities

BGCs in the IMP include:

- curated sequences (MiBIG),
- BLAST-based genomic mining hits,
- and membership in Gene Cluster Families (GCFs) defined in the PanBGC database.

Additional metadata includes taxonomic origin, biosynthetic class, confidence level and annotation provenance.

Producer organism entities

Each producer is associated with:

- hierarchical taxonomic identifiers,
- all molecules reported for that organism across databases,
- all BGCs linked to that strain or species,
- environmental and isolation context when available from (i.e. from BacDive database),
- and confidence level of reported associations.

This entity allows genotype-to-producer-to-phenotype navigation.

Fragment-level entities (GEASS building units)

The GEASS subsystem generates a second, orthogonal layer of annotation. Each fragment is stored with:

- a unique fragment identifier,
- parent–child relationships to the original compound,
- classification as exact or similarity-based annotation (KEGG/LipidMaps),
- node/edge positions within the fragment graph,

This layer provides the mechanistic interpretability that will support BGC linkage and future designs.

Predictive property entities

These entities correspond to data imported from external computation pipelines and include:

- QSPR/QSAR predictions of surfactant-like or chelating behaviour,
- GCN-based predictions (CMC, HLB, surface tension, Krafft),
- similarity-to-industrial-surfactant scores,
- solubility and partition coefficients (UNIFAC/Joback/Rackett),

Each predictive dataset is stored as a versioned, model-specific entity linked to individual compounds.

3.3. Computational Services and Integration Layer

The computational services and integration layer acts as the operational backbone of the IMP. Its purpose is to connect the harmonised datasets with the analytical and visual modules in a consistent, traceable and reproducible manner. While this layer does not execute predictive models, it ensures that all imported results—chemical descriptors, biosynthetic associations, solubility predictions and machine-learning outputs—are correctly aligned and ready for interactive exploration.

To achieve this, the layer provides several core functions:

- **Identifier synchronisation:** all molecules, BGCs, producers, GEASS fragments and predictive properties are mapped to unified internal IDs to guarantee full cross-layer consistency.
- **Descriptor and feature management:** molecular descriptors, fragments, solubility vectors and predictive values are indexed and made accessible for rapid retrieval during filtering and visualisation.
- **Similarity and structural search services:** the layer exposes fast engines for MAP4-based similarity, substructure searches and maximum common substructure (MCS) comparison, supporting interactive navigation of the chemical space.
- **Multi-criteria query processing:** property filters, text searches and cross-entity selections (e.g. “show compounds produced by these strains with CMC < X”) are resolved through a central query engine.
- **Aggregation and summarisation routines:** cluster-level profiles, BGC-level statistics and producer-level summaries are dynamically computed to support heatmaps, radar plots and network overlays.

- **Integration logic for visual modules:** this ensures that all views (compound network, BGC network, taxonomy, heatmaps) respond coherently to user actions through a unified selection and filtering mechanism.

Through these functions, the computational integration layer guarantees that all information—chemical, genetic, biological and predictive—is consistently connected and immediately available during user interaction.

3.4. Analytical and Visualisation Layer

The analytical and visualisation layer constitutes the user-facing environment of the IMP and provides the operational entry point for exploring the integrated chemical, biosynthetic and producer information. This layer translates the harmonised data model and computational integrations into an interactive analytical suite, enabling coordinated, multi-level interrogation of biosurfactants, siderophores and their associated genetic and biological contexts.

Designed as a unified system rather than a collection of isolated tools, the layer exposes all platform entities—compounds, BGCs, producers, fragment graphs and predictive properties—through fully synchronised visual modules implemented using modern interoperable libraries.

As the top tier of the system, this layer:

- operationalises the harmonised data layer into actionable scientific insights;
- exposes all integrated predictions (QSPR/QSAR, GCN outputs, solubility models, BGC associations) in a transparent and interpretable manner;
- enables complex queries that traverse chemical, genetic and producer dimensions;
- and ensures intuitive access to high-dimensional data through coordinated visualisations.

Its design is aligned with the platform's core principles of modularity, interpretability, reproducibility and industrial relevance.

Visual Modules and Their Analytical Roles

The layer incorporates a suite of complementary visual and interactive modules, each contributing a distinct analytical perspective:

• Compound Network View (Cytoscape.js)

A similarity-driven network of all molecules in the SECRETed chemical space, enabling navigation of chemical diversity and clustering via MAP4-based distances. It supports advanced filtering, multi-property exploration and side-panel detail inspection.

• BGC Network View with Clinker Integration

Displays biosynthetic gene clusters and their relationships through similarity links, GCF membership and synteny comparisons. Selecting a cluster dynamically reveals associated molecules and enables genotype-to-chemotype analysis.

• Producer Sunburst Taxonomy (D3.js)

A hierarchical taxonomic representation of all producer organisms, fully linked to their associated compounds and BGCs. Provides ecological and taxonomic context and supports strain prioritisation.

• Heatmaps and Radar Plots

Aggregate visualisations showing descriptor behaviour, predictive model outputs and cluster-level physicochemical profiles. Their coordinated structure allows rapid comparison across molecular families or property domains.

• Search, Filter and Compare Panels

Structured mechanisms for cross-category querying, supporting multi-criteria filtering of molecules, clusters, producers or BGCs. These panels enable targeted discovery workflows (e.g. molecules with low CMC and predicted chelating activity).

• Solubility and Partitioning Radar Charts

Visualisations of predicted solubility and partition behaviour across 26 solvents, derived from the UNIFAC/Joback/Rackett pipelines. These charts support downstream processability and extraction feasibility assessments.

• GEASS Fragment Graph Viewers

Provide mechanistic interpretation of molecular architecture through biosynthetically meaningful building units, enabling users to explore functional fragments, annotation status and structural logic derived from GEASS.

Coordinated Multi-Layer Interaction

A defining characteristic of the IMP is that all modules are fully synchronised. Selections performed in one module propagate automatically to the others. Examples include:

- selecting a compound highlights its associated BGCs and producer organisms;
- filtering producers updates the visible chemical space;
- choosing a BGC reveals its product molecules and displays synteny;
- selecting a cluster in a heatmap highlights corresponding molecules in the network;

- choosing a solvent property updates solubility radar charts and narrows compound sets.

This cross-linked behaviour ensures that users can reason holistically across structural, biosynthetic, ecological and predictive dimensions.

4. Predictive Models and External Computation Pipelines

A key capability of the IMP is its integration of predictive outputs generated through external cheminformatics, machine-learning and physics-based pipelines. As detailed in Deliverables **D3.2–D3.5** and **D6.1**, these models are **not executed within the IMP**; instead, they run as independent, versioned workflows whose results are imported as harmonised property tables. This approach ensures methodological transparency, reproducibility, and the flexibility to incorporate updated or alternative models in future stages.

Collectively, these external pipelines provide a multi-perspective evaluation of biosurfactants and siderophores, capturing:

- amphiphilicity and surfactant performance,
- metal-chelation propensity,
- physicochemical robustness,
- solubility and extractability,
- and biosynthetic feasibility through BGC-level inference.

These predictive layers are essential for prioritising natural and chimeric molecules for experimental validation, metabolic engineering and downstream process design.

4.1. QSAR/QSPR Models for Surfactant-Like and Chelating Behaviour

Descriptor-based modelling (Mordred)

QSAR and QSPR models developed under **D3.4** were trained on ~800 curated molecular descriptors capturing:

- topological and constitutional patterns,
- lipophilicity and polarity metrics,
- shape, flexibility and globularity,
- charge distribution parameters,
- amphiphilic head–tail separation indices.

These models classify compounds across three functional categories:

- **surfactant-like behaviour**,
- **siderophore-like behaviour**,
- **dual-function molecules** combining amphiphilicity and chelation.

Model performance was validated using feature selection, cross-validation and external test sets, ensuring robustness across the structurally heterogeneous SECRETed chemical space.

Fragment-based modelling (ISIDA fragments)

A complementary QSPR approach relied on **ISIDA fragment descriptors**, enabling fine-grained modelling centred on the binding affinity (log β) of siderophores toward iron and other metals and identify candidate siderophores and SECRETed producer strains for experimental validation. These models support:

- prediction of metal-binding propensity (log β-like behaviour),
- and mechanistic interpretability linking predicted properties to discrete chemical fragments.

This fragment resolution is critical for connecting structure to biosynthetic logic and for guiding chimera design.

4.2. Machine-Learning Models for Surfactant Performance

Graph Convolutional Networks (GCNs)

GCNs constitute the IMP's main predictive framework for surfactant performance and were developed as described in **Deliverable D3.4**. Using message-passing neural architectures adapted from state-of-the-art approaches, models were trained to predict:

- **Critical Micelle Concentration (CMC)**,
- **Krafft temperature**,
- **Hydrophilic–Lipophilic Balance (HLB)**,
- **surface tension**.

Training incorporated:

- curated natural products from the SECRETed biochemical space, and
- synthetic surfactant benchmarks,
  ensuring that predicted values fall within experimentally validated and industrially relevant ranges.

The resulting models enabled:

- comparison of natural, synthetic and **new-to-nature chimeric molecules**,
- early assessment of application potential,
- detection of favourable amphiphilic signatures across chemical families.

Industrial similarity scoring

To contextualise biological molecules within an industrial surfactant landscape, descriptor-based and GCN-latent embeddings were used to compute proximity to:

- anionic, cationic, non-ionic and zwitterionic surfactants,
- specialty detergents and formulation agents.

This "**application proximity**" module highlights biological molecules that operate within commercially relevant physicochemical envelopes.

4.3. Physics-Based Models for Solubility, Partitioning and Extractability

Predicting solubility and partitioning behaviour is critical for downstream process design and bio-based industrial substitution. Deliverable D6.1 provided a unified mathematical model whose outputs feed directly into the IMP.

Group-contribution models (UNIFAC, Joback, Rackett)
These models predict:

- boiling point, fusion point, critical properties (Tb, Tf, Tc, Pc),
- viscosity, density and molar volume,
- solubility and partitioning across **26 solvents (see deliverable 6.1)**, whose information was collected from The green solvent tool[1].
- activity coefficients and phase equilibria for extraction.

UNIFAC provides group-based activity coefficients and partition coefficients; Joback yields structural estimates of critical and thermal properties; Rackett enables molar volume and state prediction.

These results underpin the platform's **solubility radar charts**, **partition maps** and **extractability indicators**.

4.4. Biosynthetic Feasibility and BGC-Based Predictions

While the platform itself does not compute biosynthetic predictions, several external analyses feed into the IMP:

BGC inference scores

Derived from:

- BLAST-based homology,
- and PanBGC-based reconstruction of partial clusters.

[1] https://acsgcipr.org/tools/solvent-tool/

Each association is stored with a **confidence score** reflecting the strength and nature of the inference.

Structural variation potential

Using GEASS fragmentation, building-unit composition and GCF membership, molecules were analysed for:

- inherent variability in modular composition,
- potential for natural diversification,
- amenability to engineering via substrate or module swapping.

These virtual fragmentations support selection of candidates for heterologous expression or rational redesign in further steps.

4.5. Integration, Harmonisation and Version Control

All predictive outputs—QSPR/QSAR, GCN, UNIFAC/Joback/Rackett, diffusivity models, similarity indices and BGC inference—undergo the same integration workflow:

- ingestion as external tables with model version identifiers,
- alignment to internal compound IDs,
- descriptor scaling and transformation where required,
- confidence-level annotation,
- and incorporation into compound-level metadata.

This approach preserves **full transparency and scientific traceability** while ensuring interoperability across all platform modules.

5. Visualisation and Analytical Modules of the IMP

The IMP's visualisation and analytical layer provides the user-facing environment through which the harmonised data structures, predictive outputs and biosynthetic associations become accessible. Rather than acting as a static database, the platform offers a coordinated set of views that allow users to explore biosurfactants and siderophores across chemical, genetic and producer dimensions in an integrated manner.

All modules draw on the same unified data model and are fully synchronised via internal identifiers, ensuring that selections in one view are consistently reflected across the others. This design supports multi-layered interpretation of the SECRETed knowledge base and enables users to connect molecular properties, biosynthetic architecture and producer context within a single environment.

5.1. Compound Network View

The compound network is the primary entry point to the SECRETed chemical space. Molecules are represented as nodes positioned according to similarity metrics derived from MAP4 fingerprints, providing an immediate overview of structural diversity and molecular families.

From an analytical perspective, this view allows users to:

- identify clusters of structurally related biosurfactants and siderophores;
- overlay predictive properties (e.g. CMC, HLB, surface tension) to detect regions enriched in desirable surfactant and chelating agent behaviour;
- highlight molecules co-located with synthetic surfactants in the descriptor space, indicating potential industrial relevance.

The network thereby links structural relationships with predicted functional behaviour and similarity to commercial benchmarks.

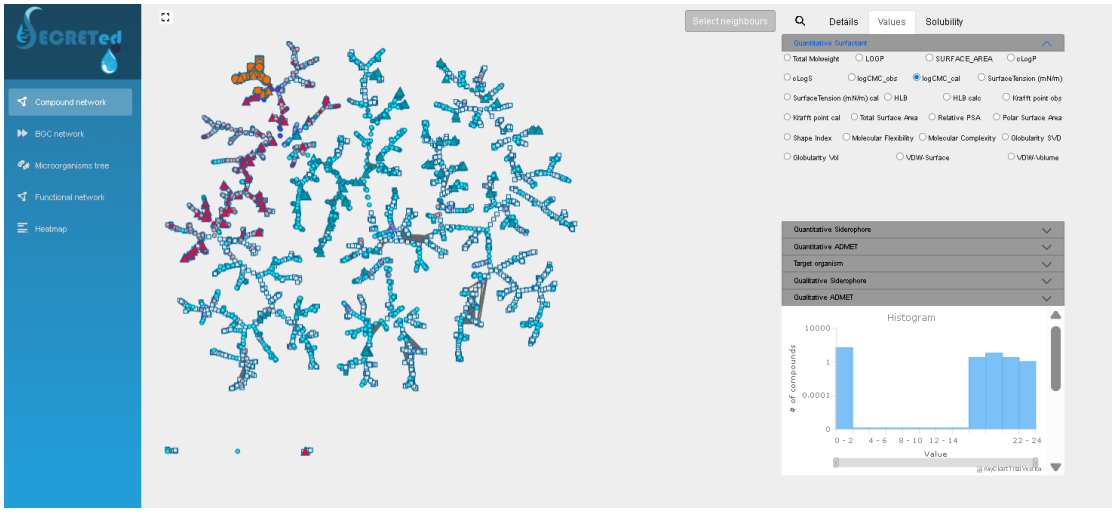

Figure 4. Compound network view with relevant parameters highlighted in the chemical space.

5.2. Heatmap and Cluster Comparison Module

Complementing the network view, the heatmap module offers an aggregated perspective on molecular properties at the level of clusters or molecular families. Rows represent chemical clusters or descriptor-based groupings, while columns correspond to selected physicochemical or predictive properties.

This module supports:

- rapid comparison of clusters in terms of amphiphilicity, chelation propensity, solubility or other surfactant-relevant indicators;
- identification of clusters where natural or chimeric molecules occupy physicochemical niches similar to synthetic surfactants;
- selection of families whose overall property profiles match specific application requirements.

Cluster-level radar plots, derived from the same underlying data, provide a concise summary of property ranges and means within each family.

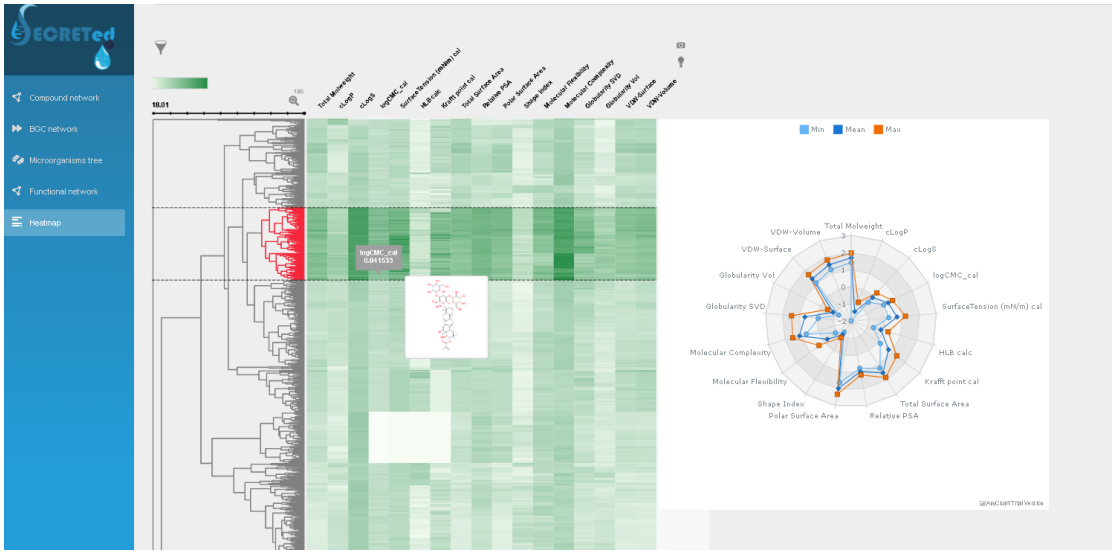

Figure 5. Heatmap view with natural compounds with synthetic compounds and their averages industrial values.

5.3. BGC Network View

The BGC network exposes the biosynthetic architecture underlying the chemical space. Biosynthetic gene clusters are represented as nodes connected by similarity relationships, including gene-content similarity, domain architecture and Gene Cluster Family membership.

Analytically, this view enables users to:

- explore how biosynthetic diversity maps onto the chemical diversity observed in the IMP;
- identify related clusters that may produce structural variants of a given amphiphile;
- connect promising molecular scaffolds to tractable BGCs suitable for heterologous expression or engineering.

Integration with Clinker-based synteny visualisation allows further inspection of gene order and domain composition for selected clusters, supporting genotype-to-chemotype reasoning.

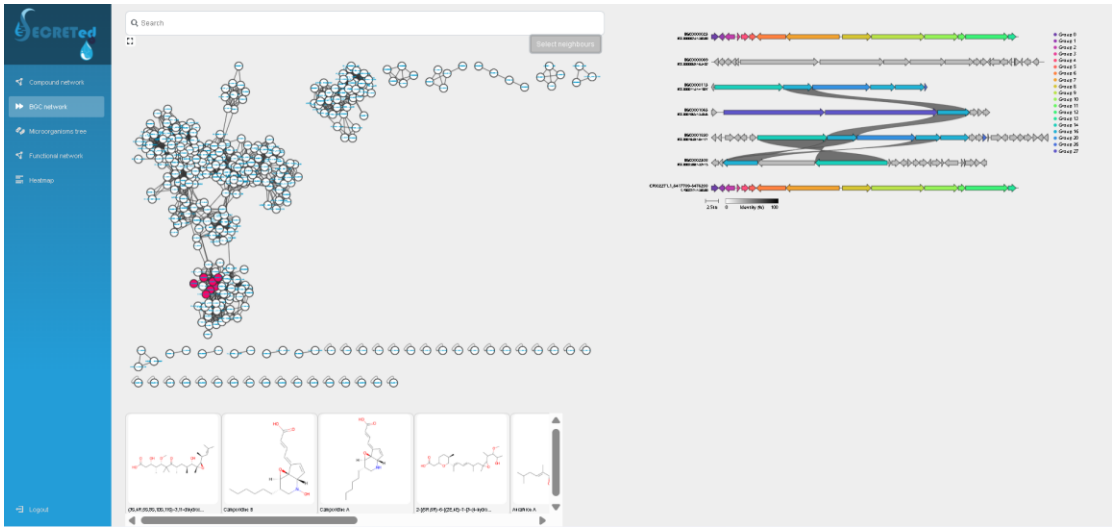

Figure 6. BGC Network view.

5.4. Producer Taxonomy View

The producer module provides a taxonomic overview of organisms represented in the IMP, capturing their relationships across major ranks and linking them to associated molecules and BGCs. It allows users to:

- navigate from high-level taxonomic groups down to specific strains;
- assess how biosurfactant and siderophore production is distributed across lineages;
- identify taxa that combine desirable chemical outputs with practical culturing potential or reduced biosafety concerns.

By coupling taxonomic structure with chemical and genetic information, this view supports strategic decisions on strain selection and portfolio expansion.

5.5. GEASS Fragmentation and Mechanistic View

The GEASS fragmentation viewer provides a mechanistic window into molecular architecture. For each molecule, the viewer exposes:

- the set of biosynthetically meaningful building units identified by GEASS;
- their connectivity within fragment graphs;
- annotations linking fragments to KEGG/LipidMaps building blocks when available.

This perspective helps users understand how complex amphiphilic structures can be decomposed into interpretable modules and how these relate to underlying biosynthetic logic. It is particularly relevant for evaluating structural variation potential and for supporting the design of chimeric molecules.

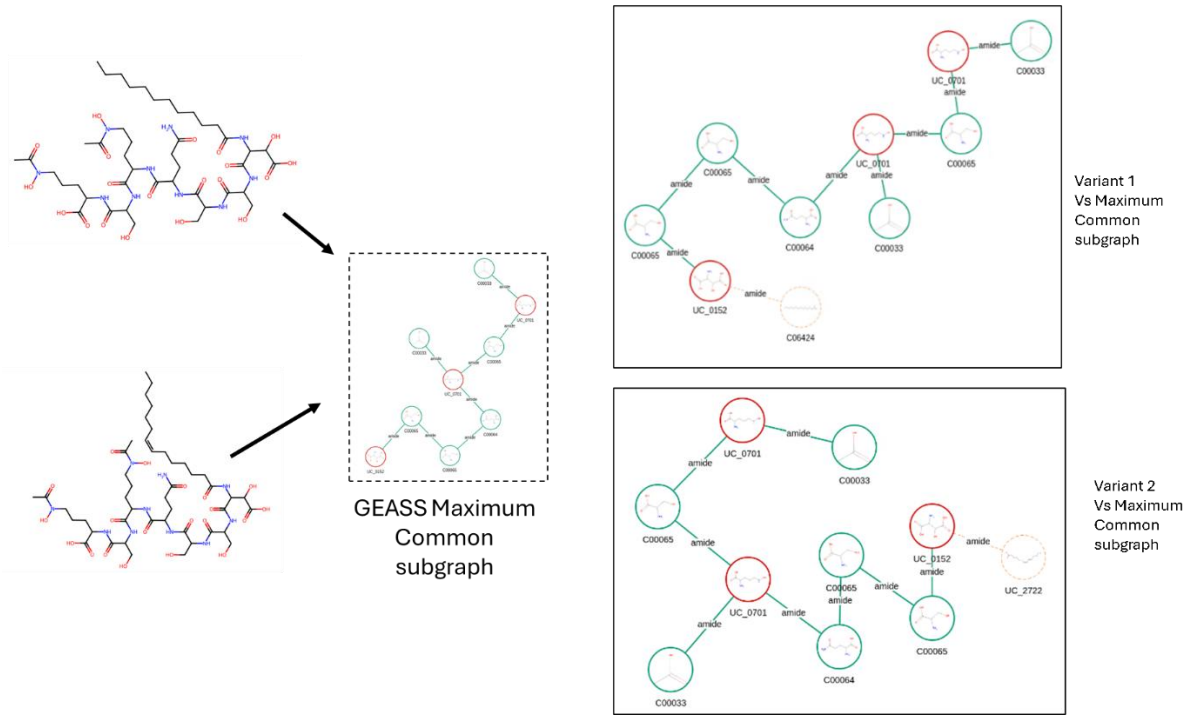

Figure 7. GEASS-derived fragmentation graph for related representative compounds. Nodes correspond to biochemically meaningful building units, and edges represent enzymatic disconnections. The modular decomposition illustrates the disassembled variants differs in their fatty acid.

5.6. Integrated, Multi-Module Exploration

A defining feature of the IMP is the tight cross-linking between all visual modules. Selections or filters applied in one view propagate through the others, enabling use cases such as:

- starting from a taxonomic group to identify its associated molecules and BGCs in the network views;
- selecting a molecular cluster with attractive surfactant properties and inspecting its biosynthetic basis and producer range;
- focusing on compounds with favourable solubility profiles and examining their position in the chemical space and their genetic accessibility.

This coordinated behaviour turns the IMP into a coherent analytical environment in which chemical, genetic and biological information is not only co-located but actively connected, providing a practical decision-support interface for both scientific and industrial stakeholders.

1. Conclusion

The IMP successfully delivers on the objectives of Task 3.2 by integrating heterogeneous chemical, genetic, biosynthetic and predictive information into a coherent, modular and extensible platform. Its harmonised data model, robust provenance tracking and multi-layer analytical environment provide a solid foundation for downstream experimental validation, metabolic engineering and process-design activities within SECRETed.

The incorporation of predictive models—GCN-based surfactant estimators, QSPR/QSAR classifiers and physics-based solubility calculations—significantly strengthens the platform's ability to prioritise candidates with both functional potential and biosynthetic tractability. Cross-linked visual modules enable users to reason simultaneously about structure, behaviour and biosynthetic origin.

In conclusion, the IMP represents a major integrative achievement that advances the scientific and industrial ambitions of SECRETed, providing a scalable and future-proof infrastructure to support the transition towards sustainable, biologically produced surfactants and siderophores.