Modelling the Environmental Fate of Petroleum Hydrocarbons During Bioremediation

Guozhong Wu and Frédéric Coulon

Abstract

This chapter provides the key steps and parameters required for three different numerical modelling approaches to predict the environmental fate of petroleum hydrocarbons in contaminated soils during bioremediation. The first approach is the molecular dynamic simulation which is used to characterise the molecular-scale adsorption, the diffusion and the distribution of the saturate, aromatic, resin and asphaltene (SARA) fractions of oil. Such approach provides insights into the microscopic aggregation, the sequestration and the collision mechanisms which are essential for a better understanding of hydrocarbon bioavailability and biodegradation. The second approach is the use of fugacity modelling to compute the equilibrium distribution of the aliphatic and aromatic hydrocarbons in an environmental matrix composed of four compartments: soil, water, air and nonaqueous phase liquid (NAPL). Further to this, the contribution of the biotic and abiotic processes to the loss of petroleum hydrocarbons including (1) biodegradation in soil and NAPL, (2) advection in air, (3) leaching from soil and (4) diffusion at the soil–air, soil–water and soil–air boundaries can be estimated during biopiling experiments. The third approach is the use of machine learning (ML), an assumption-free data mining method, to predict the changes in the bioavailability of polycyclic aromatic hydrocarbons (PAHs) in contaminated soils. The main advantage of ML models is that they are data-based technique allowing computers to learn and recognise the patterns of the empirical data and work well with highly non-linear systems without relying on prior knowledge on bioremediation processes which make their prediction more realistic than conventional statistical methods. ML outputs can be integrated into microbial degradation models to support decision making for the assessment of bioremediation end points.

Keywords: Bioremediation, Fugacity modelling, Machine learning, Molecular simulation, Petroleum hydrocarbons

1 Introduction

Despite the fact that a variety of bioremediation technologies have been successfully applied for petroleum-contaminated sites, the complicated nature of soil and oil chemistry results in a lack of a universal technology that can be the solution for all contamination. Insights into the intricate interactions between oil and soil especially at the molecular scale are essential for a comprehensive

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 165–180,
DOI 10.1007/8623_2015_121, © Springer-Verlag Berlin Heidelberg 2015, Published online: 04 August 2015

appreciation of the fate of petroleum contaminants. Molecular simulation is a versatile tool to handle such issues. It represents a category of methods that convert the microscopic-level information (e.g. the position and the diffusion velocity of atoms) to the macroscopic properties (e.g. concentration profile, diffusion coefficient and thermodynamic properties of molecules) using statistical mechanics [1]. Simulation methods include quantum chemistry, molecular dynamics (MD), Monte Carlo and mesoscale simulation with the timescale ranging from femto- to nanoseconds (Fig. 1). The fast development of these methods during the last decades improved our understanding of the environmental behaviour of petroleum contaminants in the environment. Although it remains difficult to model the chemical interactions between soil and oil using an 'average' molecular formula for the natural soil and oil from various origins, several models have been developed for predicting the influence of the inorganic mineral and the organic matter of soil on the fate of oil. For example, the structural formula of soil minerals such as gibbsite, kaolinite, pyrophyllite and montmorillonites has now all been defined [2, 3]. Since the first three-dimensional model for soil organic matter (SOM) reported by Schulten and Schnitzer [4], an increasing number of SOM models have been developed such as the optimised dissolved organic matter model [5], the Leonardite humic acid model [6] and the Temple–-Northeastern–Birmingham humic acid model [7]. The progress made with these models has allowed subsequently to refine the chemical behaviour prediction of the interlayer structure of organo-clays by considering the molecular structure of both the SOM constituents and the minerals [8, 9]. Similarly, molecular models for individual hydrocarbon compounds and hydrocarbon fractions (group of hydrocarbons such as saturates, aromatic, olefins, resins, asphaltenes) of petroleum-derived products have been

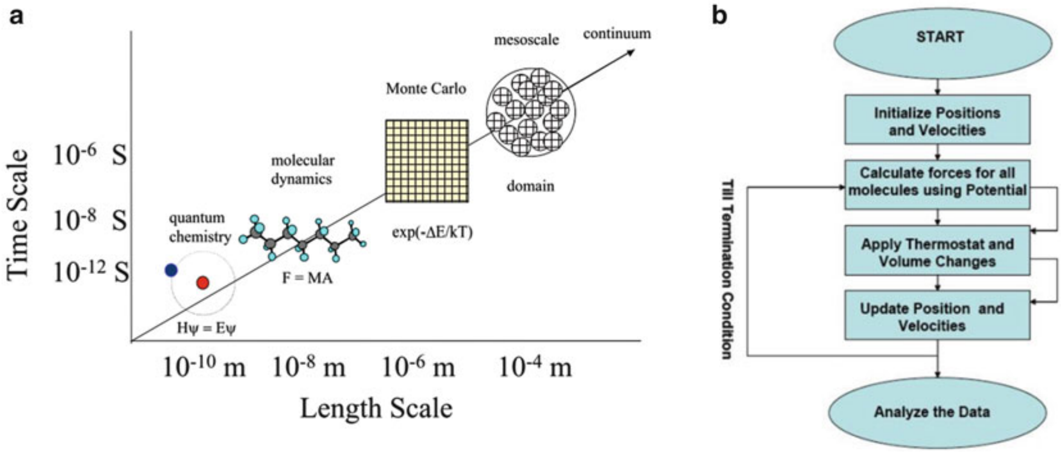

Fig. 1 Timescale (**a**) and steps (**b**) of molecular simulation methods

developed. For instance, the asphaltene molecules can be mimicked by the continental model [10] and the archipelago models [11], which have been validated by instrumental analysis (e.g. nuclear magnetic resonance, X-ray powder diffraction, Fourier-transformed infrared spectroscopy, thermogravimetric analysis) in terms of atom types, functional groups and molecular weight [12]. The development of these models makes possible to characterise the isotherm and thermodynamic parameters beyond the resolution of instrumental analysis such as the contribution of Coulomb electrostatic and van der Waals forces to the adsorption of oil contaminants on the soil mineral surface [1], the binding energy between SOM and oil pollutants [13] and the free energy for the penetration of polycyclic aromatic hydrocarbon through the bacterial membranes [14]. These physicochemical processes are highly important for bioavailability and biodegradation studies [15, 16].

However, the main challenge of using molecular simulation method is the timescale. The phenomena observed in the bioremediation process such as ageing often occurred in months or years; however, it is difficult to model such long timescale process by molecular simulation due to the extremely high computing data requirement. Our preliminary test indicated that it took about 1 month to simulate a 100 ns adsorption process of a system containing 28 asphaltene molecules and 20,000 water molecules using MD simulation on a server with 64 CPU cores (data not published). Accordingly, it would require 10^{12} months (computation time) to simulate such adsorption for only 1 day (real time). An alternative is to use fugacity modelling approach, which is a multimedia environmental fate model based on the thermodynamic theory of fugacity describing the bulk balance of a chemical substance in an ecosystem constituted by compartments [17]. The term ‘fugacity’ describes the ‘fleeing’ or ‘escaping’ tendency of a chemical species from an environmental compartment. Generally, the thermodynamic equilibrium includes chemical potential and fugacity. Chemical potential cannot be measured absolutely which is logarithmically related to the concentration of a compound [18]. By contrast, the major advantages of using fugacity are that (1) fugacity is equal to partial pressure under ideal conditions and is linearly related to concentration; (2) it is relatively convenient to transform the equations for chemical reaction, advective flow and non-diffusive transport rate into fugacity expressions with simple forms; and (3) it is easy to establish and solve sets of fugacity equations describing the complex behaviour of chemicals in multiphase and nonequilibrium environments. Various fugacity-based models have been developed such as generic models, air–water exchange models, soil–air exchange models, sediment–water exchange models, fish bioaccumulation models, sewage treatment plant models, indoor air models, plant uptake models, physiologically based pharmacokinetic models, multiple species bioaccumulation models, the EQuilibrium

Criterion model, the Level III ChemCAN model, the Level III CalTOX model, the QWASI (quantitative water, air, sediment interaction) model and the GloboPOP (chemical fate in the entire global environment) model [19]. These models have been previously applied for directing site remediation decisions [20], quantifying vapour emission from contaminated sites [21] and predicting the fate of organic compounds at landfill sites and constructed biopiles [22–24].

Both molecular simulation and fugacity modelling assume specific forms of mathematical equations, and statistical regression is only used to determine the unknown parameters in the equation. However, in some cases, the lack of knowledge on the bioremediation processes makes it difficult to accurately build these models due to the highly non-linear relationship between multiple variables. For example, there is not a universal form of equation that can be used to capture the relationship between ageing process and the bioavailability changes of different petroleum fractions in contaminated soils during composting processes. Instead machine learning is a data-based and assumption-free approach that distinguishes the data pattern and explores the relative influence of the independent variables on the dependent variables without relying on prior knowledge on the prediction process. The core of machine learning methods is allowing computers to extract general information from empirical data via 'learning' and 'recognising' the patterns of empirical data like a human brain. It has proven to be useful in modelling complex environmental processes, especially highly non-linear dynamic ecosystems such as the biosorption of metals [25–27], emissions of PAHs during combustion process [28] and bioavailability changes of PAHs in compost-amended soils [29].

This chapter provides the key steps and parameters needed when modelling the fate and transport of petroleum hydrocarbons using the above three modelling approaches. Their application can serve several purposes such as (1) improving our understanding of the chemical sources and their environmental fate and transport, (2) understanding the various factors that affect the environmental processes so that researchers can prioritise those factors that most need additional study, (3) providing tools to give a better mechanistic understanding of the degradation pattern and predict remediation end points and (4) enabling better risk-informed decisions during site remediation.

2 Molecular Dynamic Simulation

The protocol described below provides an example of MD simulation of the adsorption, the diffusion and the distribution of the different oil fractions on quartz surface (Fig. 2). The model construction, the energy minimisation and the molecular dynamic simulation can be

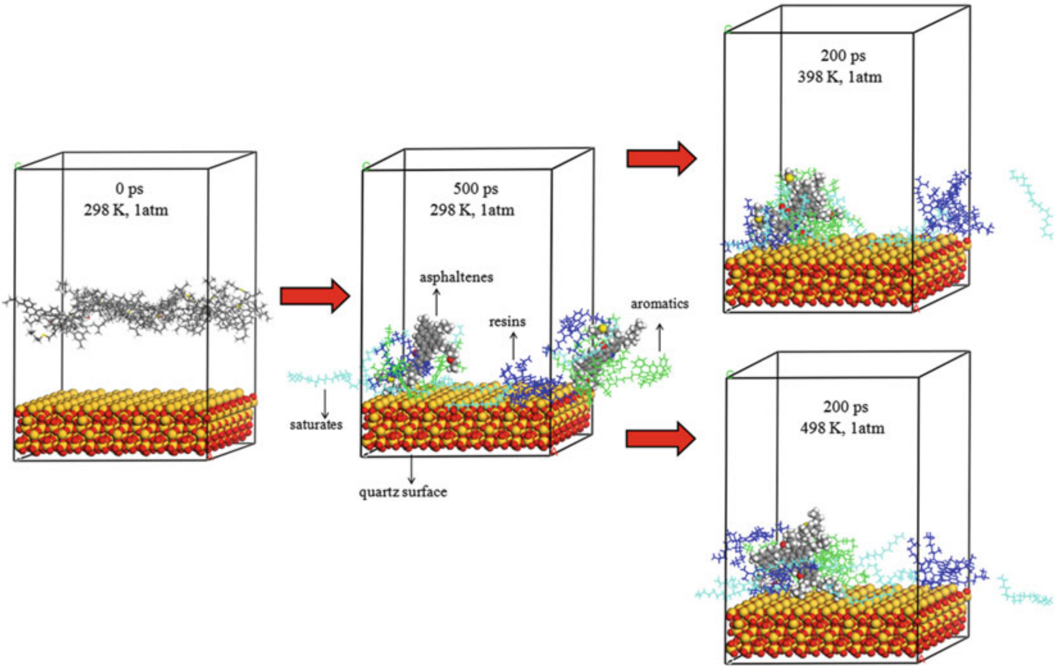

Fig. 2 Molecular modelling of the adsorption of asphaltenes, resins, aromatics and alkanes on quartz surface (adapted from [1])

carried out using a variety of software such as Materials Studio (http://accelrys.com/products/materials-studio/index.html), Gromacs (http://www.gromacs.org/), LAMMPs (http://lammps.sandia.gov/), DL_POLY (https://www.stfc.ac.uk/SCD/44516.aspx) and NAMD (http://www.ks.uiuc.edu/Research/namd/). Materials Studio was used in this protocol. Despite the differences in the program interface and computation performance among these programs, the modelling and simulation procedures are similar to that described in this protocol:

- Force field: select the condensed-phase optimised molecular potential for atomistic simulation studies (COMPASS) force field to describe the bonded and nonbonded potential of inter- and intra-molecule interactions [30] (*see* **Note** 1). Bonded potential includes quartic bond stretch, angle–bend contributions, torsion, out-of-plane angle and cross-coupling terms. Nonbonded potential consists of van der Waals interactions represented by the Lennard–Jones function and electrostatic interaction represented by the Coulombic equation (*see* **Note** 2).
- Oil layer modelling: select the following molecular compositions $C_{54}H_{65}NO_2S$ (MW: 792 g mol^{-1}) [31], $C_{50}H_{80}S$ (MW: 713 g mol^{-1}) [32], $C_{46}H_{50}S$ (MW: 635 g mol^{-1}) [33] and $C_{20}H_{42}$ (MW: 282 g mol^{-1}) to model the asphaltene, resin,

aromatic and saturate fractions of oil, respectively. Create an amorphous cell by packing these molecules with molecular number of 2, 4, 6 and 7, respectively, in a simulation box with 3D periodic boundary condition (*see* **Note** 3).

- Quartz surface modelling: create a quartz (1:1:1) surface using a quartz cell ($a = b = 0.4913$ nm, $c = 0.5405$ nm, $\alpha = \beta = 90°$, $\gamma = 120°$), build a 7 × 7 unit cell replica of the quartz surface and convert it into a 3D cell with 3D periodic boundary condition.
- Sorption system modelling: create a crystal-layered structure by adding the oil layer on the top of the quartz surface.
- Energy minimisation: use the smart minimiser approach to relax the sorption system and ensure that the system has no steric clashes or inappropriate geometry (*see* **Note** 4).
- NVT equilibration: perform an NVT MD simulation with constant number of atoms (N), volume (V) and temperature (T) to bring the system to the desired temperature for the simulation and to ensure the correct configuration of the orientation of the molecules. Assign the dynamic simulation parameters as follows: dynamics time (100 ps), time step (1 fs), temperature (298 K), thermostat (Andersen) and trajectory output (final structure) (*see* **Note** 5).
- NPT equilibration: perform an NPT MD simulation with constant number of atoms (N), pressure (P) and temperature (T) to stabilise the pressure and density of the system. Assign the dynamic simulation parameters as follows: dynamics time (100 ps), time step (1 fs), temperature (298 K), pressure (1 atm), thermostat (Andersen), barostat (Berendsen) and trajectory output (final structure).
- Production MD: perform an NVT MD simulation to the well-equilibrated system at the desired temperature and pressure. Parameters to be assigned include dynamics time (500 ps), time step (1 fs), temperature (298 K), thermostat (Andersen), trajectory output (full) and frequency of trajectory saved (every 5,000 steps).
- Data analysis: use the trajectory from the production MD simulation to analyse the mean square displacement, diffusion coefficient, root mean square deviations in structure, interaction energy, radial distribution functions and concentration profile of oil fractions on the quartz surface (*see* **Note** 6).

3 Fugacity Modelling

This protocol illustrates the application of fugacity models on a typical biopile with a volume of 624 m^3 and a mass of 750 t (Fig. 3).

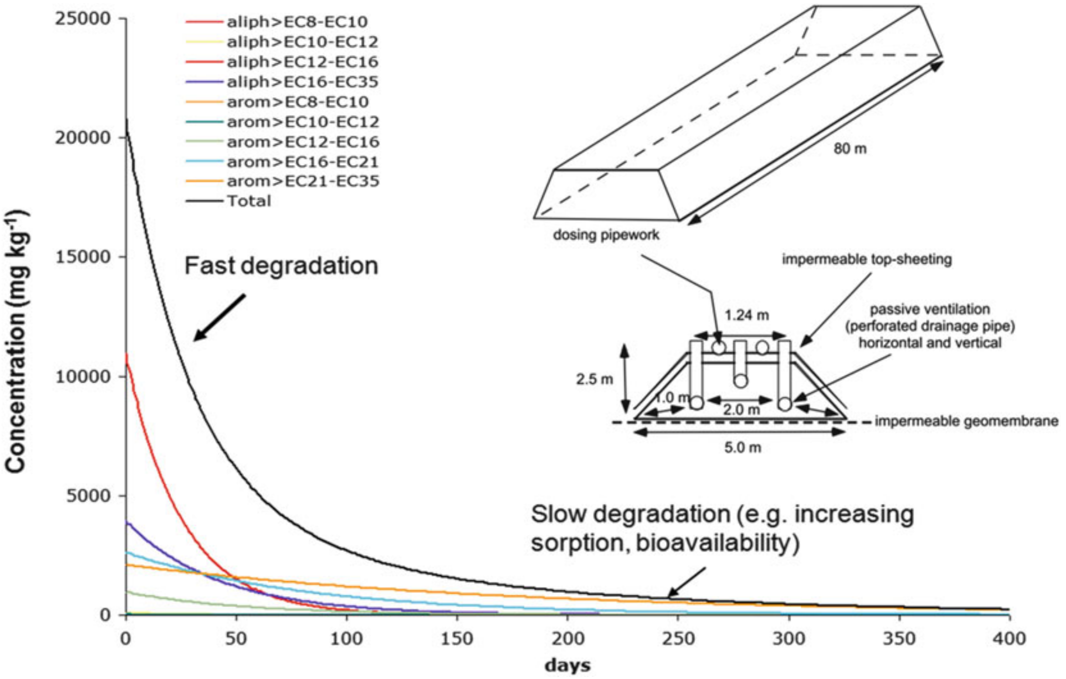

Fig. 3 Predicting the concentration of petroleum hydrocarbon fractions in contaminated soils during biopiling using fugacity modelling (adapted from [22, 23])

Level II fugacity model is used to calculate the steady-state and equilibrium distribution of petroleum hydrocarbon fractions in the contaminated soils. Level IV fugacity model is used to calculate the nonsteady-state chemical emission, advection, reaction, intermedia transport and residence time of petroleum hydrocarbons in the soil–oil matrix during biopiling. Level II and IV fugacity models can be easily coded in Microsoft Visual Basic for Applications (VBA) tool.

- Define four compartments in the biopile as follows: air, water, soil solids and nonaqueous phase liquid (NAPL). Calculate the volume of each compartment (an example of calculation is provided in the previous chapter ‘Protocol for Biopile Construction Treating Contaminated Soils with Petroleum Hydrocarbons’).
- Measure the physicochemical characteristics of soils (i.e. soil density, mass fraction of soil organic carbon) using standard methods [34]. Measure the total petroleum hydrocarbon (TPH) concentration in the soil (i.e. using the ultrasonic solvent extraction method described by Risdon et al. [35]). Assume that the measured TPH concentration is equal to the NAPL concentration.

- Divide the TPH into aliphatic and aromatic fractions. Divide each fraction into groups according to equivalent carbon number (*see* **Note** 7). Compile the physicochemical properties of petroleum hydrocarbon fractions (i.e. molecular weight, water solubility, vapour pressure, Henry's law constant, log K_{ow}, log K_{oc}, density) from books [19, 36].
- Calculate the half-lives of each petroleum hydrocarbon fraction in air, water and soil using the Estimation Program Interface Suite (http://www.srcinc.com). Assume the half-lives in NAPL to be equal to that in the bulk soil.
- Calculate the fugacity capacity (Z) for each compartment as follows:

Compartment	Fugacity capacities (mol m^{-3} Pa)	Constant definitions and units
Air	$Z_{\mathrm{AIR}} = 1/R{\cdot}T$	R = the gas constant (8.314 Pa m^3 mol^{-1} K^{-1}) T = temperature (K)
Water	$Z_{\mathrm{WATER}} = 1/H$	H = Henry's law constant (Pa m^3 mol^{-1})
NAPL	$Z_{\mathrm{NAPL}} = K_{ow}/H$	K_{ow} = octanol–water partition coefficient
Soil solids	$Z_{\mathrm{SOIL}} = K_{oc} * f_{\mathrm{OC}} * \rho_{\mathrm{SOLIDS}} * Z_{\mathrm{WATER}}/1{,}000$	K_{oc} = organic carbon partition coefficient (L kg^{-1}) ρ_{SOLIDS} = density of soil solids (kg m^{-3}) f_{OC} = mass fraction of soil organic carbon (g g^{-1})
Bulk soil (without NAPL phase)	$Z_{\mathrm{BULK}} = \phi_{\mathrm{AIR}} Z_{\mathrm{AIR}} + \phi_{\mathrm{WATER}} Z_{\mathrm{WATER}} + \phi_{\mathrm{SOIL}} Z_{\mathrm{SOIL}}$	ϕ_{AIR} = volume fraction of air (m^3) ϕ_{WATER} = volume fraction of water (m^3) ϕ_{SOIL} = volume fraction of soil solids (m^3)
Bulk soil	$Z_{\mathrm{T}} = \phi_{\mathrm{AIR}} Z_{\mathrm{AIR}} + \phi_{\mathrm{WATER}} Z_{\mathrm{WATER}} + \phi_{\mathrm{SOIL}} Z_{\mathrm{SOIL}} + \phi_{\mathrm{NAPL}} Z_{\mathrm{NAPL}}$	ϕ_{NAPL} = volume fraction of NAPL (m^3)

- Steady-state and equilibrium distribution: assume the fugacity is equal and constant in all compartments. Calculate the fugacity (f, Pa) and the concentration of each petroleum hydrocarbon fraction (C, mol m^{-3}) in each compartment using the total mass of TPH (M, mol), the fugacity capacity (Z, mol m^{-3} Pa^{-1}) and the volume of each compartment (V, m^3) as follows: $M = \sum V_i C_i = f \sum V_i Z_i$.

- Define the reaction and advection processes that occurred during biopiling. The former include the biodegradation in bulk soil and NAPL phase, which are simplified as first-order kinetic reaction (*see* **Note** 8). The latter processes include the advection in air, leaching from soil and volatilisation to the surrounding air (i.e. diffusion in soil–air, diffusion in soil–water and diffusion in the soil–air boundary).
- Calculate the D values (a transport parameter similar in principle to rate constant) for each process with the formula and parameters listed as follows (*see* **Note** 9):

Compartment process	Equation
Reaction in bulk soil	$D_R = k_R(V_A + V_S + V_W) \cdot Z_{BULK}$
Reaction in NAPL	$D_{R\_NAPL} = k_{R\_NAPL} \cdot V_{NAPL} \cdot Z_{NAPL}$
Advection in air	$D_A = G_{AIR} \cdot Z_{AIR}$
Leaching from soil	$D_L = G_L \cdot Z_W$
Volatilisation	$D_V = [(1/D_E) + 1/(D_A + D_W)]^{-1}$
Diffusion in soil–air	$D_A = A \cdot B_A \cdot Z_{AIR}/\Upsilon_D$
Diffusion in soil–water	$D_W = A \cdot B_W \cdot Z_{WATER}/\Upsilon_D$
Diffusion across the soil–air boundary	$D_E = A \cdot K_V \cdot Z_{AIR}$
Overall bulk soil D value	$D_T = D_R + D_A + D_V + D_L + D_{R\_NAPL}$

A is the surface area of biopile (m^2)*;* V_i is the volume of compartment i (m^3). A = air; S = soil solids; W = water. k_R is the first-order reaction rate constant for the fraction considered in the bulk soil, calculated as $\ln(2)/T_O$ (h^{-1}); k_{R_NAPL} is the first-order reaction rate constant for the NAPL (h^{-1}); G_{AIR} is the air flow rate through the biopile (6.24 m^3 h^{-1}); G_L is the water flow rate through the biopile (arbitrary rate of 2.1×10^{-3} m^3 h^{-1}); B_A is the effective diffusion coefficient for soil–air, derived from the Millington–Quirk equation (m^2 h^{-1}); B_W is the effective diffusion coefficient for soil–water, derived from the Millington–-Quirk equation (m^2 h^{-1}); Υ_D is the (non-tortuous) diffusion path length, set at the mean depth of the biopile (m); K_V is the partial mass transfer coefficient in the air side of the soil–air interface (m h^{-1}).

- Nonsteady-state dynamic fate: calculate the dynamic changes in fugacity (f_i) by solving the differential equation as follows –

$df_{\mathrm{i}}/dt = -D_{\mathrm{Ti}}f_{\mathrm{i}}/V_{\mathrm{T}}Z_{\mathrm{BULKi}}$. Then, predict the total concentration of all fractions in the bulk soil at each time step as follows: $C_{\mathrm{T}}(t) = \sum_{i=1}^{N} C_i(t)$.

4 Machine Learning Modelling

The protocol below described the application of ML models to predict the PAH bioavailability changes in contaminated soils during the bioremediation process via compost amendment (Fig. 4). Among the methods developed in ML, the artificial neural network (ANN)-based models, the tree-learning techniques, the rule-learning algorithms and linear regression are the most widely used and therefore are described in this protocol. This protocol provides example of using these models followed by parameterisation, optimisation and validation. It does not elaborate much on the computing theory of these models which is available from previous publications [29, 37]. All calculations were performed using the open-source software WEKA (http://www.cs.waikato.ac.nz/ml/weka/).

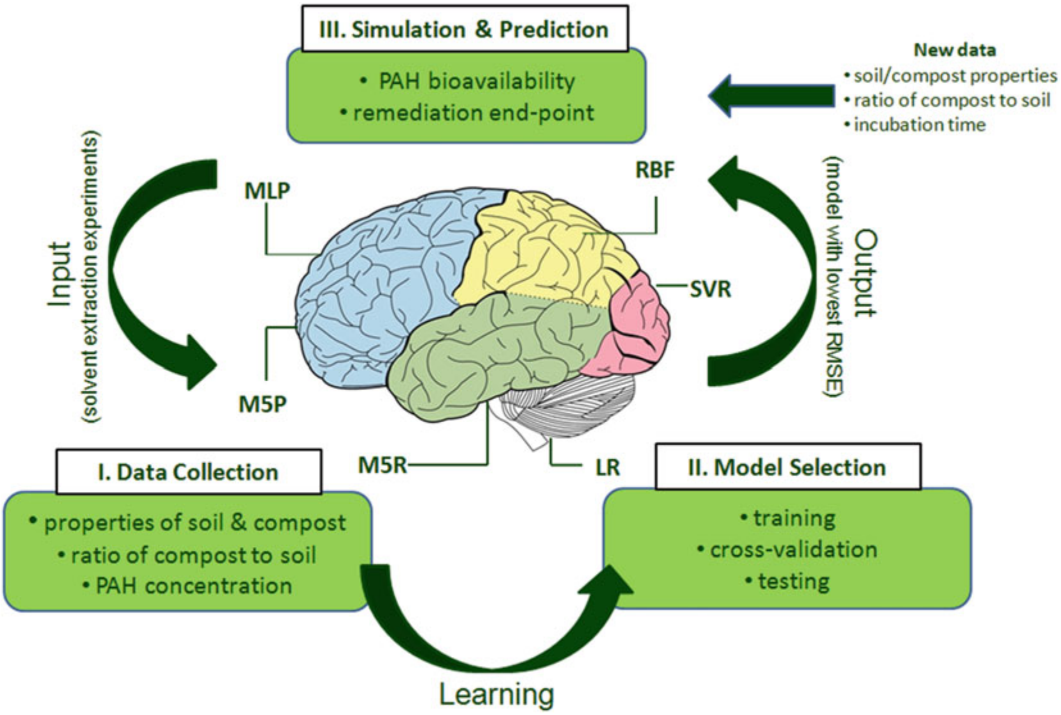

Fig. 4 Machine learning models to predict PAH bioavailability changes in soils after compost amendment. As demonstrated in the figure, the three steps included in machine learning are data collection, model selection and simulation and prediction. The input data was trained using six models including multilayer perceptron (MLP), radial basis function (RBF), support vector regression (SVR), M5 model tree (M5P), M5 model rules (M5R) and linear regression (LR). The model with the lowest root mean square error (RMSE) was finally selected for predicting PAH bioavailability and remediation end point by providing a new data set

- Incubation experiments: the data source for the ML modelling is obtained from laboratory experiments. In this example, two types of mature compost into three contaminated soils were considered, and the change in the bioavailability of the 16 US EPA priority PAHs was monitored [38]. The protocols for the physicochemical characterisation of the soils and the composts, the compost amendment regimes, the incubation time and the chemical analysis and quantification of the PAHs are available from Wu et al. [34].
- Data collection: collect the data on soil and compost properties (i.e. moisture, organic carbon of soil, total nitrogen, total phosphorus, available phosphorus, loss on ignition and soil texture), ratio of compost to soil (250 and 750 t ha^{-1}), initial concentration of total PAH, incubation time (i.e. 0, 3, 6 and 8 months) and the bioavailable concentration of PAHs at different incubation times as input parameters. Define the bioavailable concentration as dependent variable and the others as independent variables.
- Multilayer perceptron (MLP) training: build an MLP network and set the initial number of nodes in the input layer and output layer as 11 and 1, respectively. Fix the value of momentum term as 0.2. Use the CVParameterSelection module in the WEKA software to optimise two model parameters (i.e. the learning rate and the number of nodes in the hidden layer (initial value: 8–23)). The WEKA will output the root mean square errors (RMSE) and the hidden-input and hidden-output connection weights. Use the connection weights to calculate the relative contribution of input variables (i.e. soil and compost properties) to the predictive output (i.e. bioavailable concentration of PAHs) (*see* **Note** 10).
- Radial basis function (RBF) training: choose the k-means clustering algorithm to obtain RBF centres, select symmetric multivariate Gaussian function to fit the data and use CVParameterSelection module to optimise two parameters (i.e. the number of Gaussian radial basis functions (GRBFs) for the k-means algorithm and the minimum of standard deviation for the GRBFs) (*see* **Note** 11).
- Support vector regression (SVR) training: select the sequential minimal optimisation (SMO) algorithm to iteratively train the model and rearrange the non-linear data into linear data. The only parameter to optimise is the complexity parameter that determines the trade-off between the complexity of SMO model and the tolerance of errors. Use a correlation image to visualise the relationship between the input variables and the predictive output.
- M5 model tree (M5P), M5 model rules (M5R) and linear regression (LR) training: use the default parameters in the

WEKA software to train the models, which do not need optimisation.

- Cross validation: split the input data into ten groups randomly. For each model, use the instances from nine groups for model training, while use the remaining one group for model testing. Repeat this process ten times using a different group for testing at each cycle. The above validation method is termed as tenfold cross validation, which should be repeated ten times by re-splitting the data into ten groups. Average the root mean square error (RMSE) from the 100 (10 × 10) calculation to obtain the overall RMSE. Use RMSE to evaluate the performance of each model in predicting PAH bioavailability. The lower the RMSE, the better the goodness of fit.
- Simulation: overall 96 RMSE (16 PAHs × 6 models for each PAH) are obtained after training each model for each PAH. Select the best model (with the lowest RMSE) for each PAH as the final model. Use these well-trained models for predicting the bioavailable concentration of each PAH by inputting a new set of data that is not used in the model training or validation. For example, if we want to predict the bioavailable concentration after 12 months which was not measured in the incubation experiment, we just modify the variable of incubation time as 12, input it along with other variables (e.g. soil and compost properties) into the final models and run the models.

5 Notes

1. Force field is one of the most important mathematical functions to be selected prior to molecular dynamic simulation, which should be capable of predicting the structural, vibrational and thermophysical properties for substances in the model system. There are several options for force field such as COMPASS, CVFF, PCFF, GROMOS, CHARMM, AMBER, CLAYFF and OPLS-AA. Before application, the force field should be parameterised (e.g. equilibrium bond distances and angles, bond and angle force constants, dihedral angles, atom charges and Lennard–Jones parameters) via X-ray crystallography, vibrational spectra and quantum mechanics calculations. Alternatively, the force field can be empirically selected based on previous publications. For example, the COMPASS force field has been validated for a number of organic and inorganic molecules such as alkanes and benzene-fused ring compounds [30]. The CLAYFF [39] force field and the CHARMM force field optimised by Lopes et al. [40] can be effectively used for simulation of clay and silica, respectively. The GROMOS force field was initially designed for modelling biomolecules but has

been increasingly employed to model the petroleum-based system (e.g. asphaltenes and polycyclic aromatic hydrocarbons) after modification [10, 31, 41, 42].

2. The nonbonded interactions usually dominate simulations. We recommend using Gromacs (free software) for MD simulation as a lot of algorithmic optimisations have been introduced in the code, and thereby it is extremely fast at calculating the nonbonded interactions. The program is better run under Linux operating system with command line options for input and output files (although there is also a Windows version). For beginners not familiar with the Linux system, we recommend to use Materials Studio (commercial software) which has a more friendly user interface under Windows system.

3. The number of molecules in the MD simulation system should be decided on the molecular weight and concentration ratio of the oil fractions in the laboratory experiments. The periodical boundary condition (PBC) means that if any atoms leave a simulation box in one direction, then they will enter the simulation box from the opposite direction. The application of PBC can avoid problems with boundary effects caused by finite size and make the system more like an infinite one.

4. There are several algorithms to perform energy minimisation such as steepest descent, conjugate gradient and Newton–-Raphson. Convergence levels of iterations should be assigned for these algorithms. The Smart Minimizer is a module in the Materials Studio package which executes the above algorithms in sequence according to the changes in the energy potential of the model system. In this protocol, the convergence levels of iterations are set as 1,000, 10 and 0.1 kJ (mol Å)$^{-1}$ for the above three methods, respectively.

5. A full list of the parameters to set before running MD simulation can be found from the online manual of Gromacs software at http://manual.gromacs.org/online/mdp_opt.html.

6. The molecular simulation software offers a large selection of flexible tools for analysing the structural, statistic, dynamic and thermodynamic properties of the model system using the trajectories. A detail description of the data analysis principles can be found from Frenkel and Smit [43] and the online manual of Gromacs (http://www.gromacs.org/).

7. Equivalent carbon number is determined from boiling point or the retention time of the compounds in gas chromatography column, which is more appropriate than the actual carbon number of chemicals [44]. In addition to fractionating the petroleum oil into groups, specific petroleum hydrocarbons can also be selected as indicator compounds for fugacity modelling.

8. All the reaction processes are expressed in first-order form as a function of concentration. If a chemical is susceptible to several reactions in the same phase, the rate constants for the reaction are simply added to make the total rate constant. The strategy is to force first-order kinetics on systems by lumping parameters in the rate constant (i.e. express the second-order rate reactions in the form of pseudo first-order rate reaction to circumvent complex reaction rate equations). Such transformation will make the subsequent calculations much easier.

9. The general form for reaction and advection process is $D_R = k \cdot V \cdot Z$ (k, reaction rate constant, h^{-1}; V, volume, m^3) and $D_A = G \cdot Z$ (G, fluid flow rate, $m^3\ h^{-1}$), respectively.

10. The nodes in the input layer are the input-independent variables, while those in the hidden layer and output layer are processing elements with non-linear activation functions such as sigmoid function that enable the network to solve complex non-linear problems [45]. The number of nodes in the hidden layer should range from $(2n^{1/2}+m)$ to $(2n+1)$, where n and m are the number of nodes in the input layer and output layer, respectively [46]. Each node in one layer is connected with a certain weight to every node in the following layer. During model training, the predicted results are compared with the experimental results to compute the value of error. The connection weights are adjusted to reduce the error. This process is repeated for a sufficiently large number of cycles until the error is minimised.

11. RBF is another ANN model, which differs from the MLP in that it uses a special class of activation functions known as radial basis functions (e.g. Gaussian function) in the hidden layer.

References

1. Wu G, He L, Chen D (2013) Sorption and distribution of asphaltene, resin, aromatic and saturate fractions of heavy crude oil on quartz surface: molecular dynamic simulation. Chemosphere 92:1465–1471
2. Viani A, Gualtieri AF, Artioli G (2002) The nature of disorder in montmorillonite by simulation of X-ray powder patterns. Am Mineral 87:966–975
3. Teppen BJ, Rasmussen K, Bertsch PM, Miller DM, Schäfer L (1997) Molecular dynamics modeling of clay minerals. 1. Gibbsite, kaolinite, pyrophyllite, and beidellite. J Phys Chem B 101:1579–1587
4. Schulten HR, Schnitzer M (1995) Three-dimensional models for humic acids and soil organic matter. Naturwissenschaften 82:487–498
5. Sutton R, Sposito G, Diallo MS, Schulten HR (2005) Molecular simulation of a model of dissolved organic matter. Environ Toxicol Chem 24:1902–1911
6. Niederer C, Goss KU (2007) Quantum chemical modeling of humic acid/air equilibrium partitioning of organic vapors. Environ Sci Technol 41:3646–3652
7. Sein JLT, Varnum JM, Jansen SA (1999) Conformational modeling of a new building block of humic acid: approaches to the lowest energy conformer. Environ Sci Technol 33:546–552
8. Sutton R, Sposito G (2006) Molecular simulation of humic substance–Ca-montmorillonite complexes. Geochim Cosmochim Acta 70:3566–3581

9. Zeng QH, Yu AB, Lu GQ, Standish RK (2003) Molecular dynamics simulation of organic–inorganic nanocomposites: layering behavior and interlayer structure of organoclays. Chem Mater 15:4732–4738
10. Kuznicki T, Masliyah JH, Bhattacharjee S (2009) Aggregation and partitioning of model asphaltenes at toluene-water interfaces: molecular dynamics simulations. Energy Fuel 23:5027–5035
11. Alshareef AH, Scherer A, Tan X, Azyat K, Stryker JM, Tykwinski RR, Gray MR (2012) Effect of chemical structure on the cracking and coking of archipelago model compounds representative of asphaltenes. Energy Fuel 26:1828–1843
12. Sjöblom J, Simon S, Xu Z (2015) Model molecules mimicking asphaltenes. Adv Colloid Interface Sci 218:1–16
13. Wu G, Zhu X, Ji H, Chen D (2015) Molecular modeling of interactions between heavy crude oil and the soil organic matter coated quartz surface. Chemosphere 119:242–249
14. Ren B, Gao H, Cao Y, Jia L (2015) In silico understanding of the cyclodextrin–phenanthrene hybrid assemblies in both aqueous medium and bacterial membranes. J Hazard Mater 285:148–156
15. Semple KT, Morriss AWJ, Paton GI (2003) Bioavailability of hydrophobic organic contaminants in soils: fundamental concepts and techniques for analysis. Eur J Soil Sci 54:809–818
16. Reid BJ, Jones KC, Semple KT (2000) Bioavailability of persistent organic pollutants in soils and sediments – a perspective on mechanisms, consequences and assessment. Environ Pollut 108:103–112
17. Bru R, Maria Carrasco J, Costa Paraíba L (1998) Unsteady state fugacity model by a dynamic control system. Appl Math Model 22:485–494
18. Lewis GN (1901) The law of physico-chemical change. Proc Am Acad Arts Sci 37:49–69
19. Mackay D (2001) Multimedia environmental models: the fugacity approach. Lewis, Chelsea
20. Pollard SJT, Hoffmann RE, Hrudey SE (1993) Screening of risk management options for abandoned wood-preserving plant sites in Alberta, Canada. Can J Civ Eng 20:787–800
21. Mills WJ, Bennett ER, Schmidt CE, Thibodeaux LJ (2004) Obtaining quantitative vapor emissions estimates of polychlorinated biphenyls and other semivolatile organic compounds from contaminated sites. Environ Toxicol Chem 23:2457–2464
22. Coulon F, Whelan MJ, Paton GI, Semple KT, Villa R, Pollard SJT (2010) Multimedia fate of petroleum hydrocarbons in the soil: oil matrix of constructed biopiles. Chemosphere 81:1454–1462
23. Pollard SJT, Hough RL, Kim KH, Bellarby J, Paton G, Semple KT, Coulon F (2008) Fugacity modelling to predict the distribution of organic contaminants in the soil:oil matrix of constructed biopiles. Chemosphere 71:1432–1439
24. Shafi S, Sweetman A, Hough RL, Smith R, Rosevear A, Pollard SJT (2006) Evaluating fugacity models for trace components in landfill gas. Environ Pollut 144:1013–1023
25. Turan NG, Mesci B, Ozgonenel O (2011) Artificial neural network (ANN) approach for modeling Zn (II) adsorption from leachate using a new biosorbent. Chem Eng J 173:98–105
26. Giri A, Patel R, Mahapatra S (2011) Artificial neural network (ANN) approach for modelling of arsenic (III) biosorption from aqueous solution by living cells of *Bacillus cereus* biomass. Chem Eng J 178:15–25
27. Turan NG, Mesci B, Ozgonenel O (2011) The use of artificial neural networks (ANN) for modeling of adsorption of Cu (II) from industrial leachate by pumice. Chem Eng J 171:1091–1097
28. Inal F (2006) Artificial neural network predictions of polycyclic aromatic hydrocarbon formation in premixed n-heptane flames. Fuel Process Technol 87:1031–1036
29. Wu G, Kechavarzi C, Li X, Wu S, Pollard SJ, Sui H, Coulon F (2013) Machine learning models for predicting PAHs bioavailability in compost amended soils. Chem Eng J 223:747–754
30. Sun H (1998) COMPASS: an ab initio forcefield optimized for condensed-phase applications overview with details on alkane and benzene compounds. J Phys Chem B 102:7338–7364
31. Kuznicki T, Masliyah JH, Bhattacharjee S (2008) Molecular dynamics study of model molecules resembling asphaltene-like structures in aqueous organic solvent systems. Energy Fuel 22:2379–2389
32. Murgich J, Rodrı́guez J, Aray Y (1996) Molecular recognition and molecular mechanics of micelles of some model asphaltenes and resins. Energy Fuel 10:68–76
33. Verstraete J, Schnongs P, Dulot H, Hudebine D (2010) Molecular reconstruction of heavy petroleum residue fractions. Chem Eng Sci 65:304–312
34. Wu G, Kechavarzi C, Li X, Sui H, Pollard SJT, Coulon F (2013) Influence of mature compost

amendment on total and bioavailable polycyclic aromatic hydrocarbons in contaminated soils. Chemosphere 90:2240–2246

35. Risdon GC, Pollard SJT, Brassington KJ, McEwan JN, Paton GI, Semple KT, Coulon F (2008) Development of an analytical procedure for weathered hydrocarbon contaminated soils within a UK risk-based framework. Anal Chem 80:7090–7096
36. TPHCWG (1998) Total petroleum hydrocarbon criteria working group series volume 2: composition of petroleum mixtures. Amherst Scientific, Amherst
37. Mitchell T (1997) Machine learning. McGraw Hill, New York
38. USEPA (1989) Method 610-Polynuclear Aromatic Hydrocarbons, methods for organic chemical analysis of municipal and industrial wastewater. US Environmental Protection Agency, Washington, DC
39. Cygan RT, Liang JJ, Kalinichev AG (2004) Molecular models of hydroxide, oxyhydroxide, and clay phases and the development of a general force field. J Phys Chem B 108:1255–1266
40. Lopes PEM, Murashov V, Tazi M, Demchuk E, MacKerell AD (2006) Development of an empirical force field for silica. Application to the quartz-water interface. J Phys Chem B 110:2782–2792
41. Jian C, Tang T, Bhattacharjee S (2013) Probing the effect of side-chain length on the aggregation of a model asphaltene using molecular dynamics simulations. Energy Fuel 27:2057–2067
42. Bandela A, Chinta JP, Hinge VK, Dikundwar AG, Row TNG, Rao CP (2011) Recognition of polycyclic aromatic hydrocarbons and their derivatives by the 1,3-dinaphthalimide conjugate of calix[4]arene: emission, absorption, crystal structures, and computational studies. J Org Chem 76:1742–1750
43. Frenkel D, Smit B (2002) Understanding molecular simulation: from algorithms to applications. Academic, San Diego
44. Brown DG, Knightes CD, Peters CA (1999) Risk assessment for polycyclic aromatic hydrocarbon NAPLs using component fractions. Environ Sci Technol 33:4357–4363
45. Haykin S (1999) Neural networks: a comprehensive foundation. Prentice Hall, New York
46. Fletcher D, Goss E (1993) Forecasting with neural networks. An application using bankruptcy data. Inf Manag 24:159–167