Master Thesis

Modeling of pesticide biodegradation in soil

Author(s):
Loos, Martin Jürgen

Publication Date:
2010

Permanent Link:
https://doi.org/10.3929/ethz-a-006121570

Rights / License:
In Copyright - Non-Commercial Use Permitted
Modeling of Pesticide Biodegradation in Soil

Masters Thesis

of

Martin Loos

eawag Dübendorf, Switzerland

July 27, 2010

Supervisors:
Dr. Kathrin Fenner
Dr. Martin Krauss

ETH Zürich
Department of Environmental Sciences
Institute of Biogeochemistry and Pollutant Dynamics

\[\text{martin.loos@alumni.uni-potsdam.de}\]
Abstract

Quantitative Biodegradability-Structure Relationships (QSBRs) are valuable tools for estimating pesticide half-lives in soil. Advance in QSBR modeling is based on several prerequisites: (1) separation of (bio)degradation from formation of non-extractable soil residues (NER), (2) degradation data from a homogeneous soil source, (3) robust predictor selection, (4) proper control of model complexity and (5) validation of model generalization performance on data external to model construction. This study addresses (1) to (5) in two steps.

In step one, first order rate constants of both primary degradation and NER formation were calculated from inverse model optimization by Genetic Algorithms, based on 87 Speyer 2.2 soil degradation data sets. Degradation was found to be faster than NER formation for most of the compounds, although both processes were correlated.

In a second step, variation in degradation rate constants was modeled by LASSO shrinkage regression and, in addition, by a novel robust filtering technique combined with forward selection, termed Recursive Bootstrap Subsampling (RBS). RBS and shrinkage regression were applied to four predictor sets of increasing complexity and validated by (a) internal cross-validation (CV), (b) external CV and (c) external CV under proper control of model complexity. As a result, (a) overestimated generalization performance for all models and predictor sets, whereas (b) underestimated the performance achievable with a given data set. (c), on the other hand, gave more reasonable estimates. Moreover, shrinkage regression could outperform RBS under setting (b). However, when (c) control of model complexity was assured and the number \( p \) of predictors outnumbered that of the samples \( n \) (\( n << p \) problem), RBS outperformed shrinkage regression. Therefore, only RBS was utilized to assess the importance of molecular and non-molecular predictors for explaining variation in degradation rates. Of a total of 670 candidate predictors, only few were important. The most relevant ones were: 3 collinear fragments containing sulfur, 2 collinear ester functionalities, 3 collinear N-heteroaromatic moieties, nitrogen with double or triple bonds, a methyl and a methylene descriptor, double bonded oxygen and water-organic carbon partitioning coefficients.

It is concluded that a combination of robust predictor selection, external validation and control of model complexity other than internal CV is essential for reliable QSBR modeling.
# Contents

1. **Introduction** 6

2. **Methods** 9

   2.1 Pesticide fate in soil: the data source 9

   2.2 Estimation of kinetic rate parameters 10
      
      2.2.1 Inverse modeling 10
      2.2.2 Genetic Algorithm optimization 13
      2.2.3 Model evaluation 14

   2.3 Statistical modeling of kinetic rate parameters 15
      
      2.3.1 Filtering method: Recursive Bootstrap Subsampling 16
      2.3.2 Shrinkage method: LASSO regression 17
      2.3.3 Hybrid: Recursive Bootstrap Subsampling & LASSO 19
      2.3.4 Model evaluation 19
      2.3.5 Model validation 20
      2.3.6 Candidate predictors 21
      2.3.7 Predictor importance 25

3. **Results and Discussion** 27

   3.1 Kinetic rate parameters from inverse modeling 27

   3.2 Statistical modeling of kinetic degradation parameters 36
      
      3.2.1 Model validation 36
      3.2.2 Model evaluation 42
      3.2.3 Predictor importance 50

4. **Conclusion** 58

5. **Acknowledgments** 60

6. **Appendix A - List of pesticides and metabolites** 61

7. **Appendix B - List of important predictors** 73

8. **Appendix C - Statistical model coefficients** 77

9. **Appendix D - RK4 C++ code** 79

References 89
List of Figures

1. Box model representation of the four pool inverse modeling approach. .......................... 11
2. Box model representation of the five pool inverse modeling approach. .......................... 12
3. LAR solution path to the LASSO coefficients. ................................................................. 19
4. Kinetic parameters $k_{PM}$ and $k_{PN}$ from duplicate data sets. ................................. 28
5. Observed and modeled dissipation of Amidosulfuron. ................................................... 30
6. Residuals over time for the four-box model series. ......................................................... 31
7. Residuals over time for the five-box model series. ......................................................... 32
8. Inverse modeling results for kinetic rate constants of compound transformation to metabolite(s) and to NER. ................................................................. 33
9. LOO cross-validated $R^2$ as a function of model complexity. ........................................ 38
10. Result of the second iteration from recursive bootstrap subsampling (RBS). ..................... 40
11. Observed versus predicted values for the pseudovalidated RBS&forward stepwise model. 43
12. Observed versus predicted values for the tuned-validated RBS&forward stepwise model. 44
13. Pseudovalidated RBS model residues. ................................................................. 47
14. RBS model residues from tuned cross validation. ......................................................... 48
15. LOO/RBS-based predictor importance for modeling $(k_{PM}, k_{M1M2})$. ................................................. 51
16. Correlation between $(k_{PM}, k_{M1M2})$ and important predictors as a function of mean RBS rank of a predictor. ................................................................. 52
17. Scatterplot of measured and calculated log$(K_{oc})$ partitioning constants. ......................... 53

List of Tables

1. Evaluation of model fit for the four box and five box models. ........................................ 29
2. Validation of coefficients of determination ($R^2$) for predicting $k_{PM}$. .......................... 37
3. Performance measures for the various pseudovalidated modeling approaches on predicting $(k_{PM}, k_{M1M2})$. ................................................................. 45
4. Performance measures for various modeling approaches on predicting $(k_{PM}, k_{M1M2})$ for tuned LOO validation. ................................................................. 46
5. List of Speyer 2.2 pesticides and metabolites. ................................................................. 61
6. List of important predictors. ................................................................. 73
Nomenclature

AIC  Akaikes Information Criterion
AR   Applied Radioactivity
bias mean Median bias below the sample mean
bias mean Median bias above the sample mean
CV   Cross Validation
DAR  Draft assessment report
DLL  Dynamically Linked Library
GA   Genetic Algorithm
GETAWAY  Geometry, Topology, and Atom-Weights AssemblY predictors
k    1st order rate constant
kM1M2  1st order rate constant for degradation of metabolite to downstream metabolite(s)
kM1N  1st order rate constant for NER formation from first downstream metabolite
kM2N  1st order rate constant for NER formation from metabolites downstream of first metabolite
kM2V  1st order rate constant for formation of volatiles from metabolites downstream of first metabolite
kMN  1st order rate constant for NER formation from metabolite(s)
kMV  1st order rate constant for formation of volatiles from metabolite(s)
kPM1  1st order rate constant for degradation of parent compound to first downstream metabolite
kPM  1st order rate constant for degradation of parent compound to downstream metabolite(s)
kPN  1st order rate constant for NER formation from parent compound
LASSO Least Absolute Shrinkage and Selection Operator
LMO  Leave-Many-Out
LOO  Leave-One-Out
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSC</td>
<td>Liquid Scintillation Counting</td>
</tr>
<tr>
<td>M(t)</td>
<td>Total metabolite activity as a function of time</td>
</tr>
<tr>
<td>M1(t)</td>
<td>Activity of first metabolite downstream of parent compound as a function of time</td>
</tr>
<tr>
<td>M2(t)</td>
<td>Activity of total metabolites downstream of first metabolite as a function of time</td>
</tr>
<tr>
<td>MAD</td>
<td>Median Absolute Deviation</td>
</tr>
<tr>
<td>MSC</td>
<td>Model Selection Criterion</td>
</tr>
<tr>
<td>MWHC</td>
<td>Maximum Water Holding Capacity</td>
</tr>
<tr>
<td>N(t)</td>
<td>NER activity as a function of time</td>
</tr>
<tr>
<td>NER</td>
<td>Non-Extractable soil Residues</td>
</tr>
<tr>
<td>ODE</td>
<td>Ordinary Differential Equation</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary Least Squares</td>
</tr>
<tr>
<td>P(t)</td>
<td>Parent compound activity as a function of time</td>
</tr>
<tr>
<td>QSAR</td>
<td>Quantitative Structure-Activity Relationships</td>
</tr>
<tr>
<td>QSBR</td>
<td>Quantitative Structure-Biodegradability Relationships</td>
</tr>
<tr>
<td>r_spear</td>
<td>Spearman's rank correlation coefficient</td>
</tr>
<tr>
<td>RBS</td>
<td>Recursive bootstrap subsampling</td>
</tr>
<tr>
<td>RK4</td>
<td>4th order Runge Kutta method</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square prediction Error</td>
</tr>
<tr>
<td>sd</td>
<td>standard deviation</td>
</tr>
<tr>
<td>t_{1/2}</td>
<td>Half life</td>
</tr>
<tr>
<td>UM-PPS</td>
<td>University of Minnesota Pathway Prediction System</td>
</tr>
<tr>
<td>V(t)</td>
<td>Activity of volatiles as a function of time</td>
</tr>
<tr>
<td>WHIM</td>
<td>Weighted Holistic Invariant Molecular descriptors</td>
</tr>
</tbody>
</table>
1 Introduction

Structure-Biodegradability Relationships (SBRs) help identifying persistent organic compounds and are therefore of interest for exposure modeling [91] and for regulation of chemical production and use [55]. For the latter, SBRs have potentials to cut rates of down animal testing and to decrease costs of screening for pollutants of high concern [101]. On the one hand, SBR models allow predictions to external data, provided that their generalization performance has been assessed and that predictions are made for chemicals of the same domain used to build the SBR model. Thus, half lives $t_{1/2}$ can be estimated for existing chemicals or for prospective chemicals prior to synthesis [16]. On the other hand, the model building process permits valuable insights on the factors ruling the persistence of a pesticide, thus serving as explorative tool. Manifold preconditions need to be met to ensure sound model building and subsequent SBR performance.

First of all, the definition of $t_{1/2}$ must be rendered more precise. Pesticides can dissipate in at least two major ways: (1) they may truly degrade to transformation products (in the case of biodegradation: metabolites) or (2) they can form non-extractable soil residues (NER) [56][12][66]. It is questionable if the latter should be regarded as an effective pesticide dissipation step due to the fact that NER-formation may, at least on the long term, be reversible [87]. Some kinetic models for calculating $t_{1/2}$ do not distinguish degradation from NER-formation, for example those of the FOCUS work group [39], effectively assigning (1) and (2) to the same sink. On the other hand, some regulatory endpoints and/or data sources only focus on degradation [14], not on residue formation. In any case, and despite indications that (1) and (2) may be interdependent [66][56], one must distinguish both processes for SBR model construction. Logically, SBR models dealing with (1) and (2) separately are more accurate than those lumping (1) and (2), whether as predictive or as explorative tool. Henceforth, $t_{1/2}$ calculated from both (1) and (2) jointly is termed $DT_{50}$, whereas $DegT_{50}$ if referring to (1) only [66].

Another potential refinement concerns the type of model output. The output is qualitative when classifying measures of persistence, for example into “ready biodegradable” and “not ready biodegradable” [54][8]. Such binning is not entirely objective with regard to the classification threshold and eventually disposes of quantitative information about the precise values of $DT_{50}$ or $DegT_{50}$, although such values are essential inputs to multi-media fate models [55]. In short, quantitative SBRs (QSBRs) should be preferred over classification models whenever the data allows to.

No model building process whatsoever can compensate for missing or noised input information. Degradation of xenobiotics does not only depend on chemical structure alone, but also on the conditions of the environmental compartment in which degradation takes place. For example, degradation may be transport-limited in soil and sediment but much less so in well-mixed aqueous environments [36][28]. However, the soil compartment is of utmost importance for pes-
1 Introduction

ticide degradation, with degradation depending largely on microbial processes \cite{58,11,55}. Still, these two constrains do not dispel variability in degradation rates among different soils, attributable to heterogeneities in e.g. clay composition \cite{22}, microbial community \cite{62}, matrix potential \cite{28} or organic matter constitution \cite{76}. The various interactions between soil heterogeneities and molecular properties complicate matters even further. For example, interactions between individual soil organic matter constituents and pesticides depend on molecular properties \cite{17,12}, as do microbial strategies of (co)metabolism \cite{94,11}. The precise nature and the effect of such variabilities are often not resolved and can therefore not be accounted for when building a QSBR. As a consequence, mixing of degradation data from different soils can noise up molecular information that may otherwise prove relevant when derived from a homogeneous soil. In the following, homogeneity refers to data derived from one soil site only; this does not imply that variation among soil samples taken from this single soil site is eliminated. It rather implies a minimization of heterogeneities among sample properties. One can expect an ensemble of predictions from several QSBRs based on a homogeneous data set each to be more precise and meaningful than a QSBR based on the pooled data, provided that enough homogeneous data sets are available. Since homogeneous soil data sets on pesticide degradation are sparse, recent QSBRs still rely on mixing of data sources (e.g. \cite{74}).

Sparse data results in small data sets. This shortage in sample count $n$ is mismatched by the number $p$ of candidate predictors which can today be calculated for (Q)SBR model identification. To give an impression: the DRAGON software \cite{1} supports calculation of up to 1630 theoretical descriptors, some of which have found application in biodegradation modeling \cite{58,104,80}. This imbalance of $n << p$ necessitates reduced weighting or removal of predictors likely to be unassociated with degradation rates, thus improving model interpretability. Predictor selection can be based on prior knowledge or on statistical methods. Unfortunately, prior expert knowledge is subjective and thus insufficient to tackle the complex task of selection alone. As reviewed by \cite{80,55}, common statistical selection methods are univariate correlation/significance statistics, forward selection, cluster analysis, genetic algorithm search, and others. These selection processes are specific to the $n$ training sets and are not per se robust against (i) selection instability from small changes in the training set composition and (ii) selection uncertainty with regard to the collinearity of predictors \cite{61}. However, simultaneous incorporation of (i) and (ii) may be crucial when prediction to new data outside the training set is required and should therefore be incorporated into the process of model identification. To date, a thorough combination of (i) and (ii) has been neglected in QSR building and subsequent model evaluation. In the following, predictor selection incorporating aspects (i) and (ii) will be termed robust, in contrast to non-robust selection in the absence of (i) and (ii).

Furthermore, the $n << p$ setting is especially susceptible towards overfitting. That is, any predictor selection biased toward the specificity of the training data degenerates generalization performance, namely poor prediction to exter-
nal data from an overly complex model. Generalization performance of a QSBR model must be assessed via external test samples or, under sparseness of data, via cross-validation (CV) \[96, 60\]. A widespread pitfall of the latter technique is that the CV folds are not fully independent of all stages of model building and hence misleading (internal CV) \[15, 10\]. But, even under the approaches conducting CV correctly (external CV), the influence of QSBR model complexity on generalization performance is not well documented - especially not for a comparison of different modeling approaches. This leads to the following working hypotheses:

- One can expect that missing control of model complexity results in CV underestimation of the generalization performance of a training data set, even if CV is conducted correctly, i.e. externally.

- One can furthermore hypothesize that this tendency increases with the size \(p\) of the candidate predictor set relative to \(n\), as increase in \(p\) facilitates overfitting.

- Most importantly, it is hypothesized that robust predictor selection is less prone to produce overfitted models than non-robust predictor selection under constantly increasing \(p\), even if model complexity is adequately controlled for.

The last hypothesis would complement observations of \[5\], where zero CV prediction performance for too complex models was reported for gene array data, which typically show a \(n << p\) setting. In fact, this very last hypothesis has so far not been quantified within a CV-assessment of QSBR modeling (cp. \[15\]), if not even in statistical modeling in general.

To embrace the above prerequisites of sound modeling and to check for the above three hypothesis, a number of objectives are proposed. Firstly, quantitative \(\text{Deg}T_{50}\) data from a homogeneous soil source are to be assembled and shall be used for constructing a QSBR. Secondly, QSBR model building shall incorporate predictor selection using different selection techniques and predictor subsets of varying size. At least one of these techniques shall be robust towards the above aspects (i) and (ii). Thirdly, the model building process shall be embedded in an external CV estimate of generalization performance under regulation of model complexity.
2 Methods

A two-staged procedure was applied to pursue the above outlined objectives. In a first stage, homogeneous pesticide soil degradation data was collected from pesticide degradation reports (section 2.1). The assembled degradation time series were used to calculate kinetic rate parameters via inverse modeling (section 2.2). At a second stage, variance in kinetic rate parameters of compound metabolization was related to molecular and non-molecular properties within a statistical modeling framework (section 2.3). The two stages are outlined below. Modeling was executed in the R statistical environment [40], with some external function calls to customized C++ code.

2.1 Pesticide fate in soil: the data source

The European Food Safety Authority (EFSA) lists a large body of Draft Assessment Reports (DARs) on the (re)registration of substances used in plant protection. These DARs are publicly available [6] and report aerobic soil transformation/degradation time series for pesticides in different soils under controlled laboratory conditions. Although a wide range of soils were used for different compounds or by different contract laboratories, one specific loamy sand has been evaluated in a number of DARs (M. Krauss, p.c.). This one standard soil, namely Speyer 2.2 soil, allows investigating soil pesticide degradation on a homogeneous soil data basis. A second data source was provided by BAYER Crop Science (Monheim, Germany) in terms of unpublished Speyer 2.2 soil pesticide degradation reports. Speyer 2.2 degradation time series were collected from these sources and funneled into the modeling procedure presented. More information on the Speyer 2.2 soil is provided in section 2.3.6.

The Speyer 2.2 degradation reports mostly conform to OECD guideline 307 [73], and were executed in the following manner. The $^{14}$C-labeled compound was mixed with 50 to 200 g (dry weight) of sieved soil, with application concentrations reflecting recommendations from the individual pesticide use instructions. Soils were then placed in the dark in incubation flasks at temperatures of 20 to 25 °C and with soil water pressure adjusted to between 2.0 to 2.5 pF. At time intervals, soil replicates were removed from incubation and analyzed for (1) the distribution of initially applied radioactivity (AR) among soil constituents and (2) a mass balance of AR over time. With regard to (1), AR was itemized for (a) extractable AR of the parent compound initially applied, (b) extractable AR of transformation products (metabolites) of the parent compound, (c) non-extractable residues (NER) and (d) volatile compounds. (a) and (b) were extracted with different solvents, depending on the individual reports. In some cases, the structure of individual metabolites and hence the underlying reaction pathways were identified by chromatography and UV-, NMR-spectroscopy ans or MS-spectrometry. In cases where a report confirmed and itemized a single metabolite resulting via a sole reaction pathway from the parent compound, its
2.2 Estimation of kinetic rate parameters

Sets of autonomous ordinary differential equations (ODE) were established to model rates of change of the various soil AR pools over time. The differential equations were solved numerically (cp. section 2.2.1) and their rate parameters fitted to the individual time series by Genetic Algorithm (GA) optimization (cp. section 2.2.2).

2.2.1 Inverse modeling

In analogy to the inverse modeling approach of [66], temporal changes in the concentration (alias AR) of (1) the parent compound $P$, (2) the metabolite pool $M$, (3) the NER pool $N$ and (4) the volatiles $V$ were each described by the first order kinetic ODEs 2.2.1 to 2.2.4.

$$\frac{dP}{dt} = -k_{PM} * P(t) - k_{PN} * P(t)$$ (2.2.1)

$$\frac{dN}{dt} = k_{PN} * P(t) + k_{MN} * M(t)$$ (2.2.2)

$$\frac{dM}{dt} = k_{PM} * P(t) - k_{MN} * M(t) - k_{MV} * M(t)$$ (2.2.3)

$$\frac{dV}{dt} = k_{MV} * M(t)$$ (2.2.4)

with $k_{PM}$, $k_{PN}$, $k_{MN}$ and $k_{MV}$ being the first order rate constants for conversion from parent compound to metabolite, from parent compound to NER, from metabolite(s) to NER and from metabolite(s) to volatiles, respectively. The structure of this four pool model is depicted in Figure 1. It allows to distinguish NER formation from either metabolites or the parent compound, with NER formation assumed irreversible. All metabolites were subsumed into a single pool, as were their individual kinetic rate constants of conversion to
2.2 Estimation of kinetic rate parameters

NER or volatiles. First order kinetic was applied because compound application concentrations were assumed to be too low to trigger second order kinetic from e.g. bacterial growth \[^{[39]}\]. Given a single first order kinetic parameter \(k\), degradation half lives \(t_{1/2}\) can be calculated from

\[
t_{1/2} = \frac{\ln(2)}{k}
\]  

(2.2.5)

For those 14 reports with a single itemized downstream metabolite \(M_1\) the four pool model was modified for comprising an individual AR pool for \(M_1\), upstream of the other metabolites \(M_2\) (Figure 2). The underlying ODEs of this resulting five pool model read as

\[
\frac{dP}{dt} = -k_{PM1} \cdot P(t) - k_{PN} \cdot P(t)
\]  

(2.2.6)

\[
\frac{dN}{dt} = k_{PN} \cdot P(t) + k_{M1N} \cdot M1(t) + k_{M2N} \cdot M2(t)
\]  

(2.2.7)

\[
\frac{dM1}{dt} = k_{PM1} \cdot P(t) - k_{M1N} \cdot M1(t) - k_{M1M2} \cdot M1(t)
\]  

(2.2.8)

\[
\frac{dM2}{dt} = k_{M1M2} \cdot M1(t) - k_{M2N} \cdot M2(t) - k_{M2V} \cdot M2(t)
\]  

(2.2.9)

\[
\frac{dV}{dt} = k_{M2V} \cdot M2(t)
\]  

(2.2.10)

with first order rate constants \(k\) as defined in Figure 2. Both the conceptual four and the five pool model allow identifying \(k_{PM}\) and \(k_{PN}\). They do not per se allow an accurate estimation of \(k_{M1N}\) or \(k_{M2N}\) because the parent compound substructure containing the label position may not be present in some metabolites after late transformation steps. This circumstance may not be valid for \(k_{M1N}\), since transformations from \(P(t)\) to \(M1(t)\) never concerned the molecular substructures carrying the radioactive label. \(N(t)\) may consist of metabolites and/or parent compound. The setup of equations 2.2.1 to 2.2.10 allows distinguishing disappearance via NER formation from degradation to metabolites for

---

Figure 1: Box model representation of the four pool inverse modeling approach. Boxes represent pools of parent compound \(P(t)\), total metabolite \(M(t)\), NER \(N(t)\) and volatiles \(V(t)\) as a function of time. Arrows represent 1st order reaction or exchange paths among these pools.
2.2 Estimation of kinetic rate parameters

Figure 2: Box model representation of the five pool inverse modeling approach. Boxes represent pools of parent compound \( P(t) \), first downstream metabolite \( M_1(t) \), total metabolites \( M_2(t) \) downstream of the first metabolite, NER \( N(t) \) and volatiles \( V(t) \) as a function of time. Arrows represent 1st order reaction or exchange paths among the pools.

the parent compound \( P(t) \). Neither of the models embrace direct mineralization from \( P(t) \) to \( V(t) \approx [CO_2](t) \). Furthermore, none of the models separate for dissolved vs. adsorbed parent compounds and metabolites. Instead, it was assumed that sorption equilibration was fast relative to transformation or NER formation rates.

Integration of equations 2.2.1 to 2.2.4 and 2.2.6 to 2.2.10 is required to relate to the observed time series for AR pools (1) to (4). For example, take a simple equation of NER formation from a single parent compound only, i.e.

\[
\frac{dN}{dt} = k_{PN} * P(t) \quad (2.2.11)
\]

from

\[
\frac{dP}{dt} = -k_{PN} * P(t) - k_{PM} * P(t) = f(P(t), t) \quad (2.2.12)
\]

Here, analytical integration of equation \(2.2.11\) leads to

\[
N(t) = \frac{(1 - e^{-(k_{PM}+k_{PN})t}) * k_{PN} * P(0)}{k_{PM} + k_{PN}} \quad (2.2.13)
\]

For \( t \to \infty \), this equation reduces to

\[
N(\infty) = \frac{k_{PN} * P(0)}{k_{PM} + k_{PN}} \quad (2.2.14)
\]

However, several nested exponential terms form part of the solutions for integrating equations such as \(2.2.10\) containing several upstream compounds. Intermediate results from these terms can exceed computational storage capabilities for single numbers for high values of \( t \). Therefore, a numerical technique was
2.2 Estimation of kinetic rate parameters

used for solving the ODEs of equations 2.2.1 to 2.2.4 and 2.2.6 to 2.2.10, namely the 4th order Runge Kutta (RK4) method \[78\]. The RK4 method replaces the derivatives of the ODEs by finite forward difference approximations. Starting from (a) \( t = 0 \) and (b) \( P(t = 0) = \) total AR, solutions to any \( t \neq 0 \) are iterated by (in this case) fixed time steps of \( \Delta t \). RK4 shall be exemplified for solving equation 2.2.12: the value of any \( P(t + \Delta t) \) can be calculated from its predecessor \( P(t) \) by

\[
P(t + \Delta t) = P(t) + \frac{1}{6} \Delta t \times (s_1 + 2s_2 + 2s_3 + s_4)
\]  (2.2.15)

The values \( s_1 \) to \( s_4 \) represent slope section estimates along the interval of \( \Delta t \) and are calculated from

\[
s_1 = f(P(t), t)
\]  (2.2.16)

\[
s_2 = f(P(t) + \frac{1}{2} \Delta t \times s_1, t + \frac{1}{2} \Delta t)
\]  (2.2.17)

\[
s_3 = f(P(t) + \frac{1}{2} \Delta t \times s_2, t + \frac{1}{2} \Delta t)
\]  (2.2.18)

\[
s_4 = f(P(t) + \Delta t \times s_3, t + \Delta t)
\]  (2.2.19)

To ensure numerical stability in this explicit approach, \( \Delta t \) was set to a small value of 0.025 days. Time intervals at which AR measurements were taken were rounded to the first decimal place. Values of \( N(0), M(0), V(0), M1(0) \) and \( M2(0) \) were set to 0, \( P(0) \) was set to the total initial AR (see GA optimization below). For fast calculation, C++ code of RK4 was implemented, converted to a dynamically linked library (DLL) and used in the R computational environment via an external function call. The C++ code is shown in Appendix D.

The first order kinetic approach requires that \( P(t = \infty) = 0 \) for any amount of NER formation or transformation to metabolites > 0. In contradiction to this, a few of the data sets showed AR of \( P(t) \) leveling out at nonzero values at late time intervals. This suggests extractable pools of \( P(t) \) not being subject to transformation to \( N(t) \) or \( M(t) \), possibly due to decreasing soil bioactivity. Since such behavior at late time intervals can not be modeled by first order kinetics \[39\], some of the late time intervals of the concerned time series where omitted and GA fitting repeated. If this procedure did not improve the fit or if less than 5 time intervals of AR measurements remained, the time series was excluded from further investigation.

2.2.2 Genetic Algorithm optimization

A Genetic Algorithm (GA) was run to optimize values of the kinetic rate parameters \( k \) and \( P(0) \) in the above numerical setups. Optimization was done on minimizing the squared residuals between modeled and observed values. Candidate solutions were restricted to ranges of 0 to 3 \([\text{days}^{-1}]\) for all \( k \) unless fitted values reached the upper boundary of \( k = 3 \,[\text{days}^{-1}] \). In the latter case, the
upper boundary was extended and the GA procedure repeated until the concerned \( k \) undermatched the upper boundary. \( P(0) \) was restricted to in between 80 to 120\% of AR \([60]\)). It must be emphasized that \( P(0) \) states the total AR modeled in a time series.

Details of GA optimization are explained in \([63]\) and \([64]\). In short, GA mimics an evolutionary process. Therein, the parameters to be optimized are represented by a chromosome \( i \), being equal to a string \( c_{i,j} \) containing the values to be optimized, i.e. \( c_{i,j}(N(0)|k_{PM}|k_{PN}|k_{MN}|k_{MV}) \) or \( c_{i,j}(N(0)|k_{PM1}|k_{PN}|k_{M1N}|k_{M1M2}|k_{M2N}|k_{M2V}) \). The space of candidate solutions is then searched by a total number of \( n \) (with \( 1 \leq i \leq n \)) such chromosomes over a total number of \( g \) generations. At each generation \( j \) (with \( 1 \leq j \leq g \)), each \( c_{i,j} \) yields a fitness value (the reciprocal of the sum of squared residuals).

GAs then contain two phases for each generation \( j \). In an initial exploitation phase, strings are randomly paired. Thereupon, in an intermediate exploration phase, strings are recombined (i.e., some of the string values are exchanged among paired strings) and allowed to mutate (random changes are introduced to the string values) with probability \( 0 \leq m \leq 1 \). Afterwards, strings are replaced in a final exploitation phase: only a certain number (the elitism) of fittest strings are allowed to proceed to the next generation \( j + 1 \), where they replace the less fit strings omitted. Having reached \( g \), a single string \( c_{i,g} \) with the best fitness was chosen to represent the solution for fitting the box model predictions to the data. Setups were \( n = 1000 \), \( m = 0.15 \), elitism = 20\% and \( g = 500 \).

GA involves stochastic elements. For this reason, high accuracy but somewhat lowered precision of GA fits has been suggested: GA may robustly find global optima but may be weaker in specifying their precise values. However, differences in \( k \) and \( P(0) \) from repeated GA runs for one compound time series were much smaller than differences among time series of different compounds. Moreover, GA outcomes for \( k \) and \( P(0) \) correlated well with outcomes of the locally more precise quasi-Newton BFGS method \([71]\) (data not shown; based on a preliminary study on a subset of degradation time series for which optimization via BFGS was feasible).

Genetic optimization was run under the R genalg package \([105]\), combined with customized R code for visualization of the optimization progress.

### 2.2.3 Model evaluation

The goodness of model fit was calculated by the coefficient of determination \( (R^2) \), which relates the amount of variance explained by model predictions \( \hat{y}_i \) to the total variance in the observed data \( y_i \).

\[
R^2_{CV} = 1 - \frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{k}(y_i - \bar{y})^2} \tag{2.2.20}
\]

with \( \bar{y} \) being the mean of the observed \( y_i \).

The box models of section 2.2.1 are a simplification of reality; turning from less to more complex models can sometimes improve model fit. Such improvement
was investigated from fitting data sets with a single first metabolite by both four- and five-pool models. Improvement was then captured by the Model Selection Criterion (MSC) \[66\]:

\[
MSC = \ln \left( \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \right) - \left( \frac{2z}{n} \right)
\]  

(2.2.21)

Akin to Akaikes Information Criterion (AIC), MSC corrects explained variance by a theoretical estimate for the overfit produced by the number \(z\) of model parameters \[4\]. MSC allows comparing fit to the same data for different models; the higher the MSC, the better.

To check for gross deviation from homoscedasticity, residues of model fit were plotted against time across all data sets, separately for the four- and the five-box model.

Finally, relations among kinetic rate parameters \(k\) and between \(k\) and \(P(0)\) were investigated from Spearman’s rank correlation coefficients \(r_{\text{spear}}\) and from scatterplots produced for \(k\) values of (1) parent to metabolite transformation and of (2) NER formation from the parent compound.

### 2.3 Statistical modeling of kinetic rate parameters

A statistical modeling approach was deployed to indicate molecular and non-molecular predictors that may influence soil pesticide degradation. Herein, variation in the rate parameters \(k\) from GA optimization was modeled by four different sets of candidate predictors and by three different models. In turn, each model and each predictor set was subjected to a stringent validation assay.

Lacking detailed evidence for the true nature of the interdependence between a possibly complex set of predictors and the model response, a parsimonious model structure was employed, namely ordinary least squares regression neglecting predictor interactions \[13\]. \(\hat{y}\) being the vector of predicted response values, \(X\) being the matrix of predictor values (with value 1 in the first column for the regression intercept) and \(\hat{\beta}\) the vector of coefficient estimates for each predictor column in \(X\), the model is described by

\[
\hat{y} = X\hat{\beta}
\]  

(2.3.1)

Without a priori knowing the precise relation between predictors and a degradation mechanism, one must consider this linear model structure as a mere approximation of truth \[94\]. On the other hand, linear model structures are parsimonious, simple to interpret and widely used for (Q)SBR modeling \[16\].

The large number \(p\) of candidate predictors, as listed in section 2.3.6, invokes a problem of high dimensionality. This is often referred to as “large \(p\), small \(n\) problem” (\(n = \) sample size), where \(p\) may even outnumber \(n\). Including all \(p\) predictors in a full model entails several drawbacks: (1) \(X\) may be rank deficient, i.e. several adequate solutions for \(\hat{\beta}\) may exist, (2) the high number of possibly redundant predictors degrades the interpretability of the model and (3)
2.3 Statistical modeling of kinetic rate parameters

the model may contain a large number of predictors not related to the response. Aspect (3) introduces noise into the model and may lead to high prediction variance and model overfitting [38].

As summarized by [44, 51, 96], two general strategies can tackle this “large p, small n problem”: (1) feature extraction and (2) feature selection. Strategy (1) makes usage of derived inputs from dimensionality reduction of the original p predictors. For example, the response may be regressed on the principal components of a subset of $p_{\text{max}}$ most important predictors, with $p_{\text{max}}$ chosen by cross validation (supervised principal component regression [7]). Since the number of predictors in $p_{\text{max}}$ may still be of $p_{\text{max}} > n$, this approach does not necessarily lead to a sparse set of important predictors only. However, excluding unimportant predictors is desirable given that the costs and efforts to generate some non-theoretical (non in-silico) predictors are high. Therefore, the methods described in the following employ the second strategy (feature selection), and not the above first one (feature extraction). Feature selection searches for an informative subset $p_{\text{max}}$ of $p < n$ predictors, omitting predictors likely to be unrelated to the response. Such selection was either conducted by (1) robust filtering for important predictors and a subsequent stepwise predictor search, (2) via shrinkage methods or (3) a hybridization of (1) filtering and (2) shrinkage.

2.3.1 Filtering method: Recursive Bootstrap Subsampling

A novel filtering methodology is being introduced to select for informative predictors. In this methodology, linear regression was conducted on a number of bootstrap samples of the original data set, choosing the predictor with highest correlation from random subsets of predictors. For the same predictor subsets, linear regression was also applied to the residues of that first regression, as in one-step forward stagewise selection [96]. In this way, predictors were ranked by their selection frequency in any of the two regression steps. Predictors of lowermost ranks were eventually discarded and a repetition of the procedure was started on the remaining predictors. The full procedure is detailed in algorithm 1, termed recursive bootstrap subsampling (RBS). Once the predictor subset had diminished to a small size of $p_{\text{sub}} \leq 30$, a stepwise forward predictor selection was applied on the $p_{\text{sub}}$ for a final linear model and results in a maximal numbers of predictors $p_{\text{max}}$ selected into the final model. This forward selection was done with the R LARS-package [30], based on cross-validation. 

$m$ was set to 20000 repetitions to ensure sufficient sampling within the predictor space; size $q$ was randomly varied from 3 to 20.

Step 1 restrains instability of the search. Steps 2 and 3 allow competition among less important variables even in the presence of few very important (dominant) predictors in the data set; this ensures exploration of the full predictor space [92]. In addition, step 2 and 4 allow selecting predictors after the influence of more important predictors has been eliminated. Step 5 does not route the absolute relevance a predictor has for explaining the response, making the search more robust. Step 6 leads to an execution of steps 1 to 5 at constant $m$ on a decreasing size of $p_{\text{sub}}$. In such a way, RBS finally imparts higher coverage on
the predictor subspaces of successively updated higher importance. Being a novel approach, RBS awaits further validation. Confirmingly, however, RBS bears similarity with elements of commonly applied machine learning methods. For example, step 2 is applied in both Random Forest [20] and stochastic gradient boosting [41]; step 1 is applied in bagging [96]. RBS is also similar to approaches described by [86] for variable subset optimization. Step 6, discarding predictors of low relevance, is comparable to recursive feature elimination [83] and aggressive feature selection [65]. RBS must not be mixed with subsampling bootstrap [45].

The bootstrapping in step 1 may bias for selection of balanced categorical predictors instead of very unbalanced ones in steps 3 and 4. In fact, consequences of unbalanced bootstrapping for predictor importance calculation have long been documented for unbalanced responses [13], and recently for unbalanced predictors, too [93]. Further research may clarify whether this is actually of disadvantage for RBS predictor set optimization.

2.3.2 Shrinkage method: LASSO regression

The above approach of RBS & forward predictor selection is a discrete process: coefficients of all predictors except those in the final model are set to zero, i.e. dropped. Alternatively, a more continuous selection process such as shrinkage regression can be applied. According to [96], shrinkage selection can reduce model variance for a slight increase in model bias, improving overall prediction accuracy. Shrinkage regression was accomplished with the Least Absolute Shrinkage and Selection Operator (LASSO) [97], applicable to “large p, small n
2.3 Statistical modeling of kinetic rate parameters

In LASSO regression, $\hat{\beta}$ of equation 2.3.1 is estimated by

$$\hat{\beta}(\lambda) = \text{argmin}_{\beta}(n^{-1} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|)$$ (2.3.2)

In words, $\hat{\beta}$ is estimated by minimizing the usual sum of squared residuals $(Y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2$ over all predictors $p$, but with an additional penalization of the model coefficients via $\lambda \sum_{j=1}^{p} |\beta_j|$. $\lambda \geq 0$ is a penalty / complexity parameter, selected here by means of cross-validation. $\lambda = 0$ yields the OLS estimate for $\hat{\beta}$. $\lambda > 0$ can set some of the $\beta_j$ to zero and shrinks the values of the other $\beta_j$, thus conducting a smoothed form of predictors selection. The way LASSO selects variables can be emulated by the so called Least Angle Regression (LAR) algorithm that can calculate the LASSO solution path of $\hat{\beta}$ as a function of $\lambda$ (there is no closed form expression for calculating the LASSO $\hat{\beta}$). At the first LAR step, the predictor $p_1$ which has the highest correlation with the response is selected into the model, using a downsized (i.e. penalized) coefficient. Then, the value of that coefficient is stepwise moved towards its OLS estimate (decreasing $\lambda$), until a second predictor $p_2$ correlates as strongly with the residuals as the first predictor $p_1$. At this point, $p_2$ is selected into the active model and the coefficients of both predictors are increased towards their joint OLS estimate until a third predictor picks up, and so on. When a coefficient hits zero, it is removed from the active set and all remaining coefficients are recomputed. For LASSO LAR, at most $p_{max}$ predictors can join the active set. The LASSO LAR selection path is exemplified in Figure 3.

Owing to the shrinkage, LASSO is more reluctant towards overfitting than the more aggressive forward stepwise selection. Another benefit reported is that the sequence $p_{max}$ of selected predictors is somewhat more stable for shrinkage selection than for discrete forward selection with regard to small changes in the data set. Another aspect is collinearity in the data set. Two correlated predictors may be selected into a model in a stepwise, non-penalized mode. Then, an overly large negative coefficient on one variable may be compensated by an overly large positive coefficient on the other. Shrinkage alleviates this misleading effect: LASSO tends to select only one predictor from a group of strongly correlated predictors, discarding the others. In contrast, the more recent shrinkage technique of Zou & Hastie, namely Elastic Net (EN) regression, allows to select (1) groups of correlated predictors together without reversion in sign and (2) more than $n$ predictors. Since (1) and (2) were not assumed major advantages here, the simpler LASSO technique was preferred. LASSO was calculated with the R glmnet package, which is based on the coordinate descent method of Friedman et al.
2.3 Statistical modeling of kinetic rate parameters

2.3.3 Hybrid: Recursive Bootstrap Subsampling & LASSO

A third approach to tackle the “large $p$, small $n$ problem” combines predictor filtering and shrinkage. That is, predictors were RBS filtered for their effect on the response as explained in section 2.3.1. Afterwards, LASSO regression was run on the RBS-reduced predictor subset $p_{sub} \leq 30$.

2.3.4 Model evaluation

A total of six indices was utilized to assess the precision and accuracy of the model predictions [13]. Firstly, the Root Mean Square Error (RMSE) and its standard error were calculated. As a measure of precision, RMSE calculates the overall squared deviation between observed and predicted values of the Box-Cox transformed kinetic parameters $k$:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1}}$$  \hspace{1cm} (2.3.3)
2.3 Statistical modeling of kinetic rate parameters

with \( y_i \) being an observation of the \( n \) samples and \( \hat{y}_i \) being the predicted value of this observation. The standard error of the RMSE simply follows from the variance of the residuals \( y_i - \hat{y}_i \). Secondly, the Median Absolute Deviation (MAD) gives a robust estimate of the model accuracy; an increase in difference between RMSE and MAD hints at increased variance of the prediction errors. However, MAD does not detail the direction of any systematic prediction errors.

\[
MAD = \text{median}(|y_i - \hat{y}_i|)
\]  

(2.3.4)

Complementing MAD, the bias robustly estimates the direction of any systematic deviations in model predictions:

\[
bias = \text{median}(y_i - \hat{y}_i)
\]  

(2.3.5)

Finally, one additional measure was shaped to monitoring the usage of the shrinkage methods outlined in section 2.3.2. While trading bias for variance \[96\], model predictions are systematically forged towards the mean value of the sample population. This may imply that predicted values above the sample mean \( \bar{y} \) may tend to underestimate the response, whereas predictions below the mean may tend towards overestimation. Proposed indices of such deviation, \( \text{bias}_{>\text{mean}} \) and \( \text{bias}_{<\text{mean}} \), can be derived from a modification of the above described bias:

\[
bias_{<\text{mean}} = \text{median}(y_i - \hat{y}_i) \forall y_i < \bar{y}
\]  

(2.3.6)

and

\[
bias_{>\text{mean}} = \text{median}(y_i - \hat{y}_i) \forall y_i > \bar{y}
\]  

(2.3.7)

In other words, \( \text{bias}_{>\text{mean}} \) and \( \text{bias}_{<\text{mean}} \) summarize the bias above and below the sample population mean, respectively.

Finally, the measure \( R^2 \) defined in equation 2.2.3 was calculated, too.

2.3.5 Model validation

None of the indices of section 2.3.4 approximate (1) the expected prediction error for an independent data set (i.e. the model generalization performance) and (2) instability of model construction induced by changes in the data set \[96\][85]. For example, a high \( R^2 \) may indicate a good fit for the modeled samples. However, if this fit results from severe overfitting, prediction performance for independent data can still be low and the model therefore worthless \[48\]. Thus, (2) is directly related to (1), given that instability in model construction from changes in the data set can be expected to increase with increased overfitting, because overfitting reflects the specificity of a model towards a given data set.

Aspect (1) can be estimated utilizing independent data test sets. Unfortunately, the given data set of \( (k_{PM}, k_{M1M2}) \) is too small for being split into training-, validation and test sets. Hence, cross validation was conducted to assess aspect (1) within the following three settings.
Pseudovalidation (internal CV) RBS&forward stepwise (sections 2.3.1), shrinkage (section 2.3.2) and RBS&shrinkage (section 2.3.3) all utilize cross-validation (CV) to establish the final model. In a first step, the full data set or $p_{sub}$ (after RBS) is used to define either the full $\lambda$-sequence for shrinkage (cp. Figure 3) or the sequence of predictors under forward selection. In a second step, CV is conducted to yield the optimal number $p_{max}$ of predictors or the optimal coefficient value under varying $\lambda$. To this end, repeated 10-fold CV was used [60]. However, due to the fact that the $\lambda$-sequence and the sequence of selected predictors are installed based on the full data set $n$ in the first step, the CV cannot guarantee fully unbiased estimates of generalization performance [96]. For this reason, the described approach must be termed pseudovalidation (= internal CV validation).

External LOO CV Opposing pseudovalidation, a more rigorous and profound validation was applied in a second approach, in analogy to [10]. Therein, CV folds were omitted prior to any model construction steps. For instance, RBS and subsequent forward selection or shrinkage were done for data outside a CV fold, and the model performance evaluated for this fold only (= external CV validation). After repetition of this procedure over all folds, model performance was investigated by averaging outcomes of the indices of section 2.3.4 for the folds omitted from model constructions. To make efficient usage of the small data set, leave-one-out (LOO) CV was chosen for this purpose. Notably, the model construction outside the CV fold was still based on the above mentioned pseudovalidation.

Tuned external LOO CV External LOO validation gives reliable estimates of the above aspect (1), but does not regulate the complexity of the underlying pseudovalidated models outside the individual folds. Therefore, the above external LOO validation was repeated over iteratively increased model complexity, i.e. over iteratively increased values of the penalty parameter $\lambda$ (LASSO and RBS-LASSO-Hybrid) or the number of predictors $p_{max}$ (forward selection after RBS) in the underlying models outside the individual folds. Finally, the $\lambda$ or $p_{max}$ with the best external LOO CV performance over all folds was selected. This third external validation is henceforth called tuned LOO validation, because it controls the complexity the model outside a fold can attain for maximum predictive performance.

2.3.6 Candidate predictors

A wide variety of candidate predictors was considered eligible for explaining variance in the pooled response variable $k_{PM}, k_{M1M2}$. These predictors were arranged in four sets. The first set contained predictors from the EPI Suite QSAR package [3]. The second predictor set extended the first set by additional functional groups and physicochemical properties. In turn, the third set extended the second one by a selection of Speyer 2.2 soil properties and incuba-
2.3 Statistical modeling of kinetic rate parameters

The last predictor set was upgraded by geometrical and topological molecular properties. This nested approach allowed to validate model performance on predictor sets of increasing size and complexity. Moreover, the order of inclusion prioritized (1) predictors yet established and validated for compound biodegradation (set 1) and (2) molecular fragments over geometrical and topological predictors.

Fragment predictors with less than 5 fragment count classes (i.e., each class/level representing a specific count $0 \leq b < 5$ of the given fragment for the given compound) were treated as ordered categorical predictors, the other fragment predictors as continuous count data. Highly unbalance categorical predictors of less than 5 entries over all classes were excluded. The levels of the remaining categorical predictors were lumped to contain at least 5 entries per class, starting from the highest count class. In addition, binary presence/absence predictors were appended for all (i.e., both categorical and continuous) fragment predictors. Finally, all continuous predictors and the response were Box-Cox transformed \[18\] and normalized. For the Box-Cox transformation, the Box-Cox transformation parameter $\eta$ was iteratively increased and then set to the value that showed highest Pearson correlation between the axis variables of the corresponding normal probability plot. In this way, $\eta$ was found to be $-0.16$ for the Box-Cox transform of $(k_{PM}, k_{M_1 M_2})$, i.e. box$(k_{PM}, k_{M_1 M_2})$. For this small value of $\eta$, box$(k_{PM}, k_{M_1 M_2})$ is somewhat close to a log$_{10}$ transform of $(k_{PM}, k_{M_1 M_2})$. Missing predictor values were replaced by the mode (categorical data) or the median (continuous data) of the predictor. The four predictor sets are detailed below. Details on individual predictors are presented in Appendix B for predictors found important for explaining $(k_{PM}, k_{M_1 M_2})$.

Predictors set 1: BIOWIN fragments

The first predictor set consisted of 38 predictors which are included in the EPI suite BIOWIN1 to BIOWIN7 models \[3\] used for aerobic and anaerobic biodegradation modeling. One of the predictors was molecular mass, the others were fragment predictors. The BIOWIN predictors constitute a standard set established for QSAR modeling \[16\], against which additional fragment predictors shall be updated.

Predictors set 2: plus fragments and partitioning constants

Predictor set 2 updates set 1 by

1. The molecular fragments available from the KOCWIN model of the EPI Suite package \[3\],
2. 23 degradation rules from the University of Minnesota Pathway Prediction System (UM-PPS) \[32\] \[35\] triggered by the given compounds,
3. A dummy variable indicating if any of these UM-PPS rules was triggered or not,
4. The presence/absence predictors for nitrogen and sulfur in a molecule,
2.3 Statistical modeling of kinetic rate parameters

5. 12 fragment counts calculated from the ChemAxon Geometry Calculator  

6. 32 functional groups derived from the DRAGON group count descriptors  

7. 79 Ghose-Crippen atom-centered fragments of DRAGON  

8. Organic carbon-water partitioning coefficients from data included in the DARs  

9. Two organic carbon-water partitioning estimates from the KOCWIN model of the EPI Suite package [3], namely $K_{oc,MCI}$ [68] and $K_{oc,log(K_{ow})}$ [29], 

10. Molar volume and polarizability from the ChemAxon JChem Calculator  

11. Net molecular charge estimates (i.e. negative, positive, neutral) for pH 5, 6 and 7 (ChemAxon Charge Calculator  

With regard to (8), two types of organic carbon-water partitioning data are available in the DARs [72]. On the one hand, linear organic carbon normalized adsorption coefficients $K_{oc}$ [cm$^3$ g$^{-1}$] are listed as

$$K_{oc} = \frac{m_{aq}^{ads}(eq)}{m_{soil}} \frac{V_0}{m_{aq}^{ads(eq)} \%_{oc}} \times 100 \quad (2.3.8)$$

with $m_{aq}^{ads(eq)}$ and $m_{aq}^{ads(eq)}$ equal to the equilibrated compound mass in solution and adsorbed to soil [µg], respectively. $V_o = $ aqueous phase volume [cm$^3$], $m_{soil} = $ dry soil mass [g] and $\%_{oc} = $ mass percentage of organic carbon in the soil [g g$^{-1}$]. $K_{oc}$ was utilized as predictors whenever available from a DAR, which was not always the case. Instead, many DARs list nonlinear Freundlich adsorption coefficients $K_{F,oc}$ [cm$^3$ g$^{-1}$], calculated as

$$K_{F,oc} = \frac{m_{aq}^{ads(eq)}}{m_{soil}} \left( \frac{V_0}{m_{aq}^{ads(eq)}} \right)^{\frac{1}{f}} \frac{100}{\%_{oc}} \quad (2.3.9)$$

$f$ is the Freundlich regression constant describing sorption nonlinearity; that is, decreasing sorption for increasing $\frac{m_{aq}^{ads(eq)}}{V_0}$. $f$ was fitted in the DARs from the adsorption data. With $f$ typically ranging in between 0.7-1.0 [72], the effect of $f$ on $K_{F,oc}$ as a function of application concentrations was assumed small compared to overall differences in ($K_{oc},K_{F,oc}$) among the investigated compounds. To be precise, the latter are ranging in between 2.7 and 145,600.0 [cm$^3$ g$^{-1}$]. Therefore, $K_{F,oc}$ was set equal to $K_{oc}$ (i.e. $f = 1$) whenever data on the latter were missing in a compound DAR. In this manner, both mean and median $K_{oc}$ were assembled for the partitioning data available in the DAR of a compound.
2.3 Statistical modeling of kinetic rate parameters

Predictors set 3: plus soil properties  DARs list Speyer 2.2 soil and incubation data for each degradation experiment. Such data were used to build predictor set 3, from inclusion of:

1. Total organic carbon content (range: 1.3 to 2.9%, median: 2.3%),
2. Total organic matter content (range: 1.1 to 5.7%, median: 4.3%),
3. Cation exchange capacity (range: 1.0 to 21.0 meq/100g soil, median: 10 meq/100g soil),
4. Soil pH measured in (a) distilled water (range: 5.0 to 7.1, median: 6.1) and in (b) KCl or CaCl₂ (range: 5.0 to 6.8, median: 6.0),
5. Clay, silt and sand contents (lumped from DIN and USDA soil particle size distributions),
6. Microbial biomass at incubation start (range: 3.0 to 186.4 mgC/100g soil, median: 34.9 mgC/100g soil),
7. Compound application concentrations (range: 0.02 to 901.0 mg kg⁻¹, median: 0.62 mg kg⁻¹),
8. Incubation temperature (range: 20 to 25 °C, median: 20 °C),
9. Soil water content (range: 16 to 80 % maximum water holding capacity (MWHC), median: 40 %MWHC).

Predictor (6) is based on the substrate-induced respiration method [73] and treated as *missing data* for any other methods.

Predictors set 4: plus topological and geometrical descriptors  The fourth predictor set extends by a vast set of non-fragment descriptors, all being part of the DRAGON descriptor software [95]. These descriptors can be broadly categorized into the following groups:

1. Topological descriptors
2. Geometrical descriptors
3. WHIM descriptors
4. GETAWAY descriptors

Predictor group (1) encodes various molecular properties based on 2D molecular connectivity graphs, with edges interpreted as bonds and vertices interpreted as atoms. Such descriptors can further be separated into topostructural and topochemical indices. The former only encode adjacency and inter-atom distance. The latter also take into account atom identity or atom hybridization.
Group (1) was appended by the first-order Molecular Connectivity Index (MCI) of KOCWIN \[68\].

Group (2) enumerates 74 predictors which rely, in contrast to the former group, on 3D molecular structures. This 3D structure is an optimization of molecular conformation, calculated from 2D molecular graph representations with CORINA \[46\]. Examples are molecular eccentricity, spherocity or aromaticity \[95\].

Being a special subset of Cartesian geometrical indices, 99 Weighted Holistic Invariant Molecular (WHIM) descriptors compose group (3). These descriptors are statistical indices calculated from projections of individual atoms along the first three spatial principal component axes of a molecule. Such projections yield a reference frame for capturing molecular shape, symmetry, size and atom distribution. Moreover, indices were weighted by e.g. atomic mass, polarizability or van der Waals volumes \[98\].

Group (4) is yet another subset of geometrical indices, composed of 197 so-called Geometry, Topology, and Atom-Weights AssemblY (GETAWAY) predictors \[24\]. Here, a leverage matrix is calculated for all the spatial coordinates of atoms in a molecule. This matrix is then utilized to (1) measure its information content or (2) either derive indices on spatial autocorrelation alone or (3) in combination with geometric interatomic distances. Again, weighting is applied for atomic mass, polarizability, etc.

Further specification of the descriptors groups (1) to (4) can be found under \[95\] and \[98\] or in section 3.2.3 in as far as a descriptor was identified to be important.

### 2.3.7 Predictor importance

The importance of a predictor for explaining $k_{PM}$, $k_{M1M2}$ was derived solely within the RBS framework, for reasons clarified in the results section, for predictor set 4.

Within this framework, the rank $R(p_i)$ a predictor $p_i$ attains in the last RBS recursion shall signify its importance (the last RBS recursion is passed just before $p_{sub} \leq 30$ is achieved in step 6 of algorithm 1). Predictors that are discarded ahead of the last RBS recursion are, of course, not included in the last recursion set of $p_{sub}$. For these, $R(p_i)$ amounts to zero.

For good measure, the above RBS approach was repeated 87 times, with each repetition having one sample omitted (=LOO). This allows evaluating the stability of the proposed RBS calculation of predictor importance under slight alteration of the data set. In this way, an $R(p_i)$ distribution could be established over all LOO runs. The variance of such distribution may be expected to be larger for LMO than for LOO, although the bootstrapping of RBS may dampen such tendency. The method must not be confused with cross-validation, which only draws inferences on the samples omitted.

The direction of the effect a single predictor had on the response was evaluated inside the last RBS recursion, too. To this end, the signs of OLS coefficients were recorded for the predictors $p_x$ and $p_y$ over all $m = 20000$ in the last recursion of algorithm 1. The number of positive versus the number of negative signs of each
predictor over all RBS selections expressed directional uncertainty. For continuous predictors, a positive coefficient sign suggests increased \((k_{PM}, k_{M1M2})\) for increased values of the predictor. For categorical predictors, a positive sign for a level suggests \((k_{PM}, k_{M1M2})\) above the model intercept, a negative sign one below. The methodology may be unreliable for a categorical (fragment) predictor with many levels, where the mode of the signs over all the level coefficients had to be recorded at each time the predictor was selected in RBS. More precisely, this summing of coefficient signs over all levels does neither embrace the numbers of entries in each level nor the coefficients of each level. For example, a count level with many entries and a large positive coefficient may be outweighed by two other levels with few entries each and small negative coefficient signs, even though most samples belong to the first level. An overall negative sign is thus recorded even though most entries have \((k_{PM}, k_{M1M2})\) above the model intercept.
3 Results and Discussion

Section 3.1 outlines and discusses the results from inverse modeling. Based on the derived kinetic rate parameters, section 3.2 states the outcomes on statistical modeling of compound (bio)degradation.

3.1 Kinetic rate parameters from inverse modeling

92 data sets from 78 different time series were assembled. 23 time series were supplied by BAYER Crop Science, the others by EFSA DARs. 5 time series were removed because they either showed clear deviation from first order kinetics (1,2,4-triazole / label 1), questionable data (THS3995 / label 60) or poor data coverage (cycloxidim / label 8, fluazifop-P-butyl / label 31 and fluazifop-P-acid / label 32). Of the remaining 87 data sets, 13 refer to first single metabolites downstream of a parent compound. A list of all parent compounds and first downstream metabolites is given in Appendix A.

Model evaluation 9 data sets are duplicates (cp. Appendix A), offering hints at differences in $k$ values from repeated Speyer 2.2 soil pesticide incubations. Absolute differences of $\log_{10}(k_{PM,A}) - \log_{10}(k_{PM,B})$ and of $\log_{10}(k_{PN,A}) - \log_{10}(k_{PN,B})$ do not exceed 0.67 and 1.17 among these 9 duplicate data sets, respectively (with indices $A$ and $B$ indicating the two duplicate time series). In other words, these differences barely exceed one order of magnitude and are therefore not substantial when being compared to the full range of $k$ values investigated among all data sets (Figure 4).

Using the five- instead of the four-pool model slightly increases average MSC (Table 1). On an individual time series basis, turning from four- to five pool models increases MSC for 8 of the 13 series suitable for five pool modeling. Fitted and observed values for both model types are exemplified for Amidosulfuron in Figures 5.

Average $R^2$ is high for all model types (Table 1): 96% of the series have $R^2 > 0.85$, 80% have $R^2 > 0.95$. Only the degradation series for dazomet (label 14), tolyfuanid/DMST (labels 81 and 82) and cycloxidim (label 9) show $R^2$ as low as 0.73, 0.58 and 0.38, respectively. In all three cases, model residuals are largest for the first data points, having negative residuals for $P(0)$ and positive ones for $M1(0)$ (Figures 6 and 7). These three cases represent very fast degradation (see labels and rate parameters in Figure 8), with $P(0) \leq 45\%$ AR. This implies that considerable compound dissipation had occurred before the first incubation sampling. Model estimates of $P(0)$ can impossibly follow this trend, since $P(0)$ also represents the total AR to be distributed over all model time steps. Indeed, omitting $P(0)$ and $M1(0)$ raised $R^2$ values to acceptable levels, namely to 0.83, 0.93 and 0.78 for dazomet, tolyfuanid/DMST and cycloxidim, respectively.
### 3.1 Kinetic rate parameters from inverse modeling

Figure 4: Comparison of duplicate kinetic parameters for \( k_{PM} \) and \( k_{PN} \), based on four box modeling of the 9 duplicate time series data (circles). Squares show ranges over all the \( k_{PM} \) (red) and over all the \( k_{PN} \) (blue) investigated. To facilitate illustration, data were sorted so that \( k_{PM} \) (duplicate A > duplicate B) and \( k_{PN} \) (duplicate A < duplicate B). Grey lines indicate differences in ±1 on the log\(_{10} \) scale.

Figures 6 and 7 depict model residuals as a function of time for all AR pools separately. Two additional insights can be derived from residual analysis. (1) Residual variance is slightly smaller in the five- than in the four pool case. That is, the standard deviation (sd) of \( y_i - \hat{y}_i \) over all pools and data sets results to 5.4% AR in the four- and 5.0% AR in the five-box model. Standard deviation is highest for \( P(t) \) (sd = 7.17% AR) in the four pool model and highest for \( M1(t) \) (sd = 6.8 % AR) in the five pool model. sd was smallest for \( V(t) \) in both model types. (2) For both model types, NER formation is underestimated until \( t \leq 30 \) days, when some shift towards overestimation sets in. This effect has already been reported elsewhere and may be attributed to two causes [12]. On the one hand, NER formation until \( t \leq 30 \) days may be very fast and may initially deviate from first order kinetics. On the other hand, a portion of the NER produced at \( t > 30 \) may not be extractable by the solvent, the latter being incapable of
rendering this proportion of the compound-soil interactions or accessing the full micropore space.

Table 1: Evaluation of model fit for the four box and - in the case of datasets containing the single downstream metabolite - five box models. Values represent means for the individual datasets.

<table>
<thead>
<tr>
<th>Model statistics</th>
<th>Four box1</th>
<th>Five box1</th>
<th>Four box2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>MSC</td>
<td>3.2</td>
<td>3.3</td>
<td>-</td>
</tr>
</tbody>
</table>

1 Based on 13 datasets for which a single first downstream metabolite was reported. Comparison of four- and five pool model approach.

2 Based on 74 datasets without a single first downstream metabolite. Restriction to four-box model approach.

**Correlation among model parameters** In the four box model, a positive correlation is prominent between the two rate constants of (1) compound transformation to metabolite and (2) compound conversion to NER ($r_{spear}(k_{pm}, k_{pn}) = 0.72$). Moreover, the rate parameters for both NER formation and volatile formation from the metabolite pool are slightly correlated ($r_{spear}(k_{mn}, k_{mv}) = 0.48$). All other $r_{spear}$ among the $k$ and between the $k$ and $P(0)$ are $\leq 0.25$. Similar trends can be observed for the five pool model case. Again, (1) and (2) are positively correlated, both for the parent compound ($r_{spear}(k_{pm}, k_{pn}) = 0.75$) and the first downstream metabolite ($r_{spear}(k_{m1n}, k_{m1m2}) = 0.50$). Additionally, conversion from the first ($M1(t)$) to the second ($M2(t)$) metabolite pool correlates with volatilization (≡ mineralization) from $M2(t)$ to $V(t)$ ($r_{spear} (k_{m1m2}, k_{m2v}) = 0.72$). A weak correlation between $k_{pm}$ and $k_{m2n}$ was observed ($r_{spear} = 0.45$). Secondly, negative correlation is evident between the fitted $P(0)$ and both $k_{pm}$ and $k_{pn}$ ($r_{spear}(P(0), k_{pm}) = -0.68$ and $r_{spear}(P(0), k_{pm}) = -0.73$). The latter effect may arise when the numerical optimization can describe fast parent compound removal by both high $k_{pm1}$, $k_{pm}$ and - to some degree - low $P(0)$. Among the ten data sets of highest $k_{pm}$, eight belong to the five pool series of numerical modeling (cp. labels in Figure 8). For this reason, the effect may be observable only for the five- but not for the four-pool models. Other correlations among the $k$ and between the $k$ and $P(0)$ are $\leq 0.38$ in the five pool case.

Primary degradation ($k_{pm}$, $k_{pm1}$) did not correlate with mineralization ($k_{mv}$, $k_{m2n}$) in any of the models. The findings back and extend results of [66], who found correlation between $k_{pm}$ and $k_{pn}$ only.
3.1 Kinetic rate parameters from inverse modeling

Figure 5: Observed (points) and modeled (line) dissipation of Amidosulfuron on Speyer 2.2 soil. Results are given for the four-pool (upper panel) and the five-pool model (lower panel).
Figure 6: Residuals \((y_i - \hat{y}_i)\) over time for 74 four-box model time series. Labels refer to parent compounds \(P(t)\) listed in Appendix A. LOWESS polynomial smoother 23 (red line) visualizes residual deviation from zero (green line). Mind differing ordinate scales.
3.1 Kinetic rate parameters from inverse modeling

Figure 7: Residuals \((y_i - \hat{y}_i)\) over time for 13 five-box model time series. Labels refer to compounds of Appendix A, either \(P(t)\) or \(M1(t)\). LOWESS polynomial smoother \(^{23}\) (red line) visualizes residual deviation from zero (green line). Mind differing ordinate scales.
Metabolite vs. NER formation Figure 8 compiles the variation in $k$ values for (1) conversion of the compounds $P(t)$ and $M_1(t)$ to downstream metabolite(s) ($k_{PM}, k_{M1M2}$) as opposed to (2) conversion to NER ($k_{PN}, k_{M1N}$). Differences in these rate parameters among the data sets cover several orders of magnitude, with $\max(\log_{10}(k_{PM}, k_{M1M2})) - \min(\log_{10}(k_{PM}, k_{M1M2})) = 4.62$ and $\max(\log_{10}(k_{PN}, k_{M1N})) - \min(\log_{10}(k_{PN}, k_{M1N})) = 5.95$. The $k$ values for both (1) and (2) are highly right-skewed, with few very high rate parameters (cp. boxplots in the named Figure - mind logarithmic abscissa). Moreover, there is no obvious overall deviation in $k_{PM1}, k_{PM}$ versus $k_{M1M2}$ or $k_{PN}$ versus $k_{M1N}$, as indicated by the four lowermost stripcharts of Figure 8.

Despite the further above mentioned correlation between ($k_{PM}, k_{M1M2}$) and ($k_{PN}, k_{M1N}$), average rates of compound conversion to metabolites are higher than conversion to NER (cp. boxplot notches of Figure 8). The contrary, ($k_{PM}, k_{M1M2}$) ≤ ($k_{PN}, k_{M1N}$), applied to only 22% of the data sets. Degradation rates ($k_{PM}, k_{M1M2}$) are smallest for bromuconazole (label 7) and largest for haloxypol-R (label 37), whereas rates of conversion to non-extractable residues ($k_{PN}, k_{M1N}$) are smallest for oryzaline (label 58) and largest for dazomet (label 14).

Finally, not absolute but relative $k$ values of ($k_{PM}, k_{M1M2}$) as compared to ($k_{PN}, k_{M1N}$) determine whether a compound has a high tendency towards NER formation. The percentage of NER formed exclusively from $P(t)$ or $M_1(t)$ in analogy to equation 2.2.14 for $t \to \infty$ is shown on the right-hand side of the discussed figure. Thus, the percentage of NER formed can be as high as 89.0% (dichlofencarb / label 18&19) and as low as 0.0% (iprovalicarb / label 40) - in spite of none of the $k$ values equalling zero. Absolute correlation between % NER formed and $k_{PM}, k_{M1M2}$ or $k_{PN}, k_{M1N}$ is insignificant. Nonetheless, it is of note that 13 of the 19 samples with > 50% NER formed at $t \to \infty$ can be found below the median ($k_{PM}, k_{M1M2}$). To make the point clear, equation
3.1 Kinetic rate parameters from inverse modeling

![Graph showing kinetic rate parameters from inverse modeling.](image)
2.2.14 calculates the %NER formed from only one source, i.e. from either \( P(t) \) or \( M1(t) \) and not from \( (P(t), M(t)) \) (four box model) or \( (P(t), M1(t), M2(t)) \) (five box model). Thus, the above given numbers of NER formed do not cover the total %NER formed in the four- or five-box model calculations. In this context, one must further consider that \( M1(t) \) may be rapidly dissipating to NER or metabolites further downstream but that \( M1(t) \) may be slowly formed from its upstream parent compound. If this is the case, relative differences in \( k_{M1N} \) versus \( k_{M1M2} \) may express little about the concentration of \( M1(t) \) over time.

With regards to the statistical modeling approach ahead, some of the above findings need to be emphasized. Firstly, differences between two duplicate rate parameters for a single compound are (at least for the 9 duplicate sets) smaller than overall variation in \( k_{PM}, k_{M1M2} \) or \( k_{PN}, k_{M1N} \) between different compounds. It is this latter variation that the statistical modeling in the following section tries to explain. The former variation cannot be accounted for, if this variation results from differing laboratory setups or varying Speyer 2.2 soil conditions not covered here. Secondly, absence in gross deviation in the overall distribution of (a) \( k_{PM1}, k_{PM} \) versus \( k_{M1M2} \) and (b) \( k_{PN} \) versus \( k_{M1N} \) in Figure 8 backs the pooled modeling approach of parent compounds and first single downstream metabolites with regards to (1) degradation and (2) NER formation.
3.2 Statistical modeling of kinetic degradation parameters

The sections below specify outcomes for modeling variation in compound degradation, represented by the kinetic parameters $k_{PM}$ and $k_{M1M2}$. First, the model performance under (1) pseudovalidation, (2) LOO validation and (3) tuned LOO validation is compared for the three modeling strategies (a) RBS&forward selection, (b) LASSO and (c) RBS&LASSO for all predictor sets. Next, models (a) to (c) are evaluated for all predictor sets, both for (1) and (3). In a last section, important predictors are identified.

3.2.1 Model validation

A tabulation of $R^2$ over all the validation scenarios (1) to (3) and all modeling strategies (a) to (c) is provided in Table 2 for each of the four predictor sets. This table sums a number of remarkable findings which can be directly related to the working hypotheses of section 1.

First of all, the maximal variation explained is, for all models and predictor sets, as high as $R^2 = 0.80$ for the pseudovalidated approach. However, $R^2$ drops dramatically under any proper validation (external LOO CV or tuned external LOO CV). For external LOO CV (i.e., non-tuned), $R^2$ even indicates models void of any explanatory power for independent data sets, whereas the same cannot be observed for tuned external LOO CV.

An explanation for these findings is suggested by the concept of model bias-variance trade-off [96], applied to a training data set (used for building a model) and an independent test set (used to validate the model, e.g. CV folds). Increasing the complexity of a model (e.g. by the number of parameters $p_{max}$ and/or $\lambda$) can produce an ever-increasing $R^2 = 1.00$ for the training set of data. However, this is not necessarily the case for the test set of data. Here, ever increasing model complexity often triggers a decrease in the bias and an increase in the variance of the model predictions. After some degree of model complexity, a bias-variance trade-off optimal for independent data sets is surpassed. Thereafter, increasing model complexity deteriorates the $R^2$ for independent data; the model gradually loses its capability for generalization and overfits towards the training data. This circumstance is exemplified in Figure 9. Therein, the $R^2$ averaged over all LOO folds is given as a function of model complexity outside the folds. For tuned LOO CV, model complexity equals that of highest $R^2$ (to avoid confusion, one must note that model complexity in tuned LOO is not specified for each single fold, but an optimal average value over all folds). For non-tuned (“off-the-shelf”) LOO CV however, model complexity may be well off (and probably beyond) that optimum. Remarkably, for all such non-tuned LOO scenarios, LASSO showed the lowest pessimism, possibly due to its build-in penalization against model complexity. Remarkably also, for tuned LOO, this was not the case.

Model instability is intimately linked to CV estimates of poor generalization performance. Besides the above discussed shifts in model complexity, model construction can select quite different predictors and coefficients whenever a
3.2 Statistical modeling of kinetic degradation parameters

Table 2: Validation of coefficients of determination ($R^2$) for predicting $(k_{PM}, k_{M1M2})$. $R^2_{CV,-1}$ refers to leave-one-out (LOO) cross validation, $p$ denotes the total number of candidate predictors.

<table>
<thead>
<tr>
<th>Dataset &amp; model</th>
<th>$R^2_{1}$</th>
<th>$R^2_{CV,-1}$</th>
<th>$R^2_{CV,-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1: Biowin, $p = 38$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBS &amp; forward stepwise</td>
<td>0.19</td>
<td>0.03</td>
<td>0.15</td>
</tr>
<tr>
<td>LASSO regression</td>
<td>0.28</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>RBS and LASSO</td>
<td>0.28</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>Set 2: set 1 plus fragments and partitioning coefficients, $p = 214$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBS &amp; forward stepwise</td>
<td>0.56</td>
<td>0.01</td>
<td>0.27</td>
</tr>
<tr>
<td>LASSO regression</td>
<td>0.80</td>
<td>0.24</td>
<td>0.31</td>
</tr>
<tr>
<td>RBS and LASSO</td>
<td>0.61</td>
<td>0.13</td>
<td>0.20</td>
</tr>
<tr>
<td>Set 3: set 2 plus soil properties, $p = 230$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBS &amp; forward stepwise</td>
<td>0.64</td>
<td>0.05</td>
<td>0.27</td>
</tr>
<tr>
<td>LASSO regression</td>
<td>0.75</td>
<td>0.16</td>
<td>0.25</td>
</tr>
<tr>
<td>RBS and LASSO</td>
<td>0.67</td>
<td>0.13</td>
<td>0.18</td>
</tr>
<tr>
<td>Set 4: set 3 plus geometry and topology, $p = 670$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBS &amp; forward stepwise</td>
<td>0.61</td>
<td>0.00</td>
<td>0.25</td>
</tr>
<tr>
<td>LASSO regression</td>
<td>0.56</td>
<td>0.05</td>
<td>0.15</td>
</tr>
<tr>
<td>RBS and LASSO</td>
<td>0.64</td>
<td>0.02</td>
<td>0.11</td>
</tr>
</tbody>
</table>

1 Pseudovalidated approach (internal CV). The associated models can be found in Appendix C.
2 External LOO-CV.
3 Tuned external LOO-CV.
fold is omitted, especially in small data sets where the relative information contribution from each sample is large. Reducing complexity may keep model construction focused on predictors of general importance instead on such specific to a given CV data subset.

In summary, these findings demonstrate that pseudovalidated $R^2$ can be a highly misleading measure for model generalization performance. Contrariwise, non-tuned LOO CV can be uselessly pessimistic. Hence, tuned LOO CV may offer the method of choice when performance estimates for independent data need to be derived from small training sample sets.

The highest tuned LOO $R^2 = 0.31$ is found for LASSO on predictor set 2. But, such performance of LASSO steadily decreases when the predictor sets grow in size (and in noise, as it is unlikely that all 670 predictors are related to the response). Eventually, LASSO $R^2$ ranges below the one of RBS&stepwise forward. In contrast, and from predictor set 2 on, the performance of RBS&stepwise forward

Figure 9: $R^2$ of independent data (= external LOO CV folds) as a function of model complexity, exemplified for hybrid RBS&LASSO on predictor set 4 from Table 2. Models were iteratively built along the $\lambda$ sequence on the data outside the folds and predictions made to the folds at each iteration. Hence, the abscissa corresponds to that of Figure 3.
3.2 Statistical modeling of kinetic degradation parameters

forward remains stable under growing predictor sets, featuring highest LOO-validated performance under all models for predictor sets 3 and 4. Whereas the chance of LASSO to select predictors unimportant for generalization obviously increases with increasing predictor number, the more robust bootstrap filtering of RBS counteracts this tendency. Thus, for large predictor sets containing noise variables the novel Recursive Bootstrap Subsampling can outperform LASSO shrinkage, especially when considering that the information content of a data set is rarely known beforehand. As a drawback, RBS comes at higher computational costs than LASSO. Figure 10 offers further insight into the functioning of RBS.

For non-tuned LOO, hybridization of LASSO with prior RBS filtering performs better than RBS&forward stepwise but not necessarily better than LASSO alone. This may again be a result of the shrinkage LASSO performs, leading to less overfitting outside a tuned approach, in contrast to forward stepwise selection. Interestingly, RBS&LASSO always performs worse than the other two methods for tuned LOO CV. This effect may be based on lower LASSO model stability for an LOO-averaged λ once RBS has removed the majority of predictors outside the folds; though the exact cause of this phenomenon is open to future investigation.

The information the predictor sets 1 to 4 convey for explaining variance in \( k_{PM}, k_{M1M2} \) differs. Obviously, the EPI Suite covariates of predictor set 1 have much less explanatory power than set 2, which is an update of additional fragments and some physicochemical properties. Additional inclusion of varying Speyer 2.2 soil properties and incubation settings (set 3) or of the vast amount of geometrical and topological predictors (set 4) does not augment model performance, neither under pseudovalidation nor under (tuned) validation. Instead, the introduction of noise variables (especially those of set 4) attenuates model performance for the LASSO. The preponderance of molecular fragments of set 2 for explaining compound degradation is being confirmed in section 3.2.3 dealing with predictor importance.

A last paragraph shall address residual (i.e. unexplained) model variance, which may be ascribed to [81]:

1. Errors in / uncertainty of the model structure and its numerical solution
2. Errors in / uncertainty of model parameter values
3. Errors / uncertainty from non-deterministic system behavior
4. Unidentified influence factors or uncertainty thereof

Concerning (1.), a good GA fit of time series data to the model structures does not implicate correct causality. Initial deviation from first order kinetics of \( P(t) \) was observed at least for one excluded compound and in some other cases small but constant \( P(t) \neq 0 \%AR \) could be observed at the end of time series, too. In fact, there are plenty of causes evoking deviation from first order kinetics for \( P(t), M(t), M1(t), M2(t) \) and \( N(t) \). For example, varying bioactivity of the
3.2 Statistical modeling of kinetic degradation parameters

Figure 10: Result of the second iteration from recursive bootstrap subsampling (RBS) after 20000 iterations on a set of 338 predictors. The predictors consist of the algorithm subset $p_{sub}$ of predictor set 4 plus three probes P1, P2 and P3.

The ordinate gives absolute predictor selection frequencies. (1) Red dots: absolute selection frequencies on the full response (i.e. $R(p_x)$) only, cp. algorithm 1; (2) blue dots: absolute selection frequencies on the residuals after selection on the full response (i.e. $R(p_y)$) only, cp. algorithm 1 and (3) gray dots: sum of the former frequencies. Abscissa: ranking of predictors according to $R(p_x)$; vertical stacking of absolute frequencies indicates identical ranks. Predictors inside region B enter the third RBS iteration, those of region A will be discarded.

Probes $P1$, $P2$, $P3$ are noised copies of the response $[15]$. More precisely, $P1$, $P2$ and $P3$ have Pearson correlation of 0.5, 0.4 and 0.3 with the response, respectively, and refer to the gray dots. The probes are added to the predictor set to show approximate correlation regions of the sum of the selection frequencies (gray) with the response. They confirm that predictors of higher correlation with the response are selected more frequently.

Interestingly, the range in absolute selection frequencies across all 338 predictors is more pronounced for selection on the full response (red) than for selection on the residuals of the latter (blue).
3.2 Statistical modeling of kinetic degradation parameters

soil [102] or altered compound bioavailability through sorption to soil particles have been documented [22]. [12] diagnosed a direct effect of microbial activity on NER formation, too. Irreversibility of NER-formation over long temporal scales can also not be guaranteed [62][87][12], despite of little or no decrease of $N(t)$ in the DAR reports. It was merely assumed that such processes have a minor impact on compound fate. In comparison, uncertainties from numerical solution of the differential equations are typically small [81].

(2.) Uncertainties in the parameter values $P(0)$ and $k$ can result from the somewhat low precision of GA optimization, too. In addition, one may utilize alternative loss functions other than least squares (e.g. absolute loss or non-independent errors), leading to slightly different parameter fits. %AR of the individual series of $P(t)$, $M(t)$, $M1(t)$, $M2(t)$, $N(t)$ and $V(t)$ were not normalized, giving more weight to accurately fitting series of high $\Delta%AR$ (i.e., mostly $P(t)$). Other parameter uncertainties stem from the experimental setup. Replicates of compound degradation series evidence slight differences in the optimized parameters from seemingly different incubation setups or different Speyer 2.2 aliquots utilized (cp. section 3.1). Somewhat varying DAR measurement intervals give differing weight to differing time points, possibly resulting in somewhat different fits for the same compound. Most DARs had non-stationary $AR < 100\%$, signifying that the incubation setup and/or analytical techniques were not fully able to recover the initially applied radioactivity - further adding to uncertainties. Moreover, the %AR from NER vs. the %AR of soil extracts is operationally defined and the underlying extraction methods were not homogeneous among the DARs. For example, water+methanol was used in the extraction steps of the lenacil DAR (label 42), but hexane+acetone+water/acetone+sulfuric acid in that of metribuzin (label 57). As summarized by [87], sulphuric acid extracts residues from clay minerals whereas water+methanol may not - thus affecting the fraction AR regarded as non-extractable.

(3.), non-deterministic system behavior, can express itself in aggregation errors. For instance, time points of measurements are given in days, thus fixing temporal resolution. For fast degrading compounds daily resolution can be too coarse, causing $P(0) << \hat{P}(0)$, as discussed in section 3.1 The uncertainties (1) to (3) have not yet been estimated and propagated to the results of section 3.1 to 3.2.3.

(4.) constitutes ambiguities related to the predictors. Initial Speyer 2.2 soil state, temporal changes in bioactivity, characterization of soil microbial communities and the organic soil matter composition could all not be assessed. Similarly, greatly unbalanced predictors were removed, a priori excluding some underrepresented factors. Neither predictor interactions nor non-linear model structures were part of the simple linear model utilized. In addition, the covariates themselves have uncertainties. For instance, $K_{oc,median}$ and $K_{oc,median}$ are summary statistic over a set of measurements for compound partitioning, with values somewhat depending on methodological choices (e.g. organic carbon sorbate, extraction methods). Furthermore, uncertainties arise when filling data gaps by medians or modes.

Given such ample ground for model uncertainty and input errors, a somewhat
low but LOO-validated explainable variance of up to 31\% nevertheless concedes confidence to the predictors found important in section 3.2.3.

### 3.2.2 Model evaluation

The previous section demonstrated that non-tuned validation gives overly pessimistic estimates of model performance. Assuming that prediction performance to independent data must range in between pseudovalidated and tuned-validated results, these two latter approaches are evaluated in the following. Summary statistics for predicting the conversion of a compound to metabolite(s) (i.e. \((k_{PM}, k_{M1M2})\)) are given in Table 3 for all pseudovalidated models and over all predictor sets. The corresponding models are tabulated in Appendix C. In addition, Table 4 gives a complementary summary for the tuned-validated approach. The underlying models are not listed since different models were constructed over all folds. Plots of predicted versus observed \((k_{PM}, k_{M1M2})\) for RBS\&forward selection are presented in Figure 11 (pseudovalidation) and in Figure 12 (tuned validation), with figures based on the RBS-filtered predictor set 4.

Modeling on Box-Cox transformed data must at some point be linked to the relevant, non-transformed timescales. To this end, Figure 11 also shows relative deviations in \(DegT_{50}\) for a given deviation in observed versus predicted values of \((k_{PM}, k_{M1M2})\). For orientation, the median \((k_{PM}, k_{M1M2})\) of 0.02 \(\text{days}^{-1}\) corresponds to a \(DegT_{50}\) of 33 days (cp. equation 2.2.5). Predicted \(DegT_{50}\) may be over a magnitude lower (e.g. BAS-505-F, label 25) and over a magnitude larger (e.g. Tolylfluanid, label 81) than the observed one. One must mind the \(\log_{10}\) of the corresponding figure. Therein, absolute deviations of \((DegT_{50} - \hat{DegT}_{50})\) increase with increasing \((k_{PM}, k_{M1M2})\), whereas relative deviations decrease. To make this point clear, prediction for the somewhat persistent BAS-505-F (label 25, lower \((k_{PM}, k_{M1M2})\)-range) underestimates a \(DegT_{50}\) of 980 days by 95\%. In contrast, the \(\hat{DegT}_{50} = 5\) days for Tolylfluanid (label 81, upper \((k_{PM}, k_{M1M2})\)-range) overestimates the observed \(DegT_{50} = 3.8\) hours by over 3000\%. Most likely, these deviations grow further if predictions to independent data are made, as simulated by tuned LOO-CV in Figure 12.

A next essential step is a check of OLS model assumptions. Figures 13 and 14 plot model residuals against (1) predicted values and (2) observed values, for each of the above discussed models in Figures 11 and 12. With regard to (1), the graphs give no indication of heteroscedasticity or non-Gaussian distributed residuals, for none of the both models. This confirms no false assumptions from the model side. However, the residual plots reveal two clear trends for aspect (2). First of all, there is an increasing overestimation of \((k_{PM}, k_{M1M2})\) (equal to an underestimation of \(DegT_{50}\)) for samples below the sample mean and an increasing underestimation of \((k_{PM}, k_{M1M2})\) (equal to an overestimation of \(DegT_{50}\)) for samples above the sample mean. Secondly, this trend is more pronounced for the validated than for the pseudovalidated case. These two trends hold over all three models and over all four predictor
3.2 Statistical modeling of kinetic degradation parameters

Figure 11: Observed versus predicted values of \((k_{PM}, k_{M1M2})\) for the pseudovalidatated RBS forwards stepwise model, based on the full predictor set 4. Note: axes are in log scale, whereas model predictions were made on Box-Cox transformed scales. Grey lines indicate relative deviations in compound degradation halflives \(\text{DegT}_{50} = \frac{\ln(2)}{k}\) for given deviations in \((k_{PM}, k_{M1M2})\) between predictions and observations. For example, a deviation of 2 indicates an overestimation in \(\text{DegT}_{50}\) of factor 2, whereas deviations of 1/10 and 1/100 indicate underestimation in the order of two and three magnitudes, respectively.
3.2 Statistical modeling of kinetic degradation parameters

Figure 12: Observed versus predicted values of \((k_{PM}, k_{M1M2})\) for the tuned validated RBS\&forward stepwise model, based on the full predictor set 4. Note: axes are in log scale, whereas model predictions were made on Box-Cox transformed scales. Grey lines are explained in Figure I1.
Table 3: Performance measures for the various pseudovalidated modeling approaches listed in the first column of Table 2. \( p \) stands for the total number of candidate predictors, \( p_{\text{max}} \) for those selected into the final model. Individual models are listed in Appendix C.

<table>
<thead>
<tr>
<th>Candidate predictors</th>
<th>Recursive bootstrap subsampling &amp; stepwise forward</th>
<th>LASSO</th>
<th>Recursive bootstrap subsampling &amp; LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Set 1: Biowin, ( p = 38 )</strong></td>
<td>RMSE: 3.1±0.3 M AD: 2.1 bias: -0.3 bias( &gt; )mean: -2.0 bias( \leq )mean: 2.2 ( p_{\text{max}} ): 2</td>
<td>RMSE: 2.9±0.3 M AD: 2.0 bias: -0.2 bias( &gt; )mean: -1.8 bias( \leq )mean: 2.1 ( p_{\text{max}} ): 10</td>
<td>RMSE: 2.9±0.3 M AD: 2.0 bias: -0.2 bias( &gt; )mean: -1.8 bias( \leq )mean: 2.1 ( p_{\text{max}} ): 8</td>
</tr>
<tr>
<td><strong>Set 2: Set 1 plus fragments and physicochemical properties, ( p = 214 )</strong></td>
<td>RMSE: 2.3±0.2 M AD: 1.4 bias: 0.1 bias( &gt; )mean: -1.2 bias( \leq )mean: 1.0 ( p_{\text{max}} ): 8</td>
<td>RMSE: 1.6±0.2 M AD: 0.9 bias: 0.0 bias( &gt; )mean: -0.8 bias( \leq )mean: 0.8 ( p_{\text{max}} ): 48</td>
<td>RMSE: 2.1±0.2 M AD: 1.4 bias: -0.3 bias( &gt; )mean: -1.1 bias( \leq )mean: 1.3 ( p_{\text{max}} ): 18</td>
</tr>
<tr>
<td><strong>Set 3: Set 2 plus soil properties, ( p = 230 )</strong></td>
<td>RMSE: 2.0±0.2 M AD: 1.3 bias: -0.3 bias( &gt; )mean: -0.8 bias( \leq )mean: 0.8 ( p_{\text{max}} ): 16</td>
<td>RMSE: 1.7±0.2 M AD: 1.0 bias: 0.0 bias( &gt; )mean: -0.9 bias( \leq )mean: 0.8 ( p_{\text{max}} ): 35</td>
<td>RMSE: 1.9±0.2 M AD: 1.2 bias: -0.1 bias( &gt; )mean: -0.83 bias( \leq )mean: 1.0 ( p_{\text{max}} ): 21</td>
</tr>
<tr>
<td><strong>Set 4: Set 3 plus topology and geometry, ( p = 670 )</strong></td>
<td>RMSE: 2.1±0.2 M AD: 1.5 bias: -0.1 bias( &gt; )mean: -1.1 bias( \leq )mean: 1.4 ( p_{\text{max}} ): 12</td>
<td>RMSE: 2.2±0.2 M AD: 1.4 bias: 0.0 bias( &gt; )mean: -1.3 bias( \leq )mean: 1.4 ( p_{\text{max}} ): 25</td>
<td>RMSE: 2.1±0.2 M AD: 1.4 bias: 0.1 bias( &gt; )mean: -1.1 bias( \leq )mean: 1.2 ( p_{\text{max}} ): 19</td>
</tr>
</tbody>
</table>
Table 4: Performance measures for the various modeling approaches listed in the third column of Table 2. $p$ stands for the total number of candidate predictors.

<table>
<thead>
<tr>
<th>Candidate predictors</th>
<th>Recursive bootstrap subsampling &amp; stepwise forward</th>
<th>LASSO</th>
<th>Recursive bootstrap subsampling &amp; LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1: Biowin, $p = 38$</td>
<td>RMSE: 3.1±0.3 MAD: 2.0 bias: -0.3 bias$<em>{&gt;\text{mean}}$: -1.7 bias$</em>{&lt;\text{mean}}$: 2.2</td>
<td>RMSE: 3.3±0.4 MAD: 2.2 bias: -0.3 bias$<em>{&gt;\text{mean}}$: -1.9 bias$</em>{&lt;\text{mean}}$: 2.3</td>
<td>RMSE: 3.3±0.4 MAD: 2.2 bias: -0.2 bias$<em>{&gt;\text{mean}}$: -2.0 bias$</em>{&lt;\text{mean}}$: 2.3</td>
</tr>
<tr>
<td>Set 2: Set1 plus fragments and physicochemical properties, $p = 214$</td>
<td>RMSE: 2.9±0.3 MAD: 2.4 bias: -0.4 bias$<em>{&gt;\text{mean}}$: -2.0 bias$</em>{&lt;\text{mean}}$: 2.2</td>
<td>RMSE: 2.8±0.3 MAD: 1.4 bias: 0.0 bias$<em>{&gt;\text{mean}}$: -1.0 bias$</em>{&lt;\text{mean}}$: 0.5</td>
<td>RMSE: 3.0±0.3 MAD: 2.0 bias: -0.1 bias$<em>{&gt;\text{mean}}$: -1.7 bias$</em>{&lt;\text{mean}}$: 1.7</td>
</tr>
<tr>
<td>Set 3: Set2 plus soil properties, $p = 230$</td>
<td>RMSE: 2.9±0.3 MAD: 2.3 bias: -0.2 bias$<em>{&gt;\text{mean}}$: -2.2 bias$</em>{&lt;\text{mean}}$: 2.3</td>
<td>RMSE: 2.9±0.3 MAD: 1.5 bias: -0.1 bias$<em>{&gt;\text{mean}}$: -1.3 bias$</em>{&lt;\text{mean}}$: 0.6</td>
<td>RMSE: 3.1±0.3 MAD: 1.9 bias: -0.2 bias$<em>{&gt;\text{mean}}$: -1.7 bias$</em>{&lt;\text{mean}}$: 1.6</td>
</tr>
<tr>
<td>Set 4: Set3 plus topology and geometry, $p = 670$</td>
<td>RMSE: 2.9±0.3 MAD: 2.3 bias: 0.0 bias$<em>{&gt;\text{mean}}$: -2.0 bias$</em>{&lt;\text{mean}}$: 2.3</td>
<td>RMSE: 3.1±0.3 MAD: 1.8 bias: 0.1 bias$<em>{&gt;\text{mean}}$: -1.8 bias$</em>{&lt;\text{mean}}$: 2.0</td>
<td>RMSE: 3.2±0.3 MAD: 2.2 bias: 0.0 bias$<em>{&gt;\text{mean}}$: -1.9 bias$</em>{&lt;\text{mean}}$: 2.5</td>
</tr>
</tbody>
</table>
3.2 Statistical modeling of kinetic degradation parameters

Figure 13: Pseudovalidated RBS model residues $y_i - \hat{y}_i$. Upper panel: residues vs. predicted values of $(k_{PM}, k_{M1M2})$. Lower panel: residues vs. observed values of $(k_{PM}, k_{M1M2})$, with “observed” referring to kinetic parameters from inverse modeling. The grey line tags the sample mean.

sets, as summarized in Tables 3 and 4, where the trend is reflected in the indices $bias_{<mean}$ and $bias_{>mean}$. More precisely, $bias_{<mean}$ is always greater than zero and $bias_{>mean}$ always smaller than zero. In addition, $bias_{<mean}$ and $bias_{>mean}$ are each higher in absolute value in the validated than in the unvalidated case for each model and predictor set. Note also that the overall bias is unable to reflect this systematic deviation, despite its more widespread usages in statistics than $bias_{<mean}$ or $bias_{>mean}$.

A deviation from linearity may be a first guess to this systematic trend in model residuals as a function of the response. However, neither various transformations of the response nor non-Gaussian error distributions nor running non-parametric models such as Boosted Trees or Random Forest [96] resulted in de-trending. Another explanation may be based on uncovered interactions of predictors. Interactions were also checked in combination with the Boosted Trees / Random Forest settings, but without success.
3.2 Statistical modeling of kinetic degradation parameters

There may be another explanation for predictions biased towards the sample mean, i.e. the null-model intercept. Imagine (1) only a single fragment $X$ being responsible for the primary degradation of a given compound $x$ with a $(k_{PM}, k_{M1M2})$ above the sample mean. $x$ also carries another fragment $Y$ not being involved in the first degradation step of $x$. Imagine (2) a complementary compound $y$ with $X$ and $Y$ and with a $(k_{PM}, k_{M1M2})$ below the mean; primary degradation of $y$ is triggered by fragment $Y$ only, i.e. not by $X$. Under such conditions, the occurrence of $y$ in the factor level of the $X$-predictor causes $x$ to be predicted with bias towards the $(k_{PM}, k_{M1M2})$ of $y$. Similarly, the occurrence of compound $x$ in the factor level of the $Y$-predictor causes compound $y$ to be predicted with bias towards the $(k_{PM}, k_{M1M2})$ of $x$. Of course, the same holds if $x$ and $y$ both have either above- or below-average $(k_{PM}, k_{M1M2})$, as long as $y$ is much smaller than $x$. One may argue that there is probably another sample $z$ with larger $(k_{PM}, k_{M1M2})$ than $x$, balancing the effect of $y$ on $x$ given that $z$ also

Figure 14: RBS model residues $y_i - \hat{y}_i$ from tuned validation. Upper panel: residues vs. predicted values of $(k_{PM}, k_{M1M2})$. Lower panel: residues vs. observed values of $(k_{PM}, k_{M1M2})$, with “observed” referring to results from inverse modeling. The grey line tags the sample mean.
has a fragment X not triggering degradation. Such argumentation neglects that the relative frequencies of any such z with \( z(k_{PM}, k_{M1M2}) > x(k_{PM}, k_{M1M2}) \) are lower than those of any \( y(k_{PM}, k_{M1M2}) < x(k_{PM}, k_{M1M2}) \) in a single-mode distribution and given that \( x(k_{PM}, k_{M1M2}) > \text{mean}(k_{PM}, k_{M1M2}) \). The latter would also explain why the absolute bias increases the further a sample lies from the sample mean. In sum, samples above the mean may thus be underestimated and those below overestimated.

Higher bias for validated than for pseudovalidated predictions is in line with that argumentation. For example, omitting \( x \) from the sample set which is used for model construction permits \( y \) to have more influence on the coefficient of the factor level representing X, especially if the sample set is small and/or the samples for which X truly induces degradation are limited. As a result, bias increases during model validation. Interaction terms between a fragment predictor X and a dummy variable \( X_d \) indicating if X indeed triggers degradation (i.e. \( X_d(x) = 1 \) and \( X_d(y) = 0 \)) should be able to detrend the model, provided the above causations indeed apply. Surely, one must question the suitability of predictors X and Y under such circumstances, provided that \( X_d \) (or \( Y_d \)) may not be available for any external compound to predict on. One should instead establish why X affects x but not y - and assemble more suitable predictors.

Diverging costs are associated with underestimation versus overestimation of \( \text{DegT}_{50} \). For environmental risk assessment, assuming a short degradation half live when a compound is rather persistent weights heavier than overestimating the persistence of a compound. The model bias specifically underestimates \( \text{DegT}_{50} \) for compounds at the lower degradation range of the sample population, i.e. especially those samples with possible tendency towards persistence alias \( (k_{PM}, k_{M1M2} <) \) model intercept. If the latter intercept is below some critical threshold, the risks of underestimating \( \text{DegT}_{50} \) is even augmented. Further efforts must eliminate or at least quantify the outlined prediction bias and associated risks. Since bias appears to be highly systematic, it could be corrected by regressing observed on predicted values, using the resulting linear relationship for de-trending of model predictions (a similar attempt is made in the R Random Forest package [20]).

Table 3 furthermore lists the number of predictors \( p_{\text{max}} \) that were selected into the pseudovalidated models. \( p_{\text{max}} \) was lowest for RBS&forward stepwise selection among the three models. \( p_{\text{max}} \) was always highest for LASSO; up to \( p_{\text{max}} = 48 \) under predictor set 2. Of these 48 predictors, two had coefficients \( \leq 0.0 \), i.e. with negligible impact on the response (cp. Appendix C). To clarify this point, LASSO model complexity cannot be directly accessed from \( p_{\text{max}} \), but must also take account of the \( \lambda \)-penalized coefficients. Hence, LASSO may incorporate more predictors as compared to RBS&forward stepwise - under comparable model complexities. Assembling predictors can be costly, emphasizing a drawback of LASSO: a higher inclusion number of predictors for the same (or higher) explanatory power achievable with RBS&forward selection. Hybridization of LASSO with RBS diminishes \( p_{\text{max}} \), but does not outperform RBS&forward stepwise.

\( R^2 \) is directly proportional to RMSE, the latter estimating model precision.
Model accuracy is embraced by MAD, allowing two more insights from Tables 3 and 4. Firstly, LASSO has lowest MAD (highest accuracy) among the models. This may be related to the decrease in bias LASSO trades against alleviated variance under the concept of variance-bias trade-off [96]. Secondly, MAD was, over each model and predictor subset, higher for the validated than the pseudovalidated cases. Thus, predictions to independent data may not only bear lower precision, but also lower accuracy.

### 3.2.3 Predictor importance

Figure 15 depicts the importance of all utilized molecular and non-molecular predictors for explaining variation in \((k_{PM}, k_{M1M2})\). The plot is derived from validated RBS on predictor set 4. The boxplots summarize instability in the importance \(R(p_i)\) of a predictor \(p_i\) after omitting a single sample (LOO) from the full RBS run, repeated over all 87 samples. The mean \(R(p_i)\) over these 87 repetitions signifies the overall importance of a predictor. Table 6 in Appendix B lists both a short description of the concomitant predictors and the signs of predictor coefficients. Validated LASSO does not qualify for estimating predictor importance because, on the one hand, it shows poor performance for large predictor sets (cp. previous two sections) and, on the other hand, cannot handle correlated predictors in conjunction (cp. section 2.3.2).

Of all the 670 predictors considered, only 26 have a mean \(R(p_i)\) ranking above an arbitrary threshold of \(R(p_i) = 3\). The other predictors are either rarely selected during RBS recursions or do not even enter the last RBS recursion and shall therefore be eliminated from further consideration. Among the 26 remaining, lowest deviation in \(R(p_i)\) from omitting a sample can be observed for the most important predictors. One predictor stands for soil water content, 2 represent partitioning properties and another 6 predictors convey geometrical or topological information - but the majority of 17 predictors are fragment descriptors. The set of 11 predictors which never encounter \(R(p_i) = 0\) under any of the repeated 87 RBS runs is even stronger dominated by fragment descriptors: 9 predictors are fragment counts.

A common predictor selection strategy used for (QSAR/QSBR) model building is predicated on univariate statistics of correlation (1) between a predictor and the response [80] [106] and/or (2) between predictors [47]. This strategy fully neglects instability of the correlation coefficients from slight variation in the data set and instability with regards to relevance of a predictor towards external data. The same shortcomings apply to derived inputs from principal component analysis, as proposed by [7] [52]. In contrast, LOO-RBS does take these crucial aspects into account, leading to ranking different from that of simple correlation. This circumstance is illustrated in Figure 16, plotting RBS ranks against correlation between \((k_{PM}, k_{M1M2})\) and the 26 important predictors. Considerable scatter is present in this relationship; for three predictors (N.074, b00005, C.006) ranking from correlation even grossly underestimates the more robust RBS importance rank. This again emphasizes the need for robust
3.2 Statistical modeling of kinetic degradation parameters

Figure 15: Importance of individual predictors \( p_i \) from candidate set 4 for explaining variation in \( (k_{PM}, k_{M1M2}) \), based on 87 runs of recursive bootstrap subsampling (RBS), with each run having a sample omitted (=LOO). The ordinate shows mean RBS ranks \( (R(p_i)) \) over all 87 runs, giving a measure of predictor importance (red points). The abscissa ranks predictors by their importance. Boxplots show the distribution of \( R(p_i) \) over all 87 runs. In contrast to Figure 8, the boxplot whiskers extend to the extreme values. Predictors below the ranked mean \( R(p_i) = 610 \) are not shown, since they were of negligible importance. Details on the named predictors are provided in Table 6 of Appendix B.
3.2 Statistical modeling of kinetic degradation parameters

Figure 16: Absolute values of Pearson correlation between Box-Cox transformed \((k_{PM}, k_{M1M2})\) and the 26 important predictors as a function of mean RBS rank of the individual predictors. Predictor abbreviation is given in Appendix B.

methods of predictor selection to avoid consequences from misleading model building.

Three points shall facilitate the discussion of important fragment predictors. First of all, grouping of correlated predictors with potentially similar conceptual meaning seems necessary. Secondly, a check for causality can be based on matching the fragment predictors with those molecular fragments involved in any of the degradation steps listed in many of the DARs \[6\]. For this purpose, a random subset of 47 DARs was surveyed. Thirdly, the signs of RBS predictor coefficients establish the direction of effect (continuous predictors) or ranking of the associated \((k_{PM}, k_{M1M2})\) below or above the model intercept (categorical predictors).

Concerning a first grouping, the two predictors of aromatic or aliphatic esters are strongly correlated \((r_{spear}(nRCOOR, ester.1) = 0.8)\) and differ only slightly in chemical meaning. The linear coefficients of both predictors had positive
3.2 Statistical modeling of kinetic degradation parameters

Figure 17: Scatterplot of measured and calculated \( \log(K_{oc}) \) organic carbon - water partitioning constants. \( K_{oc,\text{mean}} \) and \( K_{oc,\text{median}} \) are means and medians of measured DAR values, whereas \( K_{oc,\text{MCI}} \) and \( K_{oc,\text{ow}} \) refer to EPI Suite \(^3\) estimates based on MCI \(^{68}\) and \( \log(K_{ow}) \) \(^{29}\), respectively.

Signs in 100% of the cases in which they were selected in a last RBS recursion. Confirmingly, ester moieties indeed pose sites of bond cleavage in organic xenobiotics, as reported in the DARs of acrinathrin (label 2,3), fenoxaprop-P (label 28), mepenypr-diethyl (label 43) or mesosulfuron-methyl (label 44) and elsewhere \(^{37,41,59}\). Moreover, fast hydrolysis has been observed in the DARs of haloxyp-R methyl ester (label 33) and triclopyr (label 86) at rates not typical for biologically mediated degradation \(^{88}\).

Another interrelation can be detected for descriptor moieties containing sulfur, i.e. predictors \( S, S.107, cbt0162 \) and \( cbt0259 \). Here, absolute correlation is as high as \( r_{\text{spear}}(S.107, cbt0162) = 0.8 \) and lowest for \( r_{\text{spear}}(S, cbt0162) = 0.4 \). All four predictors have positive coefficient signs in at least 99% of the RBS cases; \( S.107 \) even is the most important of all descriptors. This emphasizes the stance of S-functionalities in degradation on Speyer 2.2 soil, which may occur, for ex-
ample, via hydrolysis to sulfonic acid or oxidation to sulfoxides [88] [79]. The two UM-PPS rules cbt0162 and cbt0259 stand for degradation pathways, but their relevance may rather lie in the fragments that permit individual pathways. In fact, the reaction path represented by cbt0259 cannot be observed in the random subset of 47 DAR surveyed, the one of cbt0162 can only be confirmed for cycloxidim (label 9) and methiocarb (label 49). Other than that, sulfur-containing moieties are directly involved in at least five other compound degradation paths stated in the DARs (compound labels 14, 23, 26, 30, 37).

A third grouping arises for the organic carbon - water partitioning predictors \( K_{oc,\text{median}} \) and \( K_{oc,\text{mean}} \), which were experimentally measured and provided in the DARs [72]. The fact that partitioning properties impact degradation is no surprise. The signs of \( K_{oc,\text{median}} \) and \( K_{oc,\text{mean}} \) are both negative, suggesting decreased degradation from increased partitioning to soil constituents, i.e. diminished bioavailability [88] [8]. Other studies, e.g. those of [34] or [70], report inverse relations between \( K_{oc} \) and degradability, too. [100] draw similar inference for fenoxaprop-ethyl (label 28). On the other hand, and for the case of metamitron (label 46,47), [26] points out that calculated values of \( K_{oc} \) alone may be insufficient to represent partitioning to the bulk mineral phase. In fact, the computed partitioning estimates \( K_{oc,MCI} \) and \( K_{oc,\log(K_{ow})} \) are not confirmed important; despite some correlation, considerable scatter is present between the measured and computed \( K_{oc} \)s (Figure 17). These findings in turn stress the need for laboratory measurements and improved partitioning modeling, questioning attempts for deriving these inputs to QSBRs solely from current modeling approaches (also cp. [54] [50]). Moreover, the number of aromatic hydrogens (\( arom\_H \)) correlates positively with all the partitioning measures \( \text{spear}(arom\_H, K_{oc,\text{median}}) = 0.5 \), but only little with any of the other important predictors. The leverage of the hydrogen count on the partitioning properties of organic congeners has manifested itself in various partitioning studies and models (cp. [53] [90]), where hydrogen bonds play a major role in sorption to organic matter [89] [77]. Given the a correlation of 0.5 however, \( arom\_H \) may not be a substitute for measured \( K_{oc} \) coefficients but may rather represent the degree of saturation in a molecule. Instead, and because the utilized \( K_{oc,\text{median}} \) and \( K_{oc,\text{mean}} \) represent values over several soil types, more homogeneous partitioning coefficients specific to Speyer 2.2 soil may further improve degradation modeling.

A fourth conceptual cluster shall be proposed for descriptors on heteroaromats containing nitrogen, namely (in order of decreasing importance) \( N.073 \), \( C.042 \) and \( b\text{multiNitro}_\text{arom} \). \( C.042 \) indicates aromatic carbon bound to two heteroatoms, with correlation to \( b\text{multiNitro}_\text{arom} \) or \( N.073 \geq 0.6 \). These fragments participate in the degradation of the compounds labeled 4, 5, 11, 20, 30 and 33. In general, heteroatoms such as nitrogen can enhance reactivity via reducing aromaticity or by contributing a free electron pair to a reaction [88]. However, the fact that the level coefficients of all three fragments are < 0 points out that N-heterocyclic reactions may be somewhat slower (i.e. with lower \( k_{PM}, k_{M1M2} \)) than other degradation reactions on Speyer 2.2 soil. One must also consider that all three predictors range in the lower importance half of
3.2 Statistical modeling of kinetic degradation parameters

the 26 predictors discussed and show LOO-rank instabilities down to $R(p_i) = 0$. Other fragments descriptors do not fall into any of the above groups. For example, the $N.074$ descriptor (nitrogen with double or triple bonds, mostly non-aromatic) carries more importance than any of the predictors of the fourth group and correlates with none of them. Having a positive sign, $N.074$ was found relevant in at least eight degradation paths (compound labels 2, 3, 9, 11, 30, 39, 46 and 47), where it participates, for example, in hydrolysis.

Another important predictor fragment is double bonded oxygen $O.058$ and $O.058.1$, e.g. in ketone-, aldehyde-, sulfoxide- or carboxyl-groups. Both $O.058$ and $O.058.1$ are directly involved in 27% of the surveyed degradations (labels 2, 3, 4, 5, 9, 16, 24, 28, 48 and 49). By increasing the electrophilicity of a neighboring atom, these fragments may also influence reactions without immediate participation: exemplification is found in the DARs of iprovalicarb (label 36) or metazachlor (label 44). In addition, carbonyl groups are strong hydrogen bond acceptors and can therefore alleviate water solubility, increasing the compound presence in the soil aqueous phase [89]. While coefficient signs for the binary predictor $O.058.1$ listed in Appendix B are certainly positive, the signs from the multi-level descriptor $O.058$ are rather misleading (see explanation in methods section 2.3.7).

Next, demethylation at a heteroatom position of a parent structure can be observed in 21% of all the 47 DARs pathways surveyed (labels 4, 14, 17, 21, 22, 23, 24, 37, 44 and 45). This reaction step is represented by yet another solitary predictor, namely the methyl moiety $C.005$. The latter resides at the third position of the RBS importance ranking (Figure [15]), emphasizing its relevance for degradation. For the same reasons that apply for $O.058$, its sign cannot be unambiguously identified given the current methodology.

Another descriptor, the heteroatom-methylene fragment $C.006$, participates in 10 DAR reaction steps (labels 7, 9, 10, 20, 21, 27, 28, 39, 43 and 48). The electron-withdrawing effect of the heteroatom being part of $C.006$ may be essential to the relevance of this descriptor. Other methylene fragment predictors lacking this heteroatom have not gained importance.

Finally, the UM-PPS rule $bt0005$ cannot be grouped either and was not undergone in any of the DAR pathways; its conceptual relevance awaits further attention.

In the absence of causal validation, the importance of a predictor may occasionally result from correlation to the response or to other important predictors by chance alone. This chance surely increases with the number of predictors - and that number is mostly large in “large p, small n problem”. The theoretical DRAGON descriptors [1] may be especially susceptible to such artifacts, given that (1) the majority of 66% of all utilized predictors stems from the DRAGON extension in predictor set 4 and that (2) the conceptual relevance is harder to grasp, since relating to individual chemical reactions is less straightforward than for the above described fragment predictors. The rest of the section deals with these DRAGON descriptors.

The most important one is $BLI$, the Kier benzene-likeliness index. $BLI$ relates a first-order valence connectivity index to the number of bonds in a molecule,
normalized to benzene. The index assumes a planar and cyclic molecule and each atom in a ring participating in an overlapping orbital continuum [57]. Belonging to the group of resonance indices [98], BLI effectively estimates the resonance energy of a molecule: the higher the BLI, the more stabilization from a delocalized molecular orbital (MO) can be expected. Thus, one would expect increasing resistance to degradation from increasing BLI - but the RBS sign of this predictor suggests the contrary. Instead, and for the given data set, BLI correlates with $S$ ($r_{spear} = 0.6$) and $S.107$ ($r_{spear} = 0.5$) of the second fragment group and with the further below discussed Lop index ($r_{spear} = 0.5$).

G3p, a directional WHIM symmetry index, is the second most important EDRA-GON descriptor for the given situation. G3p is calculated by a symmetry analysis of individual principal components of a weighted covariance matrix based on atomic coordinates [98]; weighting is based on atomic polarizability. In other words, G3p represents the symmetry in the variation of polarizability in the third spatial principal direction of overall polarizability for a given molecule. The accuracy of this descriptor depends on that of calculating atomic polarizabilities and of optimizing for the 3D structure of the molecule from its topology. G3p does not correlate strongly with any of the other predictors, except for O.058 ($r_{spear} = 0.5$). WHIM descriptors have met some application in modeling of pesticide partitioning [48], toxicity [99] or atmospheric degradation [49], but there are no accounts of their direct importance for soil biodegradation to date.

The Lopping centric index, Lop, is part of the category of centric indices, which quantify the compactness of a molecule - but, in contrast to the WHIM descriptors, on molecular topology alone (no conformational 3D optimization required). To calculate Lop, the terminal vertices (atoms) of the concerned molecular topology are pruned in several steps, until only one vertex remains (the atom representing the molecular center). Based on (1) the therefore required steps and (2) the atomic pruning counts per steps, a mean information content of a molecule is derived, yielding the Lop [9]. The compactness a molecule is with regards to branching of bonds, the higher its Lop. Correlation between Lop and any other of the 26 important predictors is negligible, except for BLI. Lop was found important in structure-toxicity relationship analysis of some aromatic compounds [75]; but, again, no accounts exist with regard to pesticide degradation.

Of the DRAGON GETAWAY descriptors, only $R_{4p}$ and $R_{5m}$ had some importance, but with ranking in the lower third of the 26 predictors. Calculation is based on the spatial leverage of individual atoms on the 3D deformation of a molecule, leading to a molecular influence matrix. The latter is combined with information on topological interatomic distances (autocorrelations of certain topological lag), and weighted by atom polarizability or atomic mass in the case of $R_{4p}$ and $R_{5m}$, respectively. Neither $R_{4p}$ nor $R_{5m}$ correlate strongly with any of the other predictors. Neither have so far revealed immediate relation to pesticide degradation - despite being widely applied in QSAR modeling [25]. Another theoretical descriptor, DISPp, is of second lowest rank among the 26 predictors, correlating with the above $R_{5m}$ descriptor ($r_{spear} = 0.5$). Again, its immediate relevance for pesticide degradation remains unresolved.

The last of the 26 predictors so far not discussed is the water content of soil
incubation, $WC$, which is unrelated to any of the other predictors. Confusingly, the coefficient sign of $WC$ is negative, implying an inverse relation to variation in $k_{PM}, k_{M1M2}$. This firmly contrasts other reports of pesticide degradation in soil, which all point at a positive effect of water content on biodegradation from e.g. facilitated mass transfer or bacterial growth [28][27]. Thus, the importance of $WC$ is of doubt and may be regarded as an artifact until confirmed otherwise.

On the background that (1) some predictors cluster regarding conceptual relevance, that (2) none of these clusters can be reduced to a single predictor without uncertainty in selection and (3) some predictors may simply be artifacts which gain their importance through random collinearity with other important predictors, one may derive a final model from a tuned-validated principal component regression (PCR) on the final RBS predictors [7]. PCR makes use of major variation in the covariates in relation to that of the response, somewhat reducing the need to sort out single predictors. Plus, supervised PCR is usually preceded by variable filtering [96], for which RBS poses an entirely new strategy yet to be explored.
4 Conclusion

Pesticide QSBR modeling was conducted in two stages. At a first stage, kinetic rate parameters were derived via inverse model optimization on $n = 87$ Speyer 2.2 soil data sets. Thereupon, variation in degradation rate parameters was modeled statistically under different settings of model complexity and predictor set complexity at a second stage.

At the first stage, four- and five box models were fitted to time series of radiolabeled parent compounds, metabolites, NER and volatiles via Genetic Algorithm optimization. The resulting first order rate constants of (1) degradation and (2) NER formation were highly right-skewed and correlated to each other, both for parent compounds and some metabolites. However, degradation rates were higher than rates of NER-formation in 78% of the cases. Although these findings are in line with previous studies, further research must estimate the uncertainties associated with calculating these rate parameters.

Secondly, variability in degradation rates was modeled by (1) LASSO shrinkage regression, (2) robust filtering and forward stepwise selection and (3) a hybrid of both. Filtering of (2) was achieved by a new methodology, named Recursive Bootstrap Subsampling (RBS). In RBS, bootstrap samples of the response are regressed on random subsets of predictors; meanwhile, unimportant predictors are recursively removed.

It appeared that improper (i.e. internal) cross-validation (CV) overestimates generalization performance of all models (1) to (3), yielding too complex (i.e. overfitted) models. In turn, such overfitted models lead to zero generalization performance under external leave-one-out (LOO) CV and $n << p$ situations. Regulation of model complexity improved the latter situation. However, it was furthermore shown that, under increasing size $p$ of the predictor set, methods (1) and (3) degraded in performance, even under regulation of model complexity. In contrast, (2) RBS was robust against increases in $p$. All models (1) to (3) showed bias in prediction, that is, overestimation of degradation rates below the mean and underestimation above the mean. These results emphasize not only the need for external instead of internal validation, but also the need for confining model complexity. In addition, a benchmark example for RBS&forward selection outperforming shrinkage regression was set in the context of predictor selection in $n << p$ situations.

Finally, a combination of LOO and RBS was utilized to identify important predictors and variation in importance from slight (LOO) changes in the data set. Thus, while BIOWIN descriptors were of little importance, an extended set of (partly collinear) fragment descriptors and partitioning constants had high relevance in explaining variation in pesticide degradation rates. Causal relevance of these fragments could partly be confirmed by comparison to observed degradation paths. On the other hand, another extension by geometrical and topological predictors did not improve model performance. Given the collinear-
ity of some of these predictors, using (1) principal component regression \[96\]
on (2) RBS-filtered predictors under (3) constrained and validated model com-
plexity with (4) bias correction is suggested as a prospective QSBR model for
predicting pesticide degradation rates in Speyer 2.2 soil.
5 Acknowledgments

The study was supported by Dr. Gerhard Görlitz / BAYER Crop Science (Monheim, Germany) with additional degradation reports. Dr. Markus Kalisch and Lukas Rosinus of the ETH statistical service provided a valuable discussion on shrinkage methods.
Table 5: List of investigated pesticides and metabolites from Speyer 2.2 soil degradation experiments. Compounds are sorted by order of inclusion into the study presented, as expressed by the labels in the first column. If there were replicate data for the same compound, several labels may apply to a single table row. Labels with an asterisk indicate compounds that were excluded after inverse modeling (see beginning of section 3.1).

“Pesticide” in the second column refers to the pesticide from which the compound stems; i.e. the compound may be the parent pesticide itself or a downstream metabolite. The Speyer 2.2 soil pesticide degradation data originate from either Draft Assessment Reports (DAR) or from degradation reports provided by BAYER Crop Science.

<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>1,2,4-Triazole 288-88-0 1,2,4-Triazole BAYER</td>
<td>1,6-Dihydro-1,2,4-triazine</td>
</tr>
<tr>
<td>2,3</td>
<td>Acrinathrin 101007-06-1 Acrinathrin DAR</td>
<td>cyano(3-phenoxyphenyl)methyl-3-[(1E)-3-[(1,1,1,3,3,3-hexafluoropropan-2-yl)oxy]-3-oxo-prop-1-en-1-yl]-2,2-dimethylcyclopropane-1-carboxylate</td>
</tr>
<tr>
<td>4</td>
<td>Amidosulfuron 120923-37-7 Amidosulfuron DAR</td>
<td>1-(4,6-dimethoxypyrimidin-2-yl)-3-[methanesulfonyl(methyl)sulfamoyl]urea</td>
</tr>
<tr>
<td>5</td>
<td>o-Desmethylanidosulfuron - Amidosulfuron DAR</td>
<td>1-(4-hydroxy-6-methoxypyrimidin-2-yl)-3-[[N-methylmethanesulfonamido]sulfonyl]urea</td>
</tr>
</tbody>
</table>

Continued on next page
Table 5 – continued from previous page

<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>IN-J0290 3289-50-7 Azimsulfuron, Nicosulfuron DAR</td>
<td>4,6-dimethoxypyrimidin-2-amine</td>
</tr>
<tr>
<td>7</td>
<td>Bromuconazole 116255-48-2 Bromuconazole DAR</td>
<td>1-[4-bromo-2-(2,4-dichlorophenyl)oxolan-2-yl]-methyl-1H-1,2,4-triazole</td>
</tr>
<tr>
<td>8,9</td>
<td>Cycloxidim 101205-02-1 Cycloxidim DAR</td>
<td>2-[(1E)-1-(ethoxyimino)butyl]-3-hydroxy-5-(thian-3-yl)cyclohex-2-en-1-one</td>
</tr>
<tr>
<td>10</td>
<td>BH517-T2SO2 - Cycloxidim DAR</td>
<td>3-(4-oxo-2-propyl-4,5,6,7-tetrahydro-1,3-benzo-azol-6-yl)-1H-1,1-dione</td>
</tr>
<tr>
<td>11</td>
<td>Cyfluvenamid 180409-60-3 Cyfluvenamid DAR</td>
<td>N-[(1Z)-[(cyclopropylmethoxy)imino][2,3-difluoro-6-(trifluoromethyl)phenyl]methyl]-2-phenylacetamide</td>
</tr>
<tr>
<td>12</td>
<td>Cyprodinil 121522-61-2 Cyprodinil DAR</td>
<td>4-cyclopropyl-6-methyl-N-phenylpyrimidin-2-amine</td>
</tr>
<tr>
<td>13</td>
<td>Cyromazine 66215-27-8 Cyromazine DAR</td>
<td>2-N-cyclopropyl-1,3,5-triazine-2,4,6-triamine</td>
</tr>
<tr>
<td>14</td>
<td>Dazomet 533-74-4 Dazomet DAR</td>
<td>3,5-dimethyl-1,3,5-thiadiazinane-2-thione</td>
</tr>
</tbody>
</table>
Table 5 – continued from previous page

<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,16</td>
<td>Desamino-Metamitron - Metamitron DAR / BAYER</td>
<td>3-methyl-6-phenyl-4,5-dihydro-1,2,4-triazin-5-one</td>
</tr>
<tr>
<td>17</td>
<td>Dicamba 1918-00-9 Dicamba DAR</td>
<td>3,6-dichloro-2-methoxybenzoic acid</td>
</tr>
<tr>
<td>18,19</td>
<td>Diethofencarb 87130-20-9 Diethofencarb DAR</td>
<td>Propan-2-yl N-(3,4-diethoxyphenyl)carbamate</td>
</tr>
<tr>
<td>20</td>
<td>Difenconazole 119446-68-3 Difenconazole DAR</td>
<td>1-(2-[2-chloro-4-(4-chlorophenoxy)phenyl]-4-methyl-1,3-dioxolan-2-ylmethyl)-1H-1,2,4-triazole</td>
</tr>
<tr>
<td>21</td>
<td>Dimethachlor 50563-36-5 Dimethachlor DAR</td>
<td>2-chloro-N-(2,6-dimethylphenyl)-N-(2-methoxyethyl)acetamide</td>
</tr>
<tr>
<td>22</td>
<td>BAS 656H 87674-68-8 Dimethenamid DAR</td>
<td>2-chloro-N-(2,4-dimethylthiophen-3-yl)-N-(1-methoxypropan-2-yl)acetamide</td>
</tr>
<tr>
<td>23</td>
<td>Omethoate - Omethoate DAR</td>
<td>Dimethyl{[(methylcarbamoyl)methyl]sulfanyl} phosphonate</td>
</tr>
<tr>
<td>24</td>
<td>Dimethomorph 110488-70-5 Dimethomorph DAR</td>
<td>(2E)-3-(4-chlorophenyl)-3-(3,4-dimethoxyphenyl)-1-(morpholin-4-yl)prop-2-en-1-one</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>BAS 505 F - Dimoxystrobin DAR</td>
<td>(2Z)-2-[2-(2,5-dimethylphenoxymethyl)phenyl]-2-(methoxyimino)-N-methylacetamide</td>
</tr>
<tr>
<td>26</td>
<td>Dithianon 3347-22-6 Dithianon DAR</td>
<td>5,10-dioxo-5H,10H-naphtho[2,3-b][1,4]dithine-2,3-dicarbonitrile</td>
</tr>
<tr>
<td>27</td>
<td>Fenbuconazole 114369-43-6 Fenbuconazole DAR</td>
<td>4-(4-chlorophenyl)-2-(2,5-dihydro-1,2,4-triazin-2-yl)-2-phenylbutanenitrile</td>
</tr>
<tr>
<td>28</td>
<td>Fenoxaprop-P-ethyl 71283-80-2 Fenoxaprop-P-ethyl DAR</td>
<td>Ethyl(2S)-2-4-[(6-chloro-1,3-benzoxazol-2-yl)oxy]phenoxypropanoate</td>
</tr>
<tr>
<td>29</td>
<td>Fenoxaprop-P - Fenoxaprop-P-ethyl DAR</td>
<td>(2S)-2-4-[(6-chloro-1,3-benzoxazol-2-yl)oxy]-phenoxypropanoic acid</td>
</tr>
<tr>
<td>30</td>
<td>Fipronil 120068-37-3 Fipronil DAR</td>
<td>5-amino-1-[2,6-dichloro-4-(trifluoromethyl)-phenyl]-4-(trifluoromethane)sulfinyl-1H--pyrazole-3-carbonitrile</td>
</tr>
<tr>
<td>31*</td>
<td>Fluazifop-P-butyl 79241-46-4 Fluazifop-P-butyl DAR</td>
<td>Butyl 2-(4-[5-(trifluoromethyl)pyridin-2-yl]oxyphenoxy)propanoate</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>32*</td>
<td>Fluazifop-P-acid - Fluazifop-P-butyl DAR</td>
<td>2-(4-[5-(trifluoromethyl)pyridin-2-yl]oxyphenoxy)propanoic acid</td>
</tr>
<tr>
<td>33</td>
<td>Fludioxonil 131341-86-1 Fludioxonil DAR</td>
<td>4-(2,2-difluoro-2H-1,3-benzodioxol-4-yl)pyrrole-3-carbonitrile</td>
</tr>
<tr>
<td>34</td>
<td>4064702 - Flufenoxuron DAR</td>
<td>4-[2-chloro-4-(trifluoromethyl)phenoxy]-2-fluorophenylurea</td>
</tr>
<tr>
<td>35</td>
<td>Flurprimidol 56425-91-3 Flurprimidol DAR</td>
<td>2-methyl-1-(pyrimidin-5-yl)-1-[4-(trifluoromethoxy)phenyl]propan-1-ol</td>
</tr>
<tr>
<td>36</td>
<td>DE535acid 95977-29-0 Haloxyfop-R DAR</td>
<td>(2S)-2-(4-[3-chloro-5-(trifluoromethyl)pyridin-2-yl]oxyphenoxy)propanoic acid</td>
</tr>
<tr>
<td>37</td>
<td>Haloxyfop-R 72619-32-0 Haloxyfop-R DAR</td>
<td>Methyl (2S)-2-(4-[3-chloro-5-(trifluoromethyl)pyridin-2-yl]oxyphenoxy)propanoate</td>
</tr>
<tr>
<td>38</td>
<td>CL266066 - Imazaquin DAR</td>
<td>2-[(1-carbamoyl-1,2-dimethylpropyl)carbamoyl]quinoline-3-carboxylic acid</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>Imidacloprid 138261-41-3 Imidacloprid DAR</td>
<td>2-chloro-5-((2E)-2-(nitroimino)imidazolidin-1-yl)methylpyridine</td>
</tr>
<tr>
<td>40</td>
<td>Iprovalicarb 140923-17-7 Iprovalicarb BAYER</td>
<td>Propan-2-yl N-[(1S)-2-methyl-1-[1-(4-methylphenyl)ethyl]carbamoylpropyl]carbamate</td>
</tr>
<tr>
<td>41</td>
<td>Isoproturon 34123-59-6 Isoproturon BAYER</td>
<td>3,3-dimethyl-1-[4-(propan-2-yl)phenyl]urea</td>
</tr>
<tr>
<td>42</td>
<td>Lenacil 2164-08-1 Lenacil DAR</td>
<td>3-cyclohexyl-1H,2H,3H,4H,5H,6H,7H--cyclopenta[d]pyrimidine-2,4-dione</td>
</tr>
<tr>
<td>43</td>
<td>Mefenpyr-diethyl 135590-91-9 Mefenpyr-diethyl BAYER</td>
<td>3,5-diethyl 1-(2,4-dichlorophenyl)-5-methyl-4,5-dihydro-1H-pyrazole-3,5-dicarboxylate</td>
</tr>
<tr>
<td>44,45</td>
<td>Mesosulfuron-methyl 208465-21-8 Mesosulfuron-methyl BAYER</td>
<td>Methyl 2-(((4,6-dimethoxypyrimidin-2-yl)-carbamoyl)aminosulfonyl)-4-(methanesulfonamidomethyl)benzoate</td>
</tr>
<tr>
<td>46,47</td>
<td>Metamitron 41394-05-2 Metamitron DAR, BAYER</td>
<td>4-Amino-3-methyl-6-phenyl-4,5-dihydro-1,2,4-triazin-5-one</td>
</tr>
</tbody>
</table>

Continued on next page
Table 5 – continued from previous page

<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>Metazachlor 67129-08-2 Metazachlor DAR</td>
<td>2-chloro-N-(2,6-dimethylphenyl)-N-(1H-pyrazol-1-ylmethyl)acetamide</td>
</tr>
<tr>
<td>49</td>
<td>Methiocarb 2032-65-7 Methiocarb BAYER</td>
<td>3,5-dimethyl-4-(methylsulfanyl)phenyl N-methylcarbamate</td>
</tr>
<tr>
<td>50</td>
<td>Methiocarb-Sulfoxide - Methiocarb BAYER</td>
<td>4-methanesulfinyl-3,5-dimethylphenyl N-methylcarbamate</td>
</tr>
<tr>
<td>51</td>
<td>Methomyl 16752-77-5 Methomyl DAR</td>
<td>(Z)-[1-(methylsulfanyl)ethylidene]amino N-methylcarbamate</td>
</tr>
<tr>
<td>52</td>
<td>Metsulfuron-methyl 74223-64-6 Iodosulfuron-methyl Sodium BAYER</td>
<td>Methyl 2-([(4-methoxy-6-methyl-1,3,5-triazin-2-yl)carbamoyl]aminosulfonyl)benzoate</td>
</tr>
<tr>
<td>53</td>
<td>Iodosulfuron-methyl Sodium 144550-36-7 Iodosulfuron-methyl Sodium BAYER</td>
<td>Sodium 2-([(5-iodo-2-(methoxycarbonyl)benzenesulfonylcarbamoyl)iminoc]-6-methoxy-4-methyl-1,2-dihydro-1,3,5-triazin-1-ide</td>
</tr>
<tr>
<td>54</td>
<td>Metosulam 139528-85-1 Metosulam DAR</td>
<td>N-(2,6-dichloro-3-methylphenyl)-5,7-dimethoxy-1,2,4</td>
</tr>
</tbody>
</table>

Continued on next page
Table 5 – continued from previous page

<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>DCM-ATSA - Metosulam DAR</td>
<td>5-amino-N-(2,6-dichloro-3-methylphenyl)-1,2,4-triazole-3-sulfonamide</td>
</tr>
<tr>
<td>56</td>
<td>7-OH- Metosulam - Metosulam DAR</td>
<td>N-(2,6-dichloro-3-methylphenyl)-7-hydroxy-5-methoxy-[1,2,4]triazolo[1,5-a]pyrimidine-2-sulfonamide</td>
</tr>
<tr>
<td>57</td>
<td>Metribuzin 21087-64-9 Metribuzin DAR</td>
<td>4-amino-6-tert-butyl-3-(methylsulfanyl)-4,5-dihydro-1,2,4-triazin-5-one</td>
</tr>
<tr>
<td>58</td>
<td>Oryzaline 19044-88-3 Oryzaline DAR</td>
<td>4-(dipropylamino)-3,5-dinitrobenzene-1-sulfonamide</td>
</tr>
<tr>
<td>59</td>
<td>Pencycuron 66063-05-6 Pencycuron BAYER</td>
<td>3-[(4-chlorophenyl)methyl]-3-cyclopentyl-1-phenylurea</td>
</tr>
<tr>
<td>60*</td>
<td>THS3995 - Pencycuron DAR</td>
<td>3-[(4-chlorophenyl)methyl]-3-cyclopentyl-1-phenylurea</td>
</tr>
<tr>
<td>61</td>
<td>Phosalone 2310-17-0 Phosalone DAR</td>
<td>Diethyl [[6-chloro-2-oxo-2,3-dihydro-1,3-benoxazol-3-yl]methyl]sulfanyl(sulfanylidene)-phosphonite</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>62, 63</td>
<td>Propamocarb-Hydrochloride 24579-73-5 Propamocarb-Hydrochloride BAYER</td>
<td>Diethyl[(6-chloro-2-oxo-2,3-dihydro-1,3-benzoxazol-3-yl)methyl]sulfanyl(sulfanyl-idene)phosphonite</td>
</tr>
<tr>
<td>64</td>
<td>Propargite 2312-35-8 Propargite DAR</td>
<td>[2-(4-tert-butylphenoxy)cyclohexyl]prop-2-yn-1-ylsulfite</td>
</tr>
<tr>
<td>65</td>
<td>Propyleneurea 1852-17-1 Propyleneurea BAYER</td>
<td>1,3-diazinan-2-one</td>
</tr>
<tr>
<td>66</td>
<td>Proquinazid 189278-12-4 Proquinazid DAR</td>
<td>6-iodo-2-propoxy-3-propyl-3,4-dihydroquinazolin-4-one</td>
</tr>
<tr>
<td>67</td>
<td>IN-MM671 - Proquinazid DAR</td>
<td>2-propoxy-3-propyl-3,4-dihydroquinazolin-4-one</td>
</tr>
<tr>
<td>68,70</td>
<td>Pyrimethanil-TP - Pyrimethanil-TP DAR</td>
<td>4,6-dimethylpyrimidin-2-amine</td>
</tr>
<tr>
<td>69</td>
<td>Pyrimethanil 53112-28-0 Pyrimethanil DAR</td>
<td>4,6-dimethyl-N-phenylpyrimidin-2-amine</td>
</tr>
<tr>
<td>Label</td>
<td>Compound, CAS number, Pesticide, Data source</td>
<td>IUPAC name</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>71</td>
<td>Quinmerac 90717-03-6 Quinmerac DAR</td>
<td>7-chloro-3-methylquinoline-8-carboxylic acid</td>
</tr>
<tr>
<td>72</td>
<td>Quinoclamine 2797-51-5 Quinoclamine DAR</td>
<td>2-amino-3-chloro-1,4-dihydropyrene-1,4-dione</td>
</tr>
<tr>
<td>73</td>
<td>Spirodiclofen 148477-71-8 Spirodiclofen DAR</td>
<td>3-(2,4-dichlorophenyl)-2-oxo-1-oxaspiro[4.5]dec-3-en-4-yl 2,2-dimethylbutanoate</td>
</tr>
<tr>
<td>74</td>
<td>Enol-BAJ2740 - Spirodiclofen BAYER</td>
<td>3-(2,4-dichlorophenyl)-4-hydroxy-1-oxaspiro[4.5]dec-3-en-2-one</td>
</tr>
<tr>
<td>75</td>
<td>Tebufenozide 112410-23-8 Tebufenozide DAR</td>
<td>N-tert-butyl-N’-[(4-ethylphenyl)carbonyl]-3,5-dimethylbenzohydrazide</td>
</tr>
<tr>
<td>76</td>
<td>Tembotrione 335104-84-2 Tembotrione BAYER</td>
<td>2-(2-chloro-4-methanesulfonyl)-3-[2,2,2-trifluoroethoxy)methyl]phenylcarbonyl)cyclohexane-1,3-dione</td>
</tr>
<tr>
<td>77</td>
<td>AE0456148 - Tembotrione BAYER</td>
<td>2-chloro-4-methanesulfonyle-3-[(2,2,2-trifluoroethoxy)methyl]benzoic acid</td>
</tr>
<tr>
<td>78</td>
<td>Thiacloprid 111988-49-9 Thiacloprid BAYER</td>
<td>[(2Z)-3-[(6-chloropyridin-3-yl)methyl]-1,3-thiazolidin-2-ylidene]aminocarbonitrile</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>79</td>
<td>KKO2254 - Thiacloprid BAYER</td>
<td>[(2Z)-3-[(6-chloropyridin-3-yl)methyl]-1,3-thiazolidin-2-ylidene]urea</td>
</tr>
<tr>
<td>80</td>
<td>Thiafluamide 142459-58-3 Thiafluamide BAYER</td>
<td>N-(4-fluorophenyl)-N-(propan-2-yl)-2-[5-(trifluoromethyl)-1,3,4-thiadiazol-2-yl]oxyacetamide</td>
</tr>
<tr>
<td>81</td>
<td>Tolyfluanid 731-27-1 Tolyfluanid DAR</td>
<td>([dichloro(fluoro)methyl]sulfanyl)(4-methylphenyl)sulfamoyl)dimethylamine</td>
</tr>
<tr>
<td>82</td>
<td>DMST - Tolyfluanid DAR</td>
<td>N-[dichloro(fluoro)methyl]sulfanyl-4-methylaniline</td>
</tr>
<tr>
<td>83,84</td>
<td>Triadimenol 55219-65-3 Triadimenol DAR</td>
<td>1-(4-chlorophenoxy)-3,3-dimethyl-1-(1H-1,2,4-triazol-1-yl)butan-2-ol</td>
</tr>
<tr>
<td>85</td>
<td>Triazophos 24017-47-8 Triazophos BAYER</td>
<td>Ethyl (1-phenyl-1H-1,2,4-triazol-3-yl)ethoxy(sulfanylidene)phosphonite</td>
</tr>
<tr>
<td>86</td>
<td>Triclopyr 55335-06-3 Triclopyr DAR</td>
<td>2-butoxyethyl2-[(3,5,6-trichloropyridin-2-yl)-oxy]acetate</td>
</tr>
<tr>
<td>87</td>
<td>Triclopyr Acid - Triclopyr DAR</td>
<td>2-[(3,5,6-trichloropyridin-2-yl)oxy]acetic acid</td>
</tr>
</tbody>
</table>

Continued on next page
Table 5 – continued from previous page

<table>
<thead>
<tr>
<th>Label</th>
<th>Compound, CAS number, Pesticide, Data source</th>
<th>IUPAC name</th>
</tr>
</thead>
<tbody>
<tr>
<td>88,89</td>
<td>Triticonazole 131983-72-7 Triticonazole DAR</td>
<td>(5E)-5-[(4-chlorophenyl)methylidene]-2,2-dimethyl-1-(1H-1,2,4-triazol-1-ylmethyl)cyclopentan-1-ol</td>
</tr>
<tr>
<td>90</td>
<td>Spiroxamine 118134-30-8 Spiroxamine DAR</td>
<td>(8-tert-butyl-1,4-dioxaspiro[4.5]decan-2-ylmethyl)(ethyl)propylamine</td>
</tr>
<tr>
<td>91</td>
<td>MetB - Chloridazon DAR</td>
<td>5-amino-4-chloro-2,3-dihydropyridazin-3-one</td>
</tr>
<tr>
<td>92</td>
<td>MetB1 - Chloridazon DAR</td>
<td>5-amino-4-chloro-2-methyl-2,3-dihydropyridazin-3-one</td>
</tr>
</tbody>
</table>
## Appendix B - List of important predictors

Table 6: Details on predictors found important for explaining variance in $k_{PM}$, listed in alphabetic order.
Symbols are: R : group linked via carbon, X : heteroatom (O,N,S,P,Se,halogens), Ar : aromatic group, Al : aliphatic group, .. : aromatic single bond (e.g. C..N in pyrrole), - or - - : aromatic (de-localized) single or double bond (e.g. benzene), = : double bond, # : triple bond, bonds omitted signify aliphatic single bonds.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description, Data type</th>
<th>Source, Reference</th>
<th>Count</th>
<th>Sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>alkenyl-H.1</td>
<td>Alkenyl hydrogen, binary data</td>
<td>KOCWIN [3]</td>
<td>5</td>
<td>-1.00</td>
</tr>
<tr>
<td>arom-H</td>
<td>Number of aromatic hydrogens, count data</td>
<td>BIOWIN [3]</td>
<td>69</td>
<td>-1.00</td>
</tr>
<tr>
<td>BLI</td>
<td>Kier benzene-likeliness index, continuous data</td>
<td>DRAGON topological descriptors</td>
<td>-</td>
<td>+1.00</td>
</tr>
<tr>
<td>bmultiNitro-arom</td>
<td>Aromatic ring with $\geq$ 2 nitrogens, binary data</td>
<td>KOCWIN [3]</td>
<td>10</td>
<td>-0.99</td>
</tr>
<tr>
<td>bt0005</td>
<td>Biotransformation rule: vic-unsubstituted Aromatic $\rightarrow$ vic-Di-hydroxyaromatic, count data</td>
<td>UM-PPS</td>
<td>38</td>
<td>-0.99</td>
</tr>
<tr>
<td>C.005</td>
<td>CH$_2$X, ordinal factor data</td>
<td>DRAGON centered fragments</td>
<td>27</td>
<td>-0.52</td>
</tr>
<tr>
<td>C.006</td>
<td>CH$_2$RX, ordinal factor data</td>
<td>DRAGON centered fragments</td>
<td>40</td>
<td>+0.71</td>
</tr>
</tbody>
</table>

Continued on next page
### Table 6 – continued from previous page

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description, Data type</th>
<th>Source, Reference</th>
<th>Count</th>
<th>Sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.042</td>
<td>X- -CH..X, binary data</td>
<td>DRAGON centered fragments 103</td>
<td>5</td>
<td>- 1.00</td>
</tr>
<tr>
<td>cbt0162</td>
<td>Biotransformation rule: disubstituted sulfide → disubstituted sulfoxide, binary data</td>
<td>UM-PPS 32</td>
<td>11</td>
<td>+ 1.00</td>
</tr>
<tr>
<td>cbt0259</td>
<td>Biotransformation rule: Monoalkylthiol → Aldehyde + H2S, binary data</td>
<td>UM-PPS 32</td>
<td>6</td>
<td>+ 0.99</td>
</tr>
<tr>
<td>DISPp</td>
<td>di COMMA2-value / weighted by atomic polarizabilities, continuous data</td>
<td>DRAGON geometrical descriptors 82 98</td>
<td>-</td>
<td>+ 1.00</td>
</tr>
<tr>
<td>ester.1</td>
<td>RCOOR or HCOOR (aromatic or aliphatic), binary data</td>
<td>BIOWIN 3</td>
<td>12</td>
<td>+ 1.00</td>
</tr>
<tr>
<td>G3p</td>
<td>Weighted 3rd directional WHIM index, continuous data</td>
<td>DRAGON WHIM descriptors 98</td>
<td>-</td>
<td>- 1.00</td>
</tr>
<tr>
<td>Koc-median</td>
<td>Median of organic carbon/water partitioning coefficients, continuous data</td>
<td>DAR 6</td>
<td>-</td>
<td>- 1.00</td>
</tr>
<tr>
<td>Koc-mean</td>
<td>Mean of organic carbon/water partitioning coefficients, continuous data</td>
<td>DAR 6</td>
<td>-</td>
<td>- 0.99</td>
</tr>
<tr>
<td>Lop</td>
<td>Lopping centric index, continuous data</td>
<td>DRAGON topological descriptors 99 98</td>
<td>-</td>
<td>+ 1.00</td>
</tr>
</tbody>
</table>

Continued on next page
Table 6 – continued from previous page

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description, Data type</th>
<th>Source, Reference</th>
<th>Count</th>
<th>Sign 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>N.073</td>
<td>Ar₂NH / Ar₃N / Ar₂NAl / R..N..R, binary data</td>
<td>DRAGON centered fragments [103]</td>
<td>12</td>
<td>-</td>
</tr>
<tr>
<td>N.074</td>
<td>R#N / R=N-, ordinal factor data</td>
<td>DRAGON centered fragments [103]</td>
<td>23</td>
<td>+</td>
</tr>
<tr>
<td>nRCOOR</td>
<td>RCOOR, non-aromatic R, binary data</td>
<td>DRAGON functionnal group counts [1][95]</td>
<td>8</td>
<td>+</td>
</tr>
<tr>
<td>S</td>
<td>Miscellaneous sulfur, binary data</td>
<td>Searched by author</td>
<td>29</td>
<td>+</td>
</tr>
<tr>
<td>S.107</td>
<td>R₂S or RS – SR, binary data</td>
<td>DRAGON centered fragments [103]</td>
<td>13</td>
<td>+</td>
</tr>
<tr>
<td>O.058</td>
<td>O= , ordinal factor data</td>
<td>DRAGON centered fragments [103]</td>
<td>68</td>
<td>-</td>
</tr>
<tr>
<td>O.058.1</td>
<td>O= , binary data</td>
<td>DRAGON centered fragments [103]</td>
<td>68</td>
<td>+</td>
</tr>
<tr>
<td>R4p</td>
<td>lag 4 R autocorellation, weight: atomic polarizabilities, continuous data</td>
<td>DRAGON GETAWAY descriptors [98]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>R5m</td>
<td>lag 5 R autocorellation, weight: atomic masses, continuous data</td>
<td>DRAGON GETAWAY descriptors [98]</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Continued on next page
### Table 6 – continued from previous page

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description, Data type</th>
<th>Source, Reference</th>
<th>Count(^1)</th>
<th>Sign(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WC</td>
<td>Soil water content, continuous data</td>
<td>DAR [6]</td>
<td>-</td>
<td>1.00</td>
</tr>
</tbody>
</table>

1 Number of compounds \(P(t), M(t)\) that contain this fragment.

2 (1) Mode of the sign of predictor effect on \(k_{PM}\) and (2) relative mode sign frequency in the last RBS recursion relative to the opposite, less frequent sign.
8 Appendix C - Statistical model coefficients

This appendix lists pseudovalidated model structures for each of the three model types and each of the four predictor set on explaining variation in $k_{PM}, k_{M1M2}$.

RBS & forward stepwise selection

**predictor set 1**  
$f_{boxcox}(k_{PM}) = -5.5 - 0.8 \cdot H_{arom} + 3.4 \cdot ester_{bin}$

**predictor set 2**  
$f_{boxcox}(k_{PM}) = -5.5 + 4.2 \cdot nRCOOR_{bin} + 3.4 \cdot S107_{bin} + 0.8 \cdot C005 + 2.3 \cdot O.058_{bin} - 0.6 \cdot C006 - 5.0 \cdot alkanylH_{bin} - 0.9 \cdot H050 + 0.2 \cdot pyridine_{bin}$

**predictor set 3**  
$f_{boxcox}(k_{PM}) = -5.5 - 1.9 \cdot C042 + 0.12 \cdot bleany + 0.6 \cdot Thiocarbamate_{bin} + 1.7 \cdot S - 0.7 \cdot CH2_{cyclic,bin} - 3.8 \cdot nitrile_{bin} + 2.7 \cdot N068 + 3.1 \cdot cbt0162 - 0.4 \cdot methyl + 0.7 \cdot nArOR - 0.5 \cdot K_{oc,mean} + 6.1 \cdot nRCOOR - 0.5 \cdot C006 - 2.2 \cdot N075_{bin} - 1.6 \cdot N074_{bin} + 0.6 \cdot Cl_{arom}$

**predictor set 4**  
$f_{boxcox}(k_{PM}) = -5.5 + 3.0 \cdot S107 + 5.7 \cdot nRCOO + 0.5 \cdot C005 - 0.8 \cdot O058 + 0.1 \cdot cbt0259 + 1.0 \cdot R4p - 2.3 \cdot C042 - 0.5 \cdot WC - 1.1 \cdot K_{oc,mean} - 1.1 \cdot Aliphaticringcount - 1.3 \cdot H049 - 0.6 \cdot P2u$

**LASSO**

**predictor set 1**  
$f_{boxcox}(k_{PM}) = -5.5 - 0.5 \cdot H_{arom} - 0.0 \cdot CH2_{linear} + 2.3 \cdot ester_{bin} - 0.0 \cdot C4singlenoH_{bin} - 0.6 \cdot NH2_{arom,bin} + 0.3 \cdot pyridine_{bin} + 0.7 \cdot Thiocarbamate_{bin} + 0.6 \cdot methyl_{bin} - 0.6 \cdot CH_{cyclic,bin} - 1.4 \cdot H_{alkenyl,bin}$

**predictor set 2**  
$f_{boxcox}(k_{PM}) = -5.5 + 0.4 \cdot cbt0012 + 0.3 \cdot cbt0029 + 0.3 \cdot cb0036 + 0.7 \cdot cbt0162 - 0.1 \cdot cbt0242 - 0.2 \cdot cbt0243 + 0.6 \cdot cbt0259 + 0.2 \cdot cbt0404 + 0.2 \cdot arom._ther - 0.1 \cdot CH2_{linear} + 0.6 \cdot ester_{bin} - 1.1 \cdot C4singlenoH_{bin} + 1.1 \cdot kneone_{bin} + 0.2 \cdot Thiocarbamate_{bin} + 0.4 \cdot methyl_{bin} - 3.3 \cdot H_{alkenyl,bin} + 0.0 \cdot bN_{nonfused} + 0.8 \cdot bmiscS0_{bin} + 1.0 \cdot bquinone + 0.0 \cdot btany - 0.2 \cdot Aliphaticringcount + 4.6 \cdot nRCOOR_{bin} + 2.3 \cdot nN,C,N. + 0.5 \cdot nArNH2 - 0.6 \cdot nROO + 0.8 \cdot nRSR + 0.3 \cdot nH_{Bonds} - 0.7 \cdot C011 - 1.0 \cdot C028 - 0.9 \cdot C039 - 1.2 \cdot C042 - 0.1 \cdot H047 - 0.8 \cdot H049 - 1.5 \cdot H050 - 0.3 \cdot H051 - 0.2 \cdot H053 - 1.2 \cdot O057 - 0.3 \cdot O059 + 0.5 \cdot N068 + 2.0 \cdot nRCOOR_{bin} - 0.3 \cdot nCOn_{bin} - 0.1 \cdot C002_{bin} - 1.0 \cdot C005_{bin} - 0.8 \cdot C026_{bin} + 0.6 \cdot O058_{bin} - 1.7 \cdot N074_{bin} - 0.4 \cdot K_{oc,mean} - 0.4 \cdot K_{oc,MC1}$

**predictor set 3**  
$f_{boxcox}(k_{PM}) = -5.5 + 0.1 \cdot cbt0012 + 0.8 \cdot cbt0162 - 0.1 \cdot cbt0243 + 0.4 \cdot cbt0259 + 0.8 \cdot cbt0404 - 0.0 \cdot CH2_{cyclic} + 0.4 \cdot ester_{bin} - 0.8 \cdot C4singlenoH_{bin} + 0.4 \cdot Thiocarbamate_{bin} + 0.2 \cdot methyl_{bin} - 2.7 \cdot H_{alkenyl,bin} + 0.2 \cdot S + 0.4 \cdot bmiscS0 + 0.5 \cdot bquinone + 3.8 \cdot nRCOOR_{bin} + 0.5 \cdot nN,C,N. + 0.1 \cdot C005 -$
0.7*C.028 + 0.3*C.029 - 1.0*C.042 - 0.5*H.049 - 0.8*H.050 - 0.1*H.051 + 0.6
O.057 + 0.5*N.068 + 2.2*S.107bin - 0.2*nCconjbin - 1.1*C.005bin + 0.9*O.058bin
- 1.1*N.074bin - 0.1*temp - 0.2*bio_start - 0.2*WC - 0.4*Koc,mean - 0.1*Koc,MCI

**Predictor set 4** \( f_{boxcox}(k_{PM}) = -5.5 + 0.2*cbt0162 + 0.6*Thiocarbamate_{bin} + 
0.0*methyl_{bin} - 2.0*H_{alkenyl,bin} + 0.3*S + 3.2*nRCOOR_{bin} - 0.0*nHDon + 
0.3*C.005 - 0.2*C.028 - 1.1*C.042 - 0.3*H.050 + 2.0*S.107bin - 0.3*C.005bin + 
0.7*O.058bin - 0.5*N.074bin - 0.0*bio_start - 0.2*WC - 0.3*P2u - 0.0*G2u - 0.0*
Gu + 0.0*HATS2u - 0.0*HATS7m + 0.1*R5m - 0.3*Koc,mean - 0.1*Koc,median

**Hybrid: RBS & LASSO**

**Predictor set 1** \( f_{boxcox}(k_{PM}) = -5.5 - 0.5*H_{arom} + 0.6*methyl_{bin} - 1.4*
H_{alkenyl,bin} - 0.6*C.H_{cyclic,bin} + 0.3*pyridine_{bin} + 0.7*Thiocarbamate_{bin} - 
0.6*arom.NH2_{bin} - 0.0*C4single.noH_{bin}

**Predictor set 2** \( f_{boxcox}(k_{PM}) = -5.5 + 3.3*nRCOOR_{bin} + 0.4*C.005 + 
1.4*cbt0162 + 2.0*O.058bin + 0.6*S - 0.3*N.074 - 1.2*C.042 - 0.5*C.006 - 
0.3*N.073 - 4.0*H_{alkenyl,bin} - 0.5*C.030 + 0.8*methyl_{bin} - 0.7*H.050 - 
0.0*nCconj - 0.5*H.049 + 1.4*pyridine_{bin} - 0.5*cbt0242 - 0.1*C.H_{cyclic,bin}

**Predictor set 3** \( f_{boxcox}(k_{PM}) = -5.5 + 3.6*nRCOOR_{bin} + 0.3*C.005 - 
0.0*N.074 + 1.9*cbt0162 + 0.3*S + 1.4*O.058bin - 0.4*WC - 0.1*N.073 - 
1.0*C.042 - 0.5*C.006 + 0.0*multiNitroArom_{bin} - 0.8*Koc,mean - 3.3*
H_{alkenyl,bin} - 0.0*C.030 + 0.5*methyl_{bin} - 0.5*H.050bin + 1.0*pyridine_{bin} - 
0.3*nCconj - 0.9*H.049 - 0.3*bio_start - 0.7*nHDon

**Predictor set 4** \( f_{boxcox}(k_{PM}) = -5.5 + 4.3*nRCOOR + 0.3*C.005 - 0.3*
O.058 + 1.1*cbt0162 + 0.1*BLI - 0.7*N.073 + 0.7*cbt0259 - 0.5*H_{arom} + 
0.5*R4p + 0.4*G3p - 0.8*C.042 - 0.5*WC - 0.6*Koc,mean - 0.2*C.030 + 
0.4*R5m - 0.1*C.006 - 0.9*Aliphaticringcount - 0.7*H.049 - 0.5*P2u
9 Appendix D - RK4 C++ code

C++ code for the calculation of the 4th order Runge-Kutta (RK4) method, exemplified for the five pool model. Few parts of the code were adapted from [69].

```c++
#include <iostream>
#include <stdlib.h>
#include <math.h>
#include <R.h>

using namespace std;

void RK(int n, double x, double *y, double h, double *y1, double kpm, double kpn, double km1n, double km1m, double km2n, double km2v) {
    double c1[10], c2[10], c3[10], c4[10], yy[10], h2; int i;
    F(x, y, c1, kpm, kpn, km1m, km1n, km2n, km2v);
    h2=h/2.0;
    for (i=0; i<n; i++) yy[i]=y[i]+h2*c1[i];
    F(x+h2, yy, c2, kpm, kpn, km1m, km1n, km2n, km2v);
    for (i=0; i<n; i++) yy[i]=y[i]+h2*c2[i];
    F(x+h2, yy, c3, kpm, kpn, km1m, km1n, km2n, km2v);
    for (i=0; i<n; i++) yy[i]=y[i]+h*c3[i];
    F(x+h, yy, c4, kpm, kpn, km1m, km1n, km2n, km2v);
    for (i=0; i<n; i++) y1[i]=y[i]+h*(c1[i]+2.0*c2[i]+2.0*c3[i]+c4[i])/6.0;
}

void F(double x, double *y, double *yp, double kpm, double kpn, double km1n, double km1m, double km2n, double km2v) {
    yp[0] = -kpm*y[0]-kpn*y[0]; // Pt
    yp[1] = km1m*y[4]-y[1]*km2n-y[1]*km2v; // M2t
    yp[3] = y[1]*km2v; // Vt
    yp[4] = kpm*y[0]-y[4]*km1n-y[4]*km1m; // M1t
}

extern "C" // ... extern C call for usage in R after .dll generation
{
    void rk4m (double *P0, double *kpm, double *kpn, double *km1n, double *km1m, double *km2n, double *km2v, double *t, int *leng, double *Ftsolve)
    {
        int n=5; // size of equation system
    }
```
int i, k, w = 0; // loop counters
int x0 = 0;
int xi = t[n - 1];
int kl = xi - x0;
int finess = 40;
double h = (double) 1 / (double) finess;
double Y[5], Y1[5];
double X = (double) x0;
// initial concentrations:
Y[0] = *P0; // Pt
Y[1] = 0; // M2t
Y[2] = 0; // Nt
Y[3] = 0; // Vt
Y[4] = 0; // M1t

// numerical integration:
//(1) start with first iteration step = initialize
// equation system
RK(n, X, Y, h, Y1, *kpm, *kpn, *km1n, *km1m, *km2n, *km2v);
if (t[w] == (float) X)
{
    Ftsolve[w] = Y[0];
    Ftsolve[*leng + w] = Y[1];
    Ftsolve[2 * *leng + w] = Y[2];
    Ftsolve[3 * *leng + w] = Y[3];
    Ftsolve[4 * *leng + w] = Y[4];
    w++;
}
//(2) do integration steps after initialization (1)
for (k = 1; k <= kl * finess; k++)
{
    X = X + (double) h;
    for (i = 0; i < n; i++)
    {
        Y[i] = Y1[i];
    }
    RK(n, X, Y, h, Y1, *kpm, *kpn, *km1n, *km1m, *km2n, *km2v);
    if ((float) t[w] == (float) X)
    {
        Ftsolve[w] = Y[0];
        Ftsolve[*leng + w] = Y[1];
        Ftsolve[2 * *leng + w] = Y[2];
        Ftsolve[3 * *leng + w] = Y[3];
        Ftsolve[4 * *leng + w] = Y[4];
        w++;
    }
}; // extern "C"
References

[1] VCCLAB, Virtual Computational Chemistry Laboratory, 2005. \url{http://www.vcclab.org}


REFERENCES


[101] UFZ. Using reach as an opportunity to find alternatives to animal experiments, 2005.


[105] E. Willighagen. genalg: R Based Genetic Algorithm, 2005. \url{http://cran.r-project.org/}
