NanoCommons

Nano-Knowledge Community

The European Nanotechnology Community Informatics Platform: Bridging data and disciplinary gaps for industry and regulators

Grant Agreement No 731032

Deliverable Report 5.5

Deliverable	D5.5 A workflow and checklist of key information needed from modelling tool owners to facilitate integration into KB
Work Package	WP5: JRA3 - Analysis and Modelling tools
Delivery date	M18 - 30 June 2019
Lead Beneficiary	University of Birmingham
Nature of Deliverable	Demonstrator
Dissemination Level	Public (PU)

Submitted by	Anastasios Papadiamantis (UoB), Thomas Exner (EwC), Harry Sarimveis (NTUA) and Vladimir Lobaskin (UCD
Revised by	Philip Doganis (NTUA), Vladimir Lobaskin (UCD), and Stacey Harper (OSU)
Approved by	Iseult Lynch (University of Birmingham)

This project has received funding from European Union Horizon 2020 Programme (H2020) under grant agreement nº 731032

Table of contents

Abbreviations

AOP: Adverse outcome pathways

API: Application Programing Interface

CG: Coarse-grained

EMMC: European Materials Modelling Council

ENM: Engineered nanomaterials

GDPR: General Data Protection Regulation

GUI: Graphical User Interface

IPR: Intellectual Property Rights

IUCLID: International Uniform Chemical Information Database

KB: KnowledgeBase

KE: Key events

KNIME: Konstanz Information Miner

LOO: Leave-one-out

LMO: Leave-many-out

MD: Molecular Dynamics

MIASE: Minimum Information About a Simulation Experiment

MIE: Molecular initiating events

MIRIAM: Minimum Information Required in the Annotation of Models

MODA: Modeling Data Generalisation

NMs: Nanomaterials

NTUA: National Technical University of Athens

OSMO: Ontology for Simulation, Modelling and Optimisation

PMF: Potentials of mean force

QA/QC: Quality Assurance / Quality Control

QPRF: QSAR Prediction Reporting Format

QSAR: Quantitative Structure–Activity Relationship

RoMM: Review of Materials Models

TA: Transnational access

TP: Toxicity pathways

VIMMP: Virtual Materials Marketplace

VISO: VIMMP Software Ontology

WP: Work package

Summary

This deliverable is part of Work Package 5 (WP5), which aims to integrate state-of-the-art data mining, analysis and modelling tools into the NanoCommons Knowledge Base (KB). This will facilitate a linked data approach to integrate and exploit knowledge from publicly available sources and feed into the modelling tools for further studies. These tools, following integration, will be made available to the entire nanotechnology and nanosafety User community.

For this process to be successful, a clearly defined modelling tools integration workflow needs to be developed and implemented. This workflow, along with the relevant guidance notes, will be also used in Work Package 8 (WP8) - Networking Activity 2 – Training aligned to TA / JRA to support the integration of Users' modelling tools based on subsequent open calls for Transnational Access (TA). Specific objectives in terms of tools to incorporate into NanoCommons include, but not limited to:

- Tools for extracting knowledge from raw experimental data (such as microscopic images or spectral data);
- Tools for preprocessing data before they are sent to modelling services (normalisation, missing data handling, selection of important variables, dimensionality reduction);
- Tools for generating theoretical descriptors (such as structural descriptors or quantum mechanical descriptors);
- Tools for analysing big omics or "corona" data in terms of identifying the biological mechanisms and pathways associated with toxicity and other adverse effects and producing aggregated biologically enriched descriptors;
- Tools for harmonising and integrating diverse data and metadata originating from heterogeneous resources, so that homogeneous datasets suitable for direct import into modelling software are produced;
- Tools for semantically retrieving ontology annotated data from the project data warehouse and other data sources integrated in the knowledge infrastructure.

While the workflow presented in this deliverable is aimed at the inclusion of modelling tools into the NanoCommons KB, it can also be used with relevant modifications for other types of tools. Detailed workflows for the integration of databases and single datasets into the KB are described in detail in deliverables "D4.5: Workflow and checklist of key information needed from database/dataset owner in order to facilitate integration into KB" and "D3.3: Checklist for use in WP8 / WP9 to support integration of Users data into KB", respectively.

1. Introduction

Engineered nanomaterials (ENMs) are now being used extensively in several aspects of everyday life like consumer products (e.g. cosmetics, clothing), infrastructure (e.g. paints) and medicines (e.g. drug delivery). As a result, the safety of such materials is highly important, which has also led to new REACH guidelines regarding the registration of ENMs and the definition of the nanoform. In order for an ENM to be considered safe for biological organisms and the environment, complex and costly experiments need to take place to prove that hazards, exposures and risks are acceptable and manageable. These tests span from a full physicochemical characterisation to toxicity experiments in a range of organisms / animals over extended durations, depending on the production volume, with more data being required for high production volumes. The use of animals, however, raises significant questions regarding the ethics of such processes, which need to be addressed, adding further to the workload and cost. At the same time, the data produced from nanosafety projects remains fragmented and inaccessible hampering the identification and establishment of read-across approaches that are currently absent for ENM. Such approaches would reduce the cost of nanosafety research and regulation dramatically by removing the need for extensive laboratory and animal testing. Furthermore, the transformation of nanosafety research, due to technological advancement, to a data-heavy field and the lack of sharing or publishing of negative results, adds to the challenge of developing predictive models for use in regulation.

NanoCommons will exploit the recent technological and computational advancements through development of an e-infrastructure platform tailored for nanosafety. Using modern modelling approaches it is possible to produce robust models and simulation data and achieve substantial progress in the field of nanotoxicity, reducing the need for animal studies and the regulatory costs. At the same time, such tools can offer novel insights with the definition of new computational descriptors that would be impossible to define and calculate from simple statistical analysis.

WP5 is pursuing a linked data approach that will exploit, extract, and integrate knowledge from all available information sources (from raw experimental to modelling data with the associated metadata) captured in the NanoCommons KnowledgeBase (NC-KB). Physics- and chemistry-based materials modelling procedures will be integrated and adapted to calculate relevant NMs descriptors and complete data sets where information gaps are identified. Existing data handling and analysis tools will be further developed, extended and integrated throughout the project, taking into account existing knowledge from chemicals, and the additional needs of the nanosafety community due to the larger and more diverse data sources and ENM structures. Extracted knowledge will then be organised in formats, suitable for direct import into predictive modelling tools. The tools developed within WP5 will be implemented based on interoperable, standards-compliant modular web services maximising cross-talk and interaction between different/diverse sources of data. Five categories of modelling tools are being integrated in the NanoCommons KB, which are described in section 2:

1. Tools for calculation of theoretical descriptors;
2. Tools for generation of predictive nano quantitative structure–activity relationship (nano-QSAR) models;
3. Simulation tools for NMs transport and corona formation;
4. Modelling tools for key event prediction as part of AOPs;
5. Biokinetics models.

2. Types of modelling tools

2.1 Calculation of theoretical descriptors, image descriptors and corona formation

Physics- and chemistry-based models are being used to develop tools for the calculation of theoretical ENM descriptors. Several ENM descriptors will be obtained from electronic molecular structure representations or crystal structure representations. Examples of tools for integration into the NanoCommons e-infrastructure for calculation of theoretical descriptors include the CDK5[1] open-source software, which has been extended to include nano-specific descriptors, and the MOPAC6[2] semi-empirical quantum chemistry software. Potential ENM descriptors to be calculated include conduction band gap, ionisation potentials, heat of cluster formation, index of refraction, Hamaker constants, and hydration energy (per unit area) computed to characterise hydrophobicity of the material. Workflows for atomistic Molecular Dynamics (MD) simulations will also be developed using Gromacs MD package to evaluate the adsorption energies of water molecules at the ENM surface, corona formation and bio-nano membrane interactions. For calculation of ENM surface charge at different pH and salt concentrations, we will use the Poisson-Boltzmann equation with charge regulation. All the materials descriptor calculations and output data derived will be compatible and interoperable with the formats developed by the European Materials Modelling Council (EMMC). Once the simulation methodologies are integrated and validated, NanoCommons will produce a Materials Modeling Generalisation template (MODA, see section 3.1) descriptor for the technique and communicate it to EMMC.

Two image analysis web-based tools, namely NanoXtract (offered by NanoCommons partners NovaMechanics, Figure 1) and NanoImage (offered by NanoCommons partners NTUA, Figure 2) for processing and knowledge extraction from electronic NM images are already integrated into NanoCommons, and are extremely valuable sources of information that are currently not exploited at all in nanosafety assessment beyond a size distribution and a qualitative description of shape. The web-based solutions are providing: (i) User Interfaces to easily test the capabilities of the image analysis tools, (ii) complete access to the calculations, and (iii) easy integration to existing infrastructures through the use of web services. Various types of ENMs are supported including spherical, tube-shaped/cylindrical and plates, nanotubes as well as the complex morphologies resulting from environmental ageing of ENMs.

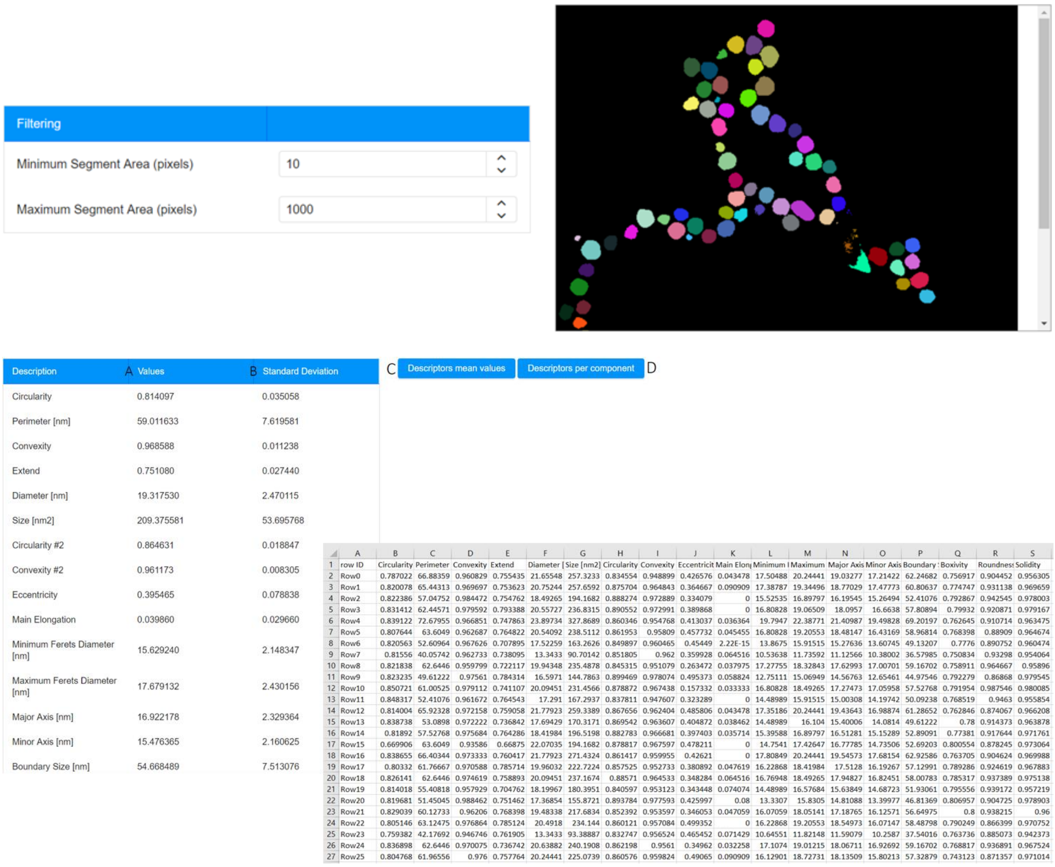
Additionally, a simulation tool for evaluation of adsorption energy of arbitrary proteins on a specific ENM surface, as a means to determine proteins coronas is presented briefly, with more detail given in Deliverable Report D5.6. The goal of the tool is to compare and rank biomolecules by their adsorption affinity and thus form a basis for producing ENM biointeraction fingerprints. The approach uses the SmartNanoTox multiscale modelling methodology, which has been developed to build coarse-grained (CG) models of lipid membranes and proteins and predict their interaction with nanoparticles. The calculation of bionano interactions includes four stages (Figure 3), which will be

[1] https://github.com/cdk/cdk

[2] http://www.ccl.net/cca/software/MS-WINDOWS/mopac6/index.shtml

integrated into a single prediction tool:

1. The models of proteins employ the united-atom scheme, i.e. replace the common groups of atoms (like amino acids or alkyl groups) by single beads. The calculation of the potentials of mean force (PMF) for amino acid beads with specified ENM surface using atomistic simulation (metadynamics with open source Gromacs MD package) is currently automated by the SmartNanoTox project and will be transferred to the NanoCommons KB.
2. A two-layer CG model of ENM represents the surface of the nanoparticle by united-atom beads, whose interaction with amino acids is parameterised using the atomistic PMFs, and the core by a continuum model using Lifshitz theory. The calculations employ ESPREsSo MD open source software.
3. A 3D structure of a protein is either retrieved from public sources like the Protein Data Bank or predicted by homology modelling using the I-TASSER freeware tool.
4. The interaction energy and entropy for a complete protein globule with the nanoparticle of specified size is calculated for representative proteins using ESPREsSo MD package. The proteins are then ranked by the adsorption affinity and, based on their concentrations in the biological fluid of interest, their abundances in the corona (the protein corona fingerprint) and other quantitative descriptors, including NM adsorption energy on lipid membrane, are calculated.

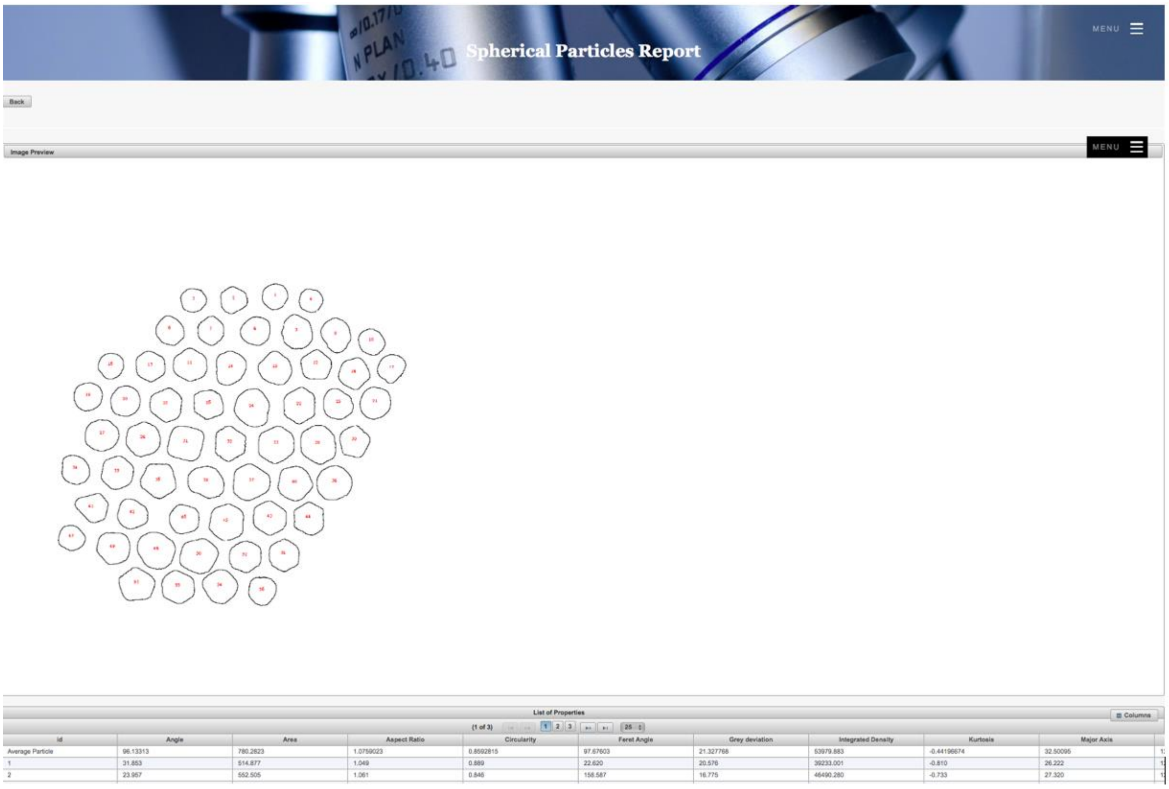


Filtering	
Minimum Segment Area (pixels)	10
Maximum Segment Area (pixels)	1000

Description	Values	Standard Deviation
Circularity	0.814097	0.035058
Perimeter [nm]	59.011633	7.619581
Convexity	0.968588	0.011238
Extend	0.751080	0.027440
Diameter [nm]	19.317530	2.470115
Size [nm2]	209.375581	53.695768
Circularity #2	0.864631	0.018847
Convexity #2	0.961173	0.008305
Eccentricity	0.395465	0.078838
Main Elongation	0.039860	0.029660
Minimum Ferets Diameter [nm]	15.629240	2.148347
Maximum Ferets Diameter [nm]	17.679132	2.430156
Major Axis [nm]	16.922178	2.329364
Minor Axis [nm]	15.476365	2.160625
Boundary Size [nm]	54.668489	7.513076

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	row ID	Circularity	Perimeter	Convexity	Extend	Diameter [	Size [nm2]	Circularity	Convexity	Eccentricit	Main Elon	Minimum I	Maximum	Major Axis	Minor Axis	Boundary :	Boxivity	Roundness	Solidity
2	Row0	0.787022	66.88359	0.960829	0.755435	21.65548	257.3233	0.834554	0.948899	0.426576	0.043478	17.50488	20.24441	19.03277	17.21422	62.24682	0.756917	0.904452	0.956305
3	Row1	0.820078	65.44313	0.969697	0.753623	20.75244	257.6592	0.875704	0.964843	0.364667	0.090909	17.38787	19.34496	18.77029	17.47773	60.80637	0.774747	0.931138	0.969659
4	Row2	0.822386	57.04752	0.984472	0.754762	18.49265	194.1682	0.888274	0.972889	0.334079	0	15.52535	16.89797	16.19545	15.26494	52.41076	0.792867	0.942545	0.978003
5	Row3	0.831412	62.44571	0.979592	0.793388	20.55727	236.8315	0.890552	0.972991	0.389868	0	16.80828	19.06509	18.0957	16.6638	57.80894	0.79932	0.920871	0.979167
6	Row4	0.839122	72.67955	0.966851	0.747863	23.89734	327.8689	0.860346	0.954768	0.413037	0.036364	19.7947	22.38771	21.40987	19.49828	69.20197	0.762645	0.910714	0.963475
7	Row5	0.807644	63.6049	0.962687	0.764822	20.54092	238.5112	0.861953	0.95809	0.457732	0.045455	16.80828	19.20553	18.48147	16.43169	58.96814	0.768398	0.88909	0.964674
8	Row6	0.820563	52.60964	0.967626	0.707895	17.52259	163.2626	0.849897	0.960465	0.45449	2.22E-15	13.8675	15.91515	15.27636	13.60745	49.13207	0.7776	0.890752	0.960474
9	Row7	0.81556	40.05742	0.962733	0.738095	13.3433	90.70142	0.851805	0.962	0.359928	0.064516	10.53638	11.73592	11.12566	10.38002	36.57985	0.750834	0.93298	0.954064
10	Row8	0.821838	62.6446	0.959799	0.722117	19.94348	235.4878	0.845315	0.951079	0.263472	0.037975	17.27755	18.32843	17.62993	17.00701	59.16702	0.758911	0.964667	0.95896
11	Row9	0.823235	49.61222	0.97561	0.784314	16.5971	144.7863	0.899469	0.978074	0.495373	0.058824	12.75111	15.06949	14.56763	12.65461	44.97546	0.792279	0.86868	0.979545
12	Row10	0.850721	61.00525	0.979112	0.741107	20.09451	231.4566	0.878872	0.967438	0.157332	0.033333	16.80828	18.49265	17.27473	17.05958	57.52768	0.791954	0.987546	0.980085
13	Row11	0.848317	52.41076	0.961672	0.764543	17.291	167.2937	0.837811	0.947607	0.323289	0	14.48989	15.91515	15.00308	14.19742	50.09238	0.768519	0.9463	0.955854
14	Row12	0.814004	65.92328	0.972158	0.759058	21.77923	259.3389	0.867656	0.962404	0.485806	0.043478	17.35186	20.24441	19.43643	16.98874	61.28652	0.762846	0.874067	0.966208
15	Row13	0.838738	53.0898	0.972222	0.736842	17.69429	170.3171	0.869542	0.963607	0.404872	0.038462	14.48989	16.104	15.40006	14.0814	49.61222	0.78	0.914373	0.963878
16	Row14	0.81892	57.52768	0.975684	0.764286	18.41984	196.5198	0.882783	0.966681	0.397403	0.035714	15.39588	16.89797	16.51281	15.15289	52.89091	0.77381	0.917644	0.971761
17	Row15	0.669906	63.6049	0.93586	0.66875	22.07035	194.1682	0.878817	0.967597	0.478211	0	14.7541	17.42647	16.77785	14.73506	52.69203	0.800554	0.878245	0.973064
18	Row16	0.838655	66.40344	0.973333	0.760417	21.77923	271.4324	0.861417	0.959955	0.42621	0	17.80849	20.24441	19.54573	17.68154	62.92586	0.763705	0.904624	0.969988
19	Row17	0.80332	61.76667	0.970588	0.785714	19.96032	222.7224	0.857525	0.952733	0.380892	0.047619	16.22868	18.41984	17.5128	16.19267	57.12991	0.789286	0.924619	0.967883
20	Row18	0.826141	62.6446	0.974619	0.758893	20.09451	237.1674	0.88571	0.964533	0.348284	0.064516	16.76948	18.49265	17.94827	16.82451	58.00783	0.785317	0.937389	0.975138
21	Row19	0.814018	55.40818	0.957929	0.704762	18.19967	180.3951	0.840597	0.953123	0.343448	0.074074	14.48989	16.57684	15.63849	14.68723	51.93061	0.795556	0.939172	0.957219
22	Row20	0.819681	51.45045	0.988462	0.751462	17.36854	155.8721	0.893784	0.977593	0.425997	0.08	13.3307	15.8305	14.81088	13.39977	46.81369	0.806957	0.904725	0.978903
23	Row21	0.829039	60.12733	0.96206	0.768398	19.48338	217.6834	0.852392	0.953597	0.346053	0.047059	16.07059	18.05141	17.18765	16.12571	56.64975	0.8	0.938215	0.96
24	Row22	0.805146	63.12475	0.976864	0.785124	20.4918	234.144	0.860121	0.957084	0.499352	0	16.22868	19.20553	18.54973	16.07147	58.48798	0.790249	0.866399	0.970752
25	Row23	0.759382	42.17692	0.946746	0.761905	13.3433	93.38887	0.832747	0.956524	0.465452	0.071429	10.64551	11.82148	11.59079	10.2587	37.54016	0.763736	0.885073	0.942373
26	Row24	0.836898	62.6446	0.970075	0.736742	20.63882	240.1908	0.862198	0.9561	0.34962	0.032258	17.1074	19.01215	18.06711	16.92692	59.16702	0.768817	0.936891	0.967524
27	Row25	0.804768	61.96556	0.976	0.757764	20.24441	225.0739	0.860576	0.959824	0.49065	0.090909	16.12901	18.72731	18.13509	15.80213	57.32879	0.743123	0.871357	0.971014

Figure 1. Calculation of image descriptors using the NanoXtract tool developed by NovaMechanics.



List of Properties

(1 of 3) 1 2 3 25

id	Angle	Area	Aspect Ratio	Circularity	Feret Angle	Grey deviation	Integrated Density	Kurtosis	Major Axis
Average Particle	96.13313	780.2823	1.0759023	0.8592815	97.67603	21.327768	53979.883	-0.44196674	32.50095
1	31.853	514.877	1.049	0.889	22.620	20.576	39233.001	-0.810	26.222
2	23.957	552.505	1.061	0.846	158.587	16.775	46490.280	-0.733	27.320

Figure 2. Calculation of image descriptors using the NanoImage tool developed by NTUA.

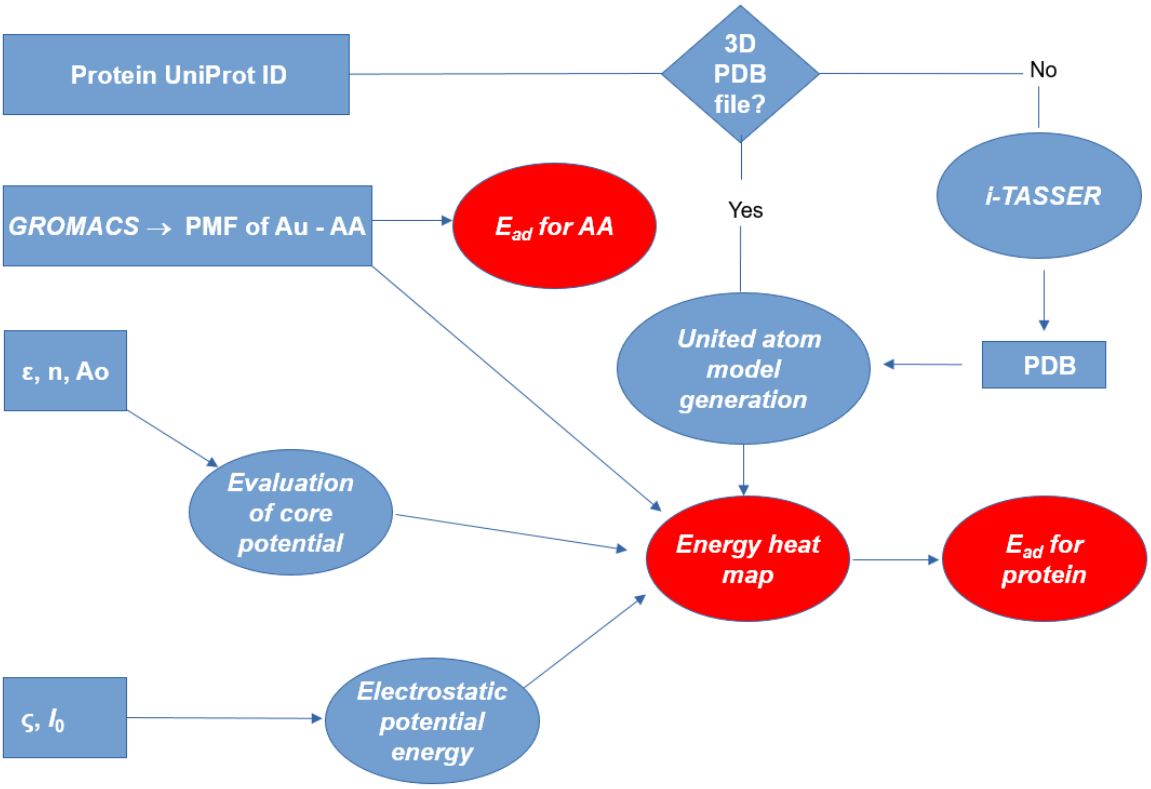
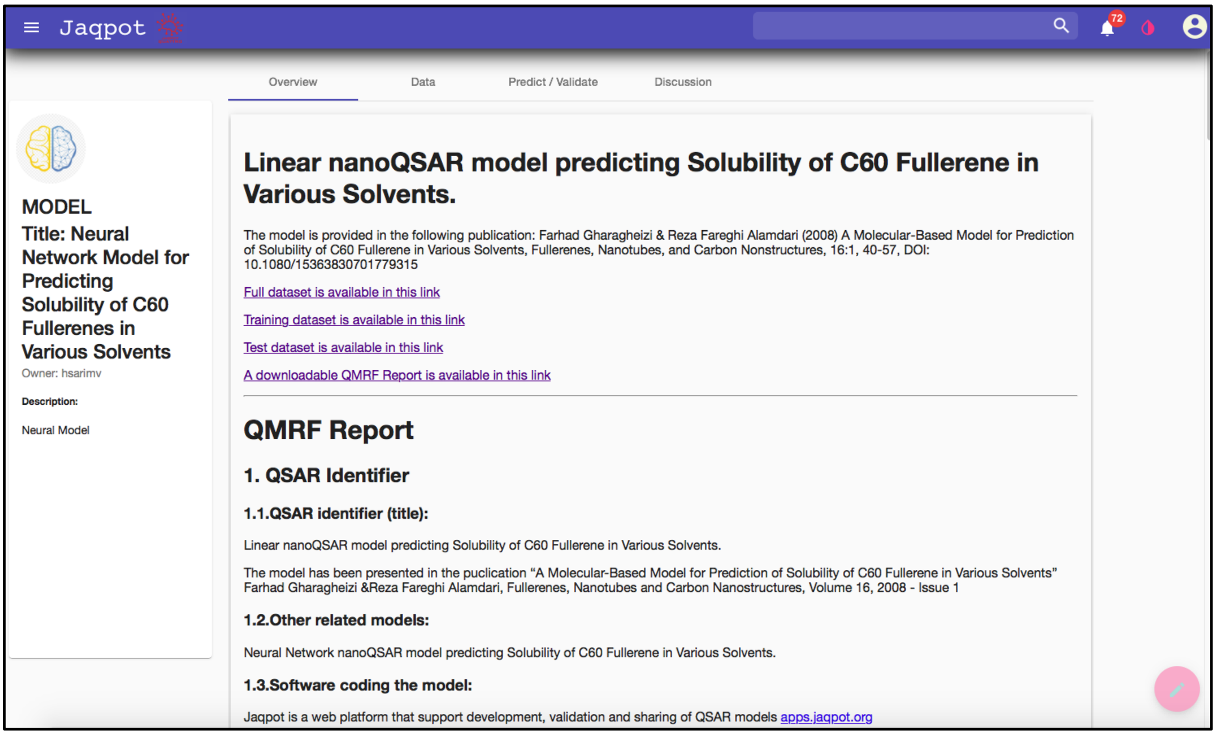


Figure 3. The workflow for the corona modelling tool.

2.2 Generation of predictive nano-QSAR models

Nano-QSAR models correlate biological responses with experimental and/or theoretically calculated descriptors using existing modelling infrastructure such as the Jaqpot (offered by NanoCommons partners NTUA, Figure 4) and the Enalos (offered by NanoCommons partners NovaMechanics, Figure 5) platforms. Nano-QSAR models allow the automatic optimal selection of variables and tuning of the statistical/machine learning algorithms based on rigorous cross-validation tests. The models will be offered in the form of public, ready-to-use web applications and will be fully and semantically integrated with the data warehouse allowing easy access to training data. The models will be adapted and generated to meet the continuously changing and emerging needs of the nanosafety and wider nanotechnology communities. As a starting point, NanoCommons has integrated popular nanoQSAR models that have been published in the literature, with the aim of producing a library of well validated and useful models. Users will have the option to apply data from the NanoCommons data warehouse or upload their own data and will receive the results and model predictions in easy to interpret and informative tables and figures. NanoCommons will also integrate tools for the automatic creation of standardised QSAR Model Reporting Format (QMRF) reports, for summarising and reporting key information on QSAR models and especially compliance with the OECD validation principles, thereby supporting regulatory acceptance of the models.



MODEL
Title: Neural Network Model for Predicting Solubility of C60 Fullerenes in Various Solvents
Owner: hsarimv

Description:

Neural Model

Linear nanoQSAR model predicting Solubility of C60 Fullerene in Various Solvents.

The model is provided in the following publication: Farhad Gharagheizi & Reza Fareghi Alamdari (2008) A Molecular-Based Model for Prediction of Solubility of C60 Fullerene in Various Solvents, Fullerenes, Nanotubes, and Carbon Nonstructures, 16:1, 40-57, DOI: 10.1080/15363830701779315

Full dataset is available in this link

Training dataset is available in this link

Test dataset is available in this link

A downloadable QMRF Report is available in this link

QMRF Report

1. QSAR Identifier

1.1.QSAR identifier (title):

Linear nanoQSAR model predicting Solubility of C60 Fullerene in Various Solvents.

The model has been presented in the puclication “A Molecular-Based Model for Prediction of Solubility of C60 Fullerene in Various Solvents” Farhad Gharagheizi &Reza Fareghi Alamdari, Fullerenes, Nanotubes and Carbon Nanostructures, Volume 16, 2008 - Issue 1

1.2.Other related models:

Neural Network nanoQSAR model predicting Solubility of C60 Fullerene in Various Solvents.

1.3.Software coding the model:

Jaqpot is a web platform that support development, validation and sharing of QSAR models apps.jaqpot.org

Figure 4. Example of a nanoQSAR model hosted in NanoCommons through the Jaqpot platform.

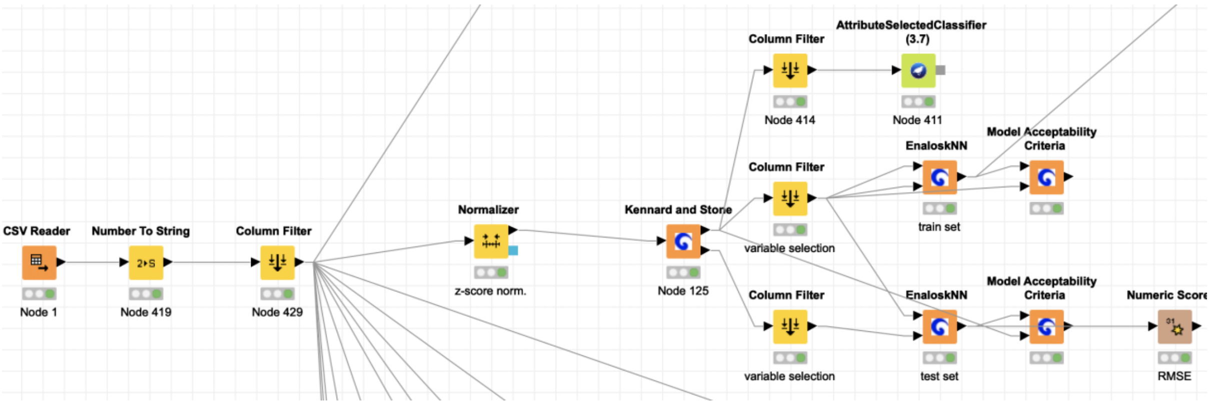


Figure 5. QSAR modeling workflow produced in Konstanz Information Miner (KNIME) using the Enalos platform.

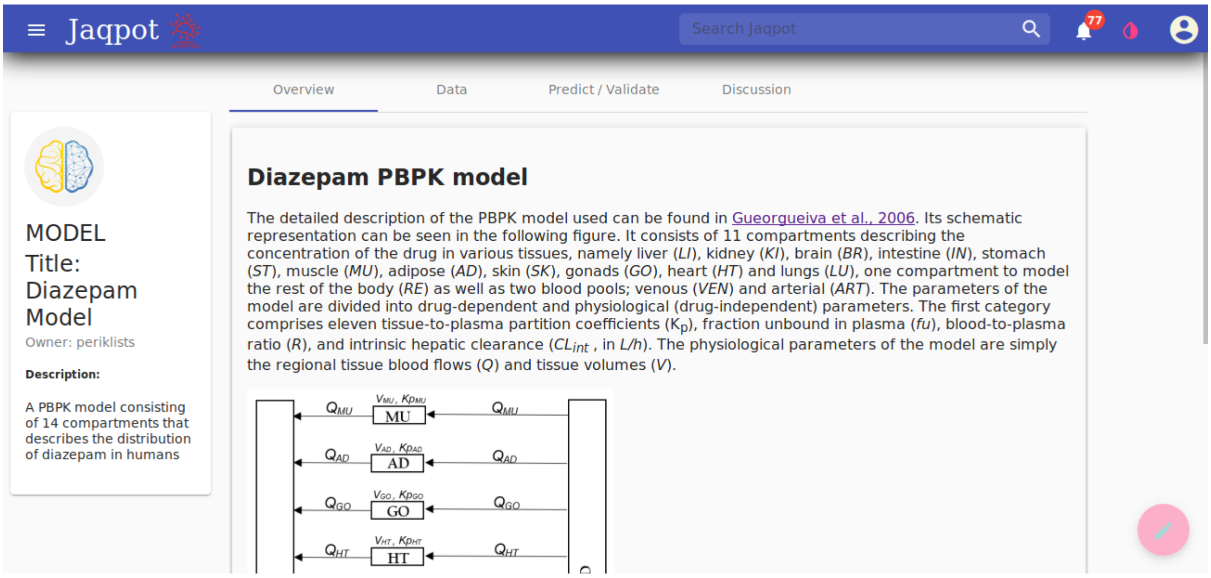
2.3 Modelling tools for key event prediction

Modelling tools for key event prediction deal with adverse outcome pathways (AOP) and toxicity pathways (TP) based on the likelihood of the Molecular Initiating Event (MIE) and Key event (KE) occurring. Through these tools, evidence relating to relevant AOP/TP to allow identification of candidate MIE/KE events at the bionano interface can be identified. Examples of such MIE/KE are: ENM cell association, cell uptake, adsorption of ENM at lipid membrane, binding of a ligand in a given biological fluid, ENM unfolding adsorbed proteins, production of reactive oxygen species, dissolution and ion release, etc. Our ability to predict these strongly depends on the understanding of bionano interactions, and requires confidence in the datasets regarding NMs coronas and how these are isolated and reported [1, 2]. The addition of information on the protein corona composition has improved the predictive performance of structure activity relationships for ENMs. This way, well-established AOPs for chemicals in the context of NM-driven toxicity using data from the AOP Knowledge base can be explored. Similarly, different levels of biological information, including omics data, can be used as a starting point to propose new AOPs for NMs. Causality analysis techniques can then be used to identify sequential chain of KEs that link a specific MIE with the observed AOP. Once pathways are described and MIE/KE identified, their description will be exported into the NanoCommons Data Warehouse from where they can be exported to the established OECD AOP Wiki (https://aopwiki.org/), as well as being utilised in the development of predictive models.

As mentioned in Deliverable Report D5.2 - ‘First big data (omics) analysis and mining tools integrated into KnowledgeBase’, omics data analysis, or Systems Biology, is a powerful tool for understanding biological mechanisms at the molecular level and such information can be used to generate predictive and mechanistic approaches to toxicity. The integrated tools allow the nanosafety community to analyse ‘omics data to identify biological responses to ENM exposure using gene set enrichment analysis (GSEA) and pathway analysis workflows to draw conclusions on the general and specific biological pathways responding under different ENM exposure scenarios. The relevant workflows, presented in D5.2, use both static analysis (biostatistics) and dynamic computational modelling to identify subsets of the multi-dimensional, information rich, ‘omics datasets that represent AOPs, i.e. mechanistically based molecular biomarker signatures that can be implemented into diagnostic screening assays to identify and characterise the impacts of chemicals and ENMs. Static methods, such as differential expression analysis, functional enrichment analysis and Network Reverse Engineering approaches, reconstruct the underlying structure of biological pathways from observational ‘omics data. The dynamical models (from ordinary differential equations to probabilistic or Bayesian models) enable *in silico* simulations of the toxicity responses to ENM, which can be tested experimentally.

2.4 Biokinetics models

Biokinetics offers a methodology for predicting the internal distribution and exposure of a NM in an organism, which can be of particular importance in a risk assessment workflow. Compartmental modeling is a concept broadly used in pharmacokinetics for describing the biodistribution of a substance inside an organism. Physiologically-based pharmacokinetic (PBPK) models represent one of the two major approaches used in compartmental modeling, with empirical models being the second one. PBPK models are mechanistic; they consist of compartments representing real organs and tissues, whose number varies based on the target substance, species, administration route and available information. Several PBPK models that describe the biodistribution of NMs can be found in literature. NanoCommons is developing the necessary infrastructure to develop, host and share PBPK models through the NTU Jaqpot computational platform (Figure 6). The user can upload or insert the physiological parameters of the specific individual where the NM is administered (gender, weight), NM related information (dose, infusion time, initial concentration in each compartment), and the duration and time step that will be used in the simulation. The result is a complete dataset containing ENM concentration -time profiled in all the compartments used in the PBPK model.



MODEL
Title:
Diazepam
Model

Owner: periklists

Description:

A PBPK model consisting of 14 compartments that describes the distribution of diazepam in humans

Diazepam PBPK model

The detailed description of the PBPK model used can be found in Gueorgueiva et al., 2006. Its schematic representation can be seen in the following figure. It consists of 11 compartments describing the concentration of the drug in various tissues, namely liver (*LI*), kidney (*KI*), brain (*BR*), intestine (*IN*), stomach (*ST*), muscle (*MU*), adipose (*AD*), skin (*SK*), gonads (*GO*), heart (*HT*) and lungs (*LU*), one compartment to model the rest of the body (*RE*) as well as two blood pools; venous (*VEN*) and arterial (*ART*). The parameters of the model are divided into drug-dependent and physiological (drug-independent) parameters. The first category comprises eleven tissue-to-plasma partition coefficients (K_p), fraction unbound in plasma (*fu*), blood-to-plasma ratio (*R*), and intrinsic hepatic clearance (CL_{int} , in *L/h*). The physiological parameters of the model are simply the regional tissue blood flows (*Q*) and tissue volumes (*V*).

Figure 6. PBPK modelling through the jaqpot platform developed by NTUA.

3. Data modeling reporting guidelines and templates

One of the key foci of NanoCommons is repeatability and reproducibility of experimental and computational approaches, and accurate and complete reporting to enable re-creation. This is why the models to be integrated into the NanoCommons KB need to be fully described and the necessary information and training materials provided so that future users will be able to accurately use and report the model details. Thus, NanoCommons has based its workflows on the Minimum Information About a Simulation Experiment (MIASE) Guidelines, as well as the QMRF (European Chemicals Agency) and the MODA (European Materials Modelling Council) templates. The detailed integration workflow is described in section 4.

3.1 Minimum Information About a Simulation Experiment (MIASE) Guidelines

One of the main requirements in scientific research is reproducibility of experimental work. This has led to an extensive debate on the amount of information needed to be reported for the experiment so that it can be reproduced. As a result, the establishment of minimum information guidelines has proven valuable for promoting reproducible science. One of the first attempts was the Minimum Information Required in the Annotation of Models (MIRIAM) [3] guidelines promoting the exchange and reuse of biochemical computational models. However, as reported by Waltemath et al. (2011) MIRIAM does not provide sufficient information for the efficient reuse in a computational setting. This is why the MIASE guidelines,[4] to describe the minimal set of information that must be provided to allow the full reproducibility of a simulation experiment, were devised. These guidelines include the list of models to use and their modifications, all the simulation procedures to apply and in which order, the processing of the raw numerical results, and the description of the final output. MIASE allows for the reproduction of any simulation experiment.

In summary, the MIASE guidelines as reported by Waltemath et al. are:

1. All models used in the experiment must be identified, accessible, and fully described.
    a. The description of the simulation experiment must be provided together with the models necessary for the experiment, or with a precise and unambiguous way of accessing those models.
    b. The models required for the simulations must be provided with all governing equations, parameter values, and necessary conditions (initial state and/or boundary conditions).
    c. If a model is not encoded in a standard format, then the model code must be made available to the user. If a model is not encoded in an open format or code, its full description must be provided, sufficient to re-implement it.
    d. Any modification of a model (pre-processing) required before the execution of a step of the simulation experiment must be described.
2. A precise description of the simulation steps and other procedures used by the experiment must be provided.
    a. All simulation steps must be clearly described, including the simulation algorithms to be used, the models on which to apply each simulation, the order of the simulation

steps, and the data processing to be done between the simulation steps.

   b. All information needed for the correct implementation of the necessary simulation steps must be included through precise descriptions or references to unambiguous information sources.
   c. If a simulation step is performed using a computer program for which source code is not available, all the information needed to reproduce the simulation, and not just repeat it, must be provided, including the algorithms used by the original software and any information necessary to implement them, such as the discretization and integration methods.
   d. If it is known that a simulation step will produce different results when performed in a different simulation environment or on a different computational platform, an explanation must be given of how the model has to be run with the specified environment/platform in order to achieve the purpose of the experiment.
3. All information necessary to obtain the desired numerical results must be provided.
   a. All post-processing steps applied on the raw numerical results of simulation steps in order to generate the final results have to be described in detail. That includes the identification of data to process, the order in which changes were applied, and also the nature of changes.
   b. If the expected insights depend on the relation between different results, such as a plot of one against another, the results to be compared have to be specified.

3.2 QSAR Model Reporting Format (QMRF)

QMRF is a harmonised template that was developed by the JRC and EU Member State authorities [5] and is used to summarise and report the key information of QSAR models and the result produced by respective validation studies. The input information is structured according to the OECD validation principles [6], and includes:

1. A defined endpoint
2. An unambiguous algorithm
3. A defined domain of applicability
4. Appropriate measures of goodness-of–fit, robustness and predictivity
5. A mechanistic interpretation, if possible.

The QMRF reports consists of ten sections [7]. The information reported in each section is described in detail in Table 1:

1. QSAR identifier (Table 1a)
2. General information (Table 1b)
3. Defining the endpoint - OECD Principle 1 (Table 1c)
4. Defining the algorithm - OECD Principle 2 (Table 1d)
5. Defining the applicability domain - OECD Principle 3 (Table 1e)
6. Internal validation - OECD Principle 4 (Table 1f)
7. External validation - OECD Principle 4 (Table 1g)
8. Providing a mechanistic interpretation - OECD Principle 5 (Table 1h)

9. Miscellaneous information (Table 1i)
10. Summary (JRC QSAR Model Database) (Table 1j)

Table 1. The 10 sections of the QMRF report. a. QSAR identifier, b. general information, c. defining the endpoint - OECD Principle 1, d. defining the algorithm - OECD Principle 2, e. defining the applicability domain - OECD Principle 3, f. internal validation - OECD Principle 4, g. external validation - OECD Principle 4, h. providing a mechanistic interpretation - OECD Principle 5, i. miscellaneous information, j. summary (JRC QSAR Model Database).

1.QSAR Identifier	
1.1 QSAR identifier (title)	Please provide a clear and concise title that allows the end user to decide whether the model is relevant for their needs. Please provide keywords which specify the endpoint modelled and the name of the expert system where appropriate.
1.2. Other related models	Some models, in particular those encoded into expert systems, might invoke the use of a sub-model or several sub models. This heading is to flag such instances.
1.3. Software coding the model	Please provide the version number of the software model! Failure to provide this information might invalidate the remainder of the QMRF as the version number determines the status of development at a given point in time. Expert systems are typically updated periodically.

(a)

2. General information	
2.1. Date of QMRF	Please provide a timeline of model development, validation and deployment. A timeline is needed to start the audit trail of the documentation of the model.
2.2. QMRF author(s) and contact details	Please provide a contact person/organisation. This is particularly useful if the QMRF author is not the same as the model developer and to provide a point of reference for further information.
2.3. Date of QMRF update(s)	This should be left blank only if the model is the first to be described. In all other instances, it provides an audit trail to track additions / modifications that have been made to an existing QSAR Model. The QMRF can be updated for a number of reasons, such as additions of new information (e.g. addition of new validation studies in section 7) and corrections of information.
2.4. QMRF update(s)	Please clearly specify any updates. Any specific changes should be noted under this field. Indicate the name and the contact details of the author(s) of the updates QMRF (see field 2.3) and list which

	sections and fields have been modified
2.5. Model developer(s) and contact details	This is particularly relevant if the QMRF author and model developer are different. It also provides another point of reference for obtaining further information. Indicate the name of developer(s)/author(s), and the corresponding contact details; possibly report the contact details of the corresponding author.
2.6. Date of model development and/or publication	Please provide a date. This ensures some indication of whether the model is leading edge science at the time of development or not. It is important information for an end user to help them determine what “value” to place on the model in a risk assessment scenario. A reference citation for the model development should also be provided in the case of models published in the peer review literature as a source of background information.
2.7. Reference(s) to main scientific papers and/or software package	Please provide key published references that describe the model development. List the main bibliographic references (if any) to original paper(s) explaining the model development and/or software implementation. Any other reference such as references to original experimental data and related models can be reported in field 9.2 “Bibliography”.
2.8. Availability of information about the model	Please specify: Does the information provided give an appreciation of the extent of information available about the model? Is the algorithm proprietary? Is the training data set available? Indicate whether the model is proprietary or nonproprietary and specify (if possible) what kind of information about the model cannot be disclosed or are not available (e.g., training and external validation sets, source code, and algorithm).
2.9. Availability of another QMRF for exactly the same mode	Please identify existing QMRF(s) for the same model, but produced by a different author. Indicate if you are aware or suspect that another QMRF is available for the current model you are describing. If possible, identify this other QMRF

(b)

3. Defining the endpoint - OECD Principle 1	
3.1. Species	Please provide the name of the species modelled
3.2. Endpoint	Please select the endpoint from the predefined drop down list. Choose the endpoint (physicochemical, biological, or environmental effect) from the pre-defined classification. If the pre-defined classification does not include the endpoint of interest, select “Other” and report the endpoint in the subsequent field 3.3.
3.3. Comment on	Please provide information of the underlying experimental data that

endpoint	has been used as the basis of developing a model. Include in this field any other information to define the endpoint being modelled. Specify the endpoint further if relevant, e.g. according to test organism such as species, strain, sex, age or life stage; according to test duration and protocol; according to the detailed nature of endpoint etc. You can also define here the endpoint of interest in case this is not listed in the predefined classification (see field 3.2).
3.4. Endpoint units	Please clearly specify units of measurement
3.5. Dependent variable	Please describe clearly whether any processing was carried out to the experimental raw data to transform the endpoint to a different form for deriving a model
3.6. Experimental protocol	Please list a test procedure or protocol that provides some background information about the raw data being used. Please provide any important experimental conditions that affect the measurement and therefore the prediction.
3.7. Endpoint data quality and variability	Please clearly specify units of measurement. Provide available information about the test data selection and evaluation and include a description of the data quality used to develop the model. This includes provision of information about the variability of the test data, i.e. repeatability (variability over time) and reproducibility (variability between laboratories) and sources of error (confounding factors which may influence testing results).

(c)

4. Defining the algorithm – OECD Principle 2	
4.1. Type of model	Explain what approach has been used to derive the model
4.2. Explicit algorithm	Please provide an explicit definition of the algorithm including definitions of all descriptors (including substructures where relevant)
4.3. Descriptors in the model	Please identify the number and the name or identifier of the descriptors included in the model. In this context, descriptors refers to e.g. physicochemical parameters, structural fragments etc.
4.4. Descriptor selection	Please provide a justification detailing how descriptors were selected. Indicate the number and the type (name) of descriptors initially screened, and explain the method used to select the descriptors and develop the model from them.
4.5. Algorithm and descriptor generation	Please provide sufficient information to enable the model to be rederived. Explain the approach used to derive the algorithm and the method (approach) used to generate each descriptor.

4.6. Software name and version for descriptor generation	If numerical descriptors are included in the model, please provide sufficient information that enables an end user to regenerate the descriptors for a new compound. Specify the name and the version of the software used to generate the descriptors. If relevant, report the specific settings chosen in the software to generate a descriptor.
4.7. Chemicals/ Descriptors ratio	Are there sufficient compounds per descriptor used in the model? This is important to judge whether the model may have been overfitted. A rule of thumb might be “5 data points per descriptor” included in the model, e.g. a linear regression model with 2 descriptors should be based on at least 10 data points (chemicals). Models with the same ratio of compounds to descriptors are questionable, due to possible overfitting. Report the following ratio: number of chemicals (chemicals from the training set) to number of descriptors , if applicable (if not, explain why).

(d)

5. Defining the applicability domain - OECD Principle 3	
5.1. Description of the applicability domain of the model	Please provide information which characterises the scope of the model such that the end user can determine whether the model is applicable for a specific chemical of interest or not
5.2. Method used to assess the applicability domain	Describe the method used to assess the applicability domain of the model.
5.3. Software name and version for applicability domain assessment	Examples of software might include AMBIT or an in-house algorithm. This can be left blank if no specific software was used to characterise the domain
5.4. Limits of applicability	Describe for example the inclusion and/or exclusion rules (fixed or probabilistic boundaries, structural features, descriptor space, response space) that defines the applicability domain. This will depend on what information has been provided in 5.1.

(e)

6. Internal validation – OECD Principle 4	
6.1. Availability of the training set	Indicate whether the training set is somehow available (e.g., published in a paper, embedded in the software implementing the model, stored in a database) and appended to the current QMRF as supporting information (field 9.3). If it is not available, explain why. This will allow the end user to inspect the underlying basis of the

	model?
6.2. Available information for the training set	Indicate whether the following information for the training set is reported as supporting information (see field 9.3): a) Chemical names (common names and/or IUPAC names); b) CAS numbers; c) SMILES; d) InChI codes; e) MOL files; f) Structural formula; g) Other structural information.
6.3. Data for each descriptor variable for the training set	Indicate whether the descriptor values of the training set are available and are attached as supporting information (see field 9.3).
6.4. Data for the dependent variable for the training set	Indicate whether dependent variable values of the training set are available and attached as supporting information (see field 9.3).
6.5. Other information about the training set	Indicate any other relevant information about the training set Give any extra information that characterises the training set in more detail?
6.6. Preprocessing of data before modelling	Indicate whether raw data have been processed before modelling (e.g. averaging of replicate values); if yes, report whether both raw data and processed data are given. Make it clear whether some processing of the data has been carried out.
6.7. Statistics for goodness-of-fit	Report here goodness-of-fit statistics (R^2, R^2 adjusted, standard error, sensitivity, specificity, false negatives/positives, predictive values etc).
6.8. Robustness - Statistics obtained by leave-one-out cross validation	Has a cross validation been carried out, if so what procedure was used (leave-one-out (LOO), leave-many-out (LMO) etc.)? Is the information sufficient to allow a judgement of the extent of model robustness to be made?
6.9. Robustness - Statistics obtained by leave-many-out cross validation	In case cross-validation was used, is the cross-validation method clearly described or referenced? For example, there are different ways of performing leave-many-out validation
6.10. Robustness - Statistics obtained by Y scrambling	Report here the corresponding statistics and the number of iterations. In case Y-sampling was applied, please add the resulting statistics.
6.11. Robustness - Statistics obtained by bootstrap	Report here the corresponding statistics and the number of iterations. In case bootstrapping was applied, please add the methodological details and resulting statistics
6.12. Robustness - Statistics obtained by other methods	Report here the corresponding statistics in case another cross-validation methods was applied, please describe this clearly and provide the resulting statistics

(f)

7. External validation – OECD Principle 4	
7.1. Availability of the external validation set	Has an external validation been carried out? If not, has an explanation provided as to why an external validation was not carried out? Is the test set available? Is information provided that allows the end-user to determine whether the representativeness of the dataset was taken into account when selecting the chemicals in the test set? Is information available about the experimental data for the test set of chemicals?
7.2. Available information for the external validation set	Please provide the test (validation) set of chemicals and identifiers (e.g. Name, SMILES, CAS#, InChI, MOL file, Formula)
7.3. Data for each descriptor variable for the external validation set	Please provide the descriptor values.
7.4. Data for the dependent variable for the external validation set	Indicate whether dependent variable values of the external validation set are somehow available and attached as supporting information (see field 9.3).
7.5. Other information about the external validation set	Has the approach for selecting test set chemicals been described?
7.6 . Experimental design of test set	Indicate any experimental design for getting the test set (In case that experimental testing was based on prior chemicals selection, make sure that the method for selecting chemicals is described clearly
7.7. Predictivity - Statistics obtained by external validation	Report here the corresponding statistics. In the case of classification models, include false positive and negative rates. Report statistics based on external validation.
7.8. Predictivity - Assessment of the external validation set	Discuss whether the external validation set is sufficiently large and representative of the applicability domain. Describe for example the descriptor and response range or space for the validation test set as compared with that for the training set.
7.9. Comments on the external validation of the model	Add any other useful comments about the external validation procedure.

(g)

8. Providing a mechanistic interpretation - OECD Principle 5	
8.1. Mechanistic basis of the model	Provide information on the mechanistic basis of the model (if possible). In the case of SAR, you may want to describe (if possible) the molecular features that underlie the properties of the molecules containing the substructure (e.g. a description of how sub-structural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region). In the case of QSAR, you may give (if possible) a physicochemical interpretation of the descriptors used (consistent with a known mechanism of biological action). If it is not possible to provide a mechanistic interpretation, try to explain why.
8.2. A priori or a posteriori mechanistic interpretation	Indicate whether the mechanistic basis of the model was determined a priori (i.e. before modelling, by ensuring that the initial set of training structures and/or descriptors were selected to fit pre-defined mechanism of action) or a posteriori (i.e. after modelling, by interpretation of the final set of training structures and or descriptors).
8.3. Other information about the mechanistic interpretation	Report any other useful information about the (purported) mechanistic interpretation described in the previous fields (8.1 and 8.2) such as any reference supporting the mechanistic basis. Give literature references that support the (purported) mechanistic basis.

(h)

9. Miscellaneous information	
9.1. Comments	Please add any additional comments to help build up an appreciation of the level of use of the model or particular scenarios where it has been successfully applied. Equally it would be useful to highlight scenarios where the model was not successfully applied so as to gain an appreciation of the limitations of the model.
9.2. Bibliography	Please add references that might provide further background information or context of use of the model.
9.3. Supporting information	Please add any other supporting information (e.g. external documents) to the QMRF.

(i)

9. Miscellaneous information	
10.1 QMRF number	A unique number (numeric identifier) is assigned to any QMRF that is published in the JRC QSAR Model Database. The number encodes the

	following information: Q YEAR-ENDPOINT-No Example: Q11-417-002 refers to a QMRF published in 2011, for the endpoint 4.17. It is the second QMRF published in 2011. The number is unique for any QMRF uploaded and stored in the JRC QSAR Model Database
10.2 Publication date	The date (day/month/year) of publication in the JRC Database is reported here.
10.3 Keywords	Any relevant keywords associated with the present QMRF are reported here.
10.4 Comments	Any comments that are relevant for the publication of the QMRF in the JRC Database (e.g., comments about updates and about supporting information) are reported here

(j)

3.3 Modeling Data Generalisation templates (MODA)

The MODA templates have been developed by the European Materials Modeling Council (EMMC) for the standardisation of the description of materials models. MODA was developed with the scope to guide Users towards a complete documentation of material models, starting from the end-user or developer through to the computational details including the underpinning theoretical basis of the model. It provides all necessary aspects for: description, reproducibility, curation and interfacing with other models. MODA also includes information about the use case, the numerical solver, and pre-and post processors, allowing full reproducibility.

The MODA templates have the benefit of being able to facilitate the reporting of complex workflows (Figure 7) where a single or multiple models are used (Figure 8), including details of the modelling process (i.e. physics-based or data-driven modeling) and any post-processing steps performed.

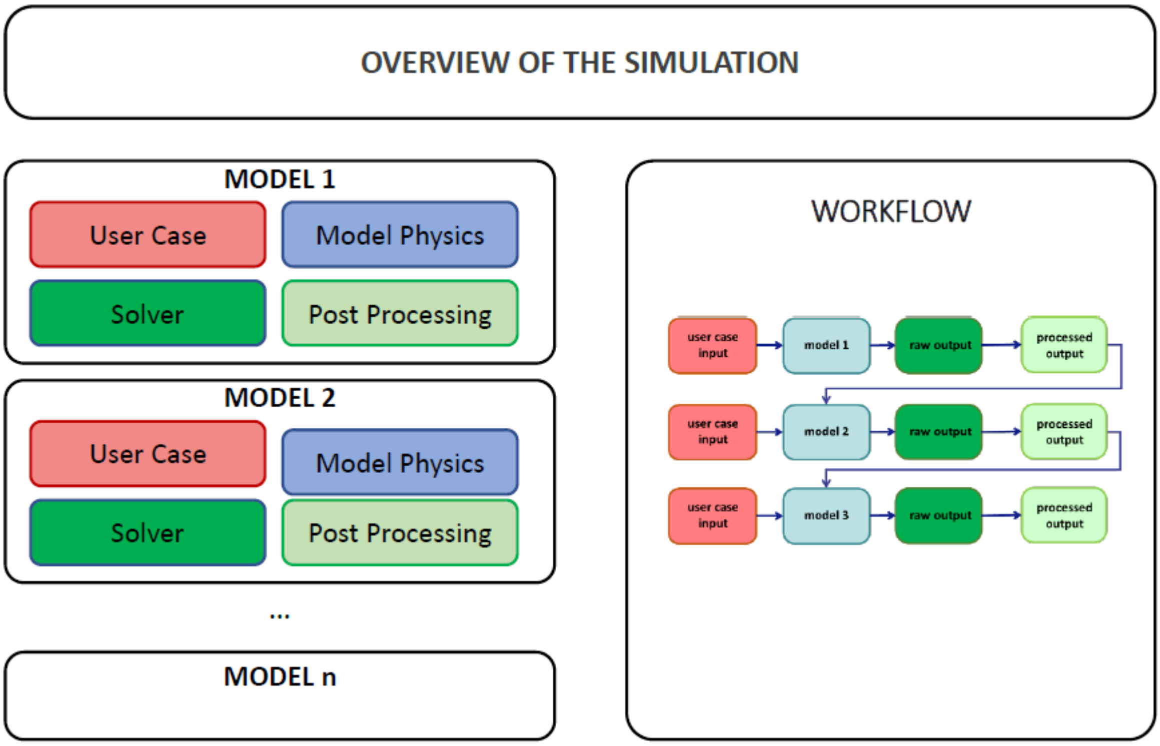


Figure 7. Modeling workflow covered by the MODA templates.
Source: https://emmc.info/moda-workflow-templates/.

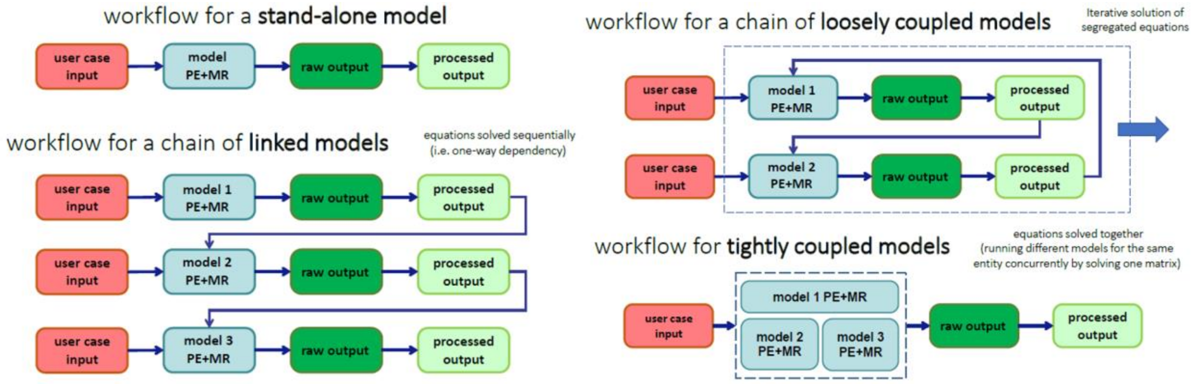

workflow for a stand-alone model
user case input
model PE+MR
raw output
processed output
workflow for a chain of linked models
equations solved sequentially (i.e. one-way dependency)
user case input
model 1 PE+MR
model 2 PE+MR
model 3 PE+MR
raw output
processed output
workflow for a chain of loosely coupled models
Iterative solution of segregated equations
model 1 PE+MR
model 2 PE+MR
workflow for tightly coupled models
equations solved together (running different models for the same entity concurrently by solving one matrix)
model 1 PE+MR
model 2 PE+MR
model 3 PE+MR

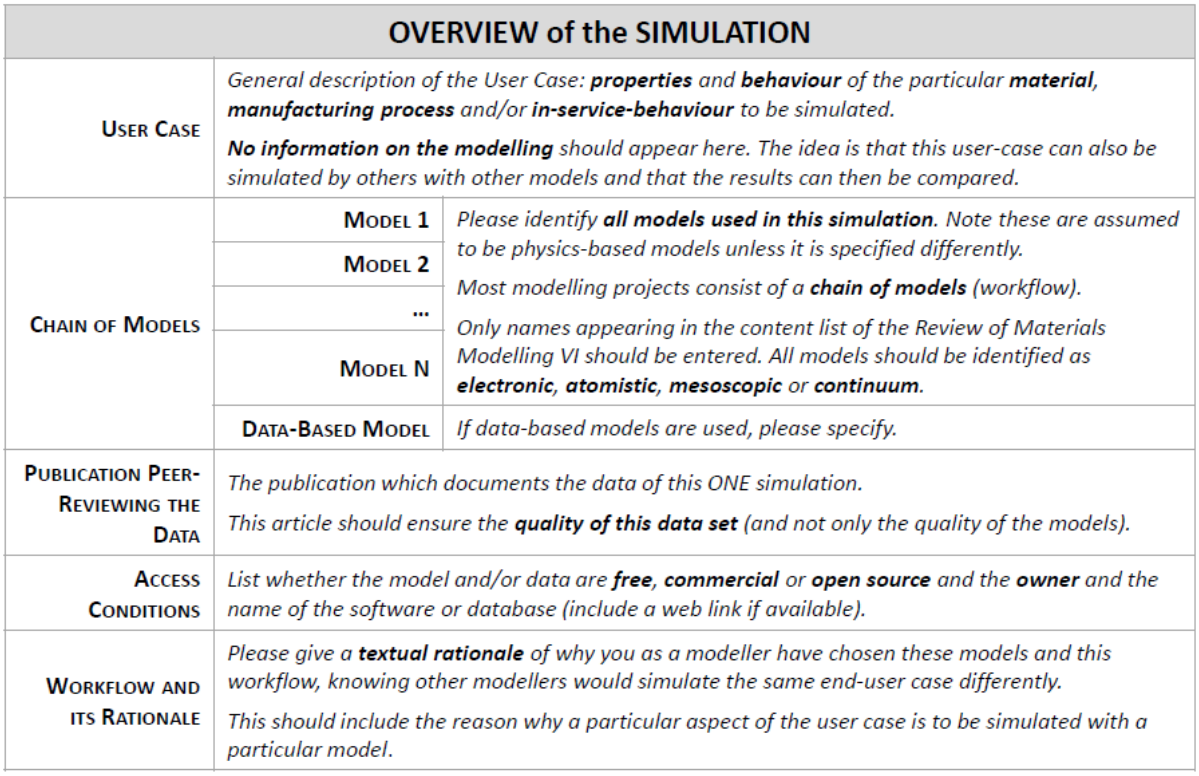
Figure 8. Modeling workflow covered by the MODA templates for different types of model relations - stand-alone (single model) and loosely or tightly coupled models.
Source: https://emmc.info/moda-workflow-templates/.

The MODA templates for physics-based modeling are divided into 5 parts:

1. Overview of the simulation (Figure 9a): A general description of the Use Case, i.e. the material to be identified, its properties and behaviour, the manufacturing process and/or in-service-behaviour to be simulated. This description should be sufficiently detailed to allow testing of other modeling approaches on this example.
2. Aspect(s) of the use case/system to be simulated (Figure 9b): This section is used to textually

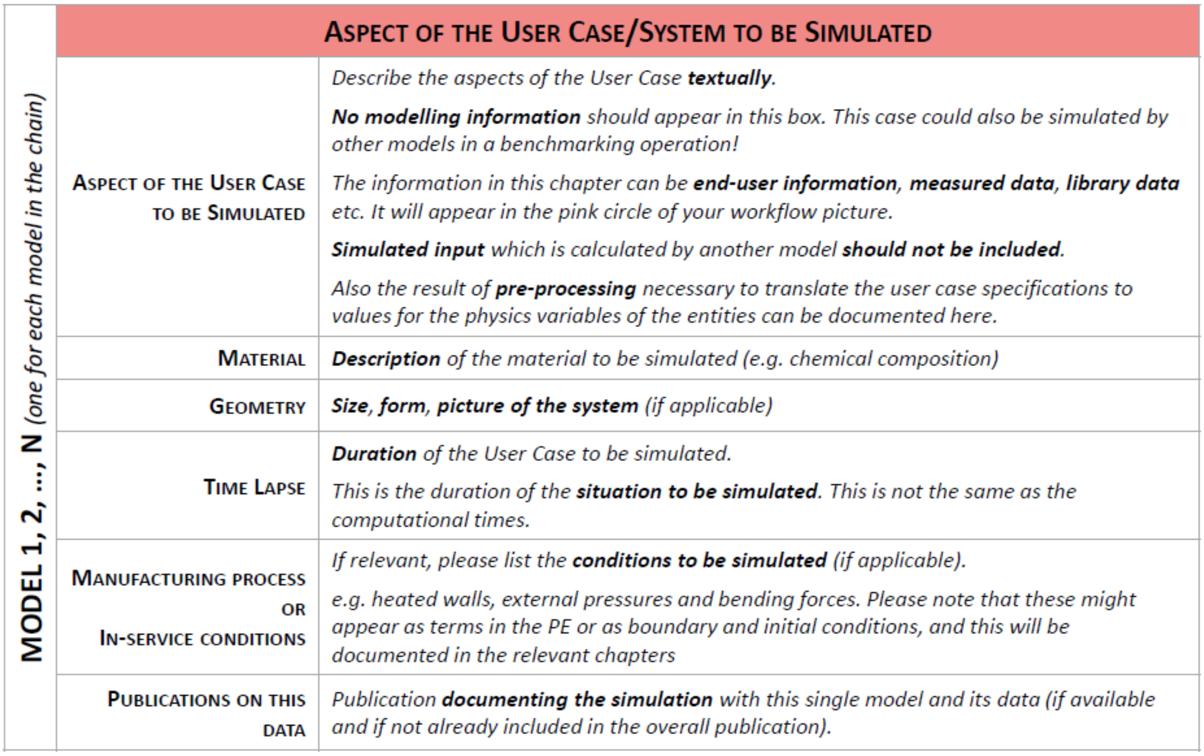
describe the material in question. Again no modeling information are to be entered and the section includes end-user information, measured data, library data etc. Results of pre-processing necessary to translate the user case specifications into values for the physical variables of the entities can be documented here.

3. Model type (Figure 9c): This section provides information on the model(s) used. The model type and name, entity, materials relations and a report of the simulated input.
4. Solver and translation of the specifications (Figure 9d): This section provides information on the numerical solver used for analysis and the specifications/settings to allow reproducibility.
5. Post processing (Figure 9e): The output obtained from the post processing (e.g., values for parameters, new MR and descriptor rules for data-based models). The entity in the next model in the chain for which this output is calculated: electrons, atoms, beads (e.g. nanoparticles, grains), volume elements needs to be specified.



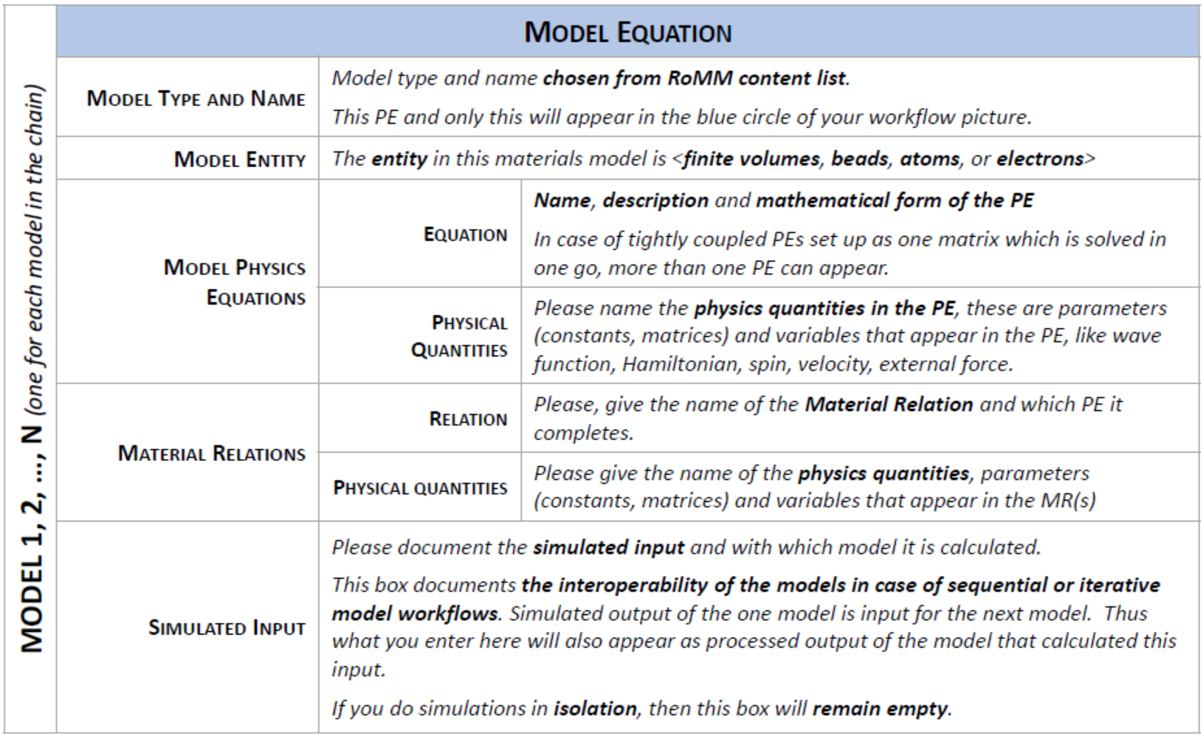
OVERVIEW of the SIMULATION		
User Case		*General description of the User Case:* ***properties*** *and* ***behaviour*** *of the particular* ***material***, ***manufacturing process*** *and/or* ***in-service-behaviour*** *to be simulated.* ***No information on the modelling*** *should appear here. The idea is that this user-case can also be simulated by others with other models and that the results can then be compared.*
Chain of Models	Model 1	*Please identify* ***all models used in this simulation****. Note these are assumed to be physics-based models unless it is specified differently.* *Most modelling projects consist of a* ***chain of models*** *(workflow).* *Only names appearing in the content list of the Review of Materials Modelling VI should be entered. All models should be identified as* ***electronic***, ***atomistic***, ***mesoscopic*** *or* ***continuum****.*
	Model 2	
	…	
	Model N	
	Data-Based Model	*If data-based models are used, please specify.*
Publication Peer-Reviewing the Data		*The publication which documents the data of this ONE simulation.* *This article should ensure the* ***quality of this data set*** *(and not only the quality of the models).*
Access Conditions		*List whether the model and/or data are* ***free***, ***commercial*** *or* ***open source*** *and the* ***owner*** *and the name of the software or database (include a web link if available).*
Workflow and its Rationale		*Please give a* ***textual rationale*** *of why you as a modeller have chosen these models and this workflow, knowing other modellers would simulate the same end-user case differently.* *This should include the reason why a particular aspect of the user case is to be simulated with a particular model.*

(a)



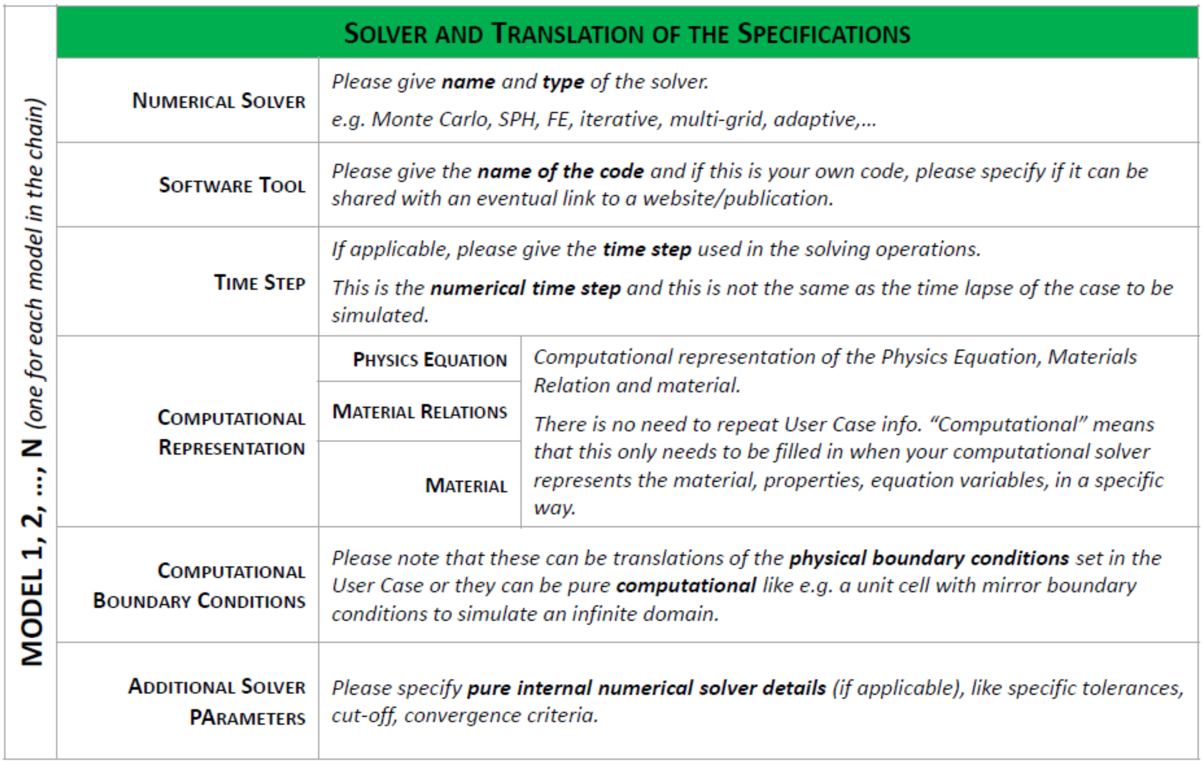
MODEL 1, 2, …, N *(one for each model in the chain)*	ASPECT OF THE USER CASE/SYSTEM TO BE SIMULATED	
	ASPECT OF THE USER CASE TO BE SIMULATED	*Describe the aspects of the User Case* ***textually****.* ***No modelling information*** *should appear in this box. This case could also be simulated by other models in a benchmarking operation!* *The information in this chapter can be* ***end-user information****,* ***measured data****,* ***library data*** *etc. It will appear in the pink circle of your workflow picture.* ***Simulated input*** *which is calculated by another model* ***should not be included****.* *Also the result of* ***pre-processing*** *necessary to translate the user case specifications to values for the physics variables of the entities can be documented here.*
	MATERIAL	***Description*** *of the material to be simulated (e.g. chemical composition)*
	GEOMETRY	***Size****,* ***form****,* ***picture of the system*** *(if applicable)*
	TIME LAPSE	***Duration*** *of the User Case to be simulated.* *This is the duration of the* ***situation to be simulated****. This is not the same as the computational times.*
	MANUFACTURING PROCESS OR IN-SERVICE CONDITIONS	*If relevant, please list the* ***conditions to be simulated*** *(if applicable).* *e.g. heated walls, external pressures and bending forces. Please note that these might appear as terms in the PE or as boundary and initial conditions, and this will be documented in the relevant chapters*
	PUBLICATIONS ON THIS DATA	*Publication* ***documenting the simulation*** *with this single model and its data (if available and if not already included in the overall publication).*

(b)



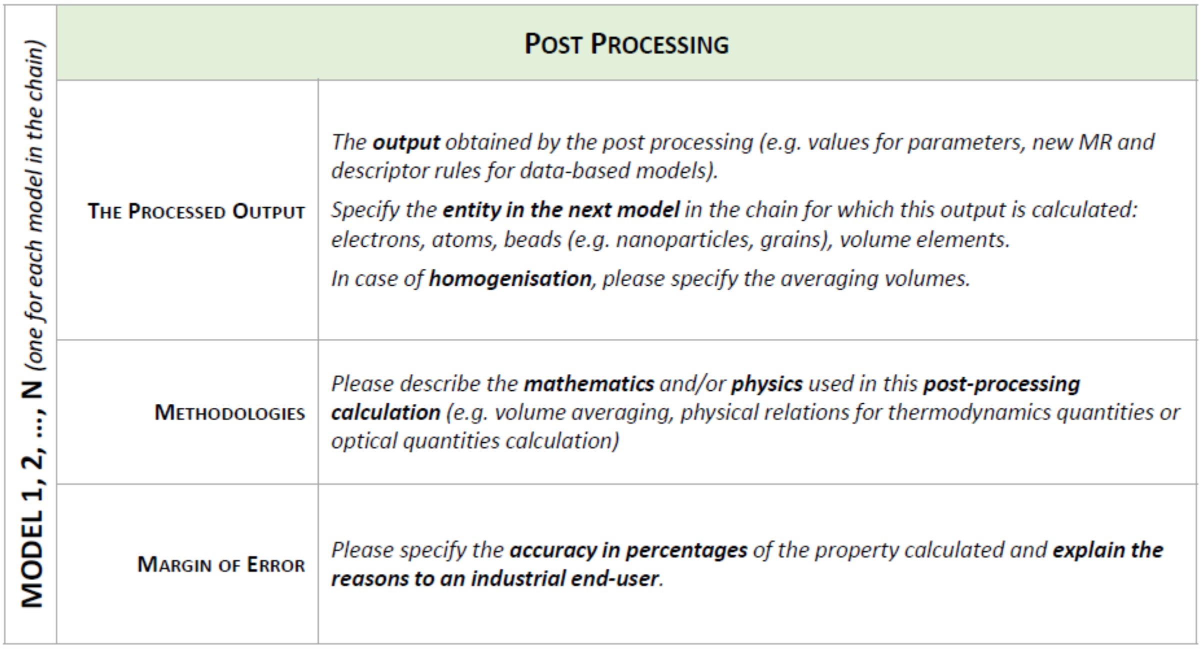
MODEL 1, 2, …, N *(one for each model in the chain)*	MODEL EQUATION		
	MODEL TYPE AND NAME		*Model type and name* ***chosen from RoMM content list****.* *This PE and only this will appear in the blue circle of your workflow picture.*
	MODEL ENTITY		*The* ***entity*** *in this materials model is <****finite volumes****,* ***beads****,* ***atoms****, or* ***electrons****>*
	MODEL PHYSICS EQUATIONS	EQUATION	***Name****,* ***description*** *and* ***mathematical form of the PE*** *In case of tightly coupled PEs set up as one matrix which is solved in one go, more than one PE can appear.*
		PHYSICAL QUANTITIES	*Please name the* ***physics quantities in the PE****, these are parameters (constants, matrices) and variables that appear in the PE, like wave function, Hamiltonian, spin, velocity, external force.*
	MATERIAL RELATIONS	RELATION	*Please, give the name of the* ***Material Relation*** *and which PE it completes.*
		PHYSICAL QUANTITIES	*Please give the name of the* ***physics quantities****, parameters (constants, matrices) and variables that appear in the MR(s)*
	SIMULATED INPUT		*Please document the* ***simulated input*** *and with which model it is calculated.* *This box documents* ***the interoperability of the models in case of sequential or iterative model workflows****. Simulated output of the one model is input for the next model. Thus what you enter here will also appear as processed output of the model that calculated this input.* *If you do simulations in* ***isolation****, then this box will* ***remain empty****.*

(c)



MODEL 1, 2, …, N (one for each model in the chain)	Solver and Translation of the Specifications		
	Numerical Solver		*Please give **name** and **type** of the solver.* *e.g. Monte Carlo, SPH, FE, iterative, multi-grid, adaptive,…*
	Software Tool		*Please give the **name of the code** and if this is your own code, please specify if it can be shared with an eventual link to a website/publication.*
	Time Step		*If applicable, please give the **time step** used in the solving operations.* *This is the **numerical time step** and this is not the same as the time lapse of the case to be simulated.*
	Computational Representation	Physics Equation	*Computational representation of the Physics Equation, Materials Relation and material.* *There is no need to repeat User Case info. "Computational" means that this only needs to be filled in when your computational solver represents the material, properties, equation variables, in a specific way.*
		Material Relations	
		Material	
	Computational Boundary Conditions		*Please note that these can be translations of the **physical boundary conditions** set in the User Case or they can be pure **computational** like e.g. a unit cell with mirror boundary conditions to simulate an infinite domain.*
	Additional Solver PArameters		*Please specify **pure internal numerical solver details** (if applicable), like specific tolerances, cut-off, convergence criteria.*

(d)

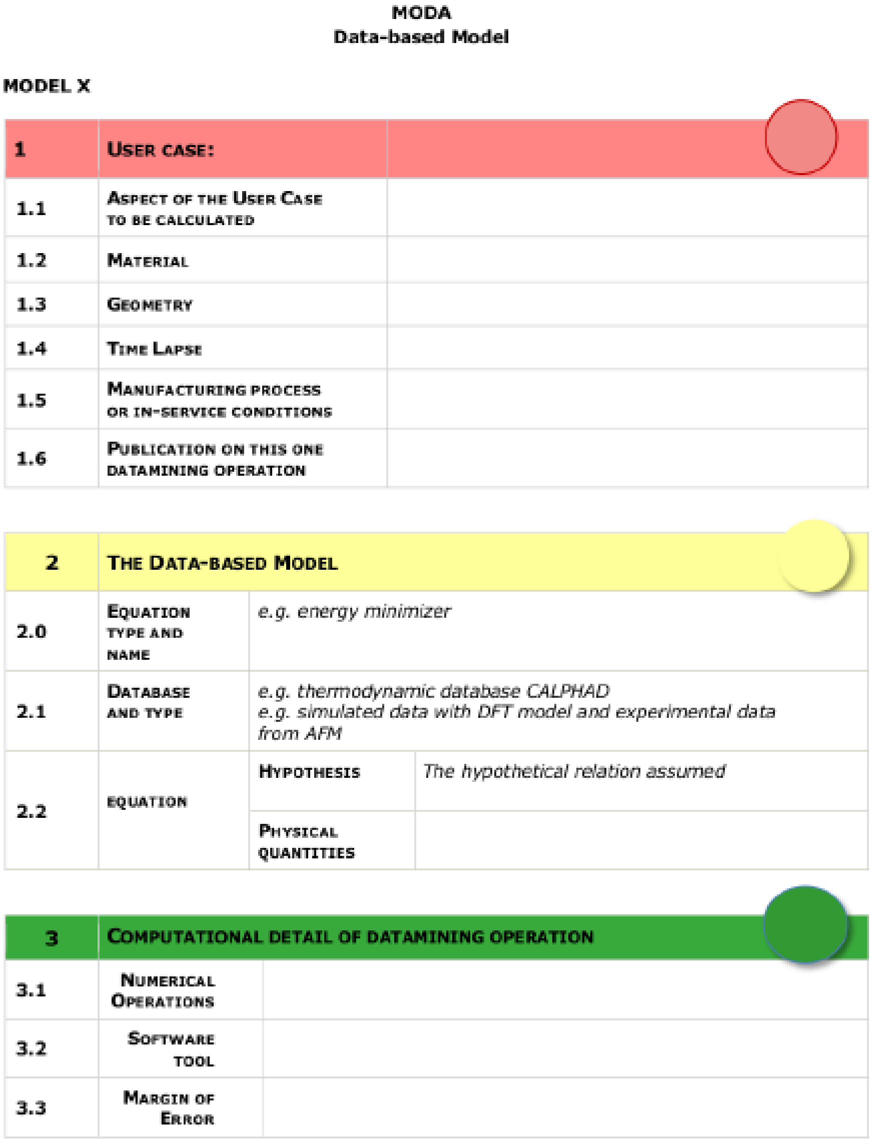


MODEL 1, 2, …, N (one for each model in the chain)	Post Processing	
	The Processed Output	*The **output** obtained by the post processing (e.g. values for parameters, new MR and descriptor rules for data-based models).* *Specify the **entity in the next model** in the chain for which this output is calculated: electrons, atoms, beads (e.g. nanoparticles, grains), volume elements.* *In case of **homogenisation**, please specify the averaging volumes.*
	Methodologies	*Please describe the **mathematics** and/or **physics** used in this **post-processing calculation** (e.g. volume averaging, physical relations for thermodynamics quantities or optical quantities calculation)*
	Margin of Error	*Please specify the **accuracy in percentages** of the property calculated and **explain the reasons to an industrial end-user**.*

(e)

Figure 9. MODA template guide for physical modeling.
Source: https://emmc.info/moda-workflow-templates/.

The MODA templates for data-driven modeling (Figure 10) are simpler since these are based on extraction/identification of relations using data-mining on simulated or experimental data. These simplified relations, when used in isolation, do not always need complicated numerical solvers as they are able to find quick answers. In this case though, the database from which these relations are extracted should always be documented.



MODA
Data-based Model

MODEL X

1	User case:	
1.1	Aspect of the User Case to be calculated	
1.2	Material	
1.3	Geometry	
1.4	Time Lapse	
1.5	Manufacturing process or in-service conditions	
1.6	Publication on this one datamining operation	

2	The Data-based Model		
2.0	Equation type and name	*e.g. energy minimizer*	
2.1	Database and type	*e.g. thermodynamic database CALPHAD* *e.g. simulated data with DFT model and experimental data from AFM*	
2.2	Equation	Hypothesis	*The hypothetical relation assumed*
		Physical quantities	

3	Computational detail of datamining operation	
3.1	Numerical Operations	
3.2	Software tool	
3.3	Margin of Error	

Figure 10. MODA template guide for data modeling.
Source: https://emmc.info/moda-workflow-templates/.

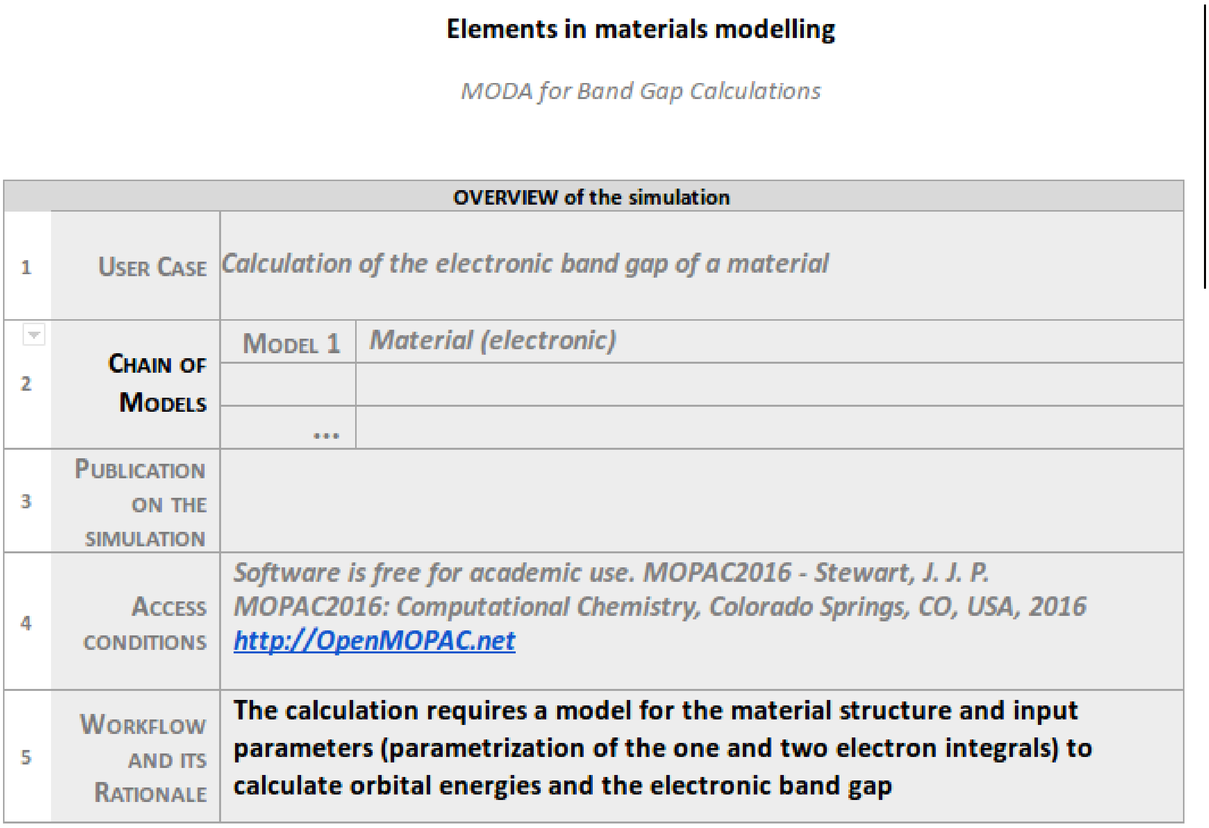
NanoCommons is developing variations of the MODA templates to accommodate the needs of its Users and to expand their usability, and is trying to automate / standardise them through, for example, use of drop-down lists to avoid multiple variants of the same concept. An example developed for models used within the Horizon 2020 project NanoFASE is presented in Table 2. These were developed in collaboration with the NanoFASE research scientists modelling environmental exposure as a NanoCommons case study, and are used to curate and upload the NanoFASE simulation data into the

NanoFASE section of the NanoCommons KB. Similarly, NanoCommons partners UCD are developing MODA templates to be used for capturing data from other sources and software packages (e.g. MOPAC, see Table 3).

Table 2. MODA template variation used in the NanoFASE project.

Part I: Overview	Preferred name of the model
	Name of the material modelled
	Language of encoding
	Time lapse
	Time step (if applicable)
	Geographical coverage
	Solver used (if applicable)
	Results expected
	Date and time of creation
	Date & Time of last modification
	Name and contact information of the creators of the model
	Related publications
	DOI
Part II: Input data	Parameters entered
	Description of how each of these parameters were obtained / estimated
Part III: Model	Structure of the model
	Step-by-step explanation of the script, with equations/algorithm
	Script
	Post-processing steps applied on the raw results (if applicable)
	Description of the changes to be made to run the model on another system

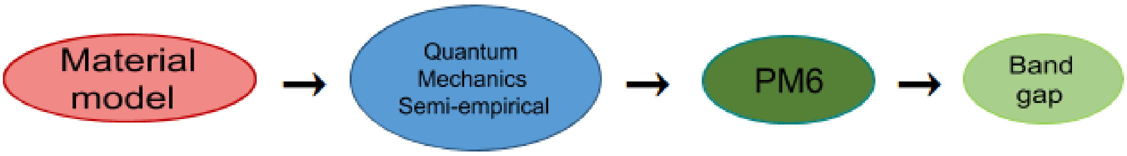
Table 3. MODA template for the MOPAC software, developed by NanoCommons partner UCD.



Elements in materials modelling

MODA for Band Gap Calculations

OVERVIEW of the simulation			
1	USER CASE	*Calculation of the electronic band gap of a material*	
2	CHAIN OF MODELS	MODEL 1	*Material (electronic)*
		...	
3	PUBLICATION ON THE SIMULATION		
4	ACCESS CONDITIONS	*Software is free for academic use. MOPAC2016 - Stewart, J. J. P. MOPAC2016: Computational Chemistry, Colorado Springs, CO, USA, 2016 http://OpenMOPAC.net*	
5	WORKFLOW AND ITS RATIONALE	**The calculation requires a model for the material structure and input parameters (parametrization of the one and two electron integrals) to calculate orbital energies and the electronic band gap**	

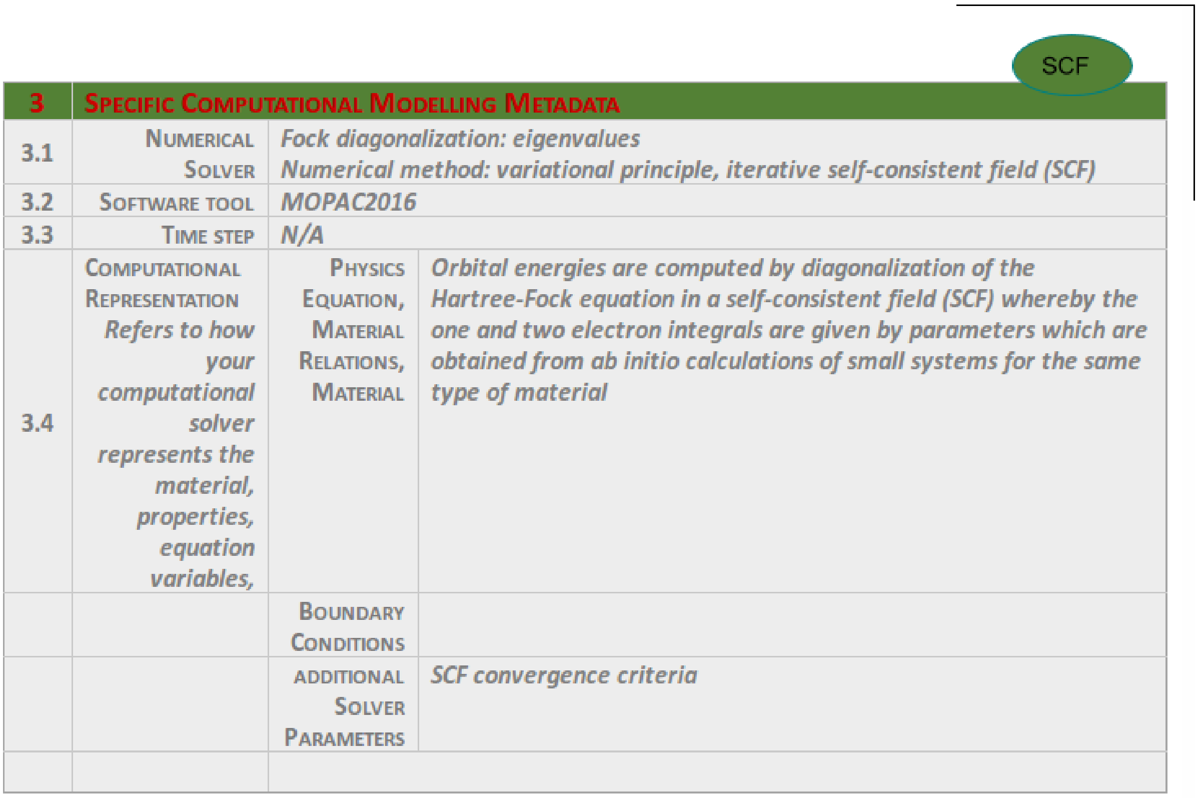


Material model

1	Aspect of the User Case/System to be Simulated	
1.1	Aspect of the User Case to be simulated and how it forms a part of the total user case	*The electronic band structure of the specified material*
1.2	Material	*user input - any dielectric material*
1.3	Geometry	*a molecular structure of the material in terms of atomic coordinates: finite size (nanoparticle) or continuous phase*
1.4	Time Lapse	*N/A*
1.5	Manufacturing process or in-service conditions	
1.6	Publication on this one simulation	

Quantum Mechanics

2	Generic Physics Of The Model Equation		
2.0	Model type and name	*Quantum mechanics - Semi-empirical*	
2.1	Model entity	*Atoms (nuclei and electrons)*	
2.2	Model Physics/ Chemistry equation PE's	Equations	*Time-independent electronic Schrödinger equation, Hartree-Fock equation*
		Physical quantities for each equation	*Wave function, Hamiltonian, orbital energy*
Materials relations		MR Equations	*Orbital wave functions which describe the electrons of the material in space; parametrization of the Hartree-Fock equation*
		Physical quantities/ descriptors for each MR	*PM6 parametrization of the one and two electron integrals*
2.4	Simulated input		



3	Specific Computational Modelling Metadata		
3.1	Numerical Solver	*Fock diagonalization: eigenvalues* *Numerical method: variational principle, iterative self-consistent field (SCF)*	
3.2	Software tool	*MOPAC2016*	
3.3	Time step	*N/A*	
3.4	Computational Representation *Refers to how your computational solver represents the material, properties, equation variables,*	Physics Equation, Material Relations, Material	*Orbital energies are computed by diagonalization of the Hartree-Fock equation in a self-consistent field (SCF) whereby the one and two electron integrals are given by parameters which are obtained from ab initio calculations of small systems for the same type of material*
		Boundary Conditions	
		additional Solver Parameters	*SCF convergence criteria*

The MODA templates are currently in development in terms of algorithmic and systematic construction of workflows and collection of metadata. Recent efforts by EMMC-CSA lead to the development of two ontologies, suitable for describing materials models: VISO and OSMO. The VIMMP Software Ontology (VISO) is designed to describe simulation software at the Virtual Materials Marketplace (VIMMP) [8], complementing OSMO (Ontology for Simulation, Modelling and Optimisation). The OSMO ontology enhances the original MODA by being machine processable, amenable to automated reasoning by semantic technology, and by which workflow semantics in materials modelling are captured in a way that is closely aligned and interoperable with the whole family of semantic assets presently under development in the context of several infrastructures and projects.

The software ontology VISO is developed to represent software packages and their features, and OSMO, an ontology for simulation, modelling, and optimisation, is introduced on the basis of MODA, a previously developed semi-intuitive graph notation for workflows in materials modelling.

The ontology for simulation, modelling, and optimization (OSMO) is based on the vocabulary and the approach from the 6th Review of Materials Models (RoMM) [9] including its representation of use cases, solvers, models, and processing is directly based on MODA, and the representation of workows is based on the LDT (logical data transfer) notation. By providing a common semantic basis for workows that were designed with different tools, OSMO can be employed to consistently integrate data provenance descriptions for materials modelling data from diverse sources. The detailed description of the four types of section entities (use cases, models, solvers, and processors) in OSMO follows the specification from MODA closely.

In OSMO, building on the terminology from RoMM, common physical equations in materials modelling are classified into 25 types, represented by subclasses of the OSMO classification_type, at four granularity levels (instances of the OSMO class granularity_level). The characterization of model granularity follows De Baas [9] where the scope of each of the RoMM vocabulary categories is discussed in great detail.

Accordingly, particle-based methods are defined to be atomistic if the particles represent single atoms and mesoscopic if they represent multiple atoms; by this categorization, e.g., molecular models following the united-atom approach are regarded as mesoscopic. This distinction between atomistic and mesoscopic physical equations, however, is only based on the role as ascribed to the discrete particles; therefore, the same equations can be applied at both levels. To ensure that the expressive capacity of OSMO matches that of RoMM, MODA, and EMMO, it is necessary to differentiate between these two levels.

4. NanoCommons modeling tools integration workflow

The scientific and technical integration workflows presented here (Figures 8 and 9, respectively) define the necessary information for a modeling tool to be integrated into the NanoCommons KB. This includes the required information to ensure that the integrated tool will be fully compatible with the NC-KB and its underpinning datasets, and is functional and secure. Users will be able to use it seamlessly, securely with the appropriate training and support. The workflow is divided in two main parts. The scientific part focuses on the type of models used in the tool and the technical part on the tool development and web integration into the KB.

The workflows presented below constitute the recommended best-practice workflows and are these that the NanoCommons project will pursue implementing. Having said that, a looser integration will also be allowed, when a relevant tool is submitted. The reason is that some of the required steps are complex and require a substantial amount of time to be completed and could lead to limiting the tools' availability.

4.1 Scientific workflow for modeling tool integration into the NanoCommons KB

The workflow (Figure 8) for the scientific validity of the modeling tool aims to make sure that the tool is properly built, is user-friendly and that the necessary training materials exist so that it will be possible for people to use it without the need for continuous expert support. As a result, the workflow steps identified are:

1. A NanoCommons modelling experts team is established to collaborate with the modeling tool owners and assist with integration. The two teams will be in continuous contact to exchange relevant information that will streamline either the integration process or develop the necessary training materials (tutorials, videos etc.).
2. Information from the modeling tool owners on the model(s) incorporated into the tool is collected. This includes the type of supported modeling (physical vs. data modeling, machine learning vs. deep learning), the equations used, the needed input parameters etc.
3. The tool owners need to provide information on the data and modelling Quality Assurance and Quality Control (QA/QC) processes, the validation methods used and how they ensure the robustness and validity of the outcomes.
4. The NanoCommons team evaluates the tool's user interface and user experience (data uploading and handling, modeling, results extraction etc.). This ensures that the tool meets the NanoCommons requirements for minimal need for computational expertise and/or expert guidance for use of the model / tool.
5. The tool owners need to provide information on the possibility of implementation of further algorithms on both the developer and user sides. With the pace at which the modelling community evolves, the ability to easily implement improved algorithms and approaches is essential to keep up with the developments in the scientific and industrial community and to fully exploit their results.
6. The NanoCommons team evaluates the results extraction and whether sufficient metadata are included to allow future interoperability and reusability.

7. The NanoCommons expert team and tool owner check the semantic annotation provisions of the tool and the FAIRness score of the extracted results.
8. The NanoCommons team receives and evaluates all existing training materials that will allow Users to use the tool independently, and integrates them into the wider set of NanoCommons training materials and tool selection decision trees to guide users.
9. The NanoCommons team identifies any gaps and works with the tool owner for the implementation of improvements to fill these in subsequent tool updates.

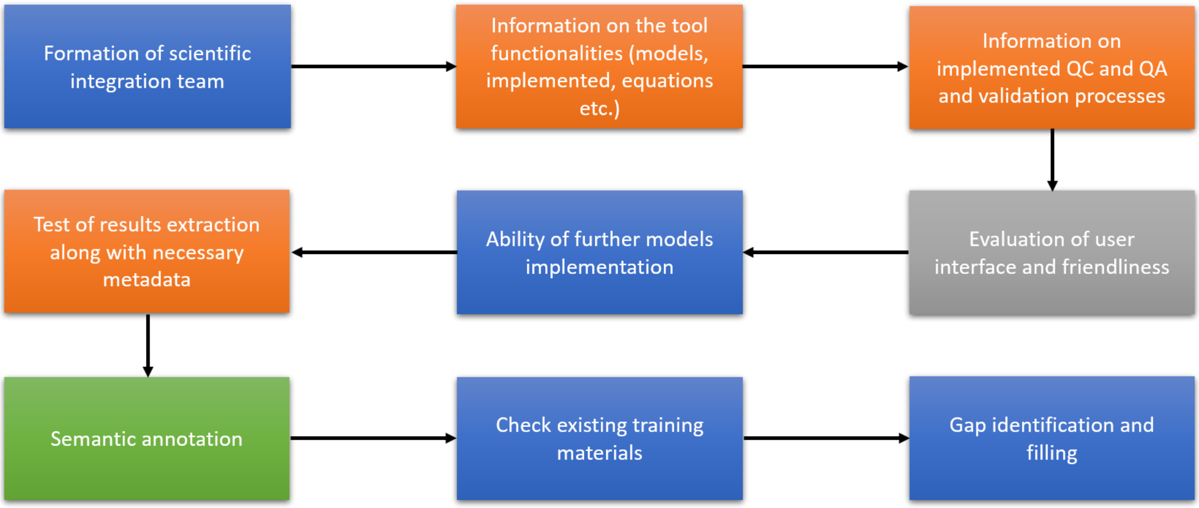

Formation of scientific integration team
Information on the tool functionalities (models, implemented, equations etc.)
Information on implemented QC and QA and validation processes
Evaluation of user interface and friendliness
Ability of further models implementation
Test of results extraction along with necessary metadata
Semantic annotation
Check existing training materials
Gap identification and filling

Figure 8. Schematic workflow for the tool scientific evaluation prior to integration into the NanoCommons KB and e-infrastructure platform.

4.2 Technical workflow for modeling tools integration into the NanoCommons KB

One of the most significant aspects for the integration of tools into the NanoCommons KB is the evaluation of the technical parameters of the tool and its compatibility with the KB. Thus, specific steps need to be taken to ensure maximum compatibility that will ensure both seamless integration into the KB and facilitate widespread use of the tool via the NanoCommons KB by the wider nanosafety community. To achieve these goals the below workflow (Figure 9) has been identified.

1. NanoCommons creates a technical team that will oversee, guide and run the integration process. This is done in close collaboration with the technical personnel of the tool developer and includes continuous exchange of information on the necessary technical details.
2. Initially the two teams exchange information on the programming language(s) used to develop the tool and any specific software/library/operating system used or needed for the tool to be functional.
3. The two teams exchange information on the use of any Application Programming Interface (API) and Graphical User Interface (GUI) that the NanoCommons KB and tool used for presentation and communication. This ensures proper data transfer, sharing and

synchronisation. Compatibility checking will allow an initial remote integration, until full integration is established.

4. The NanoCommons team then checks the requirements for deploying the tool in the NanoCommons KB. The tool development team needs to provide information on the existing Kubernetes/OpenShift setup files or Docker containers created for this purpose. If these don't exist yet, then the NanoCommons integration team supports the tool owners in developing the necessary files and creating containers.
5. The tool owners needs to provide information on the potential for the online and local (offline) use of the tool. This will allow users with sensitive data to use the tool in-house and increase the security of the application.
6. The integration team also checks the capacity and workload handling of the tool to ensure its ability to handle heavy workloads from multiple users, should the need arise.
7. Testing of the security, licensing, data handling and GDPR compliance is performed via collaboration of both teams. NanoCommons is dedicated to protecting both the personal information and the intellectual property rights (IPR) of its users and developers.
8. Finally, NanoCommons requires information on the tools sustainability, updating and bug fixing processes. This is a key step as it will ensure long term tool functionality and compatibility when updates or changes on both sides are implemented.

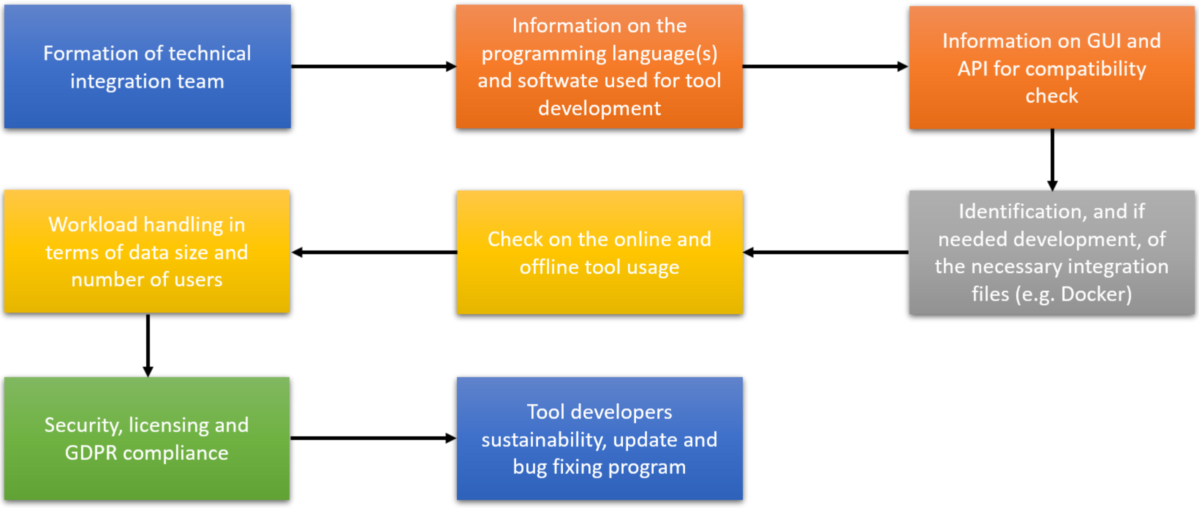

Formation of technical integration team
Information on the programming language(s) and software used for tool development
Information on GUI and API for compatibility check
Identification, and if needed development, of the necessary integration files (e.g. Docker)
Check on the online and offline tool usage
Workload handling in terms of data size and number of users
Security, licensing and GDPR compliance
Tool developers sustainability, update and bug fixing program

Figure 9. Technical workflow for tool integration into the NanoCommons KB and e-infrastructure platform.

5. Conclusions

This deliverable presented the initial set of modelling tools on offer by the NanoCommons project which were used as the basis for developing the respective workflows for full scientific and technical integration into the NanoCommons Knowledgebase. The presented workflows ensure that any model, whether physical or data driven, submitted to NanoCommons will be compatible and fully functional when it is integrated and becomes publicly available. To this end, it needs to be emphasised that the presented workflows are complex and require a certain amount of time to be completed, as well as investment of effort from both the tool developers and the NanoCommons implementation team. This means that an intermediate stage where the modelling tools submitted to NanoCommons may become available under a loose integration approach initially, and possibly remain available from outside the NanoCommons Knowledgebase. This ensures both the tools' availability, but also the full technical and functional compatibility and support to the NanoCommons Knowledgebase users.

The value of the presented workflows has been demonstrated through the integration of already available modelling tools, like the NanoImage and NanoXtract tools developed by the NanoCommons partners NTUA and NovaMechanics, respectively. In parallel, the NanoCommons consortium is actively working to implement and "impose" on tool developers interested in integration into NanoCommons (Users) a series of minimum information guidelines, such as the MIRIAM and MIASE guidelines in combination with established and community agreed templates (e.g. MODA, QMRF). This will ensure the high quality and reproducibility of the submitted models, facilitating their consideration and adoption for regulatory purposes in due course. Examples of this work are the documentation of the models developed within the NanoFASE project, developed by the NanoFASE experts in collaboration with UoB, and the MODA template for the MOPAC software developed by the NanoCommons partners UCD based on their models developed within the SmartNanoTox project. The next steps in this process will be to demonstrate the value of the presented workflows and to actively promote them to potential model developers interested to submit their models to the NanoCommons Knowledgebase. There will be a dedicated TA call for integration of modelling tools in early 2020.

6. References

1. Chetwynd, A.J., K.E. Wheeler, and I. Lynch, *Best practice in reporting corona studies: Minimum information about Nanomaterial Biocorona Experiments (MINBE).* Nano Today, 2019: p. 100758.
2. Leong, H.S., et al., *On the issue of transparency and reproducibility in nanomedicine.* Nature Nanotechnology, 2019. **14**(7): p. 629-635.
3. Le Novère, N., et al., *Minimum information requested in the annotation of biochemical models (MIRIAM).* Nature biotechnology, 2005. **23**(12): p. 1509.
4. Waltemath, D., et al., *Minimum information about a simulation experiment (MIASE).* PLoS computational biology, 2011. **7**(4): p. E1001122.
5. Triebe, J., et al., *EURL ECVAM DataBase service on ALternative Methods to animal experimentation: To promote the development and uptake of alternative and advanced methods in toxicology and biomedical sciences* Publications Office of the European Union, 2017. http://publications.jrc.ec.europa.eu/repository/bitstream/JRC107491/kjna28713enn.pdf
6. OECD, D., *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q) SAR] Models.* Organisation for Economic Co-operation and Development, Paris, France, 2007. https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm
7. *EURL ECVAM DataBase service on ALternative Methods to animal experimentation: QMRF Guideline for Authors and Editors.* Publications Office of the European Union, 2017.*http://publications.jrc.ec.europa.eu/repository/bitstream/JRC107493/kjna28714enn.pdf*.
8. Horsch, M., et al., *Semantic interoperability and characterization of data provenance in computational molecular engineering.* arXiv preprint arXiv:1908.02335, 2019.
9. de Baas, A. and R. Lula, *What makes a material function? Let me compute the ways*. 2017, Publications Office of the European Union.

Annex

Scientific checklist for modeling tools incorporation into the NanoCommons KB

R1. Create a NanoCommons modelling experts team to collaborate with the modeling tool owners and assist with integration

R2. Receive information from the modeling tools owners on the model(s) incorporated into the tool. This includes the type of supported modeling (physical vs. data modeling, machine learning vs. deep learning), the equations used, the needed parameters etc.

R3. Acquire information on the QA/QC processes the validation methods used and how they ensure the robustness and high quality of the outcomes

R4. Identify and evaluate the tool user interface, user experience and, if possible, the interoperability with other NanoCommons services (data uploading and handling, modeling, results extraction, harmonization of user guidance etc.)

R5. Information on the possibility for implementation of further algorithms on both the developer and user side

R6. Evaluate the results extraction and whether sufficient metadata are included to allow future interoperability and reusability

R7. Evaluate the semantic annotation, to ensure findability, and test the FAIRness score of the extracted results

R8. Test the existence of training materials that will allow Users to use the tool without the need of expert support

R9. Identify existing gaps, develop and implement the necessary updates.

Technical checklist for modeling tools incorporation into the NanoCommons KB

R1. Create a NanoCommons technical team to collaborate with the modeling tool owners and assist with integration

R2. Identify the programing language(s) in which the tool has been developed

R3. Test whether there is an API for fast remote integration into the KB, while full integration is ongoing

R4. Test the tool deployment requirements and whether it has already been translated using specific tools (e.g. OpenShift templates, Docker builds)

R5. Receive information on the potential for online and offline use of the modelling tool

R6. For online use, check the tools capacity and workload handling, e.g. capability to handle multiple users in parallel

R7. Collect information on the supported communication protocols for data movement, sharing and synchronisation

R8. Test the tools security, licensing, data handling and GDPR compliance

R9. Collect and evaluate information on the tools sustainability, update and bug fixes to ensure long-term usability of the tool within the NanoCommons e-infrastructure.