

Ph.D. Thesis  
Doctor of Philosophy

| DTU Compute  
Department of Applied Mathematics and Computer Science

# Characterization of absorption enhancers for orally administered therapeutic peptides in tablet formulations

## Applying statistical learning

Soren Havelund Welling

Kongens Lyngby 2016



**DTU Compute**  
**Department of Applied Mathematics and Computer Science**  
**Technical University of Denmark**

Matematiktorvet  
Building 303B  
2800 Kongens Lyngby, Denmark  
Phone +45 4525 3031  
[compute@compute.dtu.dk](mailto:compute@compute.dtu.dk)  
[www.compute.dtu.dk](http://www.compute.dtu.dk)

# Summary

---

In terms of convenience, to develop a successful oral formulation of Insulin for treatment of type-2 diabetes patients would be big mile stone. Besides protecting insulin from enzymatic cleavage, epithelial uptake is a significant barrier to overcome. Absorption enhancers are needed to ensure even small 5% of insulin is taken up. The major class of Absorption Enhancers is surfactant-like enhancers and is thought to promote absorption by mildly perturbing the epithelial membranes of in small intestine. The Caco-2 (Carcinoma Colon) cell line can grow an artificial epithelial layer, and is used to test the potency of new absorption enhancers. In Caco-2 cell all reagents are pre-dissolved, therefore this study cannot predict critical solubility issues in the final tablet formulation. This project was aimed to identify new absorption enhancers, that are both potent and sufficiently soluble. Quantitative structural activity relationship (QSAR) modeling is an empiric approach to learn relationship between molecular formulas and the biochemical properties using statistical models. Molecular formulas, which are graphs of molecules, are first encoded with different descriptor algorithms and translated into molecular descriptors. Some simple molecular descriptors simply count types of molecules or functional groups, more advanced algorithms simulate the 3D structure of the molecule to e.g. predict the dipole moment. A public data set disclosing the potency of 42 absorption enhancers in Caco-2 was used to build a QSAR model to screen for new potent compounds. A proof-of-concept and likely the first QSAR model to predict absorption enhancement was published. Likewise was another QSAR model built to predict solubility of absorption enhancers. After initial proof-of-concept modeling, focus was shifted towards more academic aspects of the project. Supervised regression was used to learn relationships between molecular descriptors and potency or solubility. The random forest algorithm was found superior to multiple linear regression in terms of cross validated prediction accuracy. However, unlike multiple linear regression, an explicitly stated random forest model is complex, and therefore difficult to interpret and communicate. Any supervised regression model can be understood as a high dimensional surface connecting any possible combination of molecular properties with a given prediction. For random forests it was discovered that a method, feature contributions, was especially useful to the decompose model structure into isolated main effects and interactions effects that could be visualized and understood more easily. A novel method forest floor and R package forestFloor was developed to improve interpretation of random forest models. Better interpretation of random forest models is an exciting interdisciplinary field, as it allows investigators of many background to find fairly complicated relationships in data sets without specifying what parameters to estimate. Forest floor was used to explain how

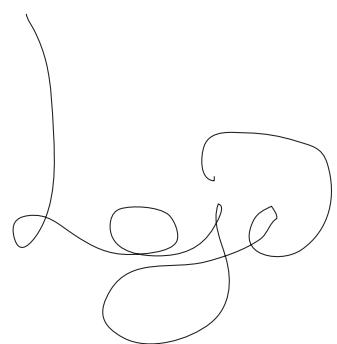
potency and solubility were predicted by the random forest model. Demonstrating how to train random forest models to predict solubility from molecular descriptors have been published several times in the scientific literature. This thesis have with forest floor contributed to actually describe the shape of these model structures. In order to develop better molecular descriptors, to understand how these are used in a complex model structure is an important step.

# Preface

---

This Ph.D. thesis was prepared at the department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfillment of the requirements for acquiring a Industrial Ph.D. degree in Applied Mathematics and Computer Science.

Kongens Lyngby, June 8, 2016

A handwritten signature in black ink, appearing to read "Soren Havelund Welling". The signature is fluid and cursive, with a large, sweeping loop on the right side.

Soren Havelund Welling



# Acknowledgements

---

Thanks to Lisette, Trine, the many pacakage writers of R



# Contents

---

<b>Summary</b>	i
<b>Preface</b>	iii
<b>Acknowledgements</b>	v
<b>Contents</b>	vii
<b>Todo list</b>	ix
<b>1 Introduction</b>	1
1.1 Torquent Arcu . . . . .	2
1.2 Luctus . . . . .	4
1.3 Sollicitudin vestibulum . . . . .	5
<b>2 Long chapter <math>\phi \wedge \sigma</math> title with <math>\pi</math>, very long title, and also <math>math = \sigma</math></b>	7
<b>3 measureAbsorption</b>	9
<b>4 predictPotency</b>	19
<b>5 interpretTheForest</b>	29
<b>6 structureOfSolubility</b>	67
<b>7 Heading on Level 0 (chapter)</b>	79
7.1 Heading on Level 1 (section) . . . . .	79
7.2 Lists . . . . .	80
<b>8 Conclusion</b>	83
<b>A An Appendix</b>	85
<b>Bibliography</b>	125



# Todo list

---

■ 1.0 (1) Make a cake . . . . .	1
■ 1.0 (2) Do it now . . . . .	1
Figure: 1.0 (3) This is some text that is with the todo and in the figure . . . . .	2
Figure: 1.1 (4) This is some text that is with the todo and in the figure . . . . .	3
Figure: 1.1 (5) This is some text that is with the todo and in the figure . . . . .	3
Figure: 1.2 (6) This is some text that is with the todo and in the figure . . . . .	5

x

---

# CHAPTER 1

## Introduction

---

- Upright shape
- *Italic shape*
- Slanted shape
- SMALL CAPS SHAPE
- Medium series
- **Bold** sereies
- Roman family
- Sans serif family
- Typewriter family

I love to write special characters like øæå inside my TEX document. Also á, à, ü, û, ë, ê, î, ï could be nice. So what about the " " character. What about ° é ® † ¥ ü | œ ‘ @ ö ä ¬ ‹ « © f ß ª Ω … ç √ ñ µ , · ¡ “ £ ∞ ™ [ ] ± ’.

Some dashes --, and the latex form ---

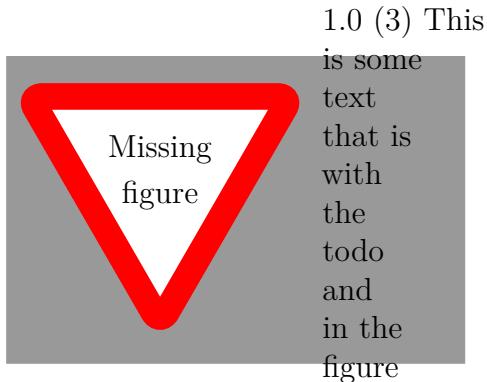
$$x = \mathbf{x}, \mathbf{x}, x, x_{1_{2_{3_4}}}^{1^2^{3^4}} \cdot \text{hello} * \text{hello worldmy world} \ddot{\text{u}} \text{third worldt}$$

1.0 (1)  
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum [Ada80].

Mauris id quam non magna fermentum malesuada id mattis lorem. In a dapibus neque. Etiam lacinia dui, malesuada ac eleifend imperdiet, imperdiet ut ipsum. Vestibulum id ultricies est. Phasellus augue mauris, semper a luctus vel, faucibus in risus. Fusce commodo augue quis elit sagittis non viverra turpis bibendum. Nunc placerat sem non sapien malesuada malesuada ullamcorper orci luctus [Ada80]. Morbi pharetra ligula integer mollis mi nec neque ultrices vitae volutpat leo ullamcorper. In at tellus magna. Curabitur quis posuere purus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Suspendisse tristique placerat feugiat. Aliquam vitae

1.0 (2)  
a cake

1.0 (2)  
it now



**Figure 1.1:** This is my special figure. Aliquam ultricies, arcu quis tempor rhoncus, tellus nisl tempus justo, condimentum tempor erat odio ac purus. Integer quis ipsum felis. Aliquam volutpat, leo ac consequat egestas, lectus lacus adipiscing quam, id iaculis dolor quam in erat. Phasellus tempor interdum arcu quis vestibulum.

est at enim auctor ultrices eleifend a urna. Donec non tincidunt felis. Maecenas at suscipit orci. See Figure 1.1.

Fusce id suscipit sem. Aliquam venenatis nibh nec nisl luctus vel consectetur neque dapibus. Nulla feugiat egestas turpis, ac viverra eros cursus sit amet. Cras tincidunt felis vel tellus ultrices condimentum. Quisque vehicula, arcu vitae interdum dignissim, purus tortor cursus libero, sit amet accumsan quam magna in neque. Phasellus luctus leo odio. Aliquam ultricies, arcu quis tempor rhoncus, tellus nisl tempus justo, condimentum tempor erat odio ac purus. Integer quis ipsum felis. Aliquam volutpat, leo ac consequat egestas, lectus lacus adipiscing quam, id iaculis dolor quam in erat. Phasellus tempor interdum arcu quis vestibulum. Pellentesque sit amet augue purus. See Table 1.1.

## 1.1 Torquent Arcu

Curabitur condimentum suscipit arcu, sit amet convallis urna pellentesque ac. Quisque fringilla tincidunt risus nec accumsan. Curabitur vel sagittis ante. Integer eget placerat leo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis risus in nulla fermentum pellentesque dictum et erat. Nulla vel pretium nunc. Integer tortor lorem, suscipit sit amet ultricies non, porta at metus. Sed pharetra, ante facilisis interdum porta, mi dolor fringilla quam, ac porttitor urna dolor

h	h	h
e	e	e

**Table 1.1:** This is a caption to the table.

quis massa. Proin viverra semper tincidunt. Vivamus pulvinar pharetra condimentum. Pellentesque rutrum mollis tellus ac scelerisque.

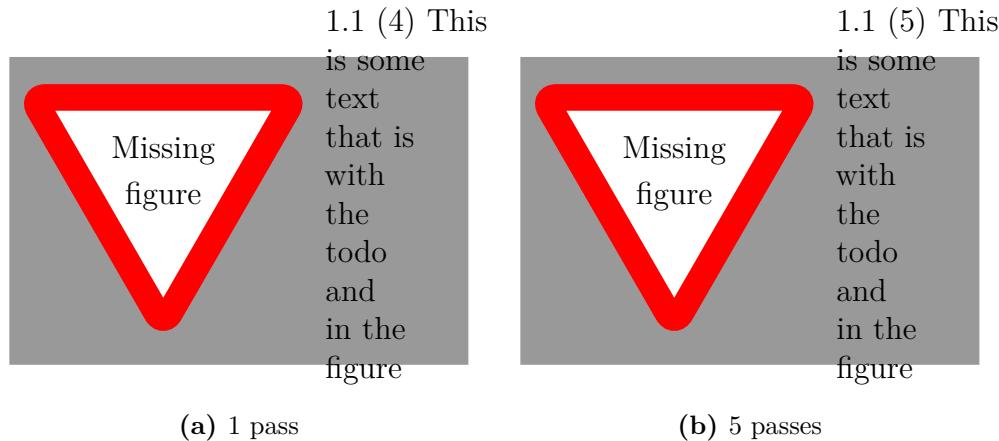
### 1.1.1 Vestibulum

Mauris luctus sollicitudin vestibulum. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos Figure 1.2(b). Duis eu nisl nec turpis porttitor bibendum eget sed orci. Aliquam consequat lorem a dui viverra porta facilisis augue rutrum. Cras luctus tellus in lectus egestas eu consequat magna cursus. Aenean aliquam neque a nibh elementum ornare. Integer eleifend imperdiet commodo. Morbi auctor, dui vel laoreet congue, purus est accumsan augue, sit amet feugiat neque nisl vel lorem. Curabitur ante sem, lacinia id adipiscing quis, viverra tristique nulla. Pellentesque ullamcorper pellentesque metus varius facilisis. Cras ac dui id odio tempor scelerisque. Curabitur a egestas risus. Pellentesque quis velit in sapien accumsan auctor. Phasellus aliquam, sapien eget lobortis volutpat, libero metus porttitor nisl, sed hendrerit urna dolor nec mi. See Listing 1.1.

```

1 # This is a comment
2 import easy
3 str = "I am a string"
4 str2 = "Now i have an awsome string with ' ' `` which are not TeX'ed"
5 str3 = "What about awsome unicode characters? Like ", , ", Ω, ¢. \" This"
6 def fib(n):
7     if n == 0:
8         return 0
9     elif n == 1:
10        return 1
11    else:
12        return fib(n-1) + fib(n-2)

```



**Figure 1.2:** loop performance comparison.

```

13 str4 = "Yes it is possible with 80 charactes. Which this string proves.  

      Wiii."  

14 str5 = "It adjusts according to the spine"

```

**Listing 1.1:** Fibonacci.

## 1.2 Luctus

Praesent et pellentesque arcu. Phasellus venenatis mi eu lorem convallis et iaculis ante aliquet. Aenean rhoncus placerat metus, vel convallis leo suscipit eu. Integer dapibus venenatis commodo. Cras laoreet faucibus sem nec luctus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Cras consectetur lacinia dolor at gravida. Phasellus ipsum arcu, vulputate fermentum ultricies eget, tempor eu odio. Aenean accumsan vestibulum risus a mattis. See it on Algorithm 1.

---

**Algorithm 1** Modified mini-batch  $K$ -means

---

```

1: Given:  $K$ , mini-batch size  $B$ , iterations  $T$ , dataset  $X$ , correlation matrix  $P$ .  

2: Initialize  $C = \{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \dots, \mathbf{c}^{(K)}\}$  with random  $\mathbf{x}$ 'es picked from  $X$ .  

3:  $A \leftarrow B \cdot T$  sorted random indexes to  $X$ , denoted  $a_1, a_2, \dots, a_{B \cdot T}$ .  

4:  $X' \leftarrow \{\mathbf{x}^{(a_1)}, \mathbf{x}^{(a_2)}, \dots, \mathbf{x}^{(a_{B \cdot T})}\}$  ▷ Cache all points  

5: size  $\leftarrow 0$   

6: for  $i = 1$  to  $T$  do  

7:    $M \leftarrow B$  examples picked randomly from  $X'$   

8:   for  $\mathbf{x} \in M$  do ▷ Assignment step  

9:      $\mathbf{d}[\mathbf{x}] \leftarrow f(C, \mathbf{x}, P)$  ▷ Cache closest center  

10:  end for  

11:  for  $\mathbf{x} \in M$  do ▷ Update step  

12:     $\mathbf{c} \leftarrow \mathbf{d}[\mathbf{x}]$  ▷ Get cached center for current  $\mathbf{x}$   

13:     $\text{size}[\mathbf{c}] \leftarrow \text{size}[\mathbf{c}] + 1$  ▷ Update cluster size  

14:     $\eta \leftarrow \frac{1}{\text{size}[\mathbf{c}]}$  ▷ Get learning rate  

15:     $\mathbf{c} \leftarrow (1 - \eta)\mathbf{c} + \eta\mathbf{x}$  ▷ Take gradient step  

16:  end for  

17: end for  

18: return  $C, \text{size}$ 

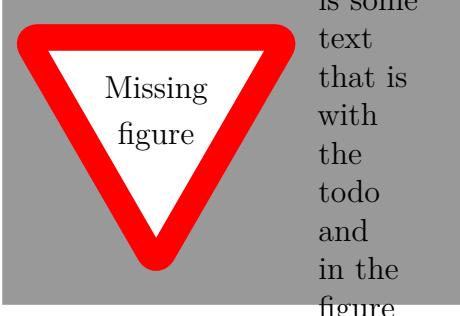
```

---

Fusce id suscipit sem. Aliquam venenatis nibh nec nisl luctus vel consectetur neque dapibus. Nulla feugiat egestas turpis, ac viverra eros cursus sit amet. Cras tincidunt felis vel tellus ultrices condimentum. Quisque vehicula, arcu vitae interdum dignissim, purus tortor cursus libero, sit amet accumsan quam magna in neque. Phasellus luctus leo odio. Aliquam ultricies, arcu quis tempor rhoncus, tellus nisl tempus justo, condimentum tempor erat odio ac purus. Integer quis ipsum felis. Aliquam volutpat, leo ac consequat egestas, lectus lacus adipiscing quam, id iaculis dolor quam in erat. Phasellus tempor interdum arcu quis vestibulum. Pellentesque sit amet augue purus. Curabitur

1.2 (6) This

is some  
text  
that is  
with  
the  
todo  
and  
in the  
figure



**Figure 1.3:** This is the caption I wrote.

condimentum suscipit arcu, sit amet convallis urna pellentesque ac. Quisque fringilla tincidunt risus nec accumsan. Curabitur vel sagittis ante. Integer eget placerat leo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis risus in nulla fermentum pellentesque dictum et erat. Nulla vel pretium nunc. Integer tortor lorem, suscipit sit amet ultricies non, porta at metus. Sed pharetra, ante facilisis interdum porta, mi dolor fringilla quam, ac porttitor urna dolor quis massa. Proin viverra semper tincidunt. Vivamus pulvinar pharetra condimentum. Pellentesque rutrum mollis tellus ac scelerisque.

## 1.3 Sollicitudin vestibulum

Mauris luctus sollicitudin vestibulum. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Duis eu nisl nec turpis porttitor bibendum eget sed orci. Aliquam consequat lorem a dui viverra porta facilisis augue rutrum. Cras luctus tellus in lectus egestas eu consequat magna cursus. Aenean aliquam neque a nibh elementum ornare. Integer eleifend imperdiet commodo. Morbi auctor, dui vel laoreet congue, purus est accumsan augue, sit amet feugiat neque nisl vel lorem. Curabitur ante sem, lacinia id adipiscing quis, viverra tristique nulla. Pellentesque ullamcorper pellentesque metus varius facilisis. Cras ac dui id odio tempor scelerisque. Curabitur a egestas risus. Pellentesque quis velit in sapien accumsan auctor. Phasellus aliquam, sapien eget lobortis volutpat, libero metus porttitor nisl, sed hendrerit urna dolor nec mi.

Praesent et pellentesque arcu. Phasellus venenatis mi eu lorem convallis et iaculis ante aliquet. Aenean rhoncus placerat metus, vel convallis leo suscipit eu. Integer dapibus venenatis commodo. Cras laoreet faucibus sem nec luctus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Cras consectetur lacinia dolor at gravida. Phasellus ipsum arcu, vulputate fermentum ultricies eget, tempor eu odio. Aenean accumsan vestibulum risus a mattis.

---

```
1 # This is a comment
2 import easy
3 str = "I am a string"
4 str2 = "Now i have an awsome string with ' ' `` which are not TeX'ed"
5 str3 = "What about awsome unicode characters? Like ", , "Ω, §. \" This"
6 def fib(n):
7     if n == 0:
8         return 0
9     elif n == 1:
10        return 1
11    else:
12        return fib(n-1) + fib(n-2)
13 str4 = "Yes it is possible with 80 charactes. Which this string proves.
14      Wiii."
15 str5 = "It adjusts according to the spine"
```

**Listing 1.2:** Fibonacci2.

# CHAPTER 2

## Long chapter $\phi \wedge \sigma$ title with $\pi$ , very long title, and also $math = \sigma$

---

My favorite RFC is [Wai99]. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. What about some inline sans serif text?.  $x = 4, y = 7$  which means that  $\sqrt{4} = 2$ .

Sans serif testing:

- $\pi$
- $\pi$
- $\pi$
- *italic*
- ***bold italic***
- **bold**
- `teletype`
- Math Sans Serif
- Text Sans Serif



# CHAPTER 3

## measureAbsorption

---

Citric acid reefnoreifoe



Contents lists available at ScienceDirect

## European Journal of Pharmaceutics and Biopharmaceutics

journal homepage: [www.elsevier.com/locate/ejpb](http://www.elsevier.com/locate/ejpb)

Research paper

## The role of citric acid in oral peptide and protein formulations: Relationship between calcium chelation and proteolysis inhibition



Søren H. Welling <sup>a,b</sup>, František Hubálek <sup>a</sup>, Jette Jacobsen <sup>b</sup>, David J. Brayden <sup>c</sup>, Ulrik L. Rahbek <sup>a</sup>, Stephen T. Buckley <sup>a,\*</sup>

<sup>a</sup> Diabetes Research Unit, Novo Nordisk A/S, Måløv, Denmark<sup>b</sup> Department of Pharmacy, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark<sup>c</sup> UCD School of Veterinary Medicine and UCD Conway Institute, University College Dublin, Dublin, Ireland

## ARTICLE INFO

## Article history:

Received 27 August 2013

Accepted in revised form 23 December 2013

Available online 31 December 2013

## Keywords:

Oral peptide delivery

Citric acid

Proteolysis inhibition

Chelation

Intestinal drug permeability

Insulin

## ABSTRACT

The excipient citric acid (CA) has been reported to improve oral absorption of peptides by different mechanisms. The balance between its related properties of calcium chelation and permeation enhancement compared to a proteolysis inhibition was examined. A predictive model of CA's calcium chelation activity was developed and verified experimentally using an ion-selective electrode. The effects of CA, its salt (citrate, Cit) and the established permeation enhancer, lauroyl carnitine chloride (LCC) were compared by measuring transepithelial electrical resistance (TEER) and permeability of insulin and FD4 across Caco-2 monolayers and rat small intestinal mucosae mounted in Ussing chambers. Proteolytic degradation of insulin was determined in rat luminal extracts across a range of pH values in the presence of CA. CA's capacity to chelate calcium decreased ~10-fold for each pH unit moving from pH 6 to pH 3. CA was an inferior weak permeation enhancer compared to LCC in both *in vitro* models using physiological buffers. At pH 4.5 however, degradation of insulin in rat luminal extracts was significantly inhibited in the presence of 10 mM CA. The capacity of CA to chelate luminal calcium does not occur significantly at the acidic pH values where it effectively inhibits proteolysis, which is its dominant action in oral peptide formulations. On account of insulin's low basal permeability, inclusion of alternative permeation enhancers is likely to be necessary to achieve sufficient oral bioavailability since this is a weak property of CA.

© 2013 Elsevier B.V. All rights reserved.

## 1. Introduction

Development of oral delivery systems for proteins and peptides offers the promise of improved patient compliance compared to conventional parenteral administration. Moreover, in the case of certain protein therapeutics (e.g., insulin), the physiological response elicited may exhibit a pharmacodynamic profile which more closely resembles the natural physiological response. However, delivery of protein therapeutics is severely hindered by poor absorption across the intestinal barrier and extensive degradation by proteolytic enzymes. Thus, to effectively overcome these impediments, a formulation strategy which can modulate both of

these processes is necessary to achieve acceptable oral bioavailability with low intra-subject variation.

Although degradation of proteins by gastric enzymes and low pH may be overcome via inclusion of an enteric coating, the approach to minimise proteolytic activity in the small intestine, while simultaneously ensuring efficient release and permeation represents a more significant challenge. In this regard, one such concept extensively explored is that of acidic inhibition of proteolysis. Luminal proteases, such as trypsin and chymotrypsin, exhibit maximum activity at pH  $\geq 6.5$  [1,2] i.e., that typically observed in the pH microenvironment of the jejunum and ileum. Via adjustment of the local pH to values corresponding to pH  $< 6.5$ , proteolytic activity of enzymes such as chymotrypsin [1], the primary luminal degrading enzyme for insulin [2], can be significantly diminished.

Indeed, acidic inhibition of proteolysis as a strategy for the oral delivery of therapeutic peptides recently gained attention following Tarsa Therapeutics (Philadelphia, PA) successful completion of a phase III trial ('ORACAL') for orally delivered salmon calcitonin (sCT) [3]. Such technology typically comprises of an enteric coated capsule or tablet, which bypasses the stomach unchanged, along

**Abbreviations:** CA, citric acid; Cit, citrate; DOC, sodium deoxycholate; EDTA, ethylenediaminetetraacetic acid; ISE, ion selective electrode; KH, Krebs-Henseleit; LCC, lauroyl carnitine chloride; sCT, salmon calcitonin; TDC, taurodeoxycholate; TEER, transepithelial electrical resistance.

\* Corresponding author. ADME Department, Novo Nordisk A/S, Novo Nordisk Park, 2760 Måløv, Denmark. Tel.: +45 3079 4609.

E-mail address: [spby@novonordisk.com](mailto:spby@novonordisk.com) (S.T. Buckley).

with a pH-lowering excipient contained in vesicles (e.g., an organic acid such as citric acid). Upon entry into the duodenum with its luminal pH range of between 5 and 6, pH-dependent disintegration of the polymer coating of the dosage form commences, followed by release from the vesicle of both co-localised API and citric acid (CA). Concomitant association of CA maintains a decrease in local pH, thus stabilising the co-released peptide. In this way, it facilitates a reduction in the luminal enzymatic activity, providing a higher concentration gradient of the API over time, which in turn promotes improved absorption and bioavailability [4,5].

Alongside pH-lowering agents, co-administration of an absorption enhancer(s) has generally been regarded as indispensable due to the inherently poor epithelial permeability properties of proteins and peptides [5,6]. Indeed, previous publications exploring this technology have employed LCC, an amphiphilic surfactant [5–7]. However, based upon the recent ORACAL sCT study, where an absorption enhancer was omitted, one may speculate that either the need for co-administration is diminished on account of the proposed permeation enhancing properties of citric acid or citrate (Cit) [3], or that enhancers might not be required for oral sCT where bioavailability of 1–3% is typical for marketed nasal versions of this particular potent molecule [8]. CA and Cit are GRAS excipients and have been widely employed in oral formulations of small molecules. Thus, despite this formulation strategy being comparatively new, a body of literature exists examining the multiple mechanisms by which CA, in its salt form (i.e., tri-sodium citrate) may promote oral absorption. Cit exhibits calcium chelating properties and evidence exists to suggest that it may increase paracellular absorption, by triggering disruption of tight junction complexes via depletion of intracellular calcium [9–11].

In this report, the potential mechanism of action of CA as both an acidic proteolysis inhibitor and calcium chelator/permeation enhancer was addressed and conclusions made as to which might be its dominant action at relevant pH values in the upper small intestine. *In silico* and *in vitro* determination of CA's calcium chelation activity and its capacity to prevent insulin degradation by peptidases across a broad range of pH values were obtained. From this data we assessed whether or not a common pH range existed over which both proteolysis inhibition and calcium chelation occurred. Finally, the capacity of CA/Cit to enhance permeability was investigated in Caco-2 monolayers and rat intestinal tissue and compared to that of lauroyl carnitine chloride (LCC), an established amphiphilic permeation enhancer [5–7] previously employed as an additional agent in pH-lowering oral peptide formulations.

## 2. Materials and methods

### 2.1. Materials

Caco-2 cells (ATCC-HTB-37) were obtained from American Type Culture Collection (ManassasVA). Cell culture media (Dulbecco's modified essential media (DMEM)) and penicillin/streptomycin were purchased from Lonza (Verviers, Belgium). All other supplements i.e., foetal bovine serum (FBS), HEPES buffer and non-essential amino acids (NEAA) as well as Hanks' balanced salt solution (HBSS) and trypsin were purchased from Gibco (Naerum, Denmark). Corning Transwell® filter inserts (1.12 cm<sup>2</sup> surface area, 0.4 µm pore diameter) were purchased from Fisher Scientific (Slangerup, Denmark). FITC-dextran 4 kDa (FD4) and D-glucose were purchased from Sigma Aldrich (Dublin, Ireland). Bovine serum albumin (BSA) was purchased from Sigma Aldrich (Copenhagen, Denmark). All other reagents were of the highest analytical grade.

The Iso-Insulin ELISA assay kit was purchased from Mercodia (Uppsala, Sweden). Lauroyl-DL-carnitine (LCC) was purchased from

Chemos (Regenstauf, Germany). [<sup>3</sup>H]-mannitol, [<sup>14</sup>C]-mannitol and Ultima Gold® scintillation fluid were purchased from Perkin Elmer (Waltham, MA). Liquid scintillation counting was carried out using a TopCount C990201 or a TriCarb 2900TR liquid scintillation counter (both Perkin Elmer). Luminescent measurements were performed using a Spectramax® 250 or Gemini® microplate reader (both Molecular Devices, Sunnyvale, CA). Fluorescent measurements were performed on a Tecan® GENios fluorescent microplate reader (Tecan, Durham, NC).

### 2.2. Cell culture

Caco-2 cells (passage numbers 40–60) were seeded at a density of  $2.5 \times 10^5$  cells/flask and grown to 70–90% confluence in DMEM (supplemented with 10% FBS, 100 U/ml penicillin and 100 µg/ml streptomycin and 1% (v/v) NEAA). For transport studies, Caco-2 monolayers were cultured on permeable Transwell® 12 mm diameter inserts with pore sizes of 0.4 µm at a density of  $10^5$  cells/cm<sup>2</sup> and used after 14–17 days in culture. Cells were cultured at 37 °C and 5% CO<sub>2</sub> atmosphere and the medium was changed every other day.

### 2.3. Modelling chelation activity of citric acid (CA)

A model to predict free calcium fraction was constructed as described in [Supplementary materials](#). The conditional pKa and citrate (Cit) calcium chelation constant, K, corresponded to previously published values in which similar ionic strengths were applied [12–14]. The model was not corrected retrospectively to take account of the experimentally determined calcium electrode measurements.

### 2.4. Calcium electrode measurements

A pH-meter (744; Metrohm, Herisau, Switzerland) was fitted with a micro pH-electrode (6.0224.100; Metrohm), a calcium selective electrode (6.0508.110; Metrohm) and an AgCl reference electrode (Dri-Ref-L; World Precision Instruments, Sarasota, FL). All titrations were performed in calcium-free transport media at room temperature (RT). To 20 ml of calcium-free HBSS 300–500 µl CA or Cit (1.5–2 M) was added to yield a final solution of CA/Cit (30 mM) and pH values of 4, 5, 6, and 7.4. The solution was titrated with 40 mM CaCl<sub>2</sub> from 0.5 µl to 1310 µl [ $5 \times 10^{-3}$ – $2.5 \times 10^0$ ] mM CaCl<sub>2</sub>. Electrical motive force (EMF, mV) and pH were concomitantly monitored during titration. Double standard curves of calcium added to transport media (without CA/Cit), assuming that free calcium concentration was equivalent to total calcium. Titrations were performed at room temperature to improve reproducibility. All solutions were maintained at room temperature for 4 h prior to titration, as the ISE was sensitive to temperature changes. Activity of the ISE was assessed within the pH range of 3–7.4.

### 2.5. In vitro inhibition of proteolysis

Cit and CA were added to zinc-free transport medium (see "Transepithelial transport studies in Caco-2"; zinc-free) to give Ca/Cit stock solutions a total concentration of 12.5 mM of CA species and a range of pH values (3.5–7.4). Subsequently, enzyme-rich washes were extracted from fasted rat duodenal lumens by rinsing 10 cm fresh duodenum with 10 ml water and instantly freezing the eluate at –80 °C until use. At time point zero, 100 mM recombinant human insulin (Novo Nordisk A/S, Copenhagen, Denmark) was mixed with duodenal extracts and CA stock solutions in a ratio of 1:1:8 respectively, yielding 10 mM insulin and 10 mM CA species. The kinetic study was performed using an autosampler robot

(Gilson 215 liquid handler; Middleton, WI) running 16 separate samples simultaneously. The reaction was sampled at six time points over a period of 120 min. Upon collection of each 20 µl sample, 50 µl trifluoroacetic acid (5% v/v) was immediately added and the samples refrigerated (4 °C) in order to stop the proteolytic degradation. Insulin content was quantified by Acquity UPLC consisting of an autosampler (Model Acq-SM), pump (Model Acq-BSM), column oven (Model Acq-SM) and detector (Model Acq-TUV; Waters, Milford, MA). RP-UPLC separation was achieved by Acquity BEH 1.7 µM C18 1 × 50 mm column (Waters), using a linear gradient of acetonitrile in 0.2 M sodium sulfate, 0.04 M sodium phosphate, pH = 7.2. Peaks were detected by UV absorption at 220 nm and quantified using a human insulin standard. Reaction rates were calculated as the slope of the linear least squares fit to the semi log-plot of concentration versus time, see Eq. (1). All reaction rates were derived by the natural logarithm, *e*.

$$C_t = C_0 \cdot e^{-kt} \Rightarrow \ln(C_t) = -k \cdot t + \ln(C_0) \quad (1)$$

where *t* (s) is a given time point, *C<sub>t</sub>* (µM) the remaining concentration of insulin at time point, *t*; *k* (min<sup>-1</sup>) is the reaction constant and *C<sub>0</sub>* (µM) is the initial concentration of insulin.

#### 2.6. Transepithelial transport studies across Caco-2 monolayers

Filter-grown monolayers were washed with warm HBSS buffer (138 mM NaCl, 5.3 mM KCl, 1.3 mM CaCl<sub>2</sub>, 0.40 mM MgSO<sub>4</sub>, 0.44 mM KH<sub>2</sub>PO<sub>4</sub>, 4.2 mM NaHCO<sub>3</sub> and 5.6 mM glucose; pH 7.4) supplemented with 0.1% BSA and 10 mM HEPES and allowed to equilibrate. Transepithelial electrical resistance (TEER) was measured with a chop-stick electrode (Millicell-ERS®, Millipore, Billerica, MA) prior to testing and monolayers with TEER values <600 Ω cm<sup>2</sup> were discarded. The buffer in the respective apical side was then replaced with a solution containing insulin (10 µM) and [<sup>3</sup>H]-mannitol (0.8 µCi/ml) alone or in combination with CA (5 mM; pH 4.5) or LCC (1 mM), and the monolayers were incubated at 37 °C for 60 min. In some studies, Cit (20 mM; pH 7.4) was added to the basolateral side. Samples containing fluxed insulin from the donor apical side were collected from the basolateral compartments every 15 min for 1 h and human insulin content, was diluted to 100–10,000 ppm (~14–1400 mU/l), and was assayed using ELISA. Flux (*J* [mol/s]) was determined from steady-state appearance rates of insulin in the receiver fluid. The apparent permeability coefficient, *P<sub>app</sub>* [cm/s], was calculated according to Eq. (2)

$$P_{app} = J/(A \cdot C_i) \quad (2)$$

where *C<sub>i</sub>* (mol/cm<sup>3</sup>) is the initial concentration of insulin in the donor fluid and *A* is the nominal surface area: 1.12 cm<sup>2</sup> for Caco-2 monolayers and 0.63 cm<sup>2</sup> for intestinal mucosae.

#### 2.7. Preparation of rat intestinal tissue for Ussing chamber studies

Studies were carried out in accordance with the UCD Animal Research Ethics Committee policy, on the use of post mortem animal tissue in research, as well as in adherence to the "Guide for the care and use of laboratory animals", (8th Edition, National Academy of Sciences, 2011. <http://www.aalac.org/resources/the-guide.cfm>). Male Wistar rats (250–300 g) (Charles River, Margate, UK) were euthanised by stunning and cervical dislocation. The lower jejunum and ileum (lower small intestine) was removed, opened along the mesenteric border and rinsed in warm oxygenated Krebs–Henseleit solution (KH; 118 mM NaCl, 4.7 mM KCl, 2.5 mM CaCl<sub>2</sub>, 1.2 mM MgSO<sub>4</sub>, 1.2 mM KH<sub>2</sub>PO<sub>4</sub>, 25 mM NaHCO<sub>3</sub> and 10 mM glucose) according to previous methods [15]. Tissue was pinned with the mucosal side down on a dissection board to expose the external muscularis, which was carefully removed with

a size 5 fine forceps. The tissue was then mounted in Ussing chambers with a circular window area of 0.63 cm<sup>2</sup>, bathed bilaterally with 5 ml KH and continuously gassed with 95% CO<sub>2</sub>/5% O<sub>2</sub> at pH 7.4 and maintained at 37 °C. The transepithelial potential difference (PD, mV) and short circuit current (*I<sub>sc</sub>*, µA) were measured across the lower small intestine. The tissue was voltage clamped to zero for 30 s and switched to open circuit configuration for 3 s by an automatic voltage clamp (EVC-4000 amplifier) and Pro-4 timer (both WPI, Hertfordshire, UK). Analogue data were digitised with a Mac Powerlab® data acquisition unit and analysed with Chart® software (AD Instruments, Oxford, UK). Following an equilibration period of 15 min, PD and *I<sub>sc</sub>* were measured and TEER was calculated at regular time points from 0 to 120 min using Ohm's law.

#### 2.8. Transepithelial transport studies in rat lower small intestinal tissue

Transport of [<sup>14</sup>C]-mannitol and FITC-Dextran 4000 (FD4), non-degradable hydrophilic flux marker, was examined across lower small intestinal mucosae mounted in Ussing chambers. Briefly, [<sup>14</sup>C]-mannitol (0.2 µCi/ml) and FD4 (1 mg/ml) were added to the apical chamber and flux was monitored periodically over 2 h by sampling the serosal chamber (200 µl) every 20 min for 2 h, and apically (200 µl) at time zero, while replenishing with fresh KH buffer at each sampling point. In some studies, CA (30 mM), Cit (30 mM) or LCC (3 mM) were simultaneously added to the apical chamber. Samples containing [<sup>14</sup>C]-mannitol were mixed with scintillation fluid and read in a scintillation counter (TriCarb 2900TR, Perkin Elmer). Where samples contained FD4, fluorescence was measured in a fluorescence microplate reader (Spectramax Gemini, Molecular Devices) with  $\lambda_{ex}/\lambda_{em}$  of 480/520 nm. The *P<sub>app</sub>* was calculated according to Eq. (2).

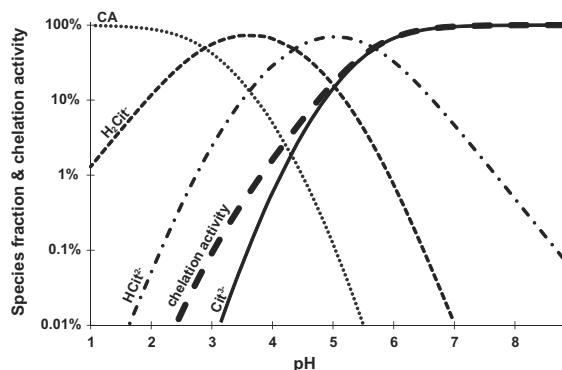
#### 2.9. Statistical data analysis

Statistical analysis was carried out using Prism-6® software (GraphPad, San Diego, USA) using two-tailed unpaired Student's *t*-tests unless otherwise stated. Results are presented as the mean ± standard deviation (SD, unbiased). The level of significance set was *P* > 0.05. Normal distribution of data was in general assumed except for apparent permeability data which elicited a relatively consistent standard deviation (%RSD) and was therefore log transformed prior to statistical analysis.

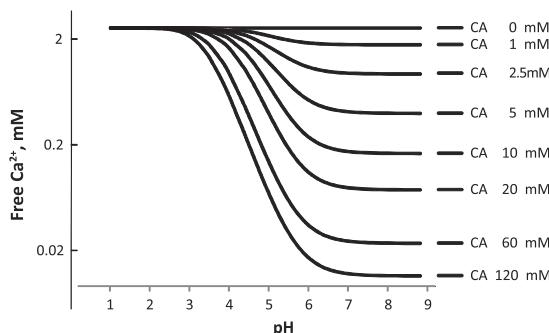
### 3. Results

#### 3.1. Prediction of the chelation activity of citric acid (CA)

To estimate the chelation activity of CA, a mathematical model was employed (see [Supplementary material](#)). Chelation activity was defined as the average of the apparent formation constants of the individual species of CA, weighted according to their presence and expressed relative to the formation constant of Cit<sup>3-</sup>, in theory the strongest chelator. The predicted relationship between each form of CA (H<sub>2</sub>Cit<sup>+</sup>, HCit<sup>2-</sup>, Cit<sup>3-</sup>) and their calcium chelation activity is depicted in Fig. 1. As illustrated, CA chelation capacity is especially high in the presence of high levels of Cit<sup>3-</sup> and that is pH-dependent. Above pH 5, total chelation activity corresponds directly to the proportion of the Cit<sup>3-</sup> species present. At values below pH 5 the fraction of Cit<sup>3-</sup> is less than 0.1. In this range (i.e., pH < 5), HCit<sup>2+</sup> becomes relatively more dominant compared to Cit<sup>3-</sup>. At pH > 5, while HCit<sup>2+</sup> exhibits some degree of chelation capacity, it is thought to be significantly less than that of Cit<sup>3-</sup> at the higher pH values. Importantly, calcium chelation activity was



**Fig. 1.** Calculated CA chelation activity related to the theoretical concentration of the individual formats (ion and salt) at various pH values. The bold dashed line represents the summarised chelation activity of all formats. The model depicted applies to solutions of ionic strength between  $I = 0.1\text{--}0.6\text{ M}$ , but does not apply to diluted solutions.



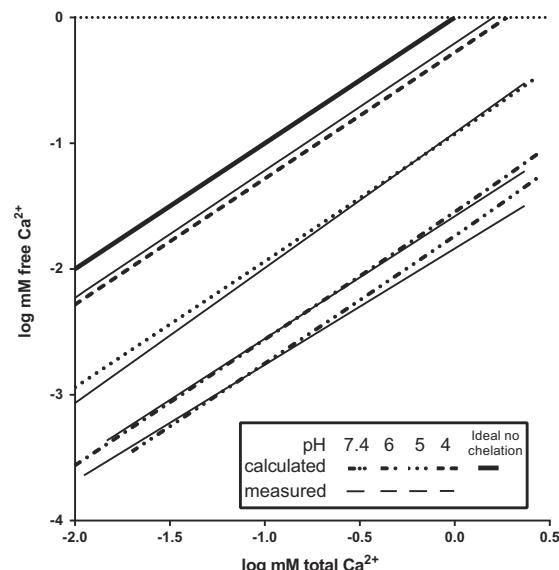
**Fig. 2.** Calculated free calcium concentration in media of 2.5 mM total calcium (KH buffer) at different concentrations of CA and pH values. The model depicted applies to solutions of ionic strength between  $I = 0.1\text{--}0.6\text{ M}$ , but does not apply to diluted solutions.

reduced ~10-fold for each unit of pH within the pH range of 3–6. As shown in Fig. 2, in order to chelate 99% of the total calcium in KH buffer at a pH of 7.4, very high concentrations of 120 mM CA (pH 5.5) or 60 mM CA (pH 7) would be required. For 90% chelation, 120 mM CA (pH 4.2) or 10 mM CA (pH 5.8) would be necessary. The prediction therefore is that at highly acidic local pH values, non-physiological concentrations of CA would be required to substantially chelate calcium, the most likely mechanism to open epithelial tight junctions for permeation enhancement.

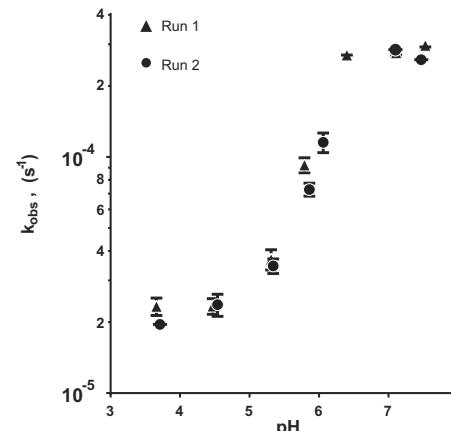
### 3.2. Correlation of 'predicted' versus 'measured' free calcium ( $\text{Ca}^{2+}$ )

CA,  $\text{H}_2\text{Cit}^-$ ,  $\text{HCit}^{2-}$  and  $\text{Cit}^{3-}$  form a buffer system in which the distribution of the species is a function of pH. Herein, all mentioned concentrations of CA or Cit are absolute and accompanied by a pH value from which the actual distribution of CA-species can be found (Fig. 1).

In order to validate the predictive model, total and free calcium were determined with a calcium-selective electrode following titration of 30 mM CA (at pH values of 4, 5, 6, and 7.4) with 10  $\mu\text{M}$ –3 mM  $\text{Ca}^{2+}$ . Fig. 3 shows predicted free  $\text{Ca}^{2+}$  levels versus corresponding experimentally determined free  $\text{Ca}^{2+}$  levels. Within the range of 0.01–2 mM total  $\text{Ca}^{2+}$ , very accurate correlations were achieved ( $R^2 > 0.98$ ). According to the conditions observed in Fig. 3, CA/Cit is at least 25-fold in excess compared to calcium. Consequently,



**Fig. 3.** Free calcium versus total calcium in solutions measured with an ion-selective electrode compared to the calculated free calcium. Solution: HBSS + HEPES (10 mM) + CA (30 mM). Dashed lines are calculated relationships of the total and free calcium. The thin black line — represents the experimentally determined calcium chelation. The thick black line — represents the ideal i.e., no chelation, whereby total calcium = free calcium.

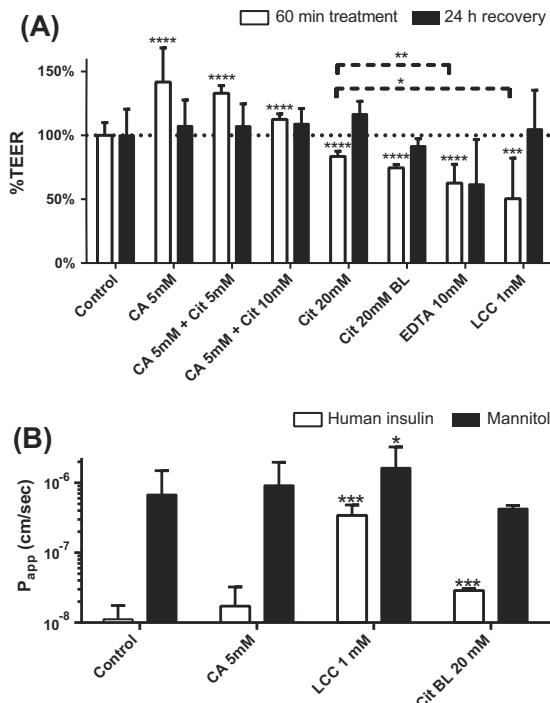


**Fig. 4.** Apparent reaction constants ( $k_{\text{obs}}$ , natural log) of the proteolysis of insulin by rat duodenal extracts at various pH values in HBSS + HEPES (10 mM) with CA/Cit (10 mM). The proteolysis rate shows an exponential pattern of increase with pH. Insulin proteolysis was fitted as 1st order decay. Data points represent an average of two measurements  $\pm \text{SD}$  from two separate runs.  $n = 28$ .

the titration slopes of free calcium versus total calcium achieved are highly linear and when plotted on a double-log scale display slopes close to 1. A minor departure from linearity was observed in the lower part of the standard slope (free calcium/mV) which fitted well with 2nd or 3rd order polynomial and was attributed to be a loss of sensitivity of the electrode in its lower detection range.

### 3.3. pH-dependent degradation of insulin

In the absence of proteolysis inhibition, insulin is readily degraded in the small intestine. However, the activity of most key



**Fig. 5.** (A) Effect of CA, Cit and LCC on TEER in Caco-2 monolayers. Normalised change of TEER (%) in monolayers after 60 min treatment with CA 5 mM (pH 4.5), CA 5 mM + Cit 5 mM (pH 5.2), CA 5 mM + Cit 10 mM (pH 5.5), Cit (basolateral) (20 mM, pH 7.4), EDTA (10 mM) or LCC (1 mM). Data are represented as means  $\pm$  SD,  $n = 7\text{--}50$ . Separate Student's *t*-tests were carried out to evaluate if treatments significantly differed from the control. (B)  $P_{app}$  of insulin and [<sup>3</sup>H]-mannitol across Caco-2 monolayers incubated with CA (5 mM, pH 4.5), Cit (20 mM, pH 4.5), LCC (1 mM, pH 7.4) for 60 min. Separate Student's *t*-tests were carried out to evaluate if treatments significantly differed from the control. Data are represented as means  $\pm$  SD,  $n = 3\text{--}12$ .

proteolytic enzymes including trypsin, chymotrypsin and elastase is strongly influenced by pH. To investigate the precise relationship between pH and proteolysis rate, an *in vitro* proteolysis assay using native extracts of rat duodenum was performed in the presence of CA/Cit (10 mM). Throughout, the degradation of insulin was characterised by first order kinetics. A plot of insulin degradation versus time is provided in the *Supplementary material*. At pH 6.5–7.4, the native pH of the small intestine, the reaction rate proceeded much faster than at acidic pH values (Fig. 4). By decreasing the pH from 7 to 4.5 however, the reaction rate was reduced markedly (10-fold) ( $p < 0.0001$ ). Hereafter from pH 4.5 to 4 the reaction rate remained unchanged. At a pH of 3.5, the reaction rate was an order of magnitude lower than that seen at pH 7.4. In order to exclude other potential confounding factors, the influence of both total Cit concentration and the quantity of NaCl added was evaluated. Although Cit (120 mM, pH 7.4) lowered the reaction rate,  $k_{obs}$ , two-fold ( $p < 0.001$ ), NaCl (75–250 mM) elicited no significant effect and both the independent factor of NaCl and Cit appeared to have no physiological significance. These data emphasise the dramatic protection against pancreatic peptidases afforded by simply maintaining the pH at 3–4 with a 10 fold lower concentration of CA than that required to chelate calcium.

#### 3.4. Effects of CA, Cit and LCC on Caco-2 monolayers: TEER and $P_{app}$ values of [<sup>3</sup>H]-mannitol and insulin

Treating Caco-2 monolayers with apical addition of LCC (1 mM) elicited a decrease in TEER of 50% in 60 min compared to that of

untreated monolayers (Fig. 5A) ( $p < 0.001$ ). Exposure to various concentrations of Cit (5–20 mM pH 7.4) apically or basolaterally however, resulted in smaller albeit concentration-dependent reductions in TEER and Cit was less potent and efficacious than the well-known chelator and permeation enhancer, EDTA and LCC (Fig. 5A). Surprisingly, CA (5 mM) at pH 4.5 elicited a significant increase in TEER of 50% relative to untreated monolayers ( $p < 0.001$ ). Comparable responses were observed with 5 mM CA in combination with 5 mM and 10 mM Cit (Fig. 5A). Following incubations, transport medium was removed and fresh DMEM introduced; monolayers were then incubated for 24 h to assess TEER recovery. All treated cultures showed a full recovery in TEER, exhibiting comparable values to that of control monolayers after 24 h. Of note, addition of excessive amounts of CA, resulted in a lowering of pH < 4 where TEER recovery was not observed due to irreversible deterioration of the monolayer barrier, as earlier reported [16]. In summary, the data reveal that CA and/or Cit only have marginal effects on TEER, but nothing like the level of decrease which would be expected from an effective permeation enhancer (i.e., >50% TEER decrease).

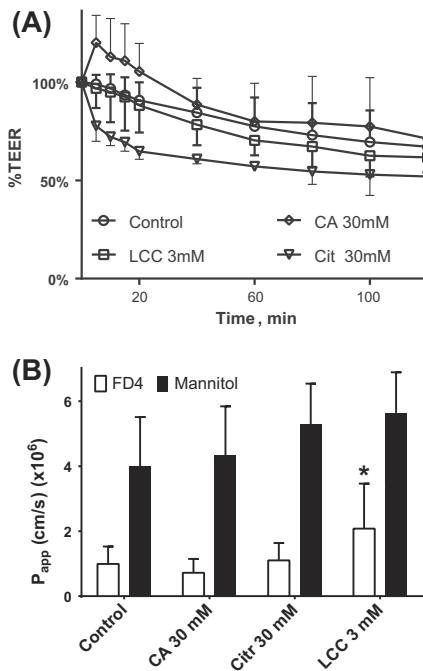
The  $P_{app}$  values of insulin across Caco-2 monolayers showed a significant increase in the presence of LCC (1 mM), generating a 40-fold increase in transport ( $8 \times 10^{-7}$  cm/s) compared to basal insulin  $P_{app}$  across untreated monolayers ( $p < 0.01$ ) while the  $P_{app}$  of [<sup>3</sup>H]-mannitol was 50% greater than in untreated monolayers ( $1 \times 10^{-6}$  cm/s) ( $p < 0.05$ ) (Fig. 5B). In contrast, CA (5 mM) had no effect on the  $P_{app}$  of either insulin or [<sup>3</sup>H]-mannitol. While high concentrations of Cit (20 mM) produced a 2-fold increase of insulin permeability ( $p < 0.001$ ), this was still significantly less than that of LCC and observed only when applied basolaterally. The effect apically was even smaller. Overall, the data confirm that, compared to the dramatic permeation enhancement induced by LCC, CA/Cit are not effective enhancers in Caco-2 monolayers and this is consistent with their nominal effect on TEER compared to positive controls.

#### 3.5. Effect of CA, Cit and LCC on rat lower small intestinal mucosae: TEER and $P_{app}$ of [<sup>14</sup>C]-mannitol and FD4

The effects of CA, Cit and LCC on the electrophysiological responses of rat lower small intestinal tissue were examined in Ussing chambers. Addition of very high concentrations of Cit (30 mM, pH 7.4) elicited a sustained decrease in TEER over 120 min, which was significantly lower (~20%) than that observed in control tissue (Fig. 6A). In contrast, incubation with CA (30 mM, pH 3) gave rise to a significant, albeit transient increase in TEER (~20%) relative to control during the initial 20 min exposure. Thereafter, TEER values aligned with those of untreated controls. Although LCC did not provoke any significant decrease in TEER, a 2-fold increase in the permeability of FD4 was observed (Fig. 6B). In contrast, neither the TEER reduction elicited by Cit nor the TEER increase observed with CA was associated with changes in mannitol permeability. Permeability of [<sup>14</sup>C]-mannitol remained statistically unaffected by the various treatments. Overall, these data do not indicate that CA/Cit is an effective permeation enhancer in rat lower small intestinal mucosae, whereas LCC was effective, at least for FD4.

## 4. Discussion

This work addressed the functional role of CA in formulation-based approaches for the delivery of peptides and proteins via the oral route. Specifically, we examined the interplay between its primary use as a pH-lowering agent and its additional function as a calcium chelator that might cause both tight junction openings and contribute to further inhibition of serine proteases. Modulation of intestinal pH via formulation is an attractive means to



**Fig. 6.** (A) Effect of CA, Cit and LCC on TEER (A) and  $P_{app}$  (B) in rat lower small intestine in Ussing chambers upon mucosal-side incubation with CA (5 mM, pH 4.5), Cit (20 mM, pH 7.4) or LCC (3 mM) for 120 min. (B)  $P_{app}$  of FD4 and [<sup>14</sup>C]-mannitol across rat lower small intestine incubated with CA (5 mM, pH 4.5), Cit (20 mM, pH 7.4) or LCC (3 mM) for 120 min. Data are represented as means  $\pm$  SD,  $n = 3\text{--}4$ .

stabilise the protein against enzymatic degradation. Indeed, it has previously been used as a strategy for small intestinal delivery of sCT. Using CA (a prototype organic acid) and LCC (an amphiphilic surfactant), there was significant enhancement in oral bioavailability of sCT in dogs [4,5,17]. However, precise elucidation of the combined effects of CA on protein stability and permeability remains undetermined.

Chelation of calcium is an efficient means by which to modulate tight junction structures. Participation of  $\text{Ca}^{2+}$  in the establishment of epithelial cell junction networks has been widely demonstrated [18,19]. CA, a hydroxy tricarboxylic acid, is an efficient chelator in its salt form (i.e., citrate), capable of sequestering multivalent cations. Conditional chelation constants for  $\text{Ca}^{2+}\text{-Cit}^{3-}$  and  $\text{Ca}^{2+}\text{-HCit}^{2-}$  are  $1880 \text{ M}^{-1}$  and  $67 \text{ M}^{-1}$ , respectively [14,12], while conditional CA p $K_{a,1,2,3}$  values applied were [2.80; 4.08; 5.33] [13,14]. Such values are dependent on ionic strength and temperature. Thus, use of the applied model to predict chelation capacity potential is suggested to be restricted to conditions whereby  $I$  ranges from 0.1 M to 0.6 M and temperature ranges from 18 °C to 45 °C. KH and HBSS have ionic strengths of  $I \sim 0.16 \text{ M}$ . In this context, our proposed model predicts that at pH values of 6–7, and above, optimal chelation activity is observed. Importantly, below pH 6 the apparent chelation constant is reduced ~10-fold for each pH unit. Therefore, at highly acidic pH values the concentration of CA/Cit necessary to chelate vast quantities of calcium dramatically increases. The pH microenvironment generated via release of CA from a pH-lowering formulation likely corresponds to a value of 4.5 or lower [17]. Under these circumstances, CA will primarily dissociate into  $\text{HCit}^{2+}$  and not  $\text{Cit}^{3-}$  the predominant chelating species. In this chemical arrangement, its potency as a chelator of calcium ions is considerably reduced. Accordingly, this formulation

approach is unlikely to engender significant chelation activity, but will be dominated by a capacity for pH-mediated peptidase inhibition.

Enzymatic degradation of proteins by luminal proteases represents a significant barrier to achieving a therapeutically relevant bioavailability. The ability of pH-lowering formulations to curtail this undesirable aspect of intestinal physiology is its primary attraction for oral peptide and protein delivery. Our *in vitro* investigations revealed that an EC<sub>50</sub> of proteolysis is achieved at pH 6. However, given that extensive proteolytic degradation arises *in vivo*, in order to effectively protect sufficient quantities of intact protein, substantial serine protease inhibition is necessary. The kinetics studies pertaining to insulin degradation indicate that proteolysis activity exhibits an apparent linear (log scale) decrease down to a value of pH 4.5 where >90% inhibition was achieved. Given that proteins are extensively and rapidly degraded especially in the duodenal and jejunal regions of the GIT, formulations employing acidic inhibition of proteolysis should therefore aim to achieve a local pH of at least 4.5. In this regard, an *in vivo* study in dog found that capsules loaded with the maximal practicable quantity of CA, corresponding to 570 mg in a 680 mg tablet, yielded the highest bioavailability, by reducing the pH to as low as 3 [18]. It should be noted that degradation studies performed *in vitro* represent a system of optimal mixing conditions, which may not be present *in vivo*. Under such circumstances, lowering the pH beyond 4.5 likely ensures a greater acidic expanse within the small intestine, thus representing a local region in which protein degradation is limited for a period of time.

Chymotrypsin is the primary enzyme responsible for the degradation of insulin [2]. Corresponding studies examining chymotrypsin-mediated degradation of casein or denatured lysozyme exhibit a slope ( $k_{obs}$  versus pH) which closely resembles that generated in our investigations examining the degradation of insulin in rat duodenal enzyme extracts [1]. Studies suggest that the activity of chymotrypsin can be lowered 2-fold when free calcium is less than 0.05 mM [20]. Similarly, the activity of chymotrypsin has been shown to be lowered when  $\text{Ca}^{2+}$  is omitted [20] or when available  $\text{Ca}^{2+}$  is chelated by EDTA [21]. Our investigations however, indicate that the function of CA/Cit as a chelator is most pronounced at high pH values (i.e., >7). Indeed, we have observed that Cit 120 mM lowers the rate constant of chymotrypsin-mediated degradation of insulin by up to 35% at pH 7.4 due to its prevalent chelating action at that pH value (see *Supplementary material*). However, for pH-lowering formulations, where the local pH at the site of release is <5, its function as a chelator does not contribute to its peptidase inhibitory capacity, which is simply due to moving the pH optimum for serine proteases away from pH 7.4. Therefore, in practice the impact of CA/Cit on reduction of enzyme activity via calcium chelation is significantly less than that due to acidity *per se*. Trisodium citrate (120 mM) will provide a considerable increase to ionic strength possibly further reducing the activity of luminal proteolysis. However, on the contrary, addition of NaCl (75, 120 and 250 mM) did not significantly reduce the degradation rate of insulin by chymotrypsin (unpublished data).

On account of its large molecular weight size, transport of insulin (5800 MW, 8 Å) across the small intestinal epithelium is severely restricted. Indeed, basal permeability results in absorption of doses far below that required to achieve a therapeutic effect [22]. Amphiphilic permeability enhancers, a class of absorption enhancers including LCC and sodium taurodeoxycholate (TDC) promote trans- and paracellular absorption via mild recoverable membrane perturbation and disruption of tight junction complexes [23–25]. For example, permeation of sCT across rat ileal tissue in Ussing chambers can be increased 5-fold and 14-fold for LCC and TDC, respectively [6]. However, their successful incorporation in a pH-lowering formulation requires that they exhibit sufficient

solubility at low pH values. In this regard, LCC satisfies this requirement. As a consequence, it has been widely co-entrapped in formulations designed for acidic inhibition of proteolysis [5,7].

Nevertheless, in the ORACAL study for orally delivered sCT, no amphiphilic enhancer was required. One interpretation was that CA/Cit might be enhancing paracellular transport presumably via modulation of tight junctions [3], in addition to their effect on inhibiting peptidases by acidifying the pH. Supporting this, work by Okada and colleagues [26] revealed that addition of organic acids facilitated vaginal absorption enhancement of the luteinising hormone releasing hormone (LH-RH)-analogue, leuproide. Furthermore, they observed a weak correlation between this effect and the chelating properties of the acids. However, the influence of pH on such effects was not addressed in this study. Collectively, our investigations indicate that even if intestinal luminal calcium chelation is associated with permeability enhancement, the pH conditions of acidic inhibition of proteolysis (i.e., pH < 4.5) will extensively reduce chelation activity. Conceivably, the positive results yielded in the Phase III trial with sCT in the absence of a recognised permeability enhancer could be ascribed to the fact that sCT is half the molecular weight of insulin and exhibits a higher baseline permeability [27–31]. Moreover, since sCT is highly potent and marketed nasal versions are associated with no more than 1–3% bioavailability for required efficacy, the hurdles for an oral sCT are therefore much lower and would not require a recognised permeation enhancer once the excipient organic acid performs the pH-lowering role. This would not be the case for insulin where a much higher oral bioavailability would be required for a commercially-viable oral formulation. Consistent with this theory, ileal instillation of insulin with soy-bean proteolysis inhibitor alone did not lower blood glucose in rats, whereas glucose levels were significantly reduced when insulin was co-administered with the bile salt, sodium deoxycholate (DOC) [32], an efficient permeation enhancing agent.

There are numerous reports which demonstrate the effect of CA as a permeability enhancer. CA (1%, pH 7) enhanced nasal absorption of oil/water (O/W) emulsions containing indomethacin by 6.5-fold [33]. However, it should be noted that the nasal clearance/dilution is smaller than that observed in the GIT [34]. Moreover, the CA partitions in water yielding a concentration of 100 mM; thus, the concentration of CA reaching the nasal mucosa is likely much higher than in the intestine.

Insulin formulated with CA (10%, pH 1.72) for vaginal administration effectively lowered blood glucose to the same level as a 10 times lower intramuscularly injection. However, although such high concentrations of CA and low pH can increase permeability, they have also been shown to lead to extensive damage of mucosal tissue [26]. In the context of chronic administration, such adverse effects are undesirable.

Our *in vitro* (Caco-2 monolayers) and *ex vivo* (rat small intestinal tissue) transport studies indicate that the ability of CA/Cit (5–30 mM) to increase permeability of insulin or FD4 is negligible, exhibiting effects which are lower than that of LCC (1 or 3 mM), regardless of pH. Examination of absorption processes *ex vivo* can be confounded by enzymatic degradation. Thus, FD4 being a non-degradable hydrophilic macromolecule-sized compound represents a useful surrogate marker compound for insulin in Ussing chamber-based transport studies, facilitating exclusive examination of the absorptive process alone. The concentrations of LCC employed correspond to those which lie below that which can adversely impact tissue viability over the course of permeation study [25]. In the case of CA, the concentrations used represent those which are anticipated to be found locally in the lumen at the site of release of such oral formulation(s). Compared to excised intestinal tissue, Caco-2 monolayers represent a more fragile model system. Thus, pH should not be reduced beyond 4 [16] and LCC

and CA concentrations can likewise be reduced to appropriate functional concentrations which ensure the cell monolayers recover following treatment.

In line with earlier findings for EDTA [19], basolateral application of Cit elicited a more marked decrease in TEER. This differential susceptibility could be attributed to compositional differences in tight junction structures at the apical and basolateral interfaces as previously shown for EDTA [18]. Application to the basolateral side ensures direct exposure of the highly calcium-dependent zonula adherens proteins – structures which are implicated in preservation of monolayer integrity. In this regard, apical addition of Cit will have little impact. However, translation of these basolateral specific effects *in vivo* is not particularly relevant; given the fact that exposure is restricted to the apical membrane of the intestinal epithelia upon release from the dosage form. Collectively, these results strongly suggest that apically applied CA, by means of significant calcium chelation at pH 7.4, is not sufficient to elicit significant augmentation of insulin permeability, notwithstanding its ability to trigger acidic inhibition of proteolysis. Nevertheless, this observation does not preclude the possibility that CA/Cit could potentially chelate calcium in a pH neutral micro-environment below the mucus layer covering the intestinal epithelium.

## 5. Conclusions

It is evident that the pH range over which CA effectively inhibits proteolysis and that whereby Cit exerts calcium chelating properties does not coincide. At pH 3–4, the capacity of CA to inhibit small intestinal serine proteases is high, and this is due to sub-optimal pH values for those enzymes rather than to calcium chelation. Moreover, *in vitro* and *ex vivo* investigations indicate that the capacity of Cit/CA to exert significant permeation enhancement on human intestinal monolayers and isolated rat small intestinal mucosae is extremely low. While oral delivery of a few potent small peptides including sCT may be successfully achieved in the presence of formulated peptidase inhibitors (e.g., CA) in the absence of permeation enhancers, larger and more impermeable peptides and proteins will require an absorption enhancing agent in the formulation.

## Acknowledgments

Signe Beck Petersen (Novo Nordisk A/S) is thanked for her helpful advice regarding Ussing chamber experiments. A calcium ion electrode was kindly provided by Jesper Østergaard (University of Copenhagen). The authors are grateful for the technical assistance of Lisette Gammelgaard Nielsen, Anette Heerwagen, Gitte Hedelund, Trine Moghaddam (Novo Nordisk A/S) and Patrick Kearns (UCD). SH Welling, F Hubálek, UL Rahbek and ST Buckley are either affiliated with or employees of Novo Nordisk A/S. This study was funded in part by Novo Nordisk A/S and Science Foundation Ireland Grant SRC/07/B1154.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.ejpb.2013.12.017>.

## References

- [1] A.D. McLaren, E.F. Estermann, Influence of pH on the activity of chymotrypsin at a solid-liquid interface, *Arch. Biochem. Biophys.* 68 (1957) 157–160.
- [2] R. Schilling, A. Mitra, Degradation of insulin by trypsin and alpha-chymotrypsin, *Pharm. Res.* 8 (1991) 721–727.
- [3] N. Binkley, M. Bolognese, A. Sidorowicz-Bialynicka, T. Vally, R. Trout, C. Miller, C.E. Buben, J.P. Gilligan, D.S. Krause, A phase 3 trial of the efficacy and safety of

- oral recombinant calcitonin: the oral calcitonin in postmenopausal osteoporosis (ORACAL) trial, *J. Bone Miner. Res.* 27 (2012) 1821–1829.
- [4] J.P.F. Bai, L.L. Chang, J.H. Guo, Effects of polyacrylic polymers on the luminal proteolysis of peptide drugs in the colon, *J. Pharm. Sci.* 84 (1995) 1291–1294.
- [5] Y.H. Lee, P.J. Sinko, Oral delivery of salmon calcitonin, *Adv. Drug Deliv. Rev.* 42 (2000) 225–238.
- [6] P. Sinko, Y.H. Lee, V. Makhey, G. Leesman, J. Suttyak, H. Yu, B. Perry, C. Smith, P. Hu, E. Wagner, L. Falzone, L. McWhorter, J. Gilligan, W. Stern, Biopharmaceutical approaches for developing and assessing oral peptide delivery strategies and systems: *in vitro* permeability and *in vivo* oral absorption of salmon calcitonin, *Pharm. Res.* 16 (1999) 527–533.
- [7] N.M. Mehta, Oral Delivery and recombinant production of peptide hormones Part I: making oral delivery possible, *BioPharm Int.* 17 (2004).
- [8] M.J. Cho, J.F. Scieszka, P.S. Burton, Citric acid as an adjuvant for transepithelial transport, *Int. J. Pharm.* 52 (1989) 79–81.
- [9] M. Grant, M.A. Leone-Bay, Peptide therapeutics: it's all in the delivery, *Ther. Deliv.* 3 (2012) 981–996.
- [10] D.P. Froment, B.A. Molitoris, B. Buddington, N. Miller, A.C. Alfrey, Site and mechanism of enhanced gastrointestinal absorption of aluminum by citrate, *Kidney Int.* 36 (1989) 978–984.
- [11] C.R. Nolan, J.R. Califano, C.A. Butzin, Influence of calcium acetate or calcium citrate on intestinal aluminum absorption, *Kidney Int.* 38 (1990) 937–941.
- [12] R.P. Singh, Y.D. Yeboah, E.R. Pambid, P. Debayle, Stability constant of the calcium-citrate(3-) ion pair complex, *J. Chem. Eng. Data* 36 (1991) 52–54.
- [13] A.J. Meyer, M. Popp, Free Ca<sup>2+</sup> in tissue 22+-selective electrodes, *J. Exp. Bot.* 48 (1997) 337–344.
- [14] J.L. Meyer, Formation constants for interaction of citrate with calcium and magnesium ions, *Anal. Biochem.* 62 (1974) 295–300.
- [15] A.W. Cuthbert, H.S. Margolius, Kinins stimulate net chloride secretion by the rat colon, *Br. J. Pharmacol.* 75 (1982) 587–598.
- [16] G. Borchard, H.L. Luefßen, A.G. de Boer, J.C. Verhoeft, C.M. Lehr, H.E. Junginger, The potential of mucoadhesive polymers in enhancing intestinal peptide drug absorption. III: Effects of chitosan-glutamate and carbomer on epithelial tight junctions in vitro, *J. Control. Release.* 39 (1996) 131–138.
- [17] Y.H. Lee, B. Perry, S. Labruno, H. Lee, W. Stern, L. Falzone, P. Sinko, Impact of regional intestinal pH modulation on absorption of peptide drugs: oral absorption studies of salmon calcitonin in beagle dogs, *Pharm. Res.* 16 (1999) 1233–1239.
- [18] M. Tomita, M. Hayashi, S. Awazu, Absorption-enhancing mechanism of EDTA, caprate, and decanoylcarnitine in Caco-2 cells, *J. Pharm. Sci.* 85 (1996) 608–611.
- [19] A.B.J. Noach, Y. Kurosaki, M.C.M. Blom-Roosemalen, A.G. de Boer, D.D. Breimer, Cell-polarity dependent effect of chelation on the paracellular permeability of confluent caco-2 cell monolayers, *Int. J. Pharm.* 90 (1993) 229–237.
- [20] T. Sasaki, H. Kise, Increase of catalytic activity of  $\alpha$ -chymotrypsin by metal salts for transesterification of an amino acid ester in ethanol, *Biosci. Biotechnol. Biochem.* 61 (1997) 1196–1197.
- [21] J.P. Bai, L.L. Chang, Transepithelial transport of insulin: I. Insulin degradation by insulin-degrading enzyme in small intestinal epithelium, *Pharm. Res.* 12 (1995) 1171–1175.
- [22] G.P. Carino, E. Mathiowitz, Oral insulin delivery, *Adv. Drug Deliv. Rev.* 35 (1999) 249–257.
- [23] A.C. Chao, M.T. Taylor, P.E. Daddona, M. Broughall, J.A. Fix, Molecular weight-dependent paracellular transport of fluorescent model compounds induced by palmitoylcarnitine chloride across the human intestinal epithelial cell line Caco-2, *J. Drug Target.* 6 (1998) 37–43.
- [24] J.A. Fix, K. Engle, P.A. Porter, P.S. Leppert, S.J. Selk, C.R. Gardner, J. Alexander, Acylcarnitines: drug absorption-enhancing agents in the gastrointestinal tract, *Am. J. Physiol.* 251 (1986) G332–40.
- [25] E.L. LeCluyse, S.C. Sutton, J.A. Fix, In vitro effects of long-chain acylcarnitines on the permeability, transepithelial electrical resistance and morphology of rat colonic mucosa, *J. Pharmacol. Exp. Ther.* 265 (1993) 955–962.
- [26] H. Okada, I. Yamazaki, Y. Ogawa, S. Hirai, T. Yashiki, H. Mima, Vaginal absorption of a potent luteinizing hormone-releasing hormone analog (leuprolide) in rats: I: Absorption by various routes and absorption enhancement, *J. Pharm. Sci.* 71 (1982) 1367–1371.
- [27] H. Ichikawa, N.A. Peppas, Novel complexation hydrogels for oral peptide delivery: in vitro evaluation of their cytocompatibility and insulin-transport enhancing effects using Caco-2 cell monolayers, *J. Biomed. Mater. Res.* 67A (2003) 609–617.
- [28] A.C. Foss, N.A. Peppas, Investigation of the cytotoxicity and insulin transport of acrylic-based copolymer protein delivery systems in contact with caco-2 cultures, *Eur. J. Pharm. Biopharm.* 57 (2004) 447–455.
- [29] K.H. Song, S.J. Chung, C.K. Shim, Preparation and evaluation of proliposomes containing salmon calcitonin, *J. Control. Release.* 84 (2002) 27–37.
- [30] P.S. Hiremath, K.S. Soppimath, G.V. Betageri, Proliposomes of exemestane for improved oral delivery: formulation and in vitro evaluation using PAMPA, Caco-2 and rat intestine, *Int. J. Pharm.* 380 (2009) 96–104.
- [31] R. Shah, M. Khan, Regional permeability of salmon calcitonin in isolated rat gastrointestinal tracts: transport mechanism using Caco-2 cell monolayer, *AAPS J.* 6 (2004) 36–40.
- [32] M. Kidron, H. Bar-On, E.M. Berry, E. Ziv, The absorption of insulin from various regions of the rat intestine, *Life Sci.* 31 (1982) 2837–2841.
- [33] E. Karasulu, A. Yavasglu, Z. Evrensanal, Y. Uyanikgil, H.Y. Karasulu, Permeation studies and histological examination of sheep nasal mucosa following administration of different nasal formulations with or without absorption enhancers, *Drug. Deliv.* 15 (2008) 219–225.
- [34] A. Allen, G. Flemstrom, Gastroduodenal mucus bicarbonate barrier: protection against acid and pepsin, *Am. J. Physiol. Cell Physiol.* 288 (2005).



# CHAPTER 4

## `predictPotency`

---

The decision tree ensemble random forest have a series of useful diagnostics which have been used in this thesis work.



## Research paper

*In silico* modelling of permeation enhancement potency in Caco-2 monolayers based on molecular descriptors and random forest

Søren H. Welling <sup>a,b</sup>, Line K.H. Clemmensen <sup>b</sup>, Stephen T. Buckley <sup>a</sup>, Lars Hovgaard <sup>a</sup>, Per B. Brockhoff <sup>b</sup>, Hanne H.F. Refsgaard <sup>a,\*</sup>

<sup>a</sup> Global Research, Novo Nordisk A/S, Novo Nordisk Park, 2760 Måløv, Denmark

<sup>b</sup> Technical University of Denmark, DTU Compute, 2800 Kgs. Lyngby, Denmark

## ARTICLE INFO

## Article history:

Received 30 January 2015

Revised 14 May 2015

Accepted in revised form 17 May 2015

Available online 21 May 2015

## Keywords:

Permeation enhancers

Caco-2

Random forest

QSAR

Surfactants

## ABSTRACT

Structural traits of permeation enhancers are important determinants of their capacity to promote enhanced drug absorption. Therefore, in order to obtain a better understanding of structure–activity relationships for permeation enhancers, a Quantitative Structural Activity Relationship (QSAR) model has been developed.

The random forest-QSAR model was based upon Caco-2 data for 41 surfactant-like permeation enhancers from Whitehead et al. (2008) and molecular descriptors calculated from their structure.

The QSAR model was validated by two test-sets: (i) an eleven compound experimental set with Caco-2 data and (ii) nine compounds with Caco-2 data from literature. Feature contributions, a recent developed diagnostic tool, was applied to elucidate the contribution of individual molecular descriptors to the predicted potency. Feature contributions provided easy interpretable suggestions of important structural properties for potent permeation enhancers such as segregation of hydrophilic and lipophilic domains. Focusing on surfactant-like properties, it is possible to model the potency of the complex pharmaceutical excipients, permeation enhancers. For the first time, a QSAR model has been developed for permeation enhancement. The model is a valuable *in silico* approach for both screening of new permeation enhancers and physicochemical optimisation of surfactant enhancer systems.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Development of oral delivery systems for proteins and peptides offers the promise of improved patient compliance compared to conventional parenteral administration. However, bioavailability is, in part, limited due to poor absorption of proteins across the

intestinal epithelial barrier. To effectively deliver a protein systemically this barrier can be modulated by the presence of permeation enhancers [1].

Quantitative Structural Activity Relationship, QSAR methods have been applied extensively for exploration of structural properties of importance for oral absorption of new chemical entities, e.g., QSAR models have been developed for permeability [2] and solubility [3–5]. To our knowledge, no QSAR model for permeation enhancement has previously been published.

Some permeation enhancers have specific mechanisms of action, e.g., modulating the function of tight junctions in the plasma membrane such as zona-occludens-toxin [6], EDTA [7] or melittin [8]. However, the majority of permeation enhancers are primarily surfactants and will non-specifically disrupt the lipid bilayer packing of phospholipids in the epithelial membrane [1]. Surfactants are molecules having segregated lipophilic and hydrophilic domains. Water soluble surfactants tend to pool in the surfaces of water/air and water/lipid, lowering the surface tension. Lowering of surface tensions of water/air surfaces and the ability to enhance the permeability across lipid bilayers correlated

**Abbreviations:** C6, sodium hexanoate; C8, sodium octanoate; c8G, octylglucoside; C10, sodium decanoate/caprate; c12PC, dodecylphosphocholine; c12GPC, dodecanoylglycerophosphocholine; c14GP, myristoylglycerophosphate; CART, classification and regression tree; CDC, chenodeoxycholate; DDM, dodecylmaltoside; EDTA, ethylenediaminetetraacetic acid; GCC, glycochenocholate; GH, glycyrrhizinate; LCC, lauroylcarnitinechloride; LOO-CV, leave-one-out cross validation; MOE, molecular operating environment; PCC, palmitoyl carnitine chloride; QSAR, quantitative structural activity relationship; SM, simomemine; RMSE, root mean square error;  $r_p$ , Pearson's correlation coefficient;  $r_s$ , Spearman rank correlation coefficient; SD, standard deviation; TEER, transepithelial electrical resistance; TDM, tetradecylmaltoside; TDS, sodium tetradecyl sulphate; TDM, tetradeceyl maltoside; TC, taurocholate;  $T_{pot}$ , TEER potency; UC, Ursocholate.

\* Corresponding author at: Insulin Pharmacology Research, Novo Nordisk A/S, Novo Nordisk Park, 2760 Måløv, Denmark.

E-mail address: [hare@novonordisk.com](mailto:hare@novonordisk.com) (H.H.F. Refsgaard).

well for a selection of surfactant-like permeation enhancers [9]. General relations between molecular structures and physicochemical properties of surfactants are thoroughly described by Rosen [10]. Several properties of surfactants, including surface pressure, have previously been modelled with a QSAR approach applying both linear regression and non-linear machine learning models as artificial neural networks, support vector machine or random forest [5,11,12]. Combining the above mentioned concepts, it seems plausible that a QSAR-model of surfactant-like permeation enhancement could be constructed.

Our modelling is based on a Caco-2 data set for 41 surfactant permeation enhancers from Whitehead [13,14] tested in cell monolayers across three concentrations. Hereby, trade-offs between potency, pathway and safety amongst a selection of mainly surfactant-like permeation enhancers were investigated. For this article only the potency data was used. *In vitro* Caco-2 monolayers are cultures of functional, differentiated enterocytes and are widely employed to evaluate permeability rates of drug candidates or pre-formulations [15]. The Caco-2 data for permeation enhancers from Whitehead [13,14] together with molecular descriptors calculated from structure of these surfactants were the basis for the QSAR model.

Non-linear machine learning models can have superior predictive capabilities compared to classical statistical explanatory modelling. However, such machine learning models are often complex “black boxes” – difficult to interpret and discuss [16]. This article presents a promising method to elucidate the interplay of features comprising good permeation enhancers within the complex non-linear model of random forest. Therefore, based on the developed model, we here can recommend ranges of the selected molecular descriptors to obtain high permeation enhancement potency.

## 2. Materials and methods

### 2.1. Materials

Caco-2 cells (ATTC-HTB-37) were obtained from American Type Culture Collection (Manassas, VA). Cell culture media (Dulbecco's modified essential media (DMEM)) and penicillin/streptomycin were purchased from Lonza (Verviers, Belgium). All other supplements (i.e., foetal bovine serum, HEPES buffer and non-essential amino acids (NEAA)) as well as Hanks' balanced salt solution (HBSS) and trypsin were purchased from Gibco, Life Technologies (Carlsbad, CA). Corning Transwell® filter inserts (1.12 cm<sup>2</sup> surface area, 0.4 µm pore diameter) were purchased from Fisher Scientific (Waltham, MA). Bovine serum albumin (BSA) was purchased from Sigma Aldrich (St. Louis, MO). All other reagents were of the highest analytical grade.

### 2.2. Cell culture and TEER measurements

Caco-2 cells (passage numbers 41–49) were seeded at a density of  $2.5 \times 10^5$  cells/flask and grown to 70–90% confluence in DMEM (supplemented with 10% FBS, 100 U/ml penicillin and 100 µg/ml streptomycin and 1% (v/v) NEAA). For transport studies, Caco-2 monolayers were cultured on permeable Transwell® 12 mm diameter inserts at a density of  $10^5$  cells/cm<sup>2</sup> and used after 14–17 days in culture. Cells were cultured at 37 °C and 5% CO<sub>2</sub> atmosphere and the medium was changed every other day. Monolayers were equilibrated in HBSS-based transport buffer 1 h prior to testing. Transepithelial electrical resistance (TEER) was measured with a chop-stick electrode (Millicell-ERS®, Millipore, Billerica, MA) prior to testing, and monolayers with TEER values <600 Ω cm<sup>2</sup> were discarded. TEER was measured after 1 h exposure to permeation enhancers.

## 3. Data processing

### 3.1. Training set

Whitehead et al., tested the ability of 51 permeation enhancers to lower the barrier integrity marker %TEER in Caco-2 cells at 1%, 0.1% and 0.01% (w/v) and published the data set as supplementary materials in two papers [13,14]. Of the 51 permeation enhancers reported, forty-two had computable molecular structures (non-mixtures) and were a wide selection of enhancers which were ascribed to 10 different categories of surfactants: Anionic surfactants, cationic surfactants, zwitterionic surfactants, non-ionic surfactants, bile salts, fatty acids, fatty esters, fatty amines, sodium salts of fatty acids, nitrogen-containing rings and others [13]. EDTA (a calcium chelator) was excluded from the training set because of a non-surfactant-like mechanism together with high potency. The remaining 41 permeation enhancers had surfactant-like structures or low potency e.g., urea could be described as an ineffective surfactant without permeation enhancement effect.

TEER-potency ( $T_{pot}$ ) was defined to concatenate measurements of TEER%-decrease (EP) at the three different concentrations (0.01%, 0.1% and 1% w/v) into one target variable.  $T_{pot}$  was simply defined as the mean TEER%-decrease across the three concentrations as given in Eq. (1).  $T_{pot} = 1$  corresponds to a permeation enhancer lowering TEER% completely at 0.01% (w/v) and  $T_{pot} = 0$  translates to no effect of a permeation enhancer on TEER% even at 1% (w/v). The TEER%-decrease EP is defined as in Eq. (2) and depends of the TEER% before and after treatment with enhancer plus TEER%<sub>+</sub> the background filter resistance.

$$T_{pot} = \frac{EP[0.01\%] + EP[0.1\%] + EP[1\%]}{3} \quad (1)$$

$$EP = 1 - \frac{TEER\%_{AE} - TEER\%_+}{TEER\%_{noAE} - TEER\%_+} \quad (2)$$

From a statistical point of view the loss of information is minimal, as the TEER%-values of the three concentrations were highly correlated. The loadings of the first principal component of a principal component analysis resembled the definition of  $T_{pot}$  and this principal component explained 71% of the variance. From a practical viewpoint  $T_{pot}$  could be seen as a linear approximation of pEC50 (−log effective concentration (w/v) of where 50% TEER-decrease is observed), see Eq. (3). pEC50 itself is dimensionless.

Thus, for a given permeation enhancer having a potency of pEC50 = 1 the corresponding value of  $T_{pot} = 0.5$ .

$$T_{pot} = \frac{pEC50 + 0.5}{3}, \quad \text{for } pEC50 \in [-0.5; 2.5] \quad (3)$$

### 3.2. Software packages, descriptors and model design

The open source R statistical software (v 3.02) was acquired freely from <http://www.r-project.org> and Rstudio integrated development environment (v 0.98.501) also acquired freely from <http://www.rstudio.com>. The R-package ‘randomForest’ (v.4.6) [17,18] was used in the random forest-QSAR model. CAS identification numbers of compounds in the training set were converted to mol-files through SciFinder [19]. Mol-files bundled in sdf-files were imported to the software application MOE [20] and sequentially pre-processed with the following functions: ‘wash’ (simulating an ideal solubilised molecular form), ‘partial charges MMFFA96x’ calculating the electron densities necessary for a number of descriptor algorithms, and finally ‘energy minimize’ relaxing the molecule in the minimum state. All 2D molecular descriptors provided by MOE were computed. The subgroup of 3D descriptors ‘vsurf’ [21] plus the single 3D descriptor ‘dipole’ were calculated as they were relatively fast to compute and therefore suitable for

screening purposes. One new descriptor carbon chain length (CCL) was implemented through R. CCL is the length of the longest saturated non-substituted aliphatic carbon chain of a given permeation enhancer. **Table 1** explains the simple implementation of CCL. After 266 molecular descriptors were acquired, a variable filtering was performed to increase prediction performance. First, fifteen descriptors were excluded for having the same value for more than 95% of the actual training-set. A descriptor having the same value for all permeation enhancers does not provide any information and is problematic for some algorithms which e.g., divide by the variance, which will be zero. Subsequently, 143 redundant descriptors were filtered off, one at a time, until no remaining descriptor pair-wise correlations exceeded  $r_p = 0.9$  (Pearson correlation). This correlation-filtering was a simplified implementation of the CORCHOP routine [22]. Lastly, the remaining descriptors were filtered by their Spearman rank correlation coefficient ( $r_s$ ) to the target variable,  $T_{pot}$ . As Spearman rank correlation utilises the target parameter ( $T_{pot}$ ), it was computed separately on training data for each fold of the cross-validations, so as to avoid latently overfitting. Nevertheless, the random forest algorithm was a robust model and the root mean square error estimated by leave one out cross-validation (RMSE<sub>LOO-CV</sub>) exhibited a variation of less than 20% for any reasonable subsets of pruning parameters. The 30 best  $r_s$ -correlating (or inverse-correlating) descriptors were included in the model. See **Table 2** gives an overview of the descriptors selected for the model. **Fig. 1** depicts the data flow from molecular formulas, computation of molecular descriptors, variable filtering, model training and cross-validation.

The default parameters of the random forest model were used as provided in the R-CRAN package 'randomForest', though the number of decision trees grown was set to 10,000 or 50,000 to ensure a conveniently high reproducibility between model-runs. Variable importance was computed for any descriptor and described the deterioration in prediction accuracy of the model, when permuting the particular descriptor. Variable importance was used to rank the importance of the descriptors and did not influence the model predictions. However, variable importance was a valuable tool to identify the molecular descriptors/features most important for predicting surfactant-like permeation enhancement.

To assess the mechanics of the random forest-QSAR, the package rffc [23,24], which is a diagnostic extension for random forest, was acquired from <https://r-forge.r-project.org/projects/rffc/>. rffc provides forest contributions which is the mean contribution of a given variable to the  $T_{pot}$  prediction of a given permeation enhancer.

### 3.3. Experimental test set

A set of 11 compounds and an additional 3 compounds from the training set were tested in Caco-2 monolayers to generate an experimental test set for validation of the developed random forest-QSAR model. Contrary to the experimental setup of the training data, the experimental test conducted for this paper differs

**Table 2**

Overview of the 30 descriptors applied in the random forest-QSAR model predicting the potency of surfactant-like permeation enhancers in Caco-2 monolayers. Descriptors were computed through MOE (18) except CCL "carbon chain length" implemented for this article.

Group of descriptors:	Amount used	Names of descriptors as available in MOE
Atom counts and bond counts	3	a_nN a_nS b_double
Kier-Hall & Kappa shape:	1	chi1_C
Adjacency and distance matrix:	8	BCUT_SLOGP_0, BCUT_SLOGP_3 BCUT_SMR_3 GCUT_PEOE_0 GCUT_PEOE_3 GCUT_SMR_0 wienerPath
Pharmacophore feature:	1	a_base
Partial charge:	10	Q_PC+ PEOE_RPC+ PEOE_PC+ Q_VSA_PPOS Q_VSA_POL Q_VSA_FPNEG PEOE_VSA_POL PEOE_VSA_FPPOS PEOE_VSA+5 PEOE_VSA+1 PEOE_VSA-1
Surface area, volume and shape:	4	vsurf_IW3 vsurf_ID8 vsurf_CP vsurf_Wp 2
Conformation dependent charge:	1	dipole
Physical properties:	1	log P(o/w)
New descriptor in this article:	1	CCL "carbon chain length"

in terms of media (HBSS versus DMEM, respectively) and incubation times (60 min versus 15 min, respectively). Likewise, morphology of Caco-2 monolayers is expected to have some inter-lab variation [25]. The most lipophilic permeation enhancers were barely soluble at 1% (w/v) at 37 °C and needed to be maintained at this temperature during the experiment at all times to avoid precipitation. Model predictions were compared to experimentally measured values of  $T_{pot}$ . The Squared Pearson correlation coefficient ( $r^2_p$ ) and the root mean square error of ordinary least square fit (RMSE<sub>OLS</sub>) were used as the validation criteria for the linear relationship between model predictions and experimental values. It is acceptable that the slope and offset deviates from 1 and 0 respectively as the absolute measured  $T_{pot}$  is method specific.

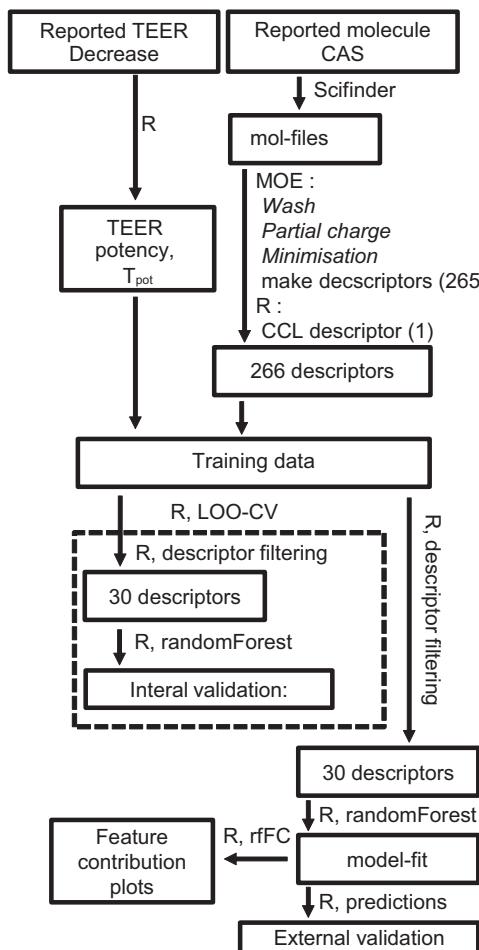
### 3.4. Literature test set

Based on a literature search, nine permeation enhancers were included as a literature test set (**Table 3**). For all included

**Table 1**

Examples of the new descriptor carbon chain length (CCL). CCL estimates the longest sequence of saturated carbon atoms by counting the longest sequence of capital C's in a corresponding SMILES representation of the structure.

Name	Structure	Smiles underscoring	CCL count
Decanoate		O=C(O)CCCCCCCC	9
3-Hydroxydecanoic acid		CCCCCC(CC(=O)O)O	8
Benzoic acid		O=C(O)c1ccccc1	1



**Fig. 1.** Scheme of the modelling process. From top, TEER-data and provided chemical identification (CAS) were processed into TEER potency ( $T_{\text{pot}}$ ) and molecular descriptors. Descriptor filtering is embedded in the leave-on-out cross-validation (LOO-CV). The final random forest model was both used for prediction of external test sets and for diagnostic interpretation through the rfFC-package. R (R) and molecular operating environment (MOE) are software applications. CCL, carbon chain length, molecular descriptor, see Table 1.

**Table 3**

Permeation enhancers from literature tested in Caco-2 monolayers. EC50% is the estimated concentration (w/V)% where the permeation enhancer will lower TEER 50%. pEC50 is the negative logarithm to EC50%. RF predicted potency (w/V)% is the average ability of the permeation enhancer to lower TEER at 1%, 0.1% and 0.01% (w/V)%.

CAS	Compound name	EC50, (w/V)%	pEC50	Refs.
6080-33-7	Simomennine (SM)	2.0	−.30	[26]
81-24-3	Tauro cholate (TC)	.80	−.096	[27,28]
474-25-9	Cheno deoxy cholate (CDC)	.50	.30	[28]
128-13-2	Ursocholate (UC)	.50	.30	[28]
29836-26-8	Glyco octyl (c8G)	.40	.40	[29]
68797-35-3	Glycyrrhizinate (GC)	.50	.70	[30]
325465-45-0	Myristoyl glycerol phosphate (c14GP)	.10	1.0	[31]
20559-18-6	Lauryl glycerol phospho choline (c12GPC)	.025	1.6	[31]
29557-51-5	Dodecyl phosphate choline (c12PC)	.021	1.7	[31]

permeation enhancers from the literature, pEC50 was estimated by interpolation to compare across various experimentally applied concentrations. pEC50 is the negative logarithm of EC50 and has an approximate linear relation to  $T_{\text{pot}}$ , as described in (Eq. (3)). The model performance was validated by the accuracy of the pEC50 prediction for the permeation enhancer in the literature test set. Again the main criteria for comparison were  $r_p^2$  and RMSE<sub>OLS</sub> between interpolated pEC50 values and predicted  $T_{\text{pot}}$  values.

#### 4. Results

A random forest-QSAR model was developed based on a 41 compound training set from literature [13,14] and permeation enhancement potency (TEER% Caco-2) values were matched with molecular descriptors. The predictability of the model was tested through validation. Three types of validation were applied: Internal leave-one-out cross validation, (LOO-CV), experimental validation and literature validation. Lastly, the mechanics from the autonomous random forest-QSAR model was extracted to provide a complimentary insight into which molecular properties there are important for permeation enhancement potency.

##### 4.1. Model validation

Internal cross-validation was used throughout the process of designing a predictive generalisable model of permeation enhancement. Table 4 summarises the validation outcome. Both the internal and experimental validation showed RMSE<sub>OLS</sub> = 0.16–0.17. This error was a sixth of the entire 0–1 range of the  $T_{\text{pot}}$  scale. As  $T_{\text{pot}}$  summarises three concentration levels 1% to 0.1% to 0.01% (w/v) with a 10-fold span between each step, the accuracy was interpreted as to confirm that the model could predict within which 10-fold concentration a given permeation enhancer was effective. Likewise, for the literature validation the RMSE was 0.39, which corresponds to less than half of one unit on the pEC50 scale. One unit of pEC50 is equal to a 10-fold change in 50% effective concentration.

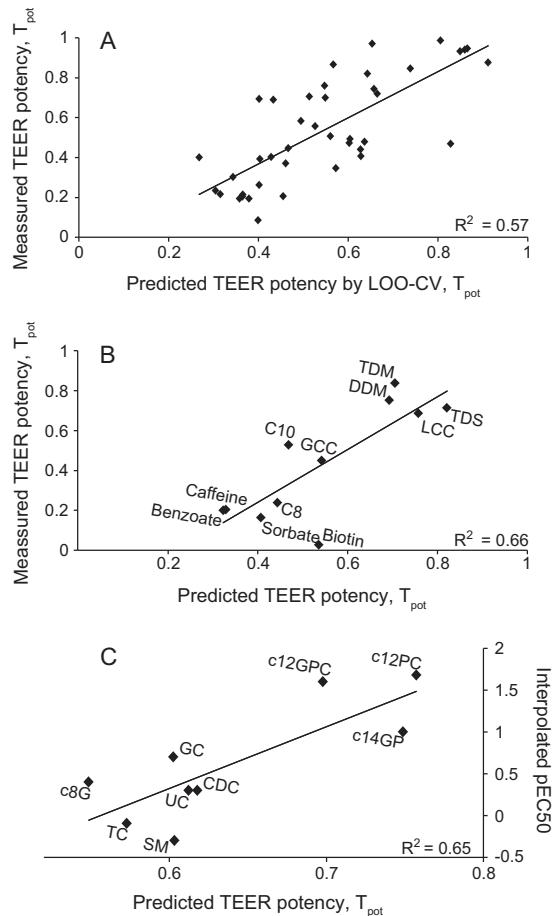
Fig. 2 shows plots of the three types of validations. In part A and B the predicted  $T_{\text{pot}}$  values are plotted against the measured values for the training set data from Whitehead et al. [13,14] and for the experimental data set. Fig. 2C depict the correlation between predicted  $T_{\text{pot}}$  potencies and the actual pEC50 values for the literature test set. The internal LOO validation correlation coefficient was lower,  $r_p^2 = 0.57$  (Fig. 2A), than for the external test-sets,  $r_p^2 = 0.65–0.66$  (Figs. 2B and 1C).

Eleven permeation enhancers were evaluated in Caco-2 monolayers as an experimental test set (Fig. 2B). Biotin and benzoate are widely used food additives and were intended as negative

**Table 4**

Summary of the validation of the random forest QSAR model predicting potency (%TEER) of permeation enhancers in Caco-2 monolayer.  $T_{\text{pot}}$ , a measure of enhancer potency defined as mean decrease of %TEER when applying 1%, 0.1% and 0.01%(w/v) in Caco-2 monolayers. pEC50 is the estimated concentration of which %TEER is decreased 50%. CV-LOO, internal cross validation – LOO. RMSE<sub>OLS</sub>, root mean square error of ordinary least square prediction fit. (a) RMSE, root-mean-square-error adjusted to compare across  $T_{\text{pot}}$  and pEC50 (Eq. (3) in Section 2).

	Training-set	Test-set, experimental	Test-set, literature
Number of enhancers	41	11	9
Data origin	1 article	Experimental	7 articles
Target value	$T_{\text{pot}}$	$T_{\text{pot}}$	pEC50(% $T_{\text{pot}}$ )
Model correlation, $r_p^2$	57% (LOO-CV)	66%	65%
Model error, RMSE <sub>OLS</sub>	0.17	0.16	0.39(0.16 <sup>a</sup> )



**Fig. 2.** Validation plots of random forest-QSAR model predicting the potency (%TEER) of permeation enhancers in Caco-2 monolayer. X-axis is the predicted target response of permeation enhancers' ability to lower %TEER in the Caco-2 monolayers. Y-axis is the true target response. (A) Internal cross-validation plot validating the predictability within the training-set, (B) validation of experimental test set, (C) validation of literature test set.  $T_{pot}$ , mean decrease of %TEER in Caco-2 monolayers of a given permeation enhancer applied at concentrations of 1%, 0.1% and 0.01%. pEC50, negated 10 base logarithm of concentration (%w/v) of 50% TEER.

controls. None of these compounds were measured to be potent permeation enhancers. All permeation enhancers in the experimental test set, with the exception of biotin, were well predicted. Biotin was predicted to elicit a moderate potency, but was devoid of any significant effects when tested in Caco-2 cells. Three compounds SDS, C6 and PCC from the training set were retested to verify, that the experimental setup applied here could reproduce findings from Whitehead et al. (data not shown).

Fig. 2C show the predicted  $T_{pot}$  potencies and the actual pEC50 values for the literature test set. The nine enhancers were surfactants with a single well defined molecular structure and sufficient data points published to estimate a pEC50 value. The range of interpolated pEC50 values from the literature data ranged from -0.3 to 1.7 corresponding to that the most potent permeation enhancer had ~100 times higher potency than the weakest. Three of the nine compounds Myristoyl glycerol phosphate (c14GP), Lauryl glycerol phospho choline (c12PC) and Dodecyl phosphate choline (c12GPC) were markedly more potent than predicted by the model. The exact predicted rankings of the second most and third most potent compounds of the experimental

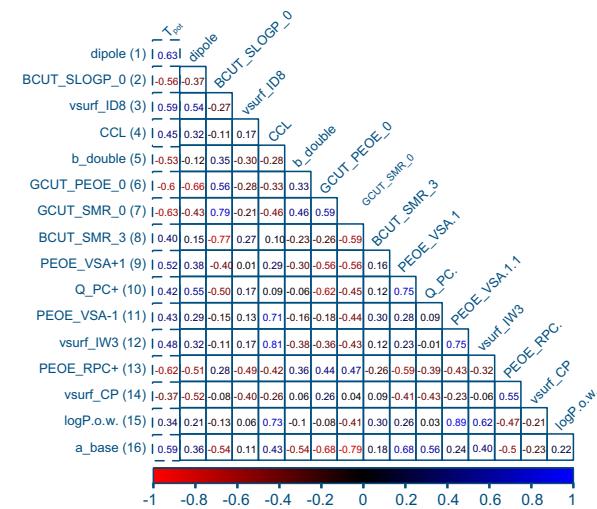
test-set were not correct, but within the expected uncertainty of the model. The same was seen for the group of low potent permeation enhancers.

#### 4.2. Reviewing descriptors useful for prediction of permeation enhancement

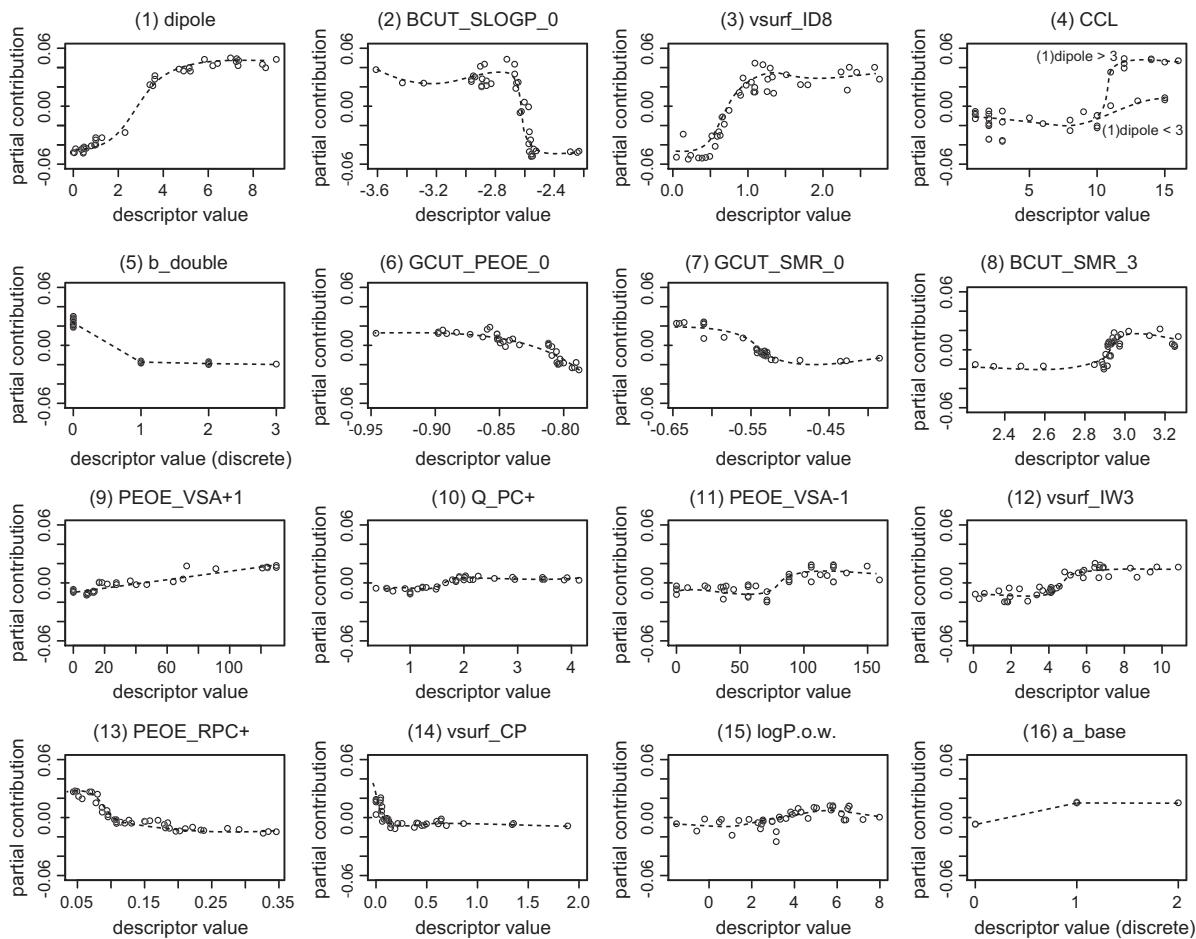
Of the 266 descriptors assessed, 30 descriptors were applied after filtering in the model. Names and grouping of the used descriptors can be seen in Table 2 in the method section. The 16 most important descriptors were included in a Spearman rank correlation matrix depicting their internal rank correlation within the training set (Fig. 3) and their rank correlation with the target parameter  $T_{pot}$ . The strongest absolute correlation coefficient, 0.89, was between variables PEOE\_VSA-1 and logP.o.W. No correlation could exceed the correlation filter limit of 0.9, as described in Section 2. All 16 descriptors were found to be rank correlated with the target value  $T_{pot}$ . The descriptors absolute rank correlations to  $T_{pot}$  ranged from  $r_s = 0.34$  to  $r_s = 0.63$ .

To interpret the precise contribution of each descriptor within the model, a diagnostic method termed 'Feature Contributions' [23,24] was used. A novel diagnostic plot of the feature contributions is presented in Fig. 4. The feature contributions of the 16 most important descriptors describing permeation enhancer-potency ( $T_{pot}$ ) in the training-set were plotted against their respective descriptor values. This provided an intuitively graphical interpretation of how features within the random forest-QSAR model context affected the  $T_{pot}$  prediction. It represents an innovative way to graphically present the computed feature contributions. This expansion of a regular random forest model summarises the total partial descriptor contribution for any permeation enhancer in the training set. The predicted  $T_{pot}$  values for a given enhancer are equal to the sum of all partial descriptor contributions, which again is dependent of the actual feature values, as outlined in Fig. 4.

Fig. 4 shows that descriptor [2, BCUT\_SLOGP\_0], and [3, vsurf\_ID8] had sharp thresholds separating the positive (i.e., beneficial) and negative contributions to the  $T_{pot}$  value of each permeation enhancer. For [1, dipole] there was also a separation between positive and negative contribution to the  $T_{pot}$  value.



**Fig. 3.** Spearman-correlation matrix of the 16 most important molecular descriptors listed by decreasing variable importance (most important first) within the random forest-model. The target variable  $T_{pot}$ , the ability of permeation enhancers to lower electrical resistance in Caco-2 monolayers across concentrations 0.1–1%(w/v), has also been included.



**Fig. 4.** Random forest feature contributions diagnostics. Scatter plots of partial descriptor contributions versus the individual descriptor values for each observation of the training set for the 16 descriptors with highest 'importance' within the random forest model. Scatter plots enumerated in descending order of 'importance'. Dashed trend lines were added to the plots. Each plot outlines a partial function of a variable as its role in the model.

Good permeation enhancers, compounds with high  $T_{\text{pot}}$  values, had high [1, dipole] values, compounds with high [2, BCUT\_SLOGP\_0] values, and [3, vsurf\_ID8] values.

[1, dipole] reflected the overall dipole moment calculated from the partial charges of the molecule. The descriptor contribution of [1, dipole] within the random forest model was well described as a function of the descriptor value itself. Thus, there was no interaction with other descriptors. On the contrary, the descriptor contributions of descriptor [4, CCL], the maximum aliphatic carbon chain length, varied for many permeation enhancers having the exact same chain length. This pointed to an interaction phenomenon between descriptors. When only emphasising permeation enhancer above the threshold value for dipole > 3, these permeation enhancers were all accredited positively for having a high CCL value. Oppositely, enhancers with dipole < 3 were accredited neutral for any CCL value. The interpretation drawn was that molecules with a high dipole moment are likely to have a hydrophilic domain and if combined with an aliphatic carbon chain of length > 10, the molecules are likely to have surfactant properties. Conversely, compounds with no significant hydrophilic groups such as oils, would not function as enhancers alone despite long carbon chains.

Descriptor [3, vsurf\_ID8] reflected the hydrophilic domains separation from the lipophilic domains which was expected to be a

central surfactant-like property. More precisely, the [3, vsurf\_ID8] reflected the distribution of hydrophobic or hydrated domains and their distance from the mass centre. It was observed that the model evaluated low [14, Vsurf\_CP] values as being beneficial. [14, Vsurf\_CP] is a micelle critical packing parameter. Cone shaped surfactants would in generally have a low [14, Vsurf\_CP] value and thereby a low critical packing number. A low critical packing number favours micellar aggregation, not liposomal. Throughout Fig. 4, the descriptors were generally declining in magnitude of feature contributions, and thus less influential.

## 5. Discussion

The random forest-QSAR model of permeation enhancement in Caco-2 cells was shown to provide reasonable estimates of permeation enhancer potency. Such a model has to the knowledge of the authors not been developed previously. Within the paradigm, that surfactant-like properties are key features of most enhancers, it was confirmed that a model could be constructed inspired by the *in silico*, *in vitro* and *in vivo* models of surfactant surface tension depression [9,11,12].

In the case of a new modelling area, a large amount of descriptors could possibly become useful. However, the relatively small

size of training examples makes such a selection process challenging. The training set of 41 permeation enhancers and 266 descriptors is an example of sparse (small  $n$  – large  $p$ ) modelling which can lead to overfitted non-generalisable models. Filtering/pruning constants, redundant and non-correlated descriptors improve model performance.

The ensemble method, random forest, is an extension of the classification and regression tree, CART. Decision tree models branch out/split data into increasingly smaller sub groups of samples having the most similar target response. Such CART decision trees are highly adaptable of many types of data, but also easily overfitted to the training data. Thus, the model becomes adaptable but highly noise sensitive. The random forest model is an extra layer to the CART, reducing noise without losing its adeptness. In short, random forest is an ensemble of many of such uncorrelated decision trees (e.g., 500). Though each tree is susceptible to random inference, the average prediction of many decision trees have been shown to be much less prone to overfit thus its conclusions/predictions are more generalisable across data sets [17]. That said, the conclusions from any model approach will always be limited by the diversity of the training set. In this example, scientific literature, the basis of this model training, tend to have a bias towards not reporting any compounds which lack an enhancing effect. In the case of this training set, all compounds elicited at least a low enhancement effect. A feature of decision tree-based models is that, they cannot extrapolate beyond the target range ( $T_{pot}$ ) of the training set. Likewise, caffeine and benzoate were predicted in absolute terms to be more potent than experimentally measured, as no learning examples could suggest such a weak potency. Nonetheless, this is not of much practical concern in terms of predicting new permeation enhancer candidates. A useful model does not have to distinguish very weak enhancers from non-enhancers.

C10, sodium decanoate, is one of the most described enhancers in literature [7,9,32,33]. Amongst the reported mechanisms of C10 are phosphorylation cascades and intracellular calcium signalling leading to tight junction opening [9,33]. Such mechanisms are far too complex to be captured from a training sample of this size. Conceivably, this may be why C10 was predicted to be a mediocre permeation enhancer, yet elicited a relatively stronger potency (Fig. 2B), caused by components not captured by the model. It is expected that doubling or tripling the size of the training set would improve prediction accuracy significantly. This would require testing another 40–80 permeation enhancers in three concentrations in Caco-2 monolayer.

Fig. 4, the feature contributions versus descriptor values of each training permeation enhancer, provides a novel and very useful way to learn from the random forest-model. For example, a molecule having a dipole > 3, a Vsurf\_ID8 > 1 and a BCUT\_SLOGP\_0 < -2.7 and CCL > 10 would appear to bear promising starting point. Furthermore, the data in Fig. 4 suggested interaction for e.g. CCL > 10, only contributing positively conditioned when dipole > 3. That carbon chain length is only conditionally advantageous matches the general understanding of surfactant-like properties. Such simple rules can help to understand what modifications of an enhancer can be made without incurring a loss of potency. The abundance of partial charge related descriptors (see Table 2) was interpreted as a consequence of, that most surfactants have one or more polar domains neighbouring carbon hydride domains and an induced dipole moment across the border [12].

The feature contributions technique represents a novel approach to data analysis and has the potential to be employed as a powerful explorative tool within many scientific areas. QSAR

models based on algorithm models such as random forest are designed to map associations (not necessarily causal) between features and the target parameters to optimise predictions. It should be noted that this is also the case for classical statistical approaches [16]. Nevertheless, as discussed above, the suggestions from the feature contributions are plausible causal from a physicochemical point of view.

Other core aspects relating to oral protein formulation such as solubility, stability and metabolism are not encompassed in the existing approach. Thus, their inclusion is necessary in order to yield a fully predictive model of protein permeation. When designing/screening for new enhancers as excipients in protein-based drug formulations, various other requirements, such as solubility, should be considered.

Thus, by applying the described *in silico* model an *a priori* prediction of the permeation enhancer potency of a surfactant can be determined based upon its structure and hence obviate the need for extensive permeability screening of novel compounds.

## 6. Conclusions

Random forest-QSAR modelling utilising molecular descriptors calculated from the molecular structure was shown useful for predicting permeation enhancer potency. Although absorption of proteins is a complex biologic phenomena, the surfactant-like properties of permeation enhancers comprise a relatively manageable component.

Sparse data combined with the biological noise (unexplained) component is a challenge to build a robust predictive model. To reduce the estimation error, the prediction challenge was alleviated in two ways:

- (1) TEER readings of three concentration levels were joined into a single value target ( $T_{pot}$ ) to create an approachable modelling question: Is the potency of a new surfactant-like enhancer high, medium or low?
- (2) Filtering of correlated descriptors to reduce redundant information and to remove descriptors with no univariate correlation to target parameter was performed to avoid too many descriptors being progressed to the random forest model with few training examples.

From the validations employed i.e., internal cross-validation, experimental validation and literature validation, the model was found to predict potency of permeation enhancers. Furthermore, it was possible to extract common structural features for high potency enhancers. Such knowledge is useful to assess the credibility of the built model and/or inspire our understanding of what makes a surfactant-like permeation enhancer potent.

Hereby, we have outlined how to robustly perform *in silico* screening for permeation enhancers with non-linear random forest, with the possibility to assess and learn from the model. The provided QSAR model forms a good basis for a systematically approach for the development of oral therapeutics formulated with potent permeation enhancers.

## Acknowledgments

Christian Vind assisted with molecular descriptors and Sten B. Christensen assisted with literature search.

This work, a part of an industrial Ph.D project for Søren Welling, was granted by The Danish Agency for Science, Technology and Innovation and the company Novo Nordisk A/S.

Søren Welling, Hanne Refsgaard, Stephen Buckley and Lars Hovgaard are employees and/or shareholders of Novo Nordisk A/S.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.ejpb.2015.05.012>.

## References

- [1] B.J. Aungst, Absorption enhancers: applications and advances, *AAPS J.* 14 (2012) 10–18.
- [2] H.H.F. Refsgaard, B.F. Jensen, P.B. Brockhoff, S.B. Padkjær, M. Guldbrandt, M.S. Christensen, In silico prediction of membrane permeability from calculated molecular parameters, *J. Med. Chem.* (2005) 805–811.
- [3] B. Fredsted, P.B. Brockhoff, C. Vind, S.B. Padkjær, H.H.F. Refsgaard, In silico classification of solubility using binary k-nearest neighbor and physicochemical descriptors, *Mol. Inform.* 26 (2007) 452–459.
- [4] D.S. Palmer, N.M. O'Boyle, R.C. Glen, J.B.O. Mitchell, Random forest models to predict aqueous solubility, *J. Chem. Inf. Model.* 47 (2007) 150–158.
- [5] L.D. Hughes, D.S. Palmer, F. Nigsch, J.B.O. Mitchell, Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and log P, *J. Chem. Inf. Model.* 48 (2008) 220–232.
- [6] A. Fassano, S. Uzzau, Modulation of intestinal tight junctions by zonula occludens toxin permits enteral administration of insulin and other macromolecules in an animal model, *J. Clin. Invest.* 99 (1997) 1158–1164.
- [7] M. Tomita, M. Hayashi, S. Awazu, Absorption-enhancing mechanism of EDTA, caprate, and decanoylcarnitine, *J. Pharm. Sci.* 85 (1996) 608–611.
- [8] S. Maher, L. Feighery, D.J. Brayden, S. McClean, Melittin as an epithelial permeability enhancer I: investigation of its mechanism of action in Caco-2 monolayers, *Pharm. Res.* 24 (2007) 1336–1345.
- [9] W.J. Xia, H. Onyuksel, Mechanistic studies on surfactant-induced membrane permeability enhancement, *Pharm. Res.* 17 (2000) 612–618.
- [10] M.J. Rosen, Surfactants and Interfacial Phenomena, third ed., Hoboken, New Jersey, 2000.
- [11] Z.W. Wang, D.Y. Huang, G.Z. Li, X.Y. Zhang, L.L. Liao, Effectiveness of surface tension reduction by anionic surfactants-quantitative structure-property relationships, *J. Dispers. Sci. Technol.* 24 (2003) 653–658.
- [12] J. Hu, X. Xhang, Z. Wang, A review on progress in QSPR studies for surfactants, *Int. J. Mol. Sci.* 11 (2010) 1020–1047.
- [13] K. Whitehead, S. Mitragotri, Mechanistic analysis of chemical permeation enhancers for oral drug delivery, *Pharm. Res.* 25 (2008) 1412–1419.
- [14] K. Whitehead, N. Karr, S. Mitragotri, Safe and effective permeation enhancers for oral drug delivery, *Pharm. Res.* 25 (2008) 1782–1788.
- [15] B. Sarmento, F. Andrade, S.B. da Silva, F. Rodrigues, J. das Neves, D. Ferreira, Cell-based in vitro models for predicting drug permeability, *Expert Opin. Drug Metab. Toxicol.* 8 (2012) 607–621.
- [16] G. Shmueli, To explain or to predict?, *Stat. Sci.* 25 (2010) 289–310.
- [17] L. Breiman, Random forests, *Mach. Learn.* 45 (2001) 5–32.
- [18] A. Liaw, M. Wiener, Classification and regression by randomForest, *R News* 2 (2002) 18–22.
- [19] SciFinder Scholar, version 2014, Chemical Abstracts Service, Columbus, OH, 2014; RN (multiple look-ups, +50).
- [20] Molecular operating environment (MOE), 2012–2013.8, Chemical Computing Group Inc., 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.
- [21] G. Cruciania, P. Crivori, P.A. Carrupt, B. Testab, Molecular fields in quantitative structure-permeation relationships: the VolSurf approach, *Theochem* 503 (2000) 17–30.
- [22] D.J. Linvingstone, E. Rahr, CORCHOP – an interactive routine for the dimension reduction of large QSAR data sets, *Quant. Struct. – Act. Relat.* 8 (1989) 103–108.
- [23] A. Palczewska, J. Palczewski, R.M. Robinson, D. Neagux, Interpreting random forest classification models using a feature contribution method, in: T. Bouabana-Tebibel, S.H. Rubin (Eds.), *Advances in Intelligent Systems and Computing: Integration of Reusable Systems*, Springer, Heidelberg, 2014, pp. 193–218.
- [24] V.E. Kuz'min, P.G. Polishchuk, A.G. Artemenko, S.A. Andronati, Interpretation of QSAR models based on random forest methods, *Mol. Inform.* 30 (2011) 593–603.
- [25] R. Hayashi, C. Hilgendorf, P. Artursson, P. Augustijns, B. Brodin, P. Dehertogh, et al., Comparison of drug transporter gene expression and functionality in Caco-2 cells from 10 different laboratories, *Eur. J. Pharm. Sci.* 35 (2008) 383–396.
- [26] Z. Lu, W. Chen, A. Viljoen, J.H. Hamman, Effect of sinomenine on the in vitro intestinal epithelial transport of selected compounds, *Phytother. Res.* 24 (2010) 211–218.
- [27] U. Werner, T. Kissel, M. Reers, Effects of permeation enhancers on the transport of a peptidomimetic thrombin inhibitor (CRC 220) in a human intestinal cell line (Caco-2), *Pharm. Res.* 13 (1996) 1219–1227.
- [28] S. Michael, M. Thöle, R. Dillmann, A. Fahr, J. Drewe, G. Fricker, Improvement of intestinal peptide absorption by a synthetic bile acid derivative, cholicysarcosine, *Eur. J. Pharm. Sci.* 10 (2000) 133–140.
- [29] P.P. Tirumalasetty, J.G. Eley, Permeability enhancing effects of the alkylglycoside, octylglucoside, on insulin permeation across epithelial membrane in vitro, *J. Pharm. Pharm. Sci.* 9 (2006) 32–39.
- [30] M. Sakai, T. Imai, H. Ohtake, H. Azuma, M. Otagiri, Effects of absorption enhancers on the transport of model compounds in Caco-2 cell monolayers: assessment by confocal laser scanning microscopy, *J. Pharm. Sci.* 86 (1997) 779–785.
- [31] D.Z. Liu, E.L. Lecluyse, D.R. Thakker, Dodecylphosphocholine-mediated enhancement of paracellular permeability and cytotoxicity in Caco-2 cell monolayers, *J. Pharm. Sci.* 88 (1999) 1161–1168.
- [32] S. Maher, T.W. Leonard, J. Jacobsen, D.J. Brayden, Safety and efficacy of sodium caprate in promoting oral drug absorption: from in vitro to the clinic, *Adv. Drug Deliv. Rev.* 61 (2009) 1427–1449.
- [33] T. Lindmark, Y. Kimura, P. Artursson, Absorption enhancement through intracellular regulation of tight junction permeability by medium chain fatty acids in Caco-2, *J. Pharmacol. Exp. Ther.* 284 (1998) 362–369.



# CHAPTER 5

## interpretTheForest

---

The decision tree ensemble random forest have a series of useful diagnostics which have been used in this thesis work.

# Forest Floor Visualizations of Random Forests

Soeren H. Welling<sup>1,2</sup>, Hanne H.F. Refsgaard<sup>2</sup>, Per B. Brockhoff<sup>1</sup> and Line H. Clemmensen<sup>\*1</sup>

<sup>1</sup>Department of Applied Mathematics and Computer Science, Technical University of Denmark, Matematiktorvet, Building 324, 2800 Kgs. Lyngby, Denmark

<sup>2</sup>Novo Nordisk Global Research, Novo Nordisk Park 1, 2760 Maaloev, Denmark

June 1, 2016

## Abstract

We propose a novel methodology, forest floor, to visualize and interpret random forest (RF) models. RF is a popular and useful tool for non-linear multi-variate classification and regression, which yields a good trade-off between robustness (low variance) and adaptiveness (low bias). Direct interpretation of a RF model is difficult, as the explicit ensemble model of hundreds of deep trees is complex. Nonetheless, it is possible to visualize a RF model fit by its mapping from feature space to prediction space. Hereby the user is first presented with the overall geometrical shape of the model structure, and when needed one can zoom in on local details. Dimensional reduction by projection is used to visualize high dimensional shapes. The traditional method to visualize RF model structure, partial dependence plots, achieve this by averaging multiple parallel projections. We suggest to first use feature contributions, a method to decompose trees by splitting features, and then subsequently perform projections. The advantages of forest floor over partial dependence plots is that interactions are not masked by averaging. As a consequence, it is possible to locate interactions, which are not visualized in a given projection. Furthermore, we introduce: a goodness-of-visualization measure, use of colour gradients to identify interactions and an out-of-bag cross validated variant of feature contributions.

## 1 Introduction

We propose a new methodology, forest floor, to visualize regression and classification problems through feature contributions of decision tree ensembles such as random forest (RF). Hereby, it is possible to visualize an underlying system of interest even when the system is of higher dimensions, non-linear, and noisy. 2D or 3D visualizations of a higher-dimensional structure may lead to details, especially interactions, not being identifiable. Interactions in the model structure mean that the model predictions in part rely on the interplay on two or more features. Thus, the interaction parts of a model structure cannot be reduced to additive scoring rules, one for each feature. Likewise, to plot single feature-to-prediction relationships is not a sufficient context for visualizing any interactions. Often a series of complimentary visualizations are needed to produce an adequate representation. It can be quite time consuming to look through any possible low dimensional projection of the model structure to check for interactions. For-

est floor guides the user in order to locate prominent interactions in the RF model structure and to estimate how influential these are.

For RF modeling, hyper parameter tuning is not critical and default parameters will yield acceptable model fits and visualizations in most situations [10, 23]. Therefore, it is relatively effortless to train a RF model. In general, for any system where a model has a superior prediction performance, it should be of great interest to learn its model structure. Even within statistical fields, where decision tree ensembles are far from standard practice, such insight from a data driven analysis can inspire how to improve goodness-of-visualization of a given model driven analysis.

Although the RF algorithm by Breimann [3] has achieved the most journal citations, other later decision tree ensemble models/algorithms such as ExtraTrees [14], conditional inference forest [8], Aborist [21], Ranger [26] and sklearn.random森林 [17] will often outperform the original RF on either prediction performance and/or speed. These models/algorithms differ only in their software im-

\*lkhc@dtu.dk

plementation, split criterion, aggregation or in how deep the trees are grown. Therefore all variations are compatible with the forest floor methodology. Another interesting variant, rotation forest [19], does not make univariate splits and is therefore unfortunately not directly compatible with forest floor visualizations. To expand the use of feature contributions and forest floor, we also experimented with computing feature contributions for gradient boosted trees [6]. This is possible, as splits still are univariate and trees contribute additively to the ensemble prediction. A proof-of-concept of computing feature contributions on gradient boosted regression trees and visualizations are provided in supplementary materials.

Decision trees, as well as other machine learning algorithms, such as support vector machines and artificial neural networks can fit regression and classification problems of complex and noisy data, often with a high prediction performance evaluated by prediction of test sets, n-fold cross validation, or out-of-bag (OOB) cross validation. The algorithms yield data driven models, where only little prior belief and understanding is required. Instead, a high number of observation are needed to calibrate the adaptive models. The models themselves are complex black-boxes and can be difficult to interpret. If a data driven model can reflect the system with an impressive prediction performance, the visualization of the model may deduce knowledge on how to interpret the system of interest. In particular, a good trade-off between generalization power and low bias is of great help, as this trade-off in essence sets the boundary for what is signal and what is noise. The found signal is the model fit, which can be represented as the mapping from feature space to prediction space (output, target, response variable, dependent variable,  $y$ ). The noise is the residual variance of the model. The estimated noise component will both be due to random/external effects but also lack of fit.

## 1.1 Overview of the article

In this article we introduce the forest floor methodology. The central part is to define a new mapping space visualization, forest floor. Forest floor rely on the feature contributions method [9][16], rather than averaging many projections (partial dependence) [6] or projecting the average (sensitivity analysis) [5]. In Section 1.2 these previous mapping space visualizations are introduced and the challenges to overcome are discussed. In the theory section, 2.1, we discuss the feature space, prediction space and the joined mapping space for any regression or classification model and define local increments as vectors in the prediction space. Properties of the RF algorithm by Breimann [3] and the fea-

ture contributions method by Kuz'min *et al* [9] and Palczewska *et al* [16] are highlighted and illustrated in section 2.2. In section 2.3 we argue that the prediction of any node in any tree is a point in the prediction space and the local increments are the vectors that connect the nodes of the trees. Any prediction for any observation is basically the sum of a series of local increments plus the grand mean / base rate. Since local increments are vectors and not a tree graph, the sum of vectors is not dependent on the order of the sequence. In Section 2.4 we show how that feature contributions, a particular reordering of local increments by splitting feature, can be used to decompose the model structure 2.4. We also introduce a new cross-validated variant of feature contributions and provide an elaborated definition of feature contribution to also account exactly for the bootstrapping process and/or stratification.

The materials and methods sections, 3.1 and 3.2, provide instructions on how to reproduce all visualization in this paper. The result section 4 is dedicated to three practical examples of visualizing models with forest floor. The three examples are a simulated toy data set, a regression problem (white wine quality) and a classification problem (contraception method choice). A low-dimensional visualization is not likely to convey all aspects of a given RF mapping surface. For all practical examples, we describe how to find an adequate series of visualizations that do.

## 1.2 Representations of random forest models

A RF model fit, like other decision tree based models, can be represented by the graphs of the multiple trees. Few small tree graphs can be visualized and comprehended. However, multiple fully grown trees are typically needed to obtain an optimal prediction performance. Such a representation cannot easily be comprehended and is thus inappropriate for interpretation of model fits. A random forest fit can be seen as a large set of split rules which can be reduced to a smaller set of simpler rules, when accepting a given increase in bias. This approach has been used to reduce the model complexity [13]. But if the minimal set of rules still contains a large number, e.g. hundreds or thousands, then this simplified model fit is still incomprehensible. It is neither certain which rules have influence on predictions nor which rules tend to cancel each other out. We believe that the rule-set or tree-structure representations are mainly appropriate to understand how a RF algorithm possibly can model data. On the other hand, these representations are indeed inappropriate for interpreting RF model fits and conveying the overall model structure. For that pur-

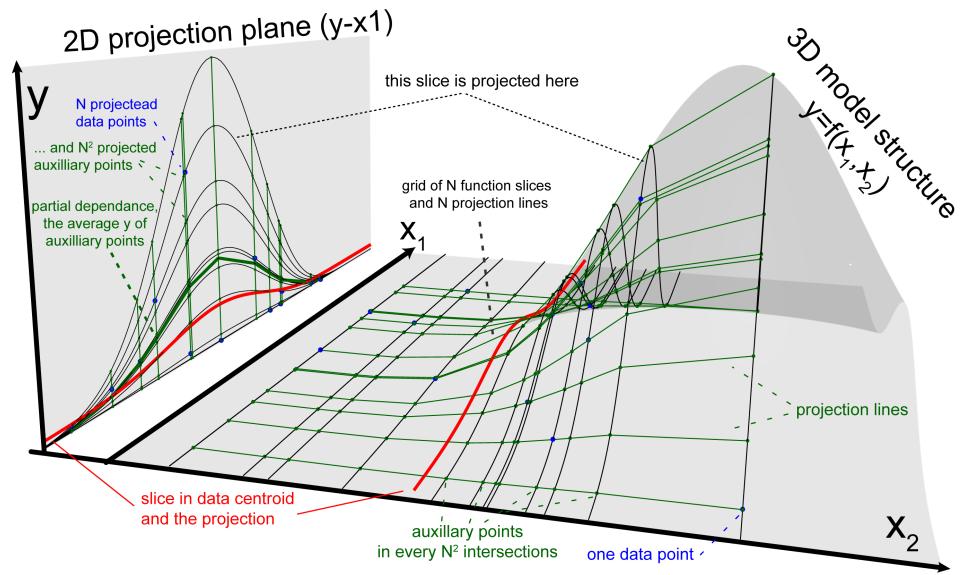


Figure 1: Illustration of sensitivity analysis and partial dependence plots. The grey response surface depicts a given learned model structure of two input features ( $X_1$  and  $X_2$ ) and one prediction axis ( $\hat{y}$ ). 11 data points vs. predictions are depicted as blue dots. 1D-sensitivity analysis (fat red lines): one partial function slice intersects the centroid where  $X_2 = \bar{X}_2$  and is projected to the  $X_1-y$  plane. d-ICE plot: Multiple function slices (black lines) all parallel to  $X - 1$  intersect each one data point and all slices are projected to the  $X_1-y$  plane. Partial dependence plots: Each data point intersected by one black line is projected to any black lines (green points). The green point outline a grid. All green and blue points are projected into the  $X_1-y$  plane, and the fat green line connects the average prediction values as a function of  $X - 1$ . This illustration can be generalized to any dimensional reduction.

pose, a mapping space visualization is superior in terms of visualization and communication.

If we join the feature space and prediction space, this function will be represented as a geometrical shape of points. Each point represents one prediction for a given feature combination. This geometrical shape is the model structure and is an exact representation of the model itself. Nevertheless, for a given  $d$ -dimensional problem where  $d > 3$ , this is still difficult to visualize or even comprehend. Instead, one may project/slice or decompose the high-dimensional mapping into a number of marginal visualization where small subsets of features can be investigated in turns. This allows us to comprehend the isolated interplay of one or a few features in the model structure.

Following, we will introduce previous examples of mapping space visualizations to specify what forest floor aims to improve. Different types of sensitivity analysis (SA) were used by Cortez and Embrechts to make such investigations [5], we will here discuss sensitivity analysis and data based sensitivity analysis. First a supervised machine learning model is trained. Next the model is probed. That means to input a set of simulated feature observations (points in feature space) into the model fit and record the output (target predictions). Instead of probing the entire high-dimensional mapping space, only one confined slice of fewer dimensions is probed in order to make feasible visualizations.

The simplest visualization in SA is one dimensional (1D-SA), where a single feature is varied in a range of combinations, and this range will span the X-axis of the visualization. When two features are varied (2D-SA), the resulting grid of combinations will span the XY-plane. All other features must be fixed at e.g. the mean value, the feature centroid of the training set. The model fit is probed with these observations and the resulting predictions will be plotted by the Z-axis. The obtained line/surface will now visualize one particular 2D or 3D slice of the full mapping structure.

In figure 1, a non-linear regression model structure ( $y = \sin(X_1)^8 \sin(X_2)^8 + \epsilon$ ) is represented by the grey transparent surface. The model has two feature axes in the horizontal XY-plan and the prediction axis by the vertical Z-axis. Thus, the mapping space has 3 dimensions and the model structure is some curved 2D-surface which connect any given feature combination with one prediction. The red line/slice in the model structure is the example of an 1D-SA visualization. This single slice is projected into the  $X_1$ -Z plane. This 1D-SA projection portrays the partial effect of feature  $X_1$  in the special case, where other features are set to mean observed value. Notice that the red line almost completely misses the local hill in the model structure.

A single low dimensional slice of the mapping structure can easily miss prominent local interactions, when number of model dimensions is high.

A 2D-SA slice can explain a main effect and/or the possible interaction within two selected features. Figure 1 only illustrates a 1D-SA slice projection, but represents the idea of any projection. The depicted model structure itself could infact be a 2D-SA projection of a higher dimensional model structure. Whether a given slice is a good generalization of the full mapping structure is unknown. A good generalization means that any parallel slice, where the fixed features are set to another combination, yield the same XYZ-visualization, with only perhaps a fixed offset in the prediction axis (Z) [7]. We will for now term that such visualization has a high goodness-of-visualization. In section 2.4 we will propose a metric for goodness-of-visualization. For a data structure with only additive effects and no interactions, the obtained model mapping structure is likely to have no interactions as well as any slice will be identical to its many parallel counterparts. In Figure 1, all the black parallel slices to the red slices give different projection lines in the mirror plane which could not be corrected by a simple offset. Therefore the model structure must have an interaction which cannot be seen in this projection alone. The iceBOX package displays multiple projection lines to search for masked interactions and is a good alternative to the forest floor approach [7].

A second concern is whether a given slice or slices extrapolate the training data. For a RF model with a satisfactory cross validated prediction performance, the mapping structure will represent the underlying data structure, but only within the proximity of the training data. Extrapolated areas of the mapping structure are far from guaranteed to represent an underlying data structure. Several different non-linear learners (RF, SVM, ANN, etc.) may easily have comparable model structures in the proximity to training data points, whereas far from the training set the models will heavily disagree. For RF models containing dominant interaction effects, the mapping structure on the borders of the training data becomes noise sensitive, as decision trees only can extrapolate parallel to feature axes, as the splits only are univariate. RF models only containing additive main effects have stable and smooth mapping structure at the borders of the training data. Model extrapolation of random forests with dominant interaction effects have been illustrated in supplementary materials.

SA plots remain a useful tool. When forest floor yield plots of similar structure, these plots generally represents the model mapping well. Visualization of multiple parallel projections, the so called d-ICE plots (individual conditional expectation) with the

ICEbox package, can also reveal interactions [7]. However multiple projection lines cannot directly filter out main effects by other features. These will tend to offset the projection lines on the prediction axis.

A frequently used visualization method proposed by Friedman is the partial dependence plot (PD) which is the same as what Cortez and Emblechts later have termed data-based sensitivity analysis (DSA)[5, 6]. In Figure 1, the green fat line in the mirror plane represents a partial dependence projection. Whereas 1D-SA and 2D-SA only project the slice intersecting e.g. the training data centroid, the partial dependence plot projects multiple slices. Each projected slice intersects one data point. The partial dependence line is the average prediction values of all slices. Thus, the obtained PD visualization summarizes all parallel slices of the mapping structure by averaging. To summarize, SA averages and then projects, whereas PD projects and then averages. ICE-plot projects many slices and do not aggregate lines. The PD approach may improve generalization across slices as it up-weights the parts of mapping structure, that are well represented by data points. Still, interactions between varying and fixed features will be lost by averaging. Furthermore, the PD projections form a regular data grid spanned by the data observations. See the grid of black and green lines on the model structure surface in Figure 1. However, for data sets with high feature collinearity, data points will mainly be positioned in one diagonal of the grid, whereas the remaining part of the grid will span extrapolated parts of the model structure. This extrapolation occur for both SA, PD and d-ICE-plots.

Feature contributions was introduced by Kuz'min [9] for RF regression and elaborated by Palczewska *et al* [16] to also cover RF multi-classification. Feature contributions are RF predictions split into components by each feature. Feature contributions are essentially computed utilizing information from the tree networks of a RF model. Feature contributions have not before been used or understood in conjunction with the idea of function mapping structure. The contribution of this paper, is to show that feature contributions can be understood as a different way of slicing the mapping structure. From this insight the methodology, forest floor, was developed.

We have developed a number of tools to increase the usefulness of the forest floor methodology. These are: Out-of-bag cross validated feature contributions to increase robustness without increasing computation time, goodness-of-visualization tests to evaluate how well slices generalize the mapping structures and color gradients traversing mapping space to visually identify latent sources of interac-

tions. Furthermore, the methods have been implemented as a freely available R-package, from which all mapping visualizations of this paper originate. The R-package forestFloor [25] aims to assist the user visualizing a given RF model fit through a serious of appropriately chosen slices.

## 2 Theory and calculation

Here is provided a new notation for RF regression and classification to combine a mapping space representation with the feature contributions method developed by Kuz'min [9] and Palczewska *et al.* [16]. Moreover to obtain an exact decomposition of the model structure, we expand the previous notion of feature contribution to also cover the initial bootstrap and/or stratification step for each decision tree. For RF multi-classification we describe a probabilistic (K-1)-simplex prediction space, to improve the interpretation of feature contributions. Lastly we introduce how to calculate out-of-bag cross-validated feature contributions.

### 2.1 Defining regression and classification mappings

Any regression model  $f_r$  can be seen as a mapping between a  $d$ -dimensional feature space  $X \in \mathbb{R}^d$  and a prediction scale  $\hat{y} \in \mathbb{R}^1$

$$\hat{y} = f_r(X), \quad (1)$$

where  $X$  represents the infinite set of points in the feature space. A subset of points in  $X$  can be notated as e.g.  $X_t$  where  $t$  is a defined set. Single value entries of a countable subset of  $X$  is notated as  $x_{ij}$  where  $i \in \{1, \dots, N\}$  ( $N$  points) and  $j \in \{1, \dots, d\}$  ( $d$  features).  $\hat{y}$  represents the entire prediction scale, where  $\hat{y}_s$  could be a subset, if countable with point entries  $\hat{y}_i$ .

The entire mapping can be represented as a  $d$ -dimensional (hyper)surface  $S$  in a  $d+1$ -dimensional mapping space  $V$ .  $S$  can be understood as a learned model structure trained on a set of training observations. Obviously, if  $d \in \{1, 2\}$ , then  $S$  can conveniently be plotted by Cartesian axes as a 2D function plot or a 3D response surface (prediction as function of two features). Each label of a categorical feature can be assigned an integer value from 1 to  $K'$  categories and thus also be plotted.

A classification model can be seen as a mapping from  $X \in \mathbb{R}^d$  to  $\hat{y} \in \{1, 2, \dots, K\}$ . Some models, as RF, provides a probabilistic prediction (pluralistic voting) of class membership  $\hat{p}_k$  for any class  $k \in \{1, 2, \dots, K\}$  and assign the class membership hereafter. Thus, the probabilistic classification model  $f_c$  is a mapping from  $X$  to the probability space  $P$ ,

$$f_c(X) = P. \quad (2)$$

Any point in  $P$  is a possible prediction  $\hat{p}$  with a unique probability distribution over  $K$  mutually exclusive classes, such that  $\hat{p} = \{\hat{p}_1, \hat{p}_2, \dots, \hat{p}_K\}$ . As class memberships are mutually exclusive, the sum of the class probabilities is always one,  $|\hat{p}|^1 = 1$ . Therefore the probability space is a  $K-1$  dimensional simplex [15], which contains any possible combination of assigned probabilities to  $K$  mutually exclusive classes, see Figure 2. The  $K$  axes, which assign probability of 0 to 1, are not orthogonal, meaning it is not possible to modify the assigned probability of one class without affecting at least one other.

The classification mapping can be represented by simply joining the simplex-space with the feature space, but this would only allow a 2D or 3D visualization when  $(d + K - 1) \in \{2, 3\}$ , thus either maximally a 2 feature problem for 2 classes, or a 1 feature separation for 3 classes. Instead, this mapping can also be represented as  $K$  separate  $d$ -dimensional surfaces  $S_k$  in a  $d + 1$ -dimensional space  $V$  with  $d$  axes representing features and one axis ( $p$ ) representing the probability of either of the  $K$  classes. Thus, we align the directions of all  $K$  probability axes to reduce the dimensionality of the mapping space with  $K - 2$  dimensions. Then, any line parallel to the probability axis  $p$ , will intersect every  $S_k$  surface, describing the predicted probability of the  $k^{th}$  class at this point of input features. The sum of predicted probabilities of all intersections for any such line will be equal to one. To summarize, multi classification model structures are more difficult to visualize, as each class adds another dimension to the mapping space. It is possible to plot the individual predicted probability of each class and overlay these plots. Figure 2 summarizes the mapping topology for regression, for binary classification, and for multi classification.

RF mapping for both regression and classification can jointly be defined as

$$\hat{y} = f(X) \quad (3)$$

Here  $\hat{y}$  is the  $c$ -dimensional prediction space. For regression,  $c = 1$ ,  $f$  maps to a 1-dimensional prediction scale. For classification,  $c = K$  classes, and  $f$  maps to a prediction vector space, where the  $k^{th}$  dimension predicts the probability of class  $k$ . For classification the predictions  $\hat{y}$  can be any point within the  $(K - 1)$ -simplex. On the other hand, the training examples  $y$  can only be of one class each, which are the  $K$  vertices (corners) of the  $(K - 1)$ -simplex.

We define a local increment vector,  $L$ , pointing from  $\hat{y}_i$  to  $\hat{y}_j$  in a prediction space of  $c$  dimensions, such that

$$L_{ij} = \hat{y}_i - \hat{y}_j = \{\hat{y}_{i1} - \hat{y}_{j1}, \dots, \hat{y}_{ic} - \hat{y}_{jc}\}, \quad (4)$$

For regression, where  $(c = 1)$ , the local increment is a scalar with either a positive or negative direction. For classification,  $(c > 1)$ , the local increment is a vector with  $c$  elements, one for each class. Each node of a RF model fit is a prediction, which is a specific point in the prediction space. Local increments are the connections between nodes, describing the change of prediction. Computing the thousands or millions of local increments for trees and nodes, and sum these individually for each observation and feature is essentially the feature contributions method.

## 2.2 Properties of random forest related to feature contributions

RF is an ensemble of bootstrapped decision trees for either regression or classification. Figure 3 illustrates how the RF algorithm operates for regression. For each of the trees (1 to  $ntree$ ) the training set is bootstrapped (random sampling with replacement). In average  $(\frac{N-1}{N})^N \approx 0.37$  of the observations will not be included in each bootstrap. These observations are called out-of-bag (OOB). Thus for any tree, a selection of observations will be 'inbag' and used to train/grow the tree starting from the root node. Any node will have a node prediction which is defined by inbag observations in that node.

$$\hat{y}_j'' = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \quad (5)$$

For a regression tree, the node prediction of the  $j^{th}$  node  $\hat{y}_j''$  is equal to the mean of inbag predictions in the  $j^{th}$  node. Where  $y_{ij}$  is the prediction value of the  $i^{th}$  observation in the  $j^{th}$  node.  $n_j$  is the number of observations in the  $j^{th}$  node. Thus we are only computing a node prediction from inbag elements.

For classification, the probabilistic node prediction  $p_{jk}$  of the class  $k$  of the node  $j$  is equal to the number of inbag observations of class  $k$  divided with total number of inbag observations in the node:

$$\hat{p}_{jk} = \frac{n_{jk}}{n_j} \quad (6)$$

A node prediction  $\hat{y}_j''$  can also describe all class probabilities at once as a vector corresponding to a point in the  $(K - 1)$ -simplex space.

$$\hat{y}_j'' = \{\hat{p}_{(j,1)}, \dots, \hat{p}_{(j,K)}\} \quad (7)$$

For classification  $c > 1$ , the class probabilities of any node will always sum to 1 for any node:

$$|\hat{y}_j''|^1 = \sum_{k=1}^K p_{jk} = 1 \quad (8)$$

Therefore, the elements of any local increment vector for classification, see Equation 4 will always

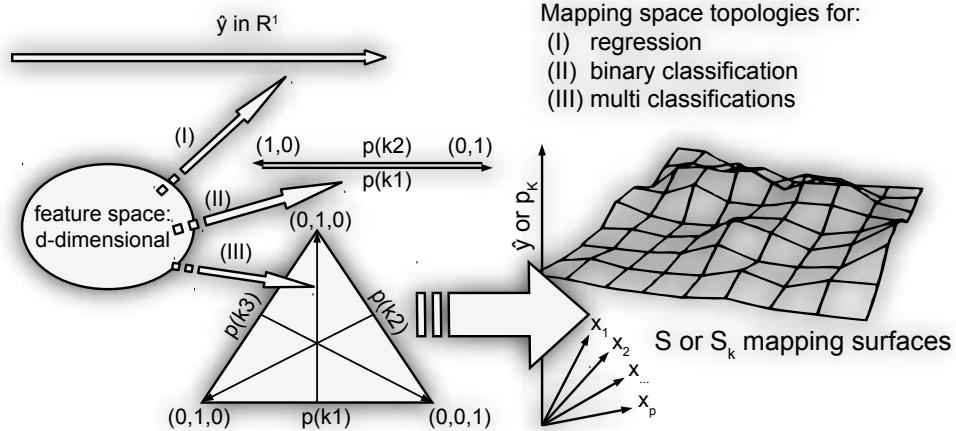


Figure 2: Topologies of random forest model represented as a function mapping from  $d$ -dimensional feature space to one of the following prediction spaces: (a) regression, 1-dimensional scale; (b) binary classification,  $K = 2 - 1$  probability simplex reducible to a 1-dimensional probability scale; (c) multi-classification, probabilistic  $(K - 1)$ -simplex. The mapping can be represented as a high-dimensional surface  $S$ , in a joined feature and prediction space linking any combination of features to a given prediction. For multi-classification  $S$  can be split into multiple  $S_k$  surfaces describing predicted probability for each of  $K$  individual classes.

sum to zero. This is not true for the local increment scalars of regression,  $c = 1$ .

For an original RF implementation [10], predictions of terminal nodes of classification trees are reduced to a single majority vote. Other implementations such as `sklearn.randomForestClassifier` [17] would rather pass on the probabilistic vote from terminals nodes and only on the ensemble level perform reduction by majority vote or just keep the full probabilistic average. In practice, implementations of feature contributions usually have to re-estimate node predictions. A feature contributions implementation such as forest floor should match the specific rule of terminal node predictions of the specific model algorithm.

A node is by default terminal if there are 5 or less inbag observations left for regression or a single inbag observation for classification. Any non-terminal node will be split into two daughter nodes to satisfy a loss-function. For regression the loss function is a typical sum of squared residuals.

For classification, a Gini criterion is used as the loss function. That is to select the split yielding the lowest node size weighted Gini impurity. Gini impurity ( $g$ ) is 1 minus the sum of squared class prevalence ratios in nodes,  $g = 1 - \sum_{k=1}^K \hat{p}_{jk}^2$ . Gini impurity is in fact the equation of a  $K$ -dimensional hypersphere, where  $\sqrt{1-g}$  is the radius and all  $\hat{p}_{jk}$  are the coordinates. The  $(K-1)$ -simplex space intersects this hypersphere where all prevalences sum to one,  $1 = \sum_{k=1}^K \hat{p}_{jk}$ . Therefore for a  $K = 3$  classification, a Gini loss function isobar appear as a 2D-circle, when visualized in the  $(K-1)$ -simplex

space. One circular isobar is drawn in Figure 4, the Gini loss function chooses the split placing two daughter nodes the furthest from the center of the  $(K-1)$ -simplex.

Splitting numerical features of ratio-, ordinal- or integer-scale is all the same for RF. A break point will direct observations lower or equal to the left node. Splitting by categorical features is to find the best binomial combination of categories designated for either daughter node. A feature with 8 categories will have  $2^{8-1} - 1 = 63$  possible binary splits. Any available break point are evaluated by the loss-function, but the RF algorithm is constrained to only access a random selection of the features in each node. The amount of features available,  $mtry$ , can e.g. be a third of the total amount of features. This *random variables subspace* and bootstrapping will ensure decorrelation of trees and feature regularization without overly increasing the bias of each fit. Each fully grown tree is most likely highly overfitted as the individual predictions of each terminal node are dictated by 5 or less observations. Combining the votes of many overfitted but decorrelated trees forms an ensemble with lowered variance and without increased bias. Out-of-bag(OOB) predictions are calculated for each terminal nodes. As OOB observations are not used actively in growing the trees of the forest, they can serve as an internal cross validation which yields similar results as a 5 fold cross validation [23]. The prediction of individual trees are written as  $\hat{y}'_{ij}$  for  $i \in \{1, \dots, N\}$  observations predicted by  $j \in \{1, \dots, ntree\}$ . The

ensemble predictions are computed as

$$\hat{y}_i = \frac{1}{ntree} \sum_{j=1}^{ntree} \hat{y}'_{ij} , \quad (9)$$

and the OOB cross validated ensemble predictions  $\tilde{y}_i$  are computed as

$$\tilde{y}_i = \frac{1}{n_{OOB,i}} \sum_{j \subseteq \tilde{J}_i} \hat{y}'_{ij} , \quad (10)$$

where  $\tilde{J}_i$  is the subset of  $\{1, \dots, ntree\}$  trees, where  $i^{th}$  observation is OOB.  $n_{OOB,i}$  is the size of the subset  $\tilde{J}_i$ . Thus let any training observation  $i$  iterate through the  $\tilde{J}_i$  subset of trees, defined as those trees where  $i$  was not inbag, and find the mean of terminal node predictions.

To obtain value/class predictions of new observations, the observations will be forwarded through all trees according to the established split rules. A tree prediction is dictated by the terminal node a given observation ends up in. The ensemble prediction of a RF model fit will by default be the average for regression and the majority vote for classification. Figure 3 explains graphically the structure of a single regression tree by feature  $x_1$  and  $x_2$ . First all bootstrapped observations exist within the node n1. The mean prediction value of n1 is in this example 0.14 a slight offset compared to the training set prediction mean of 0. The first split is over a break point in  $x_2$ , dividing n1 into n2 with low prediction value and n3 with a high prediction value. Both n2 and n3 are further split by  $x_1$ . Interestingly, n2 and n3 have almost opposite splits by  $x_1$ . In n2, high  $x_1$  leads to a lower prediction, while reversely in n3. This illustrated tree have only grown 7 nodes. Nonetheless, the tree contains an interaction term, where high  $x_1$  only contribute positively to the prediction  $\hat{y}$  when conditioned by high  $x_2$ .

### 2.3 Local increments and feature contributions

This section explains how feature contributions are computed. This paper expands the feature contributions defined by Palczewska *et al* [16] to also account for bootstrapping and/or stratification and to allow OOB cross validation. Feature contributions summarize the pathways any observation (a given combination of input features) will take through the many decision trees in a RF model. Each sub node of the trees holds a prediction, which is average observed target of observations populating it, see Equations 5 & 6. The sum of the many steps from node to node (local increments) is for regression exactly the resulting large step from the grand mean of the training set to the given numeric target prediction. Likewise for classification, the large step

is from base rate to a probabilistic target prediction. A proof hereof is provided in supplementary materials. As these many small steps towards the final prediction is an additive process, it is possible to reorder the sequence of steps and end up by the same prediction. The important implication hereof is that the RF model structure can be decomposed into additive sub models, each with the same dimensionality. As each sub model structure is the sum local increments of decision splits by one specific feature, each sub model structure tend to only describe the main effect of this one specific feature plus perhaps interactions with other features.

In order to efficiently describe how variations of feature contributions are computed, a notation of how to access any local increment in a given RF model fit is formulated. We define  $L$  as a list of lists of lists containing all local increments.  $L$  is defined in the following three levels (observations, increments):

1.  $L_i$  is a list with  $i \in \{1, \dots, N\}$ , and  $N$  is the number of observations predicted by the forest.  $i$  is the  $i^{th}$  observation.
2. Each element of  $L_i$ , called  $L_j$  is a list with  $j \in \{1, \dots, ntree\}$ , and  $ntree$  is the number of trees in the ensemble.
3. Each element of  $L_j$ , called  $L_k$  is a list with  $k \in \{1, \dots, n_{increment,i,j}\}$ , and  $n_{increment,i,j}$  is the number of increments encountered by the  $i^{th}$  observation in the  $j^{th}$  tree.

Note that  $L$  can be ordered as a 2-dimensional array ( $i$  observation,  $j$  tree) where each element is a sequence of local increments specific for the  $i^{th}$  observation in the  $j^{th}$  tree. Overall, we can access any local increment in  $L$  with  $L_{ijk}$ . Depending on the model type,  $L$  will contain local increments as scalars for regression or as vectors for classification. The first local increment  $k = 1$  for any tree and observation in  $L_{ijk}$  is the step from node 0 (training set) to node 1 (root node of tree). Thus the  $k^{th}$  local increment steps from the parent node  $k - 1$  to a daughter node  $k$ . The local increment  $L_{ijk}$  is the change of node prediction  $\hat{y}_{ijk}'' - \hat{y}_{ij(k-1)}''$

Equation 11 describes how any prediction can be computed from  $L_{ijk}$  as the sum of all local increments plus grand mean or base rate. A proof hereof can be found in the supplementary materials.

The target prediction  $\hat{y}_i$  is computed as

$$\hat{y}_i = \frac{\sum_{j=1}^{ntree} \sum_{k=1}^{n_{increment,i,j}} L_{ijk}}{ntrees} + \bar{y}, \quad (11)$$

where  $L_{ijk}$  is a local increment and where  $\bar{y}$  is the grand mean or base-rate. The numerator is a scalar for regression and a vector for classification. The denominator,  $ntree$ , is always a scalar.

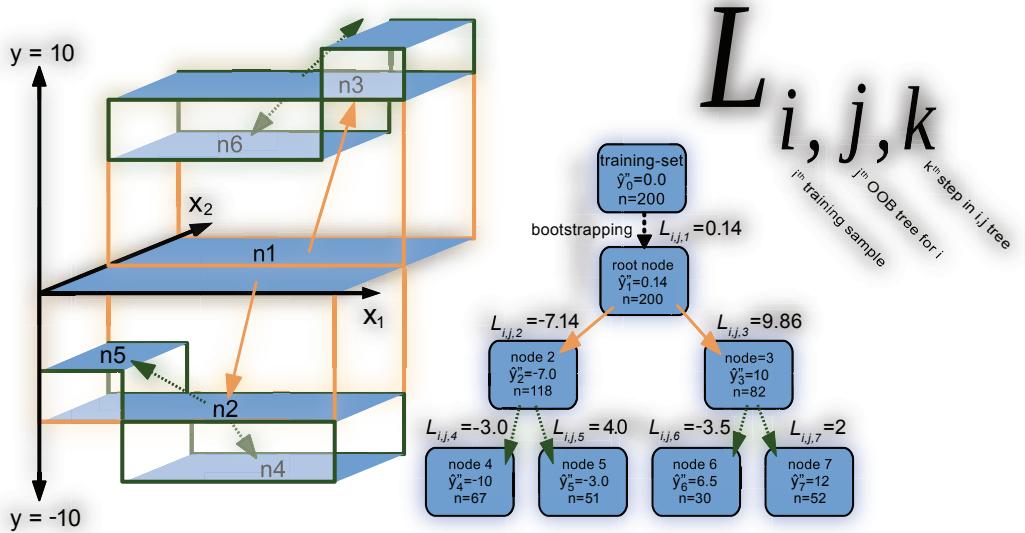


Figure 3: Random forest and local increments explained. Left, an 3D illustration of a small regression tree of 7 nodes. Right, the same tree described by node means( $\bar{y}$ ), node size( $n$ ) and local increments  $L_{ijk}$ .  $L$  is subsetted by observation, tree, node and feature. A observation falling in e.g. node 4, will have a prediction as the sum of the local increments in its path plus the grand mean of the training set.

So far the prediction of the  $i^{th}$  observation is the grand mean (regression) or the base-rate (classification) plus the sum of all local increments  $L_{ijk}$  encountered by this  $i^{th}$  observation divided by  $n_{trees}$ .

Figure 4 is a new geometrical representation of local increments for a 3-class classification. Figure 4 is not intended as a model structure visualization, but rather as a representation of how decision trees branch out in the prediction space. Each node in the classification tree can be seen as a probabilistic prediction defining a point in a probabilistic ( $K - 1$ )-simplex. Figure 4 depicts node predictions and local increments for a small tree with four terminal nodes. To this tree graph is appended a node (T) for training set to the root node of the tree. This train node represents the class distribution of the training set. The bootstrap increment leads to the root node. This step is often small and a result of random uniform sampling w/o replacement. If applying class stratification, the length and direction of this step can be controlled. Stratification corresponds to defining a prior expected class distribution, which will be the position of the root nodes in the prediction space. From here all trees will branch out from this point. The following local increments and nodes comprise the entire tree. Any split produces two nodes and two local increments of opposite direction. If not of equal node size, there will be one shorter local increment defined of many in-bag observations and one longer local increment defined of fewer in-bag observations. This is a consequence of that class distributions of

daughter nodes multiplied by the node sizes and added together is exactly equal to class distribution of parent node multiplied by its node size. This symmetry effect can be found in Figure 11 in section 4.3. For the unbalanced binary features *wives' religion*, *wives working* and *media exposure* the prediction is offset a lot for a few observations, while the prediction of remaining many observations will only change a little in the exact opposite direction. For regression and binary classification such a direction is essentially one-dimensional and can be positive or negative. For multi classification the direction is a vector of  $K$  elements with the restriction that the sum of elements is zero. In Figure 4, the circle represents a Gini loss function isobar. The further away (euclidean distance) nodes are placed from uniform class distribution the better a split according to RF Gini loss function. The best kind of split is one placing both daughter nodes onto two of the  $K$  vertices of the ( $K-1$ )-simplex.

For the training set, a cross validated OOB-prediction  $\tilde{y}$  can be formulated as

$$\tilde{y}_i = \frac{\sum_{j \subseteq \tilde{J}_i} \sum_{k=1}^{n_{increments,i,j}} L_{ijk}}{n_{OOBtrees,i}} + \bar{y} \quad (12)$$

where  $\tilde{J}_i$  is the subset of trees where  $i^{th}$  sample is OOB. One can reason, that if Equation 11 is true for any set of trees, then Equation 12 must also be true for a given subset of any trees, such as the OOB subset  $\tilde{J}_i$ , see supplementary materials.

When predicting the training set with an RF

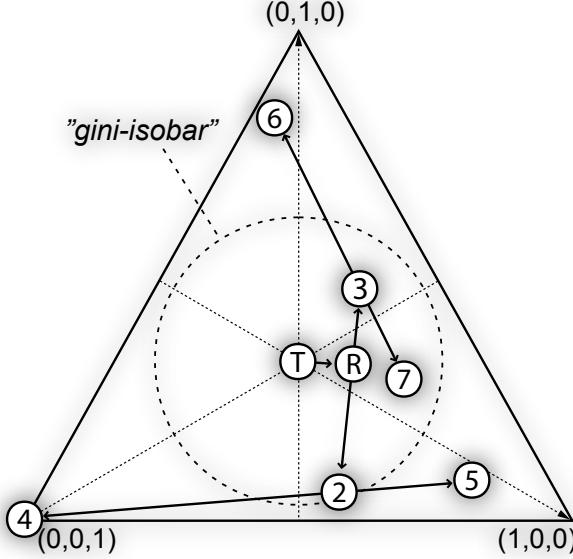


Figure 4: A representation of how node predictions and local increments for a small classification tree with four terminal nodes. The first node in center represents the class distribution of a balanced training set (T). The bootstrap increment leads to the root node of the tree (R). The following local increments and nodes comprises the entire tree. Any split produces two local increments of opposite direction. The circle represents Gini loss function isobar. The further the two nodes (weighted by size) are from uniform class distribution the better a split according to the Gini loss function.

model, any training observation  $i \in \{1, \dots, N\}$  will have a high proximity to itself, that is, it will in any in-bag tree both define the in-bag node predictions of the terminal node and be predicted by the very same terminal node. For data sets with a high noise levels this becomes a problem and the points  $S_i$  of model structure  $S$  will overfit the sampled training set observations  $T_i$ , and visualizations hereof will look more noisy. If the RF training parameter minimum terminal node size is increased and/or bootstrap sample size is lowered then training observation  $i$  will have a lower influence on its own prediction and visualizations will not look noisy.

To compute feature contributions, the sum local increments over each feature, it is necessary to keep a record of splitting features in each parent node. Let  $H_{ijk}$  be a list of lists of lists with the exact same structure as  $L_{ijk}$ . For every local increment the corresponding element of  $H_{ijk} \in \{0, \dots, n_{vars}\}$  is an integer index pointing to feature used to split the parent node. Notice the first local increment  $L_{ijk}$  where  $k = 1$  for every tree  $j$  and every observation  $i$  is due to random bootstrapping. The local increments of bootstrapping are assigned to *feature 0*. Therefore  $H_{ijk} = 0$  for any  $i$  for any  $j$  where  $k = 1$ .

This distinction between OOB-predictions  $\tilde{y}$  and regular test predictions  $\hat{y}$  of training set now becomes important as how to feature contributions are defined. Previously [16, 9] feature contributions

have been defined for regression and classification analogous to this:

$$F_{il} = \frac{\sum_{j=1}^{n_{tree}} \sum_{k=0}^{n_{increment},i,j} L_{ijk} \psi(H_{ijk}, l)}{n_{tree}} \quad (13)$$

Here  $F_{il}$  the feature contribution of the  $i^{th}$  observation for the  $l^{th}$  feature is the sum over all local increments  $L$ , where observation  $i$  was split by feature  $l$  divided by  $n_{tree}$  trees of the forest. The binary equality function  $\psi$ , ensures only to sum local increment over splits by one specific feature. For any integers  $a, b$ , then  $\psi(a, b) = 1$  if  $a = b$  and  $\psi(a, b) = 0$  if  $a \neq b$ .

This definition of feature contributions is fine if: (a) the noise level is low or (b) if feature contributions  $F$  only is computed for some test set different from training set or (c) if the user is confident, that the model structure is not over fitted. It would be possible to cross validate by segregating the data set in a training set and test set to avoid over fitted visualizations. To discard data points is not desirable for a data set with limited observations. It would be possible to perform an n-fold cross validation, but n-fold random forests would be necessary to train.

We propose to compute feature contributions for the OOB cross validated predictions. OOB cross validated predictions are only the sum of local increments over trees where  $i^{th}$  observation was OOB,

see Equation 12. Analogously, we OOB feature contributions  $\tilde{F}_{il}$  as

$$\tilde{F}_{il} = \frac{\sum_{j \subseteq \tilde{J}_i} \sum_{k=0}^{n_{increments,i,j}} L_{ijk} \psi(H_{ijk}, l)}{n_{OOBtrees,i}} \quad (14)$$

$j$  only iterates the subset of tree  $\tilde{J}_i$  where  $i^{th}$  observation was OOB.  $n_{OOBtrees,i}$  is the total number of times the  $i^{th}$  observation was OOB and the size of the subset  $\tilde{J}_i$ . Equation 14 is used in forest floor visualizations to compute cross validated feature contributions of the training set predictions.

## 2.4 Decomposing the mapping surface with feature contributions

We can compute the OOB cross validated set of points  $\tilde{S}_i = \{X_i, \tilde{y}_i\}$  for  $i \in T$  the training set. That is the combination by training features  $X_i$  and the cross validated predictions  $\tilde{y}_i$ , where  $c = 1$  for regression and  $c > 1$  for classification. To decompose  $\tilde{S}_i$ , then  $\tilde{y}_i\}$  is expanded with  $\tilde{F}_{il}$ , such that:

$$\tilde{y}_i = \sum_{l=0}^d \tilde{F}_{il} + \bar{y} \quad (15)$$

Likewise non cross-validated  $\hat{y}_i$  is a sum of non cross-validated  $F$ .

$$\hat{y}_i = \sum_{l=0}^d F_{il} + \bar{y} \quad (16)$$

The ensemble prediction  $\hat{y}$  or  $\tilde{y}$  is equal to sum of local increments + grand mean / base rate, see Equation 11,12. As sequences of additive vectors can be rearranged, it is possible to compute sub totals of local increments of the full prediction. Feature contributions is just the subtotal of encountered local increments for the for the  $i^{th}$  observation where the parent node was split by the  $l^{th}$  feature.

Notice feature 0 ( $l = 0$ ) is included to accurately account for the normally small and negligible feature contribution of random bootstrapping. For an increasing number of trees, this bootstrapping feature contribution will approach zero. However, if the bootstrapping is stratified  $F_{i0}$  and  $\tilde{F}_{i0}$  is equal to local increment from training set base rate  $\bar{y}$  to the chosen stratification rate in every root node.

Figure 5 illustrates OOB cross validated feature contributions and regular feature contributions. A so called “one-way feature contribution plot” is a single feature contribution column plotted against the values of the corresponding feature. In Figure 5 the ”one-way feature contribution plot” can be seen as projections of  $\tilde{F}$ . Conveniently, the main effects of either feature  $x_1$  and  $x_2$  have been separated with feature contributions before the projec-

tion into the 2D plane. In Figure 5, the goodness-of-visualization fit to the projected feature contributions can be seen for both  $\tilde{F}_{i1}$  and  $\tilde{F}_{i2}$ . If it is possible to re-estimate the set feature contributions e.g.  $\tilde{F}_{i1}$  with some estimator  $f$  only by the feature context of the visualization, it is certain, that no interactions have been missed. Thus the model structure do not contain any interaction effect with feature  $x_1$ . To quantify this we use a leave-one-out cross validation.

$$GOV(\hat{f}_\lambda) = \text{cor}(\hat{g}_{il}, \tilde{F}_{il})^2 \quad (17)$$

Here, the goodness-of-visualization ( $GOV$ ), is the pearson correlation between LOO predicted feature contributions. Where  $\hat{g}_{il} = \hat{f}_i^{-i}(X_{i\lambda})$  is the leave-one-out prediction of the  $\tilde{F}_{il}$  feature contribution of the  $i^{th}$  observation for the  $l^{th}$  feature.  $\lambda$  is the features which are used to fit the estimator. When  $\lambda = l$ ,  $GOV$  quantifies how well feature contribution of the  $l^{th}$  feature  $\tilde{F}_{il}$  is explained as a main effect. In Figure 5  $\tilde{F}_{i1}$  is predicted by  $X_{i1}$  and  $\tilde{F}_{i2}$  is predicted by  $X_{i2}$ .  $GOV$  can also quantify other visualization contexts than main effect plots. E.g. in Figure 7 of result section the goodness of a visualization context of two features  $x_3$  and  $x_4$  is quantified, where  $\lambda = \{3, 4\}$ .

## 3 Materials and methods

### 3.1 Data and software

The real datasets *contraceptive method choice* (cmc) and *white wine quality* (wwq) were acquired from the UCI machine learning repository [4, 11]. All algorithms were implemented in R (3.2.4) [18] and developed in Rstudio (0.99.892) [20]. The main functionality is available as the R-package, *forestFloor* (1.9.3) [25], published on the repository CRAN. If not stated otherwise all RF models was trained with the CRAN package *randomForest* [10] by default parameters except *keep.inbag*=TRUE in order to reconstruct the individual pathways of observations through the trees. To reproduce result section, R scripts for each data example have been included in the package.

### 3.2 Simulating toy data

To demonstrate that the visualizations in the result section 4 provide correct representations of the data structure, it is beneficial to use simulated (toy) data from a given hidden function. Such functions as Friedman#1 and ‘Mexican hat’ are known examples [1]. To illustrate the principal functionality of *forestFloor* a new hidden function,  $G$  is defined.  $G$  is the ideal hidden structure, which cannot be observed directly. The toy function was defined as  $G(X) + \epsilon = G^*(X) = y = x_1^2 + \frac{1}{2}\sin(2\pi x_2) + x_3 x_4 +$

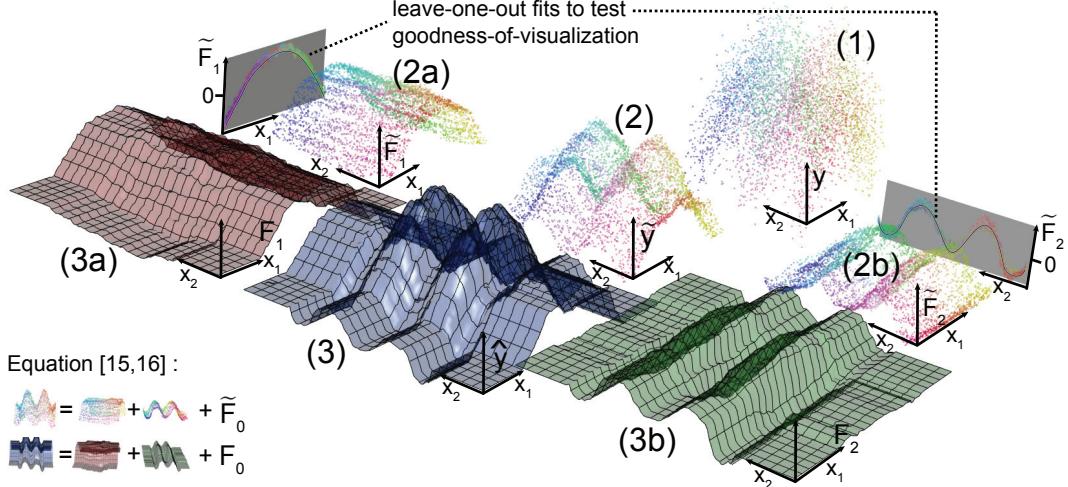


Figure 5: (1) Simulated data set of 5000 observations,  $y_i = f(X_i) = -(X_{i1})^2 - \cos(X_{i2}) + \epsilon_i$  where  $X_{i1}$  and  $X_{i2}$  are drawn from a uniform distribution such that  $X_{i1} \in [-\frac{\pi}{2}; \frac{\pi}{2}]$ ,  $X_{i2} \in [0; 8\pi]$ . For all plotted points, a colour gradient (hue color wheel) is used to mark different combinations of  $X_{i1}$  and  $X_{i2}$ . (2) Out-Of-Bag cross-validated predictions  $\tilde{y}$  are plotted. (3a/3b)  $\tilde{y}$  is decomposed into feature contributions  $\tilde{F}_1$  and  $\tilde{F}_2$  and projected into a 2D plane, see Equation 14 and 15. Either contain almost only variance from the two main effects  $-(X_{i1})^2$  or  $\cos(X_{i2})$ . (3) Blue surface depict the full model structure,  $\hat{y}$ . To either side (3a/3b)  $\hat{y}$  is decomposed into  $\hat{F}_1$  and  $\hat{F}_2$ , see Equation 13. The sum of cross-validated feature contributions by each observation is equal to the cross-validated predictions, and vice versa for non-cross validated.  $F_0$  is the corrections for random bootstrapping. If no stratification,  $F_0$  will be negligibly small. This illustration also generalizes more input features/dimensions and probabilistic classification.

$\epsilon k$  and was sampled 5000 times.  $x_i$  were sampled from a uniform distribution  $U(-1, 1)$ . The noise variable  $\epsilon$  was sampled from a normal distribution  $N(0, 1)$  and  $k$  was set such that the Pearson correlation  $\text{cor}(G(X), G^*(X)) = 0.75$ . Thus the true unexplainable variances component is roundly 25% of the total variance. The level of detail, RF can capture from hidden structure  $G$ , declines as the noise increases.

## 4 Results

Three data sets were modeled with RF regression or RF classification and subsequently explored with forest floor. The examples demonstrate how feature contributions can be used to visualize the data structure and how to identify unaccounted interactions in a visualization.

### 4.1 Random forest regression of toy data

A default RF regression model was trained on the toy data set with a hidden structure,  $y = x_1^2 + \frac{1}{2}\sin(2\pi x_2) + x_3x_4$ . Figure 6 plots feature contribution of all six features against the training set feature values of the toy data. This type of plotting

illustrates the main-effects, as feature contributions by each feature were plotted against their respective feature values. Hereby, the mapping surface  $S$  was visualized as the sum of  $d$  partial functions (black-lines), one for each feature. As the feature contributions retained any variance (main effects + interactions) associated with the node splits by each feature, it was possible to visually verify and test the goodness-of-visualization. Notice that main effect plots of  $x_1$  and  $x_2$  form nonlinear patterns representing the underlying additive  $x_1^2$  and  $\frac{1}{2}\sin(2\pi x_2)$  contributions to the target  $y$ . Therefore, the leave-one-out  $R^2$  goodness-of-visualization was  $> 0.95$  for both these plots. As the explained variance of feature contributions of  $x_1$  and  $x_2$  was more than 95% when fitted as main effects, there was no considerable unaccounted interactions. On the other hand, feature contributions of  $x_3$  and  $x_4$  were poorly explained in main effect plots. The GOV was poor, less than  $R^2 < 0.1$ . It was hence concluded that plotting the one-way feature contributions of  $x_3$  and  $x_4$  did not assist to explain the structure of  $S$ . Feature contributions of  $x_5$  and  $x_6$  were also poorly explained but contained no large variance and were therefore not interesting to explore further. The features  $x_5$  and  $x_6$  could also be identified as unrelated to the target  $y$  for having a very low variable importance (not shown). To include such

uncorrelated/unrelated features illustrated the base line of random fluctuations in the mapping structure. This helped to assess whether a given local structure only was a random ripple.

As the feature contributions of  $x_3$  and  $x_4$  were inadequately accounted for, a broader context was needed to understand the hidden structure. To identify interactions relevant for the feature contribution of  $x_3$  a color gradient (red-green-blue) was applied in mapping space  $V$  along the  $x_3$  axis. The color of any other observation in any other plot was decided by its projected position on the  $x_3$  axis. Low values were assigned red and high values blue. Figure 6 depicts the main effects feature contribution plot of  $x_1, \dots, x_6$  with the applied color gradient to  $x_3$ . Any main effect feature contribution plot of features who neither correlate and neither interact with  $x_3$  will show a random color pattern. Such features were  $x_1, x_2, x_5$  and  $x_6$ , which neither correlated nor interacted with  $x_3$ . Plots of only correlated features would reproduce the same horizontal color pattern. In the extreme case, a feature identical to  $x_3$  would reproduce the exact same horizontal color pattern. Plots of only interacting features would reproduce the color gradient vertically along the feature contribution axis. A combination of correlation and interaction would make the color gradient reappear diagonally. In Figure 6 the color gradient suggests, that  $x_3$  interacted with  $x_4$  due to the vertical color gradient in the plot of  $x_4$ . In Figure 7 their combined feature contributions were plotted in the context of both feature  $x_3$  and  $x_4$ . In this 3D plot it was observed, that the 2D rule of color gradients of interacting features was a basic consequence of perspective. Both color patterns of  $x_3$  and  $x_4$  could be reproduced by rotating the 3D plot. In this 3D plot, there was no large deviation of feature contributions from the fitted grey. Thus, it was evident that any structure of  $S$  related to  $x_3$  and  $x_4$  were well explained in the joined context of both features  $x_3$  and  $x_4$ . The GOV of this fit was  $R^2 > .9$ . Therefore, this second order effect plot was an appropriate representation of how  $x_3$  and  $x_4$  contribute to the target  $y$ . The depicted saddle-point structure of Figure 7 was expected, as the product of  $x_3$  and  $x_4$  contributed additively to the target  $y$ . Overall, the model surface  $S$ , could be represented by two one-way plot of  $x_1$  and  $x_2$  and one two-way plot of  $x_3$  and  $x_4$ . Hereby the hidden structure of the toy data was fully recovered.

## 4.2 Random forest regression of white wine quality (wwq)

The previous example of forest floor visualization was an idealized example with uncorrelated features and either representing clear main effect or clear interaction effects. The white wine quality

data set (wwq) is an example of mixed main effects and interactions by most features. The target, consumer panel ratings(1-10) of wines, was predicted on basis of 11 chemical features. A default RF model was trained and explained 56% of variance and the mean absolute error was 0.42 rating levels matching the previous best found model performance [5]. To explore the model structure of  $S$ , first all main effect plots were inspected. Figure 8 depicts all plots by all 11 features. Features were sorted in reading direction by variable importance to present most influential feature first. A color gradient along the most influential feature, *alcohol*, was applied to search for interactions. Hereby it was observed that *density* was negatively correlated with *alcohol*, that *volatile acidity* interacted with *alcohol* and that *residual sugar* both correlated and interacted with *alcohol*. The observed correlation between *residual sugar*, *density* and *alcohol* is trivial, where low-density *alcohol* linearly lowers *density* while high-density *residual sugar* increases *density*. Close to 98% of the scaled variance of these three features can be described by two principal components. This information redundancy was expected to affect variable importance of the three implicated features and to lower the general variance of the respective feature contributions. Although the overall structure suggested that alcohol content in general was associated with higher preference scores, there was a local cluster identified as low *alcohol*, high *residual sugar* and low *pH* which was associated with high preference scores also. Figure 8 suggested that wines could achieve a high preference score when *residual sugar* $\approx$ 17, *pH* $\approx$ 2.9, *citric acid* $\approx$ .35 and *fixed acidity* $<$ 7 despite a low alcohol content. Such white wines was perhaps by the consumer panel attributed fruity and fresh. Any found interaction could be investigated with several color gradients and two-way forest floor plots. It was chosen to investigate the interactions of *volatile acidity*, as this feature was the third most important feature, whereas the goodness-of-visualization of the one-way forest floor plot was only  $R^2 = 0.69$ . Two-way forest floor plot was therefore a more suitable representations of this effect. The color gradient along alcohol content already suggested a notable interaction between *volatile acidity* and *alcohol*. Figure 9 depicts the two-way forest floor plot of feature contributions of *volatile acidity* in the context of itself and the feature *alcohol*. The goodness-of-visualization was then  $R^2 = 0.94$ . Therefore, the residual variance of feature contributions not explained by this plot was low. For wines with alcohol content more than 10% (blue area) *volatile acidity* appeared slightly positively to preference score. For wines with lower than 10% *alcohol* (red area) *volatile acidity* appeared to contribute negatively to preference score.

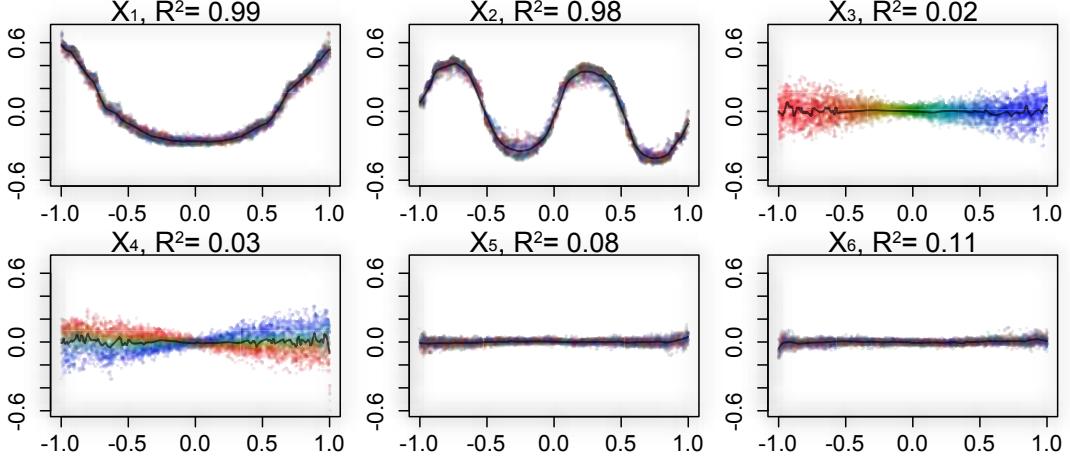


Figure 6: Forest floor main effect plot of a RF mapping structure trained on hidden function  $y = x_1^2 + \frac{1}{2}\sin(\pi x_2) + x_3x_4 + \kappa\epsilon$ .  $x_5$  and  $x_6$  have no relation to  $y$  and were included only to illustrate a base line signal. A color gradient parallel to  $x_3$  is applied to identify latent interaction with  $x_4$ . Leave-one-out k-nearest neighbor gaussian kernel estimation provides goodness-of-visualization(black line &  $R^2$  correlation) to evaluate how well each feature contribution can be explained as a main effect.

### 4.3 Random forest multi-classification: *Contraceptive method choice (cmc)*

To illustrate the capabilities of forest floor for multi-classification the data set cmc was chosen. The data set originates from a survey of 1473 non-pregnant wives in Indonesia in 1987 comparing current choice of contraception with socioeconomic features. These features were, *wives' age* (16-49), *wives' education level* (1-4), *husbands' education* (1-4), *n\_children* (0-16), *wives' religion* (0 (not islam), 1 (islam)), *wives working* (0 (yes), 1 (no)), *husbands' occupation* (I,II,III,VI), *standard-of-living index* (1-4), *media exposure* (0=Good, 1=not good) and the target *contraceptive method choice* (1=no-use (629), 2=long term(333), 3=short term (511)).

In the *cmc* data set the choice of contraception was far from fully described by the available features [12]. The OOB cross validated RF model error-rate was .44. Assuming wives did not use contraception (the most prevalent case) yielded a  $\frac{629}{1473} = .57$  error rate. Anyhow, if the RF model performance would be regarded as good by domain specialists, the model structure could possibly provide insights to the socioeconomic mechanisms in play. Hyper parameters *Sample size* and *mtry* were tuned to yield the best OOB cross validated performance. Optimal parameters was found to be bootstrap *sample size*= 100 and *mtry* = 2. A lower *sample size* can increase robustness by tree decorrelation but also introduce more bias. To lower *sample size* of trees can be advantageous, when

explained variance component is less than 50%. Thus a RF model different from default settings, was chosen to slightly improve predictions and to simplify/smooth the mapping structure to explore. Hereby the mapping structure may better represent the underlying social/economic mechanisms, that the specific data structure of survey reflects.

Three types of plots were constructed to investigate the mapping structure. As the number of features was  $d = 9$  and number of classes was  $c = 3$ , a full dimensional mapping space visualization would require 12 dimensions. As shown in Figure 2, probability axes can be aligned along the y-axis, to reduce the number of dimensions to represent prediction space to only one. Also, when the cross validated predictions were decomposed into cross validated feature contributions, only 2 dimensions were needed to plot any main-effect. These plots resembled one-way forest floor regression plots although coloring was reserved to identify class of predicted probability. Otherwise each class by each feature would need to be plotted separately. Black assigns no usage. Red assigns long-term usage and green assigns short-term usage. Figure 10 illustrated the main effects of each feature of a RF-fit, the y-axis describes the additive change of predicted probability for each observation for each each class. The actual feature value for each observation was depicted by the x-axis. Thus any observation were placed three times in each plot by the same feature value in three colors once for each three classes. The sum of changed probability over classes for any observation must be zero, see Equation 8. Overall, Figure 10 showed that main effects were dominant, as most

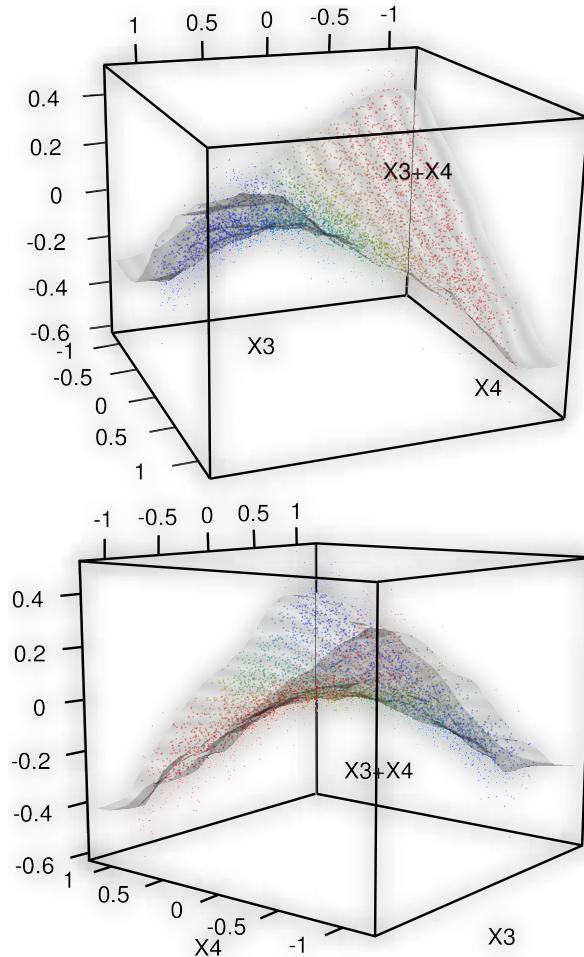


Figure 7: One forest floor interaction plot. XY-plan represent feature values  $x_3$  and  $x_4$  and Z-axis is the summed feature contributions of  $\tilde{F}_{i3} + \tilde{F}_{i4}$ . goodness-of-visualization is evaluated with leave-one-out k-nearest neighbor gaussian kernel estimation (grey surface,  $R^2 = .90$ ). This indicates no remaining latent interactions related to features  $x_3$  and  $x_4$ .

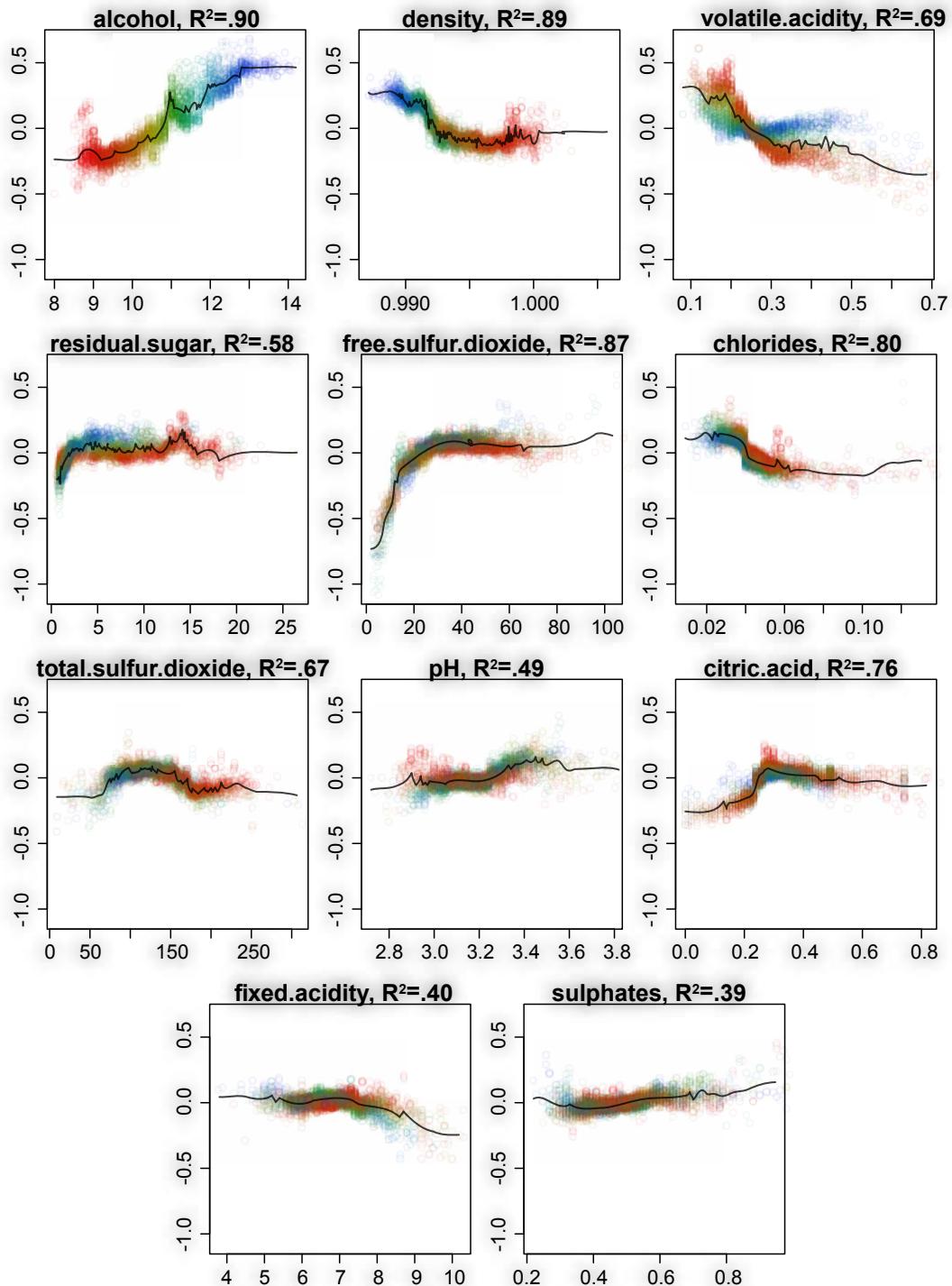


Figure 8: Forest floor main effect plots of random forest mapping structure of model predicting panel ratings of 4900 white wines on basis of chemical properties. The plots are arranged according to variable importance. X-axis are variable values and Y-axis the corresponding cross validated feature contributions. Color gradient in all plots are parallel to the feature *alcohol* (content w/w). goodness-of-visualization is evaluated with leave-one-out k-nearest neighbor estimation (black line ,  $R^2$ values)

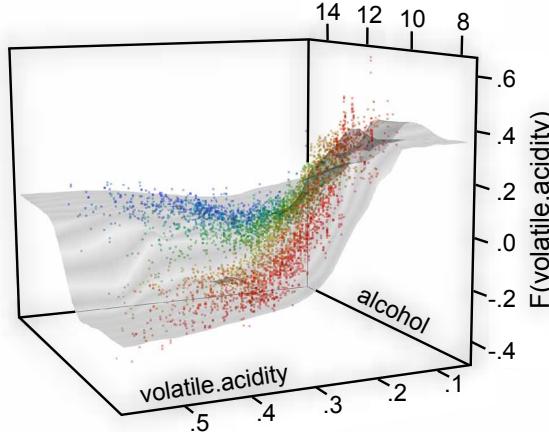


Figure 9: Forest floor interaction plot: Feature contribution of *volatile acidity* versus feature values of *volatile acidity* and *alcohol*. Color gradient is parallel to *alcohol* axis. goodness-of-visualization is evaluated with leave-one-out k-nearest neighbor estimation (grey surface and  $R^2 = 0.93$ )

variance was explained by the respective features. *n\_children* was the most important feature strongly predicting (probability change up to +/- .30) that wives with 0 or 1 child tended not to use contraception. On the other hand, more than 4 children predicted a slight increase in either type of contraception. Except for a preference separation for long-term contraception over short-term for wives with more 7 children, the *n\_children* feature was not found useful to predict the choosing between the two types of contraception. *Wives's education* especially separated between no-use of contraception and long-term use, where lowest level predicted up to +/-10% probability change. With more education the wives tended to use long-term contraception over no usage. The use of short-term contraception was comparably unchanged as a function of *wives' education*. *Wives' age*, the third most important feature, favored short-term contraception for wives younger than 30, while long-term and no contraception for wives elder than 30. After 40 years, either use of contraception declined. *Husbands' education* elicited same pattern as *wives' education* though size of effect was half. A small subgroup of 7% was reported to have a not good *media exposure* and this predicted a probability increase in no contraception of 8%. Type of *Husband' occupation* favored for category *I* long-term by 5% over short-term, whereas category *III* predicted an opposite 3% effect. Standard of living predicted a pattern much similar to *husband's education*. A small subgroup (15%) of wives were not muslim, and this predicted a 5% increase in short-term contraception over long-term usage and no usage. Lastly for a subgroup of 25% working wives was predicted a very slight increase (2%) of no-use over short-

term.

The main effects for this 3-class problem could also be depicted as a series of  $(3 - 1)$ -dimensional simplexes, where the position in the triangle depicts the predicted probability distribution for any observation. Colors can either depict true class (black: no-use, red: long-term and green: short term) or colors can depict a feature (low value (red), middle (green), high(blue)). Figure 11 depicts all main effects in bi-simplex plots, with left simplex colored by cross-validated true class separation, and right simplex colored by feature value distribution across the simplex space. Figure 11 depicts 10 pairs of simplexes. Lines were added to the simplexes to illustrate majority vote. Only 17% of wives were predicted to use long-term contraception even though 22% of the sample population did so. Because RF models effectively used the sampled base rate as prior (marked as a blue cross) and the effective separation was weak, predictions tended to be skewed towards largest class away from smallest class. A different prior than the sampled base rate could be set by stratified bootstrapping of each tree in a random forest model. E.g. to stratify sampling by target class would move the blue cross to the middle of the simplex, and roughly a third of predictions would fall into either class. Stratified bootstrapping would e.g. be reasonable if the preferred contraception is expected to be different in the full population than in the training population.

In the second total separation simplex, to present an overview of any differences in socioeconomic status, principal component analysis was used to reduce the full feature space to two principal color components. Here a purple cluster indicated no-use, a green cluster was shifted towards long-

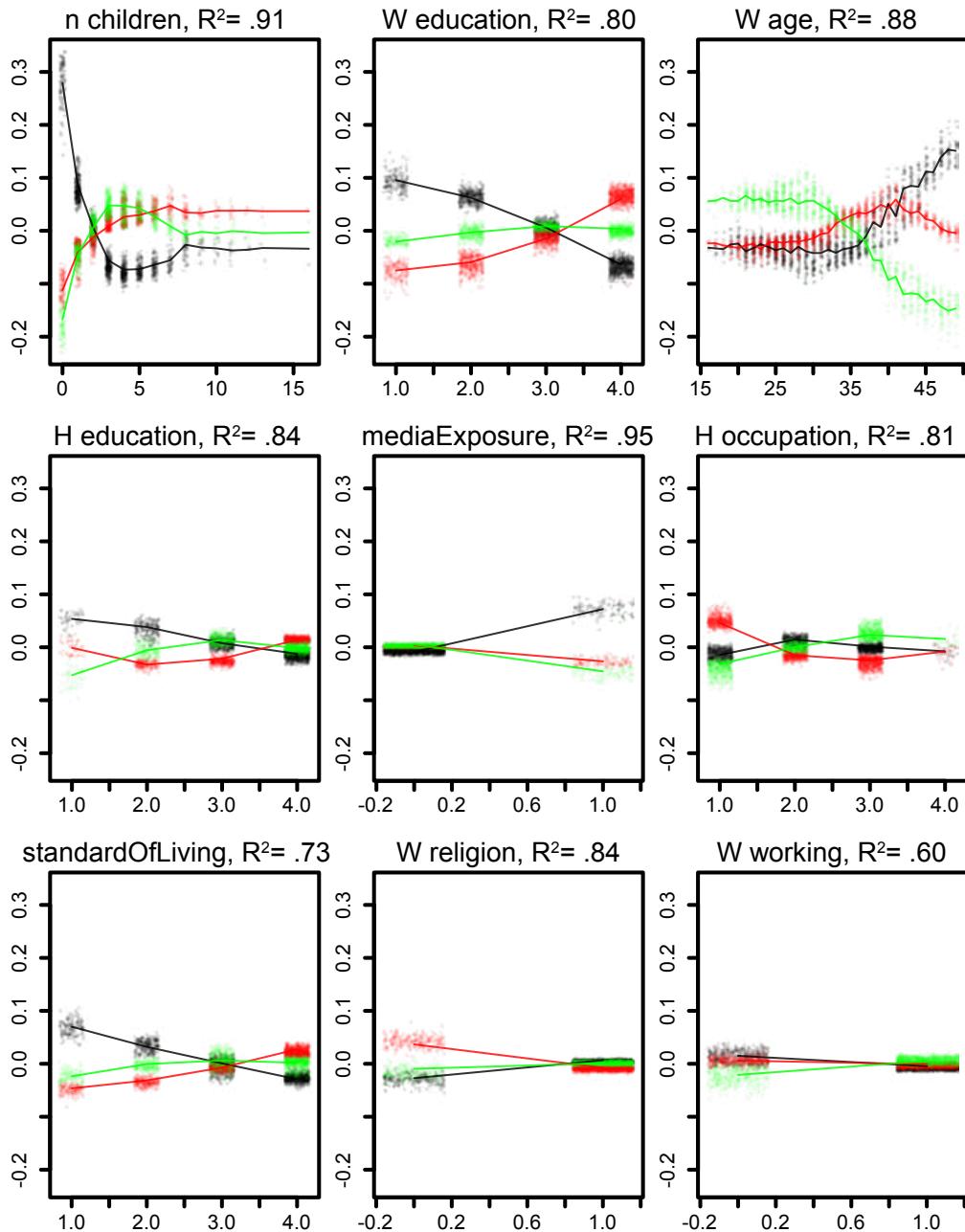


Figure 10: Cross validated feature contributions for each feature for each class (black, red, green) and for all training observations plotted against the corresponding feature values. Categorical features are coded with integers. Feature contributions can be understood as change of predicted class probability attributed to a given feature.

term usage, light blue cluster predicted short-term usage, and a dark-blue cluster predicted short-term or no usage. The color separation was not perfect, partly because the separation problem was difficult and partly because PCA cannot fully characterize a potential nonlinear mapping surface of random-forest. To colour be several features at the same time, seemed to be most useful for data sets with high linear feature collinearity.

The left of following bi-plots of simplexes depicted the effective separation of true class separation by any feature contribution. The right simplex depicted the separation as a function of the corresponding feature (by color). This second simplex could be used both to illustrate the main effect of each feature and to assess whether higher order effects were present. For features with small set of levels such as womans education, a separation in four clusters (red(1), brown(2), pale blue(3), deep blue(4)) could be seen. Education level 1 and 2 were partly joined. The local centroids of these cluster was interpreted as the main effect, and the deviation from the centroids as higher order effects + unfiltered noise. For all simplexes the global centroid and prior is the (blue cross).

The series of bi-plot simplexes of Figure 11 could illustrate with finer detail the predicted probability distribution for any observation, whereas the precise feature value was depicted with less fidelity than in Figure 10.

The three features *media exposure*, *wives' religion* and *wives working* were binary and showed the largest change of predicted probability in the smallest subgroups. This observation was regarded trivial, as the group size weighted probability change across a binary feature split must have equal size. Thus few observations can change prediction a lot, if many observations only change prediction a little in a opposite direction. This was regarded a property for all binary decision tree models and Figure 4 in Section 2.3 depicted a similar pattern of how local increments would propagate in a probability simplex.

To search for higher order effects, similar to forest floor regression, simplex plots can in turn be colored by other features. In Figure 12 the simplex plots of *wives' age* and *wives' education* was printed 3 times each. From left to right, color gradients illustrated respectively *wives' age*, *wives' education*, and lastly *n\_children*. The simplexes in the diagonal reproduced the main effect coloring from Figure 11, whereas other depicted simplexes possibly would detail 2<sup>nd</sup> order interactions. E.g. *wives' education* of Figure 12 showed the four clusters, one for each education level. The distance from any point to its local cluster as a mix of higher order effects and a small noise component. It was found that wives with highest education aged 20 were pre-

dicted more likely to use contraception than when aged 25. Wives' with highest education and few children (red) preferred short term contraception over long term. As the features *n\_children* and *wives' age* are correlated, these will both interact with *wives' education*, not only one.

## 5 Discussion

Forest floor is a methodology to visualize the mapping structure of a RF model using feature contributions. RF can be termed a predictive algorithmic model, designed to have a high predictive accuracy on the expense of model transparency [22, 3]. RF could also be termed as data driven, as the model can adapt itself to the data with little guidance. The opposite is a theory driven model where the user manually choose an explicitly and clearly stated model to capture the data structure. A practical advantage of using RF, is when the user have little prior knowledge or theory on the subject. The majority of nonlinear machine learning algorithms models have in common, that the resulting model stated as an equation is fairly complex in the eyes of a human user. The complexity may be difficult to avoid if the model should be able to capture an unknown structure. But exactly when little prior theory is given, that is when the model should inspire the interpretation of the data structure. A dualistic approach is to choose both a perhaps linear explanatory model to interpret the system and a machine learning algorithm to get the most accurate predictions [22]. Such an approach may leave a gap between users comprehension and the actual structure of the nonlinear model. If the user is far from understanding a certain data-structure any optimization cannot hardly evolve from brute trial-and-error searches such as grid search or ant-colony-optimization methods.

For nonlinear high-dimensional multivariate models, it is not straight forward to visualize the trained mapping function. The provided visualizations can be understood as slices or projections of the mapping structure. It appears that a given series of 2D and/or 3D projections can jointly explain the structure of a RF mapping surfaces. The quantifiable goodness-of-visualization measure describes how well the variance of the full structure can be explained in the context of the provided feature axis. If a large component of feature contribution variance remains unexplained, there is likely an unaccounted interaction pattern associated with this feature. Thus an advantage of forest floor is, that it aids the user to learn what local interaction effects are not yet visualized. With feature contributions it is possible to make an interpretation of what variance is attributed main effects, second order effects or higher order effects. Feature

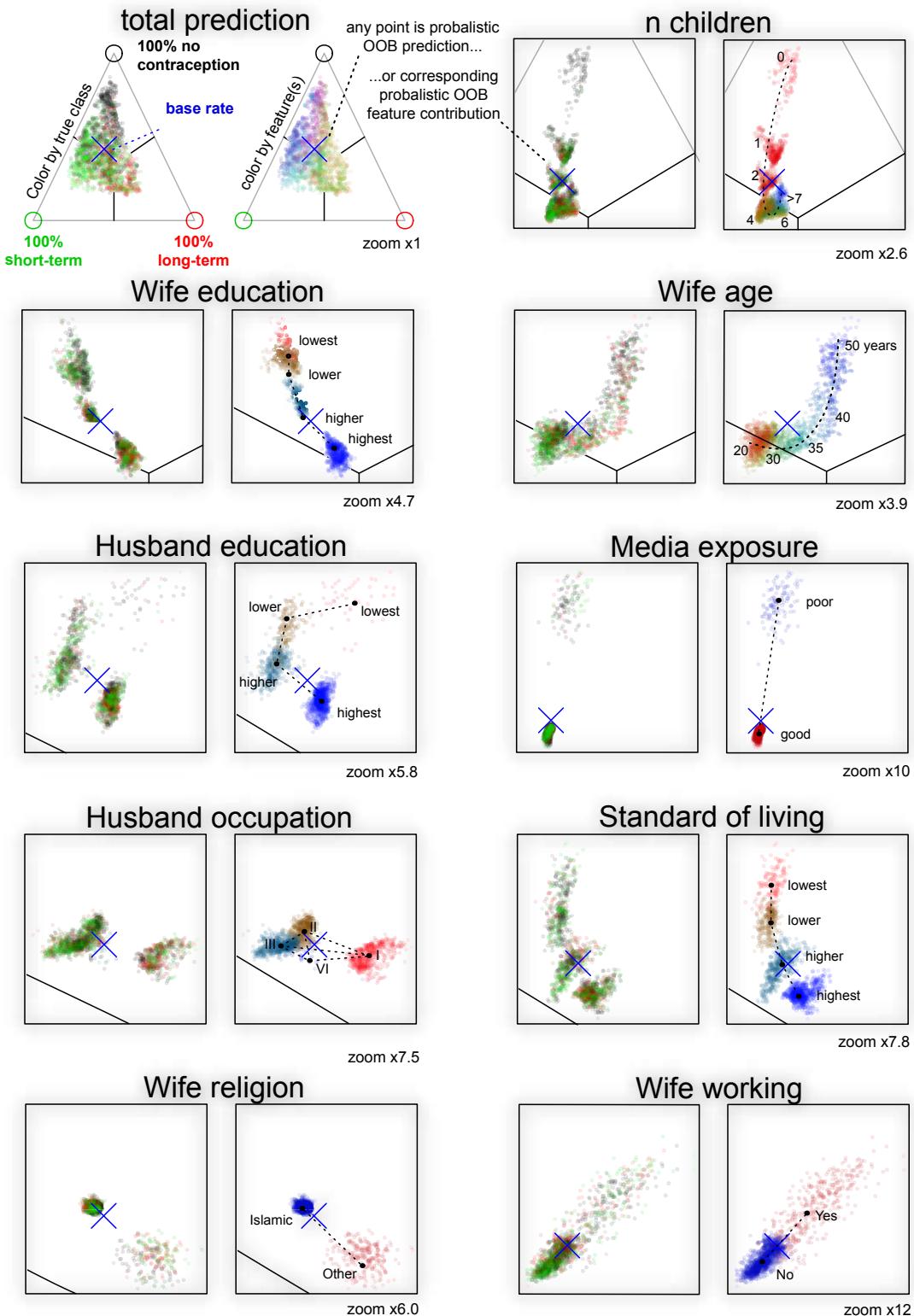
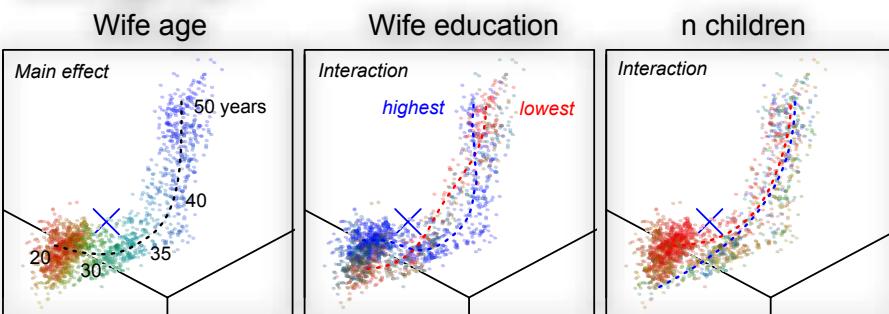


Figure 11: From top left: Cross validated predicted class probability colored by true class and a PCA color gradient describing observation diversity. Following pairs of plots, were the predicted probability decomposed into feature contributions. Left colored by true class, right colored by corresponding feature value. Red is minimal value, blue is maximal value. Blue cross is class base rate of training set. Dashed lines are drawn manually to assist interpretation of main effects.

### Wife age by...



### Wife education by...

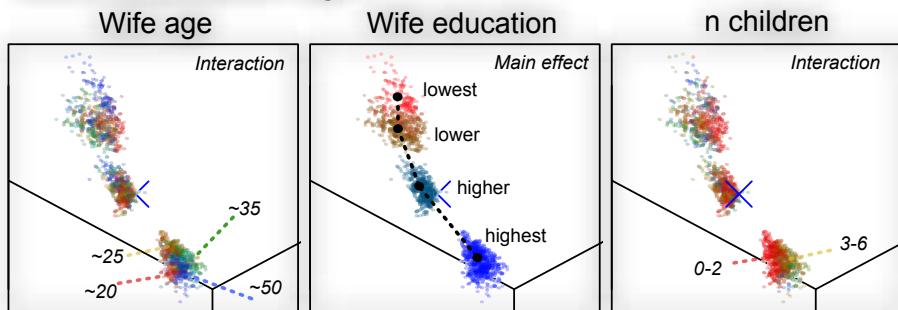


Figure 12: Feature contributions for the three most important features plotted row-wise. Each plot is colored column-wise by corresponding feature values. Dash lines are drawn manually to assist interpretation of interactions.

contributions can be computed from the training set itself and thus do not extrapolate the training set. The training set is used to set boundaries for model structure, such that extrapolated and unrelated model structures are not visualized. Feature contributions can be combined with the out-of-bag concept allowing cross validation to avoid presenting an overfitted mapping structure. Visualizations of cross validated feature contributions appear less noisy.

Color gradients allowed to include one or two extra dimensions in an illustration thus otherwise limited of three dimension. Color gradients traversing entire mapping space was used to highlight selected latent dimensions in a series of main effect plots to pinpoint missing interactions. We perceive colors as a combination of three channels red, green and blue. Thus, it may seem possible to visualize three additional dimensions in colors. Nonetheless, the ranges of color saturation and brightness should be constrained to avoid indistinguishable grey color tones and to ensure a minimal contrast to the background. Such considerations, limited color gradients to provide only two additional dimensions at maximum. It was possible to summarize a high-dimensional structure with e.g. principal component analysis and apply color gradients along the first 2 loading vectors, such as in Figure 11. In practice, we found a sequence of 1-dimensional color gradients best suited to uncover latent interaction structures in a RF model fit.

Feature contributions were first described in the context of RF regression, where a given feature can contribute either positively or negatively to a given prediction [9]. Next, the concept of feature contributions has previously been extended to classification, where the categorical majority vote labeling were replaced with numeric probability predictions [16]. We have argued that these probabilistic predictions are confined in a prediction space defined the  $(K - 1)$ -simplex, for model with  $K$  classes. Any node in any tree will itself be a prediction and have a position in this space. We argue local increments are in fact vectors connecting nodes in the  $(K - 1)$ -simplex space. The first local increment (the bootstrap increment) of any tree will be the vector connecting the class distribution of the training set to the class distribution of the root node. As the bootstrap increments will point randomly in any direction, the sum of a large number of such will approach the zero vector if no stratification is chosen. For stratification by true class, the bootstrap increments will connect the training set class distribution point in the  $(K - 1)$ -simplex to the point in the  $(K - 1)$ -simplex chosen by stratification.

The Gini loss function can be understood as maximizing the squared distance of node positions to the center of  $(K - 1)$ -simplex (equal class prob-

ability). Therefore any split by Gini will place the daughter nodes the furthest from the center, weighted by node size. As the classification trees are fully grown, the terminal nodes of one pure class can only be positioned on the vertices of the simplex. In Figure 11 was shown that the distribution of classes in the training set will function effectively as the prior of the RF model. If the user do not expect to find the same class distribution in future predictions as in training set, this prior can be moved in the simplex by stratification during the bootstrap process. In Figure 11 the center blue cross marked that the average root node center was skewed towards class 1 (no contraception) as 42% of the wives did not use any contraception. As class separation by the RF model was not strong the majority of predictions fall close to this prior base rate. In supplementary materials a RF model was trained with bootstrap stratification by true class such that the average root node is positioned in the center of the  $(K - 1)$ -simplex and following predicted class probabilities were also centred around this point. Figure 4 depicted how any node-split will produce two new nodes with local increments in perfectly opposite direction. Thus, training set predictions will always be centred around this point.

Direct plotting of  $K$  class probabilities requires  $K - 1$  dimensions. This is possible for 3 or 4 classes with 2D plot or 3D plot respectively. The context of feature values can only be included as one extra axis or as color gradients. We have shown that the axis of the  $(K - 1)$ -simplex can be aligned such that only one axis is needed to visualize the feature contributions as seen in Figure 10. This frees 1 or 2 axis to provide an adequate feature value context. In such visualization each observation will be plotted one time for each predicted class probability. Colors can be used to distinguish the classes.

In a previous article we trained a molecular descriptor model with RF to predict protein permeation enhancement in an epithelial cell model (Caco-2) [24]. A diagnostic tool was missed to address why such a model would be credible and to communicate intuitively the found pattern to fellow chemists/biologist with little knowledge of machine learning. We first stumbled upon feature contributions in the two articles [16, 9] and experimented to plot these feature contributions against the feature values. The R package rFC [2] provided the first computations of feature contributions and was an inspiration to the design of the forestFloor package [25]. Hereafter we discovered partial dependence plots and sensitivity analysis [5, 6]. Now in hindsight we can report the set of advantages to forest floor, especially the tracking of unaccounted interactions such that no strong interaction will be overlooked when visualizing the mapping structure.

The following citation by Friedman [6] origi-

nates from an article from 2001 discussing the usefulness of partial dependence plots on nonlinear functions: *"Given the general complexity of these generated targets as a function of their arguments, it is unlikely that one would ever be able to uncover their complete detailed functional form through a series of such partial dependence plots. The goal is to obtain an understandable description of some of the important aspects of the functional relationship."* [6]

Indeed the structure of RF models can be highly complex and visualizations are unlikely to present every detail at once. Therefore a visualization tool-set should assist the user to navigate the mapping structure. This has been done by isolating the part of the model structure related to the data structure, by evaluating the goodness-of-visualization of a given plot, and by pointing to where locally in the model structure a sizable latent interaction is not yet visualized. Our goal is to present complex models as adequately detailed visualizations. In a RF model there will likely always be a baseline of random ripples in the mapping structure, that we do not expect to be able to reproduce. These ripples are partly filtered of by using the out-of-bag cross validated feature contributions. Other ripples occur due to biases of the RF algorithm. Especially does the RF model structure surface contain wave like curvature parallel to the feature axes due to the univariate step functions of RF, see RF surfaces in Supplementary Materials.

We predict that 4D projections of a third order interaction rarely would be needed for the RF algorithm. In supplementary materials we have provided a simulation suggesting that RF only poorly can fit interactions higher than second order even when trained on 10.000 observations without any noise. This can be explained as the RF algorithm is limited in its potential complexity as the algorithm only can perform univariate splits decided by an immediate loss function. Another algorithm such as rotation forest [19] is not limited to perform univariate splits and therefore better on such simulated tasks with higher order interactions. What initially was an interaction effect can be rearranged into a main effect by new combined features. Multivariate split methods are not compatible with forest floor, but they are compatible with the generic methods partial dependence plots and sensitivity analysis [6, 5].

## 6 Conclusion

Forest floor has extended the tool-box to visualize the mapping structure of RF models. The geometrical relationship between random forest models and feature contributions has been described. For RF multi-classification it was useful to understand

the prediction space as a  $(K - 1)$ -simplex probability space. Hereby the feature contributions can be interpreted as changes of predicted probability due to a given feature. A  $(K - 1)$ -simplex prediction space can also visualize how the training set stratification affect RF predictions. Target class stratification is effective to modify the prior for the RF model.

We have emphasized that parts of a mapping structure which extrapolates the training set are irrelevant. To extract only the relevant mapping structure, feature contributions are computed only from the training set itself. Two new variants of feature contributions have been introduced to avoid inherent overfitting when using training set predictions. These variants of feature contributions are out-of-bag cross validated feature contributions, and n-fold cross validated feature contributions.

Feature contributions from a single feature can contain variance from main effects and/or interaction effects. A measure of goodness-of-visualization has been introduced to evaluate if the feature contributions of a given feature alone can be explained in the context of itself. If not, color gradients traversing the mapping space can be used to pinpoint overlooked interactions within feature contributions and features. Sizable interactions can be visualized in two-way interaction plots in the context of two features and perhaps even a third feature as color gradient. Again a goodness-of-visualization can be computed and evaluated for such a visualization.

Ultimately, it is difficult to communicate a context of more than 2 or 3 dimensions + target dimension(s). Thus fourth order interactions would be difficult to visualize and communicate. Anyhow, such visualizations are likely not missed, as the random forest algorithm could not fit fourth order interactions well and had a poor efficiency already with third order interactions.

As forest floor can break down a RF model fit into effects attributed to each feature and assist to find adequate context to understand these effects. It is intended that RF no longer should be seen as a non interpretable model. Learned associations between features and targets should inspire new ideas of the underlying possible causality structure.

## References

- [1] Monther Alhamdoosh and Dianhui Wang. Fast decorrelated neural network ensembles with random weights. *Information Sciences*, 264(0):104 – 117, 2014. Serious Games.
- [2] Richard Marchese Robinson Anna Palczewska. *rFFC: Random Forest Feature Contributions*, 2015. R package version 1.0/r6.
- [3] Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):pp. 199–215, 2001.
- [4] Paulo Cortez. UCI machine learning repository, 2009.
- [5] Paulo Cortez and Mark J. Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225(0):1 – 17, 2013.
- [6] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [7] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.
- [8] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.
- [9] Victor E. Kuz'min, Pavel G. Polishchuk, Anatoly G. Artemenko, and Sergey A. Andronati. Interpretation of qsar models based on random forest methods. *Molecular Informatics*, 30(6-7):593–603, 2011.
- [10] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
- [11] Tjen-Sien Lim. UCI machine learning repository, 1987.
- [12] Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40(3):203–228, 2000.
- [13] Sheng Liu, Shamitha Dissanayake, Sanjay Patel, Xin Dang, Todd Mlsna, Yixin Chen, and Dawn Wilkins. Learning accurate and interpretable models based on regularized random forests regression. *BMC Systems Biology*, 8(Suppl 3):S5, 2014.
- [14] Raphael Maree, Pierre Geurts, Justus Piater, and Louis Wehenkel. Random subwindows for robust image classification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 34–40. IEEE, 2005.
- [15] Deirdre B. O'Brien, Maya R. Gupta, and Robert M. Gray. Cost-sensitive multi-class classification from probability estimates. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 712–719, New York, NY, USA, 2008. ACM.
- [16] Anna Palczewska, Jan Palczewski, Richard Marchese Robinson, and Daniel Neagu. Interpreting random forest classification models using a feature contribution method. In Thouraya Bouabana-Tebibel and Stuart H. Rubin, editors, *Integration of Reusable Systems*, volume 263 of *Advances in Intelligent Systems and Computing*, pages 193–218. Springer International Publishing, 2014.
- [17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

- [19] Juan José Rodriguez, Ludmila I Kuncheva, and Carlos J Alonso. Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1619–1630, 2006.
- [20] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015.
- [21] Mark Seligman. *Rborist: Extensible, Parallelizable Implementation of the Random Forest Algorithm*, 2015. R package version 0.1-0.
- [22] Galit Shmueli. To explain or to predict? *Statistical science*, pages 289–310, 2010.
- [23] † Vladimir Svetnik, \*, † Andy Liaw, † Christopher Tong, ‡ J. Christopher Culberson, § Robert P. Sheridan, , and Bradley P. Feuston‡. Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003. PMID: 14632445.
- [24] Soeren H. Welling, Line K.H. Clemmensen, Stephen T. Buckley, Lars Hovgaard, Per B. Brockhoff, and Hanne H.F. Refsgaard. In silico modelling of permeation enhancement potency in caco-2 monolayers based on molecular descriptors and random forest. *European Journal of Pharmaceutics and Biopharmaceutics*, 94(0):152 – 159, 2015.
- [25] Soeren Havelund Welling. *forestFloor: Visualizes Random Forests with Feature Contributions*, 2015. R package version 1.8.6.
- [26] M. N. Wright and A. Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *ArXiv e-prints*, August 2015.

## Suplementary materials for: "*Forest Floor Visualizations of Random Forests*".

Soeren H. Welling, Line K.H. Clemmensen, Hanne H.F. Refsgaard, & Per B. Brockhoff

May 31, 2016

### 1 Proof for Equation [11] and [12] of article.

**Part 1 - Any sequence of  $d$ -dimensional vectors:** Denote a sequence of  $n+1$  real vectors (or scalars) describing points in a  $\mathbb{R}^d$   $d$ -dimensional space as  $\hat{y}_k''$  for  $k \in \{0, \dots, n\}$ . The difference between any two adjacent vectors is defined as  $L_k = \hat{y}_k'' - \hat{y}_{k-1}''$  for  $k \in \{1, \dots, n\}$ .

**Lemma 1:**

$$\hat{y}_n'' = \sum_{k=1}^n L_k + \hat{y}_0'' \quad (1)$$

**Proof 1:**

For  $k \in \{1, \dots, n\}$ ,  $\hat{y}_k''$  is the additive part of  $L_k$  and  $\hat{y}_{k-1}''$  the substractive part.

When summing every  $L_k$ , all intermediary vectors of the sequence cancel out.

$$\sum_{k=1}^n L_k = (\hat{y}_1'' - \hat{y}_0'') + (\hat{y}_2'' - \hat{y}_1'') + (\hat{y}_n'' - \hat{y}_{n-1}'') = \hat{y}_n'' - \hat{y}_0''$$

Replacing  $\sum_{k=1}^n L_k$  with  $\hat{y}_n'' - \hat{y}_0''$  in stated Lemma 1, one obtain

$$\hat{y}_n'' = \hat{y}_n'' - \hat{y}_0'' + \hat{y}_0''$$

**Part 2 - a single tree:** A tree is a hirachial graph. The first node, node 0, is connected to node 1. Every node from node 1 is either terminal and only connected to one parent node or

an intermediary node and has two daughter nodes. Every node of a tree has a prediction  $\hat{y}_k''$  which is a real vector/scalar with exactly  $d$  dimensions.

### Notes for part 2

*For regression, a node prediction is a real scalar and computed as the target mean of inbag samples passing through the node. For classification a vector of  $d$  dimensions, where  $d$  is the number of classes in the training set, and each element from 1 to  $d$  describe the prevalence ratio of inbag samples by a given class. Notice some random forest implementation use majority voting in terminal nodes. Here majority class element will be 1 and the remaining 0. Virtually any other prediction rule for nodes in classification trees outputting real valued vectors of length  $|\hat{y}_k''|=1$  would be acceptable. Virtually any other prediction rule for nodes in regression trees outputting real values would be acceptable.*

An observation is an entity which will take one direct path of steps through the tree, starting from node 0 and ending in a terminal node. Observations are enumerated for  $i \in \{1, \dots, N\}$ . Each observation will attain a sequence of predictions, one for each node it passes through. Each prediction is a real vector/scalar and written  $\hat{y}_{ik}''$ , where  $k$  sequentially enumerates the  $n$  nodes of the path for observation  $i$ . As  $n$  may differ for each observation  $i$ , it is thus written  $n_i$ . In one tree, any observation step sequence share the same first node 0 and node 1 also called the root node of the tree. A local increment ( $L_{ik}$ ) is defined as a vector describing the prediction difference from  $(k-1)^{th}$  to the  $k^{th}$  node for observation  $i$ .

Therefore we write  $L_{ik} = \hat{y}_{i,k}'' - \hat{y}_{i,k-1}''$  for  $k \in \{1, \dots, n\}$  for  $n_i \geq 1$ .

The first node  $y_0$  of one tree contain all observations and the prediction is the training set base rate / grand mean.  $y_0$  can also be written as  $\bar{y}$ . The tree prediction of the  $i^{th}$  observation  $\hat{y}'_i$ , is defined as defined as the terminal node  $\hat{y}'_i = \hat{y}_{i,n_i}''$  where  $k = n_i$ .

### Lemma 2

$$\hat{y}'_i = \sum_{k=1}^{n_i} L_{ik} + \bar{y} \text{ for any } i \in \{1, \dots, N\} \quad (2)$$

**Proof 2:** As a given sequence of node predictions  $\hat{y}_{ik}''$  for a given observation  $i$  are real

vectors/scalars, then the local increments of this sequence must be a part of any sequence postulated in lemma 1. Replacing  $\bar{y}$  with  $y_0$  and  $\hat{y}'_i$  with  $\hat{y}''_n$  we obtain lemma 1. Thus lemma 2 must be true also.

**Part 3 - the test set prediction of any ensemble of trees** The tree prediction of the  $i^{th}$  observation of the  $j^{th}$  tree is written  $\hat{y}'_i$ . An ensemble prediction  $\hat{y}_i$  of  $ntree$  decision trees is equal to the mean of the tree predictions  $\hat{y}'_{ij}$  for each  $i$  observation.  $\hat{y}_i = \frac{1}{ntree} \sum_j^{ntree} \hat{y}'_{ij}$  for A local increment of the  $j^{th}$  tree  $L_{ijk}$  can be written  $L_{ijk}$  the number of local increments/steps for the  $i^{th}$  sample in the  $j^{th}$  tree can be written  $n_{ij}$ .

### Lemma 3

$$\hat{y}_i = \frac{\sum_{j=1}^{ntree} \sum_{k=1}^{n_{ij}} L_{ijk}}{ntree} + \bar{y}, \quad (3)$$

### Proof 3:

$$\begin{aligned} \hat{y}_i &= \frac{\sum_{j=1}^{ntree} \sum_{k=1}^{n_{ij}} L_{ijk}}{ntree} + \bar{y} \\ \hat{y}_i &= \frac{\sum_{j=1}^{ntree} \sum_{k=1}^{n_{ij}} (L_{ijk} + \bar{y})}{ntree}, \text{ use Lemma 2 to replace } L_{ijk} \text{ with prediction of } j^{th} \text{ tree } \hat{y}'_{ij} \\ \hat{y}_i &= \frac{\sum_{j=1}^{ntree} \hat{y}'_{ij}}{ntree}, \text{ this is the definition of the ensemble prediction} \end{aligned}$$

**Part 4** For any  $j^{th}$  tree, any training observation  $i$  is either be designated as inbag or out-of-bag (OOB). The OOB prediction  $\tilde{y}_i$  computed from a subset of all trees  $\{1, \dots, ntree\}$  where  $i$  is OOB, we call this subset for  $\tilde{J}_i$  and this set will have  $n_{OOB_{tree,i}}$  members. The OOB ensemble prediction is defined as the mean prediction of OOB tree for the  $i^{th}$  observation.

$$\tilde{y}_i = \frac{1}{n_{OOB_{tree,i}}} \sum_{j \in \tilde{J}_i} y_{ij} \text{ where subset } \tilde{J}_i \subseteq \{1, \dots, ntree\}.$$

### Lemma 4

$$\tilde{y}_i = \frac{\sum_{j \in \tilde{J}_i} \sum_{k=1}^{n_{ij}} L_{ijk}}{n_{tree}} + \bar{y} \quad (4)$$

Proof 4: As lemma 3 was shown for any set of trees in an ensemble, and as lemma 4 is just the special case for particular subsets of trees, then lemma 4 must be true also.

### 1.1 How to highlight the mapping structure of a local cluster

In the white wines quality (wwq) data set. A local interaction was identified among wines with the lowest alcohol content (< 9.3%). In spite of low alcohol content in general lead to lower preference predictions, a subgroup of low alcohol wines deviated from this main effect. It was possible to further characterize this local interaction in the mapping structure of the trained RF model. Any wine of alcohol content more than than 9.3 was colored transparent grey. Remaining low alcohol wines were colored by the feature contribution of alcohol, such that wines with a relatively positive impact of low alcohol content were marked blue and wines with a relatively negative impact were marked red. Intermediate wines were green. Main effect plots by all features were colored by these scheme as depicted. Hereby it was possible to visualize the local interaction. It was possible to observe that the most clear differences between wines marked blue and wines marked red was the content of chlorides, citric acid and residual sugar. This observation characterized a certain cluster of fruity wines (acidic and sweet) of high preference despite low alcohol content.

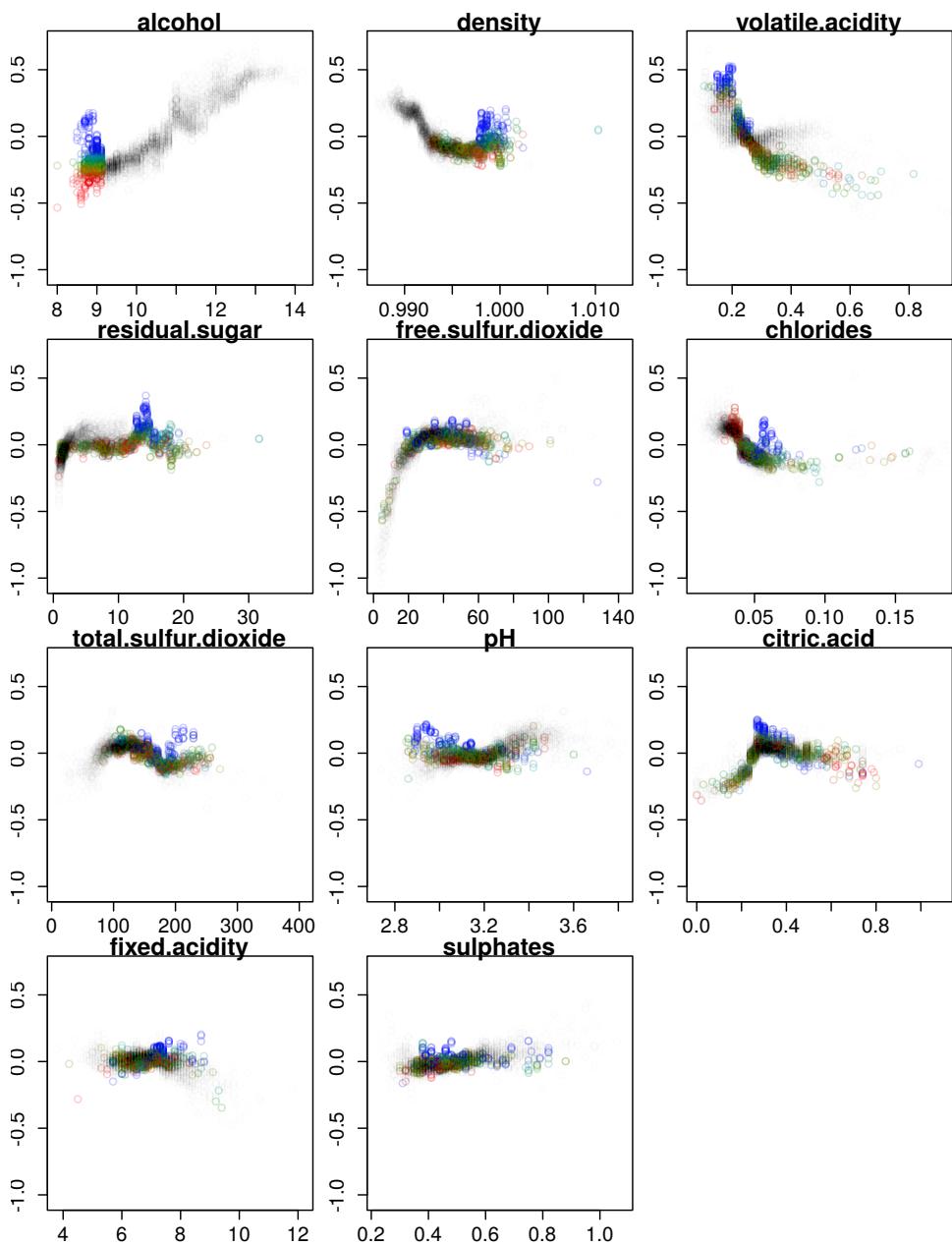


Figure 1: Cross-validated main effect feature contributions of predicted preferences of 4900 white whines. The color gradient along feature contributions of alcohol characterizes the specific interaction pattern between low alcohol content and remaining features.

## 1.2 RF mapping unrelated to data structure

Two normal distributed( $N(0,1)$ ) variables  $x_1$  and  $x_2$  is related to a target  $y$  by either  $G_1(X) = y_1 = (x_1)^2 + 2\sin(2x_2)$  or  $G_2(X) = y_2 = x_1x_2$ . 3000 samples were drawn and a default RF-model was trained. A grid of 300 grid lines and  $300^2$  grid points was formed. Each grid point represented a combination of  $x_1$  and  $x_2$  from  $-7$  to  $7$  such that the entire grid extended the range of sampled values 3 times. Any grid point of  $x_1$  and  $x_2$  was predicted by the RF model. The predicted  $\hat{y}$  was plotted as a function of  $x_1$  and  $x_2$  in a 3D plot. The mapping structure was represented as a surface outlined by the grid points and colored by high  $\hat{y}$  (red/high, green/low). The mapping of the training set is represented by the set blue points on the mapping surface. For the data structure  $G_1$  there is no unstable boundary effect as the partial quadratic function of  $x_1$  and the partial sine function of  $x_2$  do no interact and simply intersect additively in the region of the training set (blue points). The saddle-point structure of  $G_2$  is not the sum of two additive partial functions. In a rectangular boundary of were training set was observed a series of ripples in the mapping structure was observed. Here predictions alternated between high and low values. This boundary mapping structure do not reflect the data structure of  $G_2$ . Likely as RF only performs univariate splits, it can only capture interaction effects by splitting data into sub groups. As these sub group becomes less populated at the boundaries of the data set the fit becomes markedly unstable.

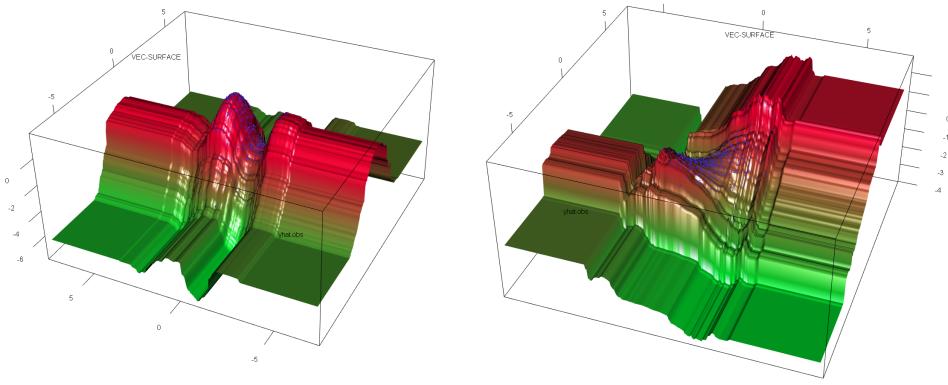


Figure 2: RF regression model structure of two hidden functions  $y_1 = (x_1)^2 + 2\sin(2x_2)$  (left) or  $y_2 = x_1x_2$  (right). Red-green color gradient is parallel to the vertical target axis,  $\hat{y}$ . Positions marked blue are the training examples used to train the mapping structure. The visualized surface extrapolates the training set 100% in each direction. Left plot( $y_1$ ) depicts a stable main effect only structure. Right plot( $y_2$ ) depicts an unstable interaction effect only structure.

### 1.3 Shallowness of Random forest

Although splits of nodes in RF is performed univariately, RF can still capture interactions due to the many local rules applied. Presumably as the sequential decisions performed by RF satisfy only an immediate loss function of each split and splits are only univariate, RF cannot grow decision trees to capture 4<sup>th</sup> order interactions or higher. To test the ability of RF to captivate data structures of various complexity, three hidden structures were designed. A series of  $i$  variables  $x_i$  were drawn from a distribution and multiplied. The structure have no error component. Figure 3 depicts from  $d = 1$ (light green) to  $d = 6$ (red) the ability of random forest models to fit a training set of  $N$  train samples. A single main effect is modelled with almost no error already from 100 observations. A second order interaction needs 100-200 samples to explain 75% of the variance when cross validated. A third order interaction in a feature space of continuous variables ("saddle" & "sineprod") requires 10,000 samples to explain 75% variance cross validated.

”saddle”

$$y_d = \prod_{i=1}^d x_i, x_i \in N(0, 1) \quad (5)$$

”sineProd”

$$y_d = \prod_{i=1}^d \sin(x_i), x_i \in U(-\pi/2; \pi/2) \quad (6)$$

”binaryProd” (notice only -1 or 1 is sampled)

$$y_d = \prod_{i=1}^d x_i, x_i \in U\{-1, 1\} \quad (7)$$

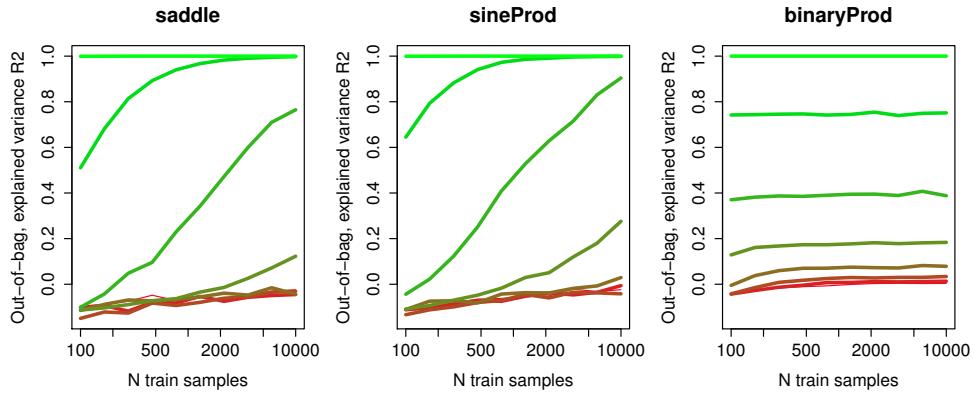


Figure 3: How many orders of interactions can RF capture? Three structures saddle, sineProd and binaryProd, ranging from main effect(light green) 6th order of interaction(red line). RF already becomes an poor estimator at 3rd order interactions.

#### 1.4 The effect of stratification

Stratified bootstrapping by target variable moves weighted centroid of cross validated training predictions to the center of the simplex. Hereby, highly prevalent classes are down-sampled, but every sample will likely participate at least in a small number of trees. Appendix Figure 5 depicts such a stratified RF model, where root node is balanced in respect of target classes. Besides the centroid of prediction were moved to the center of K-1 probability simplex, the general structure of the model structure seemed similar to the non-stratified version in manuscript.

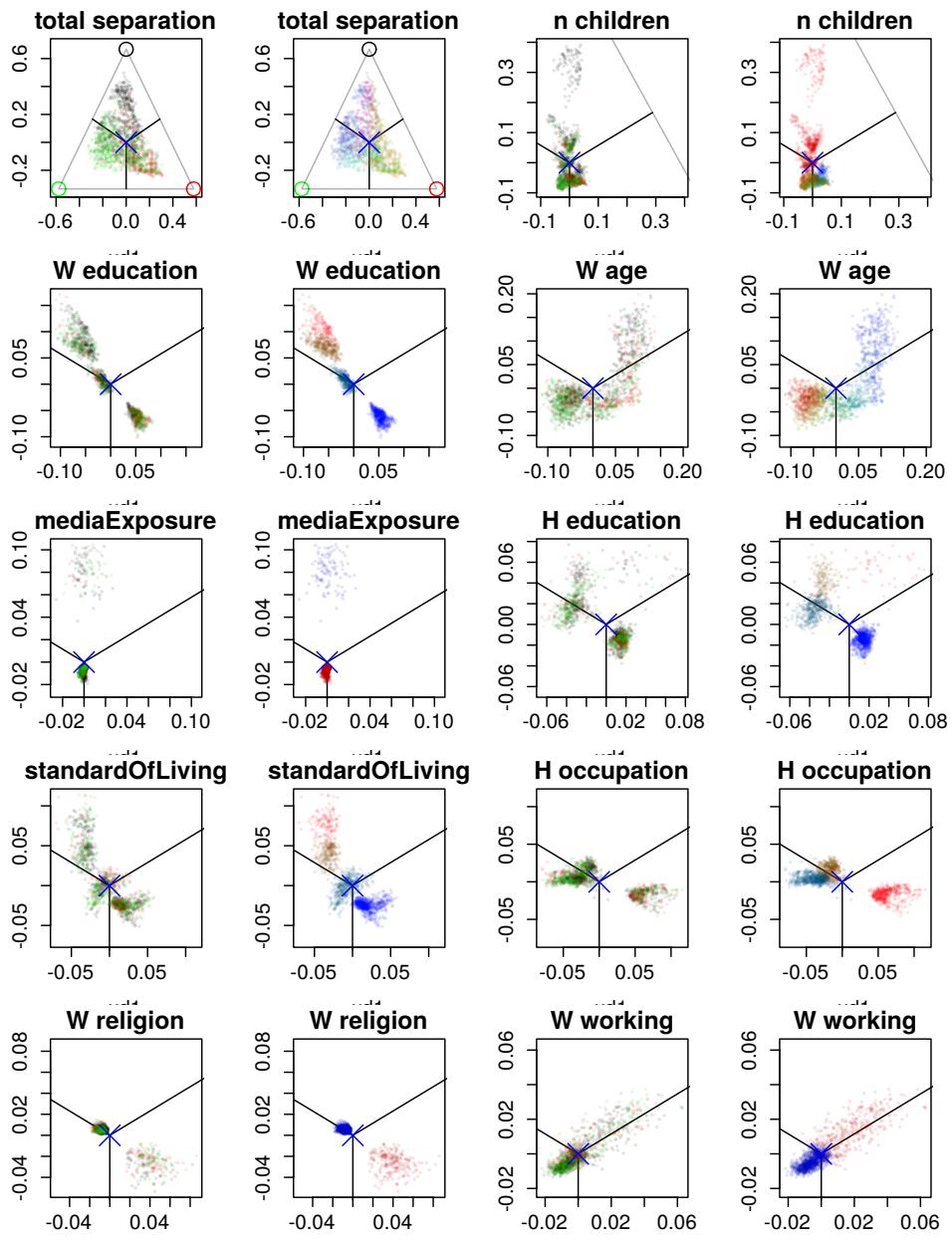


Figure 4: Feature contributions for contraceptive method choice (cmc) data set when RF was trained with target class stratification. Blue cross marks average root node which is also the center of the average cross validated prediction.

## 1.5 forest floor visualizations of gradient boosted tree

Gradient boosted trees suggested by Friedman is a boosted ensemble, where each new tree is fitted to the residuals of the current ensemble of trees [2]. Nonetheless, all grown trees in the ensemble are regular decision trees similar to trees of random forest ensembles. The gradient boosted ensemble prediction is the sum of votes, whereas for a random forest ensemble it is the average vote. In either case, both boosted trees and bagged trees contribute additively to the ensemble prediction. Therefore can every prediction be split into local increments and the feature-wise subtotals, named feature contributions can be computed. Presently, the perhaps most popular gradient boosting algorithm is XGBoost [1]. To make a fast proof-of-concept we preferred not to write an entirely new adaptor for XGBoost, but rather to write a wrapper around the randomForest implementation [3], making it behave as a gradient boosted ensemble and retain compatibility with forestFloor. This short wrapper is printed below and included in the forestFloor package as an example script **ffGradientBoost.R**.

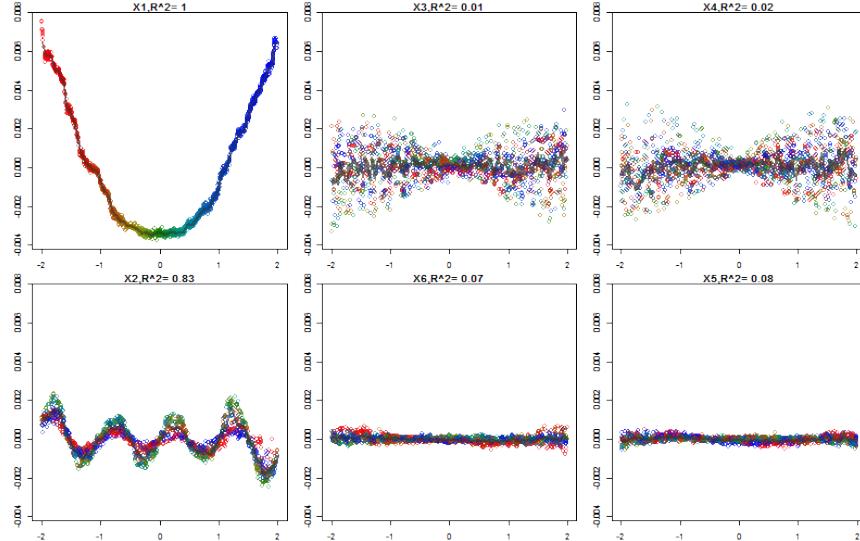


Figure 5: forestFloor visualization of a simpleBoost model. simpleBoost is a gradient boosted tree ensemble, implemented as a simple wrapper of the CRAN randomForest algorithm.

```

1 library (randomForest);library (forestFloor)
3 #simulate data
4 X      = data.frame(replicate(6,4*(runif(3000)-.5)))
5 Xtest = data.frame(replicate(6,4*(runif(1500)-.5)))
6 y      = with(X,X1^2+sin(X2*2*pi)+X3*X4) + rnorm(3000)/3
7 ytest = with(Xtest,X1^2+sin(X2*6*pi)+X3*X4) + rnorm(3000)/3

9 #define boosted tree wrapper
10 simpleBoost = function(
11   X,y,      #training data
12   M=100,    #boosting iterations and ntrees
13   v=.1,     #learning rate
14   ...) {  #other parameters passed to randomForest
15   y_hat = y * 0  #latest ensemble prediction
16   res_hat = 0    #residuals hereof...
17   Fx = list()    #list for trees
18   for(m in 1:M) {
19     y_hat = y_hat + res_hat * v  #update prediction, by learning rate
20     res = y - y_hat            #compute residuals
21     hx = randomForest(X,res ,ntree=1,keep.inbag=T,...) #grow tree on residuals
22     res_hat = predict(hx,X)           #predict residuals
23     cat("SD=",sd(res), "\n")       #print
24     hx$forest$nodepred = hx$forest$nodepred * v #multiply nodepredictions by learning rate
25     Fx[[m]] = hx  #append tree to forest
26   }
27   Fx = do.call(combine,Fx) #combine trees with randomForest::combine()
28   Fx$y = y #append y
29   Fx$oob.times = apply(Fx$inbag,1,function(x) sum(!x)) #update oob.times
30   class(Fx) = c("simpleBoost","randomForest") #make simpleBoost a subclass of randomForest
31   return(Fx)
32 }
33
34 predict.simpleBoost = function(Fx,X) {
35   class(Fx) = "randomForest"
36   predMatrix = predict(Fx,X,predict.all = T)$individual
37   ntrees = dim(predMatrix)[2]
38   return(apply(predMatrix,1,sum))
39 }

40 plot.simpleBoost = function(Fx,X,ytest ,add=F,...) { #plots learning curve
41   class(Fx) = "randomForest"
42   predMatrix = predict(Fx,X,predict.all = T)$individual
43   ntrees = dim(predMatrix)[2]
44   allPreds = apply(predMatrix,1,cumsum)
45   preds = apply(allPreds,1,function(pred) sd(ytest-pred))
46   if(add) plot=points
47   plot(1:ntrees ,preds,...)
48   return()
49 }
```

```

51 }
52
53 #build gradient boosted forest
54 rb = simpleBoost(X,y,M=300,replace=F,mtry=6,sampszie=500,v=0.005)
55
56 #make forestFloor plots
57 ffb = forestFloor(rb,X,Xtest)
58 #correct for that tree votes of gradient boosts are summed, not averaged.
59 #forestFloor will as default divide by the same number as here multiplied with
60 ffb$FCmatrix = ffb$FCmatrix * c(rb$oob.times,rep(rb$ntree,sum(!ffb$isTrain)))
61
62 #plot forestFloor for OOB-CV feature contributions and regular feature contributions
63 plot(ffb,plotTest=T,col=fcol(ffb,3,plotTest = TRUE))
64 plot(ffb,plotTest=F,col=fcol(ffb,1,plotTest = FALSE))
65
66 #validate model structure
67 pred = predict(rb,X)
68 predtest = predict(rb,Xtest)
69 plot(y,pred,col="#00000034")
70 plot(rb,Xtest,ytest,log="x")
71 vec.plot(rb,X,i.var=1:2)
72
73 #export plot
74 png(file = "ffGradientBoost.png", bg = "transparent",width=800,height = 500)
75 plot(ffb,plotTest=T,col=fcol(ffb,1))
76 rect(1, 5, 3, 7, col = "white")
77 dev.off()

```

## References

- [1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016.
- [2] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [3] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

# CHAPTER 6

## structureOfSolubility

---

The decision tree ensemble random forest have a series of useful diagnostics which have been used in this thesis work.

DOI: 10.1002/minf.200((full DOI will be filled in by the editorial staff))

# Learning the structure of random forest models in QSAR modelling: Predicting molecular Solubility

Søren H. Welling(1,2), Line KH Clemmensen(1), Per B. Brockhoff(1,2) and Hanne HF Refsgaard(1,2).

Dedication((optional))

**Abstract:** Random forest (RF) models are used in QSPR models to predict solubility by the molecular structure. Non-linear models, such as RF, have been difficult to interpret as the model of many trees each of many nodes is far too complex to comprehend. Instead a model can be understood as a high-dimensional mapping structure which can be decomposed into a series of main effects and interactions. With feature contributions, and a newly developed tool it is possible to produce 2D and 3D visualizations to browse the model structure. We have built a model

of 12 standard molecular descriptors on a very cited data set of 1200 molecules and illustrated how a RF model fit weigh the information to produce predictions of solubility. It appears that interactions between used descriptors have a minor contribution on solubility prediction accuracy. The exemplified particular RF model fit can be boiled down to a series non-linear transfer functions, one for each descriptor, and some minor interactions. Moreover, the error of making such a specific generalization can be quantified. The proposed tool will likely be useful to interpret many other RF based QSAR models.

**Keywords:** keyword 1, keyword 2, keyword 3, keyword 4, keyword 5

## 1 Introduction

Quantitative structural activity/property relationship (QSAR/QSPR) models have been used to perform solubility predictions, and have e.g. been used in the pharmaceutical industry to select drug candidates for oral delivery. Insufficient solubility is likely lead to lower bioavailability [10]. Related, QSAR models have been used to estimate impact of pesticides on aquatic environment as a function estimated molecular octane/water partition coefficients ( $\log P$ ) [11].

QSAR models represent an empirical approach to establish a relationship between measured properties such as molecular solubility and a numerical description of molecules. Molecular formulas, SMILES or connection tables are graph representations of connected atoms by different types of bonds[cite]. These representations can be encoded to produce numerical descriptions, such as *molecular weight* or *ratio of rotatable bonds*. Molecular descriptors can make use of physicochemical theoretical calculations to estimate internal partial charges between atoms to predict e.g. polarity of the molecule [8,PEOE]. Prediction by other empirical derived models of  $\log P$ (*SlogP*) or molar refractivity(*SMR*), can be reused to predict solubility [4(*SlogP/SMR*)]. Molecular descriptors, based on atomic contributions or functional group contributions, will naively view the molecule as a simple sum of its atoms or functional groups. Scores for each type of atom or functional group are fitted to explain a data set of measured  $\log P$  or molar refractivity. Other descriptors

such as KierHall can quantify how branched the molecular graph is[cite]. Finally encodings can perform a 2D or 3D force field simulations predicting an energy favourable conformation of the molecule (MFA,dipole[cite]).

[section on the previous models and descriptors from ESOL, huuskonen, Delaney,]

Multiple linear regression (MLR) has been used to find a linear relationship between molecular descriptors and the predicted property. Often within a narrow selection of related molecules [cite sulphonate prediction] or when the molecular descriptors are well designed, linear models will perform well[zheng]. The last couple of decades, non-linear models such as support vector machines, neural nets and random forest have improved the prediction performance[cite some review]. These algorithmic models do not rule out unspecified non-linear relationships and neither interactions.

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Matematiktorvet, Building 324, 2800 Kgs. Lyngby, Denmark

[2] Novo Nordisk Global Research, Novo Nordisk Park 1, 2760 Maaloev, Denmark  
 \*e-mail:HARE@NOVONORDISK , +45 3075 0367

Supporting Information for this article is available on the WWW under [www.molinf.com](http://www.molinf.com)

[Continue with new models, new palmer, laura, bergstrom]

Ideally by improving molecular descriptors a complex non-linear regression model would not be needed. In practice it is difficult to adapt to an unknown non-linearity, especially when high. A successful RF model structure should not only be considered as a black-box structure, but it should inspire to new and better feature engineering. A deeper insight of the trained model structure of RF could improve our understanding of predicted molecular solubility and point to further improvements of molecular descriptors.

## 1.2 Article, aim, goal, approach

We will demonstrate how a RF model structure can be systematically deconstructed and visualized to describe the learned QSAR between molecular descriptors and solubility. In a previous article[cite], we used the concept of feature contributions[kuzmin, anna] to illustrate how a random forest model was able to predict a specific biological activity of molecules in a cell-culture model. Hereafter we investigated the topology of feature contributions and made a new tool, forest floor, to visualize and understand the structure of random forest models[cite]. In this article we build a conventional QSPR solubility model based on a highly cited and reused data set((huskonen)), Delaney, palmer, bergstrøm, laura] to discuss how RF utilize this information of molecular descriptors to produce predictions of solubility. [specify athours contribution, hertil gør palmer normalt så langt går vi videere]

[Beskriv composition af artikel]

## 2 Method

### 2.1 Introduction to random forest regression

A random forest model[leo] is a bootstrap aggregate ensemble model (bagging). It consists of hundreds or thousands of individual decision trees, whom are aggregated to form a joint robust, yet adaptive, ensemble model. Growing each decision tree starts with drawing  $N$  samples from the training set with replacement. Hereby, in average a .631 fraction out of  $N$  training set molecules are sampled to the root node of a tree at least once. These samples for a given tree are the inbag samples. The root node has a node prediction defined as the average measured solubility of the molecules in the node. The mean square error (mse) of the root node prediction can be reduced by splitting into two daughter nodes. Molecular descriptor are used to search for a splitting rule. A splitting rule (larger or equal to a value by one molecular descriptor) will split the node into two daughter nodes. One daughter node will have a higher molecular solubility and one with a lower than the parent node. The best split will lower mse of predicted solubility in the daughter nodes the most. By default, only a random third of descriptors are evaluated in any node to ensure not only one dominant molecular descriptor greedily is used first. Every node is recursively split until node size reaches 5 or less. Then the node is designated as terminal node. To perform predictions, new samples are passed down the tree according to the split rules. The terminal nodes will make up the possible solubility predictions of the tree. These almost

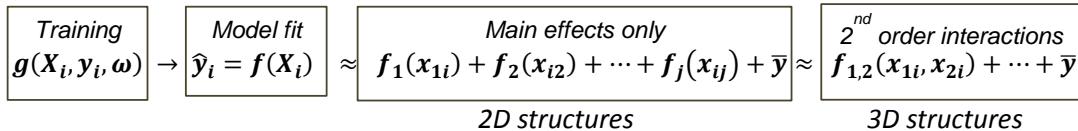
fully grown trees are likely low biased, as the potential model structure is very flexible. Though individually, a tree has a high variance as each prediction is based on only 5 or less samples. To counter the instability of each tree, the bootstrapping and only evaluating a random third of descriptors in every node ensure low correlation between trees. When trees are less correlated and the variance is random and symmetrical, the learned structure will be amplified and the variance will be averaged out [breiman].

For every tree, a set of molecules will be out-of-bag (OOB) in contrary to inbag, when used to grow the tree. To estimate the accuracy of the model fit, it must be cross validated. Any sample will be OOB in roundly one third of the trees in the model and thee tree can independently predict this sample. The cross validated prediction error is the expected performance of the model if new molecules were predicted, assuming these molecules were drawn independently from the same population as the training set. OOB cross validation is faster and yields comparable estimates as 5-fold cross validation [wessivik]. Variable importance (VI) can be used to order molecular descriptors by usefulness to the model. VI is the decrease of OOB cross validation performance (mse) if a given molecular descriptor, after growing trees, but before predicting OOB samples, was permuted (random shuffled) [caroline strobl]. VI can be used for variable selection or as in the visualizations of this paper, to bring the attention to the most useful variables first. Random forest and feature contributions can also be used for probabilistic classification[mig, anna]. In a QSAR context mainly regression is used.

### 2.2 Decomposing a RF model with feature contributions

The applied methodology, forest floor, does not visualize directly the decision trees of the random forest. With hundreds or thousands of trees, it is intractable for a user to comprehend the overall structure of a trained RF model by inspecting the trees. Instead the model can be understood as the learned mapping function ( $f$ ), that maps from a feature space of molecular descriptors ( $X$ ) to a physicochemical target ( $\hat{y}$ ).  $X$  has as many dimensions as features in the model. The geometrical shape of the model mapping can neither be visualized nor comprehended directly, as the mapping is likely non-linear and high dimensional. Instead, projections or decompositions are needed to visualize the structure with only 2-3 dimensions. Feature contributions [kuzmin, anna] serve as a particular useful decomposition of the prediction for each descriptor, which assist to choose the optimal visualization of the model structure.

A random forest algorithm ( $g$ ) when trained on a data set of  $N$  solubility measurements  $y_i, i \in \{1, \dots, N\}$  and encoded molecular features ( $X_i$ ) adjusted with a set of parameters ( $\omega$ ) will yield a model fit ( $f$ ). This model fit maps from any point in feature space ( $X$ ) of molecular descriptors to a predicted solubility scale ( $\hat{y}$ ). This mapping can be understood as a high dimensional geometrical structure. A decomposition is used to visualize and navigate what model structure connects  $X$  and  $\hat{y}$  in 2D or 3D visualizations. The simplest and perhaps adequately correct decomposition splits the solubility prediction into separate effects with one unique function to explain each molecular descriptor.



**Figure x:** The Random forest algorithm is a function that when given a data set and training parameters will output a model fit. This model structure can as a start be interpreted as consisting of main effects only and visualized in 2D. Any deviation from a main effect only can be visualized as a 2<sup>nd</sup> order interaction.

[up to 2<sup>nd</sup> order interactions, include. Make dash line boxes to indicateconcept formula]

Hereby the model fit  $f$  can be simplified to series of additive functions  $f_1 + f_2 + \dots$ , which separately can be plotted in 2D. Feature contributions are used to estimate such additive functions and allows an isolated interpretation of each molecular descriptor.

### 2.3 Computation of feature contributions?

Every root-, intermediary- or terminal node of a decision tree is an individual prediction. When a parent node is split by a given variable, the daughter nodes will each receive some of the inbag samples and hereby construct two new predictions. A local increment is the change of node predictions from a parent node to a daughter node. For any sample, the RF prediction is simply the sum of all its encountered local increments divided by number of trees plus the grand mean of the training set. Feature contributions are constructed by the same local increments divided by number of trees, but feature contributions are summed separately for each sample by each variable. Thus a feature contribution can be understood as the average change of prediction for one sample molecule due to the information of one specific molecular descriptor - given all other molecular descriptors. *Given all*, means in practice that any interaction structure is preserved in the feature contributions.

When computing feature contributions for a training set, the yielded feature contributions can be arranged as a matrix with same dimensions as the molecular descriptor training matrix  $X_{ij}$ . Feature contributions can be denoted  $F_{ij}$ . Any prediction  $\hat{y}_i$  can be split into separate contributions attributed each of the molecular descriptors plus the grand mean of all solubility measurements ( $\bar{y}$ ).

$$\hat{y}_i = \sum_{j=1}^p F_{ij} + \bar{y}$$

To estimate the most accurate RF model structure it is most efficient to use any available training sample. To visualize the model structure it is also preferable to use all training predictions to compute feature contributions. Just as training predictions of a RF model can be out-of-bag cross-validated, so can feature contributions. Cross validated feature contributions yields fewer random ripples in the visualized

model structure. These random ripples arise from the inherent overfitting of individual decision trees. [forestFloor]

### 2.4 Plotting, quantifying goodness-of-visualization and identifying latent interactions.

The first way to plot feature contributions for a given molecular descriptor is as a function of the corresponding descriptor values, and this function can be fitted with an estimator. For this purpose, we suggest an estimator based on leave-one-out k-nearest neighbour Gaussian distance weighting, as it can fit most RF model structures and produces a fast cross-validation.

$$E(F_j, X_j) \rightarrow f_j(X_{ij}) = \hat{F}_{ij}$$

Hereby is obtained, a 2-axes plot of feature contributions (y-axis) versus the corresponding molecular descriptor values (x-axis) plus a fitted line describing the trending main effect not considering any interactions. See Figure 1 of the result section as an example. In Figure 1 the y-axis is feature contributions for any molecule in training set by a specific molecular descriptor (x-axis).

The fitted line may be an inadequate description, as a random forest model possibly may also have captured one or more interaction effects related to this molecular descriptor. The cross-validated explained variance of the feature contributions ( $R^2$ ) by the fitted estimator quantified how well the 2D visualization describes the descriptor effect as a main effect only.

$$R_{f_j}^2 = 1 - \frac{\sum_{i=1}^N (F_{ij} - \hat{F}_{ij})^2}{\sum_{i=1}^N (F_{ij})^2}$$

If the explained variance is e.g. only 50%, one may choose to find a better context to understand the feature contribution. A broader context can be plotted as a 3D plot where the feature contribution e.g. can be plotted as a function of the two descriptors, e.g. by the first and second descriptor

$j=1,2$ . Again the feature contributions can be fitted with an estimator and the goodness-of-fit can be quantified.

$$E(F_{t,(1,2)}, X_{t,(1,2)}) \rightarrow f_j(X_{t,(1,2)}) = \hat{F}_{ij}$$

In the 3D plot the estimated fit will no longer be a line but a surface, see the fitted surfaces in Figure 2. Unexplained variance of the estimated surface may remain; perhaps a 4D visualization is needed to explain an interaction between 3 molecular descriptors. Fortunately, we observe for random forest models in several data sets, that main effects tend to dominate over second order effects, which tend to dominate over higher order effects[cite me]. Thus, visualizing a model structure in 2D and 3D is likely adequate for most practical purposes.

Colour gradients can be used to provide one extra dimension. The molecule samples in a visualisation can be assigned to a colour gradient reflecting a latent variable to visually identify possible local or global interactions. A local interaction is understood as an interaction effect only learned in a smaller confined part of the model structure. A local interaction for a group of molecules can be highlighted with a colour pattern. In Figure 4 in result section such a highlighting is used to visualize the local model structure for 57 polychlorinated bi-phenyl molecules .

## 2.5 Software implementations

All visualizations in this article were produced with the R package forestFloor (1.8.9) [cite forestFloor cran]. The supplementary file of this article contains scripts to reproduce the model and visualizations of this paper. The forestFloor package depends on the rgl[Duncan, version] package to produce 3D visualizations, the kknn[cite] package for function estimators and the Rcpp [eddelbuettel, versoion] package to integrate functions implemented in C++ with the R environment. The RF models were trained with randomForest packae [liaw, version]. All packages are available from the CRAN repository [cite cran].

## 2.6 Data set and molecular descriptors

A public data set by Huuskonen *et al*[3] was chosen because it is well cited and as it has been reused in many other datasets [palmer,Delaney,bergstrøm,wiisinger,Laura]. Training set and test set were merged in to on single data set

of 1256 molecules. SMILES were imported to the software with the application MOE [cite] and sequentially pre-processed with the following functions: 'wash' (simulating an ideal solubilised molecular form), 'partial charges MMFFA96x' calculating the electron densities necessary for a number of descriptor algorithms, and finally 'energy minimize' relaxing the molecule in the minimum 3D state as suggested by [palmer]. To limit the scope of this article, only a small selection of 12 common and useful descriptors identified by Palmer *et al*[palmer] were used. The full data set with descriptors is provided in supplementary materials.

## 3 Results

### 3.1 Visualising main effects

A default random forest model of 2000 trees and mtry=4 was trained on the data set. Mean test error of 20 repeated 10-fold was  $0.636(+/-0.004)$  sd? and  $r^2_s = .903(+/-0.001)$ . Which was a similar performance as [palmer, huskonnen, laura?, hou?]. With the default RF model, out-of-bag feature contributions for every molecule were plotted as a function of the respective features/descriptors (main effect plots). *SlogP* was the most important descriptor by variable permutation importance and plotted first in upper left corner, followed by other descriptors in a decreasing order. A negative linear relation with solubility contribution was observed. High *SlogP* yielded negative contribution to solubility. A flattening of the main effect curve was observed in both ends. Fitted lines and calculated explained variance hereof described how well each molecular descriptor could be regarded as a main effect. The explained feature contribution variance by fitted main effect lines ranged from 90%-87% for the first 7 most important descriptors. Hereafter declined the explained main effect to range .71 to .48 explained variance. And the least important descriptor was only explained 15% as a main effect. Hence, the latter 5 variables were poorly described as main effects, where at the same time less influential for the model prediction deemed on the variable importance and as seen in Figure 1 the absence of feature contribution variance. Thus, overall to visualize the entire random forest model fit as strictly additive explained by the sum 12 main effect estimators explained 89% of the cross validated predictions. Thus to view these descriptors as contributing individually additively to the prediction of solubility would be a fair generalization of this particular instance of a random forest model fit.

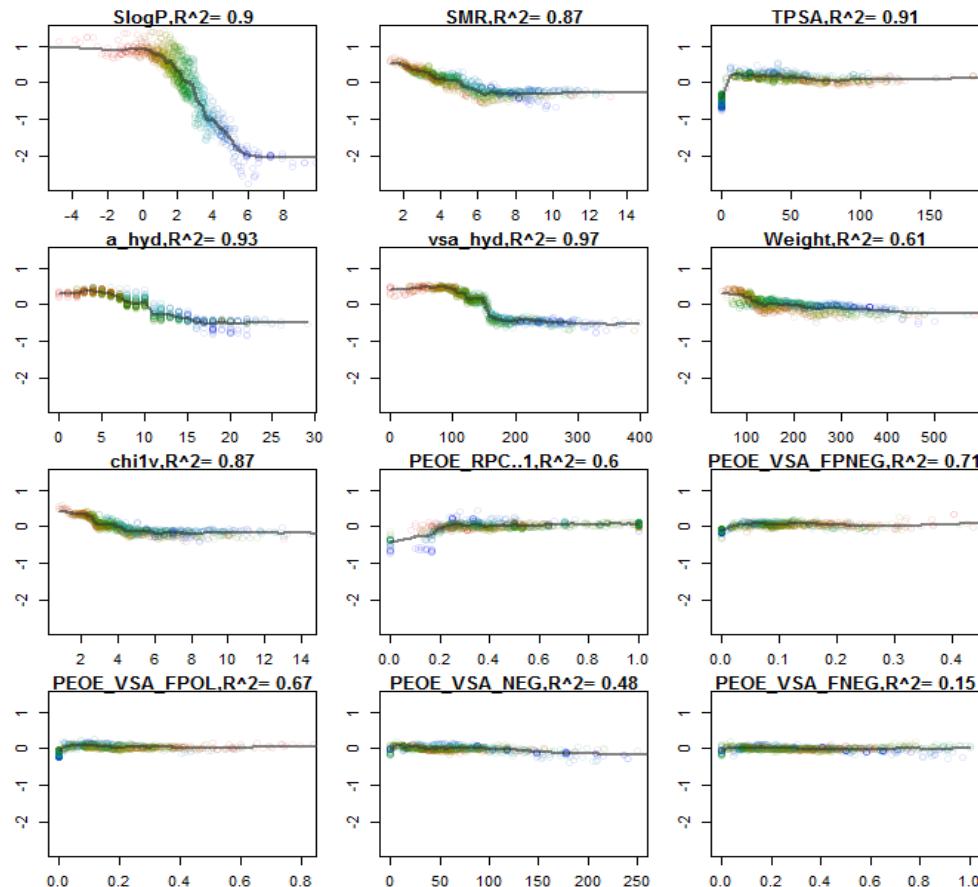


Figure 1. Main-effect illustration of the 12 descriptors ordered by variable importance. Each molecule is represented once in each plot as a point of a specific colour. Point colour is by defined *SlogP* descriptor value of each molecule, corresponding to horizontal colour gradient in *SlogP* plots. A horizontal/diagonal gradient indicates local interactions with *SlogP*. Black lines +  $R^2$  values are estimated fits, a strictly non-interaction interpretation of molecular descriptor effect, as described in equation 1.

### 3.2 Identifying and visualizing interactions

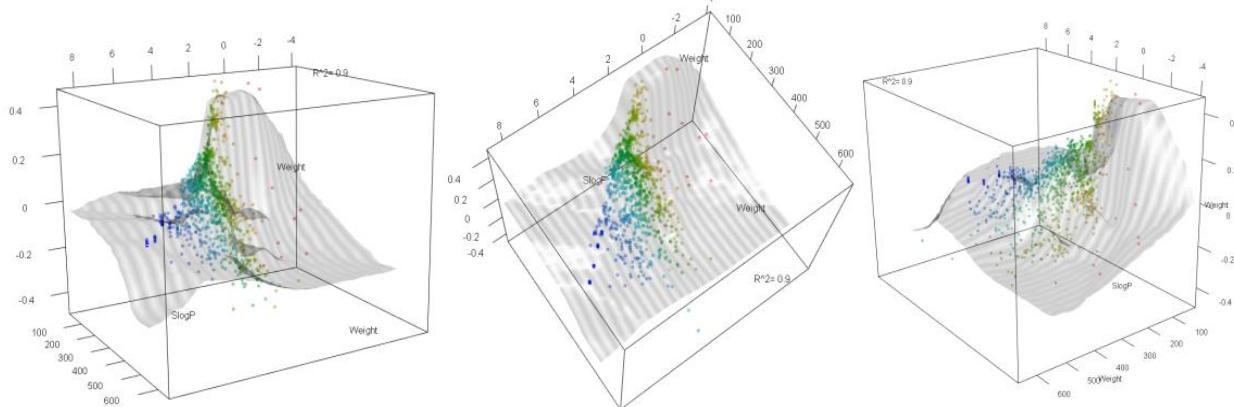
To step beyond a strictly main effect interpretation, interaction effects must be identified. A colour gradient (red-yellow-green-teal-blue) horizontally aligned with the *SlogP* axis was used to characterize *SlogP* value of molecules in all other plots. Each molecule will have the exact same colour in all plots. Correlations and interactions with *SlogP* were visually highlighted with this colour gradient. Molecular descriptors correlating with *SlogP*, reproduced the colour gradient horizontally as observed for *SMR*, *PEOE\_VSA\_NEG*, *vsa\_hyd*, *a\_hyd*, *Weight*, *chi1v*). Other descriptors *TPSA* and *PEOE\_VSA\_FPOL* showed a reversed horizontal colour gradient as these descriptors negatively correlated with *SlogP* within the data set  $R_p \sim 0.5$ . For all descriptors, deviations from fitted main effect lines were observed. Thus, the variance of each individual feature contribution could not entirely be explained by the descriptor alone. Molecules with specific *SlogP* values indicated by colour gradient were observed to deviate from the fitted lines in specific patterns. Hence, such deviations from a pure main

effect could be explained by the many upstream decision splits by the *SlogP* or other correlated descriptors. In Figure 1 a low *Weight*(<120 Dalton) was attributed to a positive contribution to solubility, only when *SlogP*<1.5 (red/yellow). Molecules with high (*SlogP*>4, blue) had a feature contribution near zero for any molecular weight. Only 61% of the feature contribution variance of *Weight* was explained by the fitted main effect line. The remaining variance was thus attributed to interactions, such as the interaction with *SlogP* identified with the colour gradient. *Weight* was a descriptor with medium importance, yet poorly explained as a main effect. Hence, it was found as needed to elucidate the model contribution of *Weight* further. Figure 2 depicts in 3D the feature contributions of *Weight* for every molecule plotted by *Weight* and *SlogP*.

Again the interaction effect between *SlogP* and *Weight* could be observed. The fitted surface, explains the contribution of *Weight* (z-axis) as a main effect by *Weight* (x-axis) itself and as an interaction by *SlogP* (y-axis). This fit increased the explained feature contribution variance to 90%. In figure 2, it

was observed that there were no examples of molecules with low *Weight* and low *SlogP*. Thus this part the RF model structure is extrapolated and the model structure is less likely to be predictive for any such molecule. That the boiling point

of small apolar molecules (e.g. propane, halothane etc.) is far below room temperature likely explains no such learning examples exist.



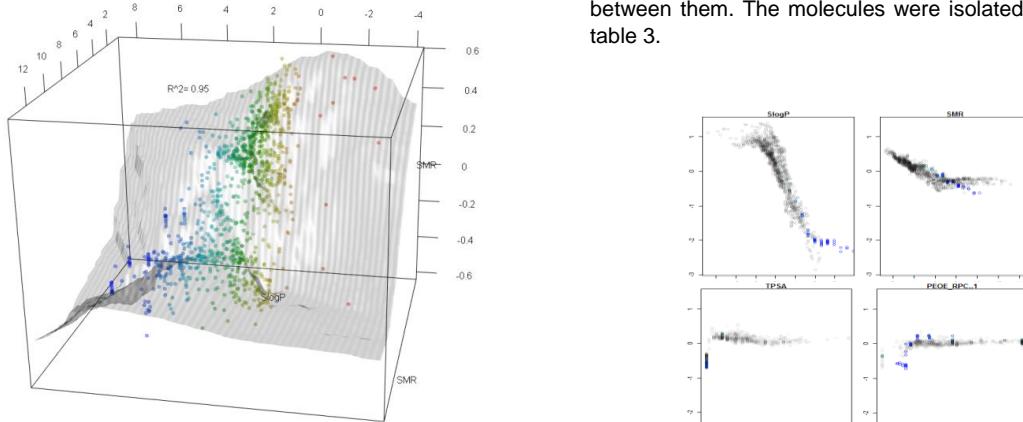
**Figure 2.** Feature contributions of *Weight* (z-axis) versus feature values *Weight* (X-axis) and feature values *SlogP* (Y-axis). Surface visualizes the fitted estimator, which describes 90% of the variance. Colour gradient parallel to *SlogP* axis as in figure 1. Image visualizes an interaction where *Weight* contributes most to solubility prediction when *SlogP* is negative.

The *SMR* feature was the second most important feature. The main effect of *SMR* feature contribution (molecular refraction by atom contributions) was explained 87%. When viewed as an interaction with *logS*, 95% of the feature contribution variance was explained. *SMR* is intended to approximate the polarizability of molecules, such that these e.g. can form induced dipoles in polar solvents and obtain an energy favourable charged interaction with water[cite]. Such an effect may have been anticipated to contribute in general positively to solubility, but in fact as main effect *SMR* contribute negatively to solubility. As molar refractivity is the ‘polarizability per molecule’, this measure was highly correlated with *Weight* ( $r_p = .93$ ). If the *SMR* feature was divided by *Weight* and the RF model was refitted. The *SMR* feature dropped to the 11<sup>th</sup> most important feature and the main effect was flat.

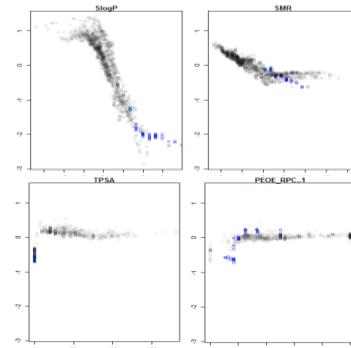
### 3.2 Identifying local effects

In main effect plot figure 1, a distinct group of 57 molecules with low *logS* showed distinct interactions in *SlogP*, *SMR*, *TPSA* and *PEOE\_RPC..1*. The group of molecules can be identified in figure 2 middle plot, as having a perfect linear relationship between *logS* and *SMR* ( $r_p=1$ ). In figure 4, the position of these molecules in the model structure was highlighted by colouring any other molecule black. The observed interactions was for *SlogP* a flattening of the negative contribution to solubility of molecules with *SlogP* above 5 whereas ~15 non PCP molecule with *SlogP* >5 were predicted decreasing soluble as a function of *SlogP*. For *SMR* a linear reduction in solubility as contrary to the general main effect.

Furthermore these molecules were not only on a line but only placed on 10 different steps with equal distance between them. The molecules were isolated and showed in table 3.



**Figure 3.** Interaction plot of *SMR* feature contribution as function *SMR* and *SlogP*. This fitted estimator describes 95% of variance of the feature contributions of *SMR*.



**Figure 4.** Highlighted feature contributions of PCB molecules.

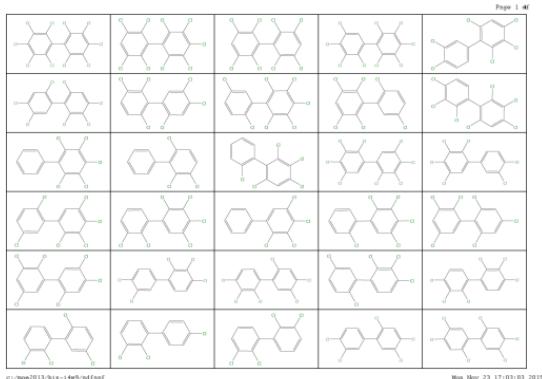


Table 1: Depiction of 36 PCB molecules. What kind of table would be fine here?

It showed that all molecules were polychlorinated biphenyl compounds (PCB). As both *SlogP* and *SMR* are defined by atomic contributions, all PCB with the same amount of

substituted chloride atoms will have same *SlogP* and *SMR* values. In fact only two features *chi1v* and *PEOE\_RPC..1* produced unique feature values for PCB's with same amount of chloride atoms. But these differences in values were minute, and they were more likely to arise from a non-deterministic convergence algorithm estimating the partial charges[cite method]. Also there appeared to be no obvious relationship between these two features and solubility beyond the number of chloride atoms. Moreover the random forest model fit did not seem to capture any relationship related to substitution pattern, as the OOB cross-validated predictions for these PCB with equal amounts of chloride atoms did not correlate with the actual solubility. Predictions ranged only 0.12 logS units for PCB with same amount of chloride atoms, where the predictions ranged 1.3 logS units. Thus the random forest model was unable with the 12 selected features to predict the relationship between PCB substitution pattern and solubility.

### 3.2 Model structure is affected by training parameters

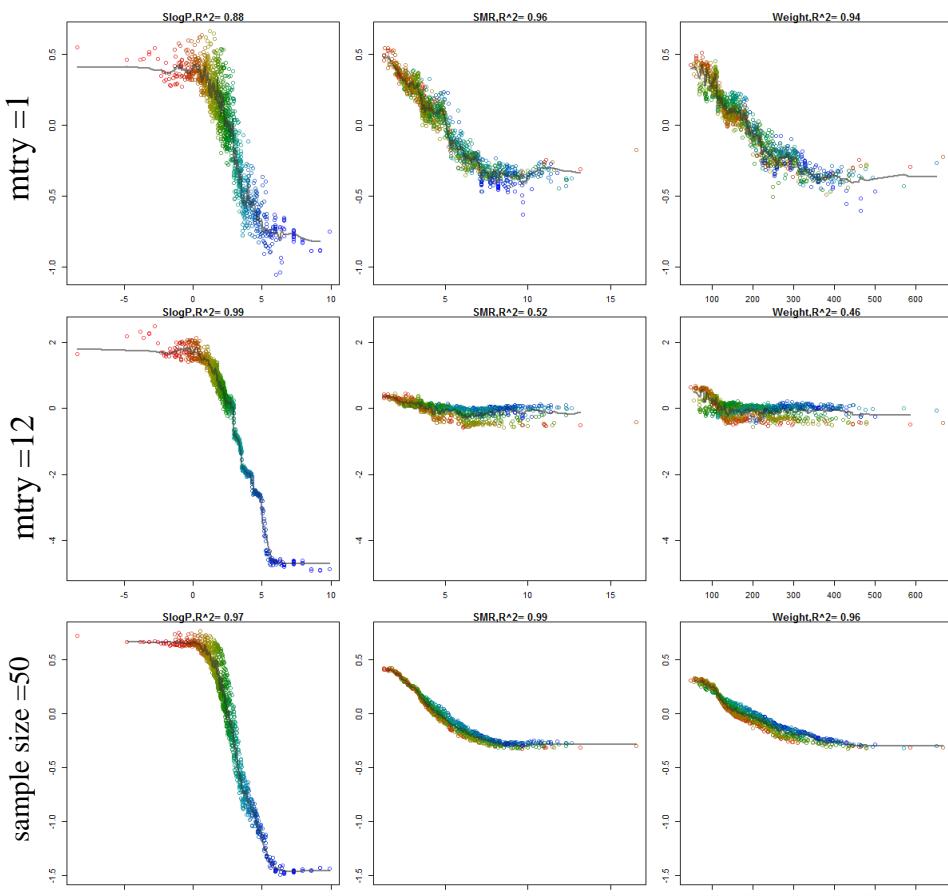


Figure 5: Model structure varies with the parameters. Low *mtry*(1), uniform use of features. All variables have main effect and interaction effects. High *mtry*(12,all), algorithm will greedily use best feature first, other features are mainly used for interaction effects. Low sample size(50), smoothens model structure, interactions reduced, model approaches strictly additive model. Sample size is by default 1250 and *mtry* is by default 4.

## Discussion [unfinished]

Choosing a correct set of low dimensional visualizations to account for a complex model structure is not necessarily fully attainable [friedman]. Forest floor can identify and quantify the residuals of any visualization, such that the depiction of the RF model structure can iteratively be elaborated until a sufficiently correct depiction has been attained. Any high dimensional structure cannot be visualized in two or three dimension. In a regression context, a main effect requires 2 dimensions, a 2<sup>nd</sup> order (interaction) effect requires 3 dimensions and a 3<sup>rd</sup> order requires 4 dimensions. That said it is possible to understand the 3D structure of a DNA helix from a 2D drawing, and likewise the 4D Kleinbottle structure in form a 3D representation. RF is a relatively shallow model and 3<sup>rd</sup> or higher order interactions or seems almost absent.

This presented methodology of decomposing effects by descriptors, estimating main effects and interactions effects is one representation of the model structure. Another representation such as the actual trained ensemble of decision trees is concise but is too complex to lend itself to a clear interpretation. Another representation, such partial dependence plots can e.g. in 3D describe an interaction effect between two variables. But classic PD plots are not guaranteed to well generalize the overall high dimensional structure, nor do they point to the location of potential sizeable latent interactions. Thus, the forest floor is a methodology that provides the investigator means to browse the model structure of a random forest model and quantify how well a given low dimensional representation, as a series of visualisations, describe the overall structure.

### With

#### 3.1 discussion of other methods

With another method to visualize a mapping such partial dependence plots, to uncover hidden interactions and avoid to extrapolation is more difficult.

Today, mainly variable importance [palmer, laura, others] is used in conjunction with random forest models to interpret the model. Variable importance describes the loss of cross-validated predictive performance when each variable in turns

were permuted. VI only approximates the usefulness of each molecular descriptor. VI does not outline how each descriptor is used by the model.

[insert in result section] A group of PCB molecules were identified as to elicit a distinctive interaction pattern. With the 12 selected molecular descriptors, was the chloride substitution pattern of this PCB molecules not learned. *SlogP* and *SMR* the most important descriptors are e.g. themselves based predictions on *logP* and molar refractivity for 10.000 measured molecules. Predictions are based summing empirical derived scores for each atom in the molecule. Atoms are categorized by atom number and type bonding to neighbouring atoms. Thus for PCB molecules having the same number of substituted chloride atoms all scores will be exactly alike. [maybe two extra sentences of why neither other descriptors has any clue of this effect.] Ghavami et al. [6] produced a regression model only to predict solubility of PCB molecules and found that 90 percent of the variance of PCB log solubility can be attributed the number chloride atoms in a linear regression model. Introducing counts of ortho-, meta- and para configuration contributed to explain up to 97% cross-validated variance of the log solubility of PCB molecules. As the PCB molecules collapse to only extending a string of connecting points in the feature space, where each point consist of PCB molecules with same amount chloride atoms, the sampling density around these PCB molecules is high. Thus, is the random forest model able to fit a very specific structure accounting for the solubility variance related to chloride atoms in PCB molecules. If predicting the solubility of a random molecule, it would be unlikely to fall within the small sub feature space of PCB molecules. If it did fall within this subspace, the learned relationship from PCB's would dominate the prediction of the RF model.

First Main Text Paragraph----without indentation.  
Main Text Paragraph----with indentation.

((Insert schemes above the captions. Note: Please do not combine scheme and caption in a textbox or frame))

**Scheme 1.** Scheme Caption.

Main Text Paragraph---with indentation.

((Insert figures above the captions. Note: Please do not combine figure and caption in a textbox or frame))

**Figure 1.** Figure Caption.

Main Text Paragraph---with indentation.

**Table 1.** Table Caption. ((Note: Please do not include the table in a textbox or frame))

Head 1 <sup>a</sup>	Head 2	Head 3 <sup>b</sup>	Head 4 <sup>c</sup>	Head 5
Column 1	Column 2	Column 3	Column 4	Column 5
Column 1	Column 2	Column 3	Column 4	Column 5

<sup>a</sup> Table Footnote.

<sup>b</sup> ...

### 3 Conclusions

First Main Text Paragraph---without indentation.

Main Text Paragraph---with indentation.

### Acknowledgements

Acknowledgements Text.

### References

- [1] ((Reference 1, Example for Journals))a) A. Author, B. Coauthor, *Mol. Inf.* **2009**, 1, 1-10; b) A. Author, B. Coauthor, *Angew. Chem.* **2006**, 118, 1-5; *Angew. Chem. Int. Ed.* **2006**, 45, 1-5.
- [2] ((Reference 2, Example for Books))J. W. Grate, G. C. Frye, in *Sensors Update*, Vol. 2 (Eds: H. Baltes, W. Göpel, J. Hesse), Wiley-VCH, Weinheim **1996**, pp. 10-20.)
- [3]

[1]  
ESOL: Estimating Aqueous Solubility Directly from Molecular Structure  
John S. Delaney\*  
Syngenta, Jealott's Hill International Research Centre,  
Bracknell, Berkshire, RG42 6EY, United Kingdom  
Received October 29, 2003

[2]  
Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules

Christel A. S. Bergström ,† Carola M. Wassvik ,† Ulf Norinder,‡ Kristina Luthman,§\* and Per Artursson †  
Center for Pharmaceutical Informatics, Department of Pharmacy, Uppsala University, Uppsala Biomedical Center, P.O. Box 580, SE-751 23 Uppsala, Sweden, Department of Medicinal Chemistry, AstraZeneca R&D, SE-151 85 Södertälje, Sweden, and Department of Chemistry, Medicinal Chemistry, Göteborg University, SE-412 96 Göteborg, Sweden  
*J. Chem. Inf. Comput. Sci.*, 2004, 44 (4), pp 1477–1488  
DOI: 10.1021/ci049909h  
Publication Date (Web): June 23, 2004  
Copyright © 2004 American Chemical Society

[3]  
Random Forest Models To Predict Aqueous Solubility  
David S. Palmer , Noel M. O'Boyle ,† Robert C. Glen , and John B. O. Mitchell \*  
Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom  
*J. Chem. Inf. Model.*, 2007, 47 (1), pp 150–158  
DOI: 10.1021/ci060164k  
Publication Date (Web): December 2, 2006  
Copyright © 2007 American Chemical Society

[4] Wildman, S.A., Crippen, G.M.; Prediction of Physiochemical Parameters by Atomic Contributions; *J. Chem. Inf. Comput. Sci.* 39 No. 5 (1999) 868–873.  
[6] Ertl, P., Rohde, B., Selzer, P.; Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties; *J. Med. Chem.* 43 (2000) 3714–3717.  
[7] [Cruciani 2000]Cruciani, G., Crivori, P., Carrupt, P.-A., Testa, B.; Molecular Fields in Quantitative Structure-Permeation Relationships: the VolSurf Approach; *J. Mol. Struct. (Theochem)* 503 (2000) 17–30.  
[8] Gasteiger, J., Marsili, M.; Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges; *Tetrahedron* 36 (1980) 3219.,

[9] Prediction of drug solubility from structure. William L. Jorgenson, , , Erin M. Duffyb . doi:10.1016/S0169-409X(02)00008-X  
[10] Yu, L. X.; Amidon, G. L.; Polli, J. E.; Zhao, H.; Mehta, M. U.; Conner, D. P.; Shah, V. P.; Lesko, L. J.; Chen, M.-L.; Lee, V. H. Biopharmaceutics classification system: the scientific basis for biowaiver extensions. *Pharm. Res.* 2002, 19 (7), 921–925  
[11] OVERVIEW OF DATA AND CONCEPTUAL APPROACHES FOR DERIVATION OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS FOR ECOTOXICOLOGICAL EFFECTS OF ORGANIC CHEMICALS STEVEN P. BRADBURY,† CHRISTINE L. RUSSOM,\*† GERALD T. ANKLEY,† T. WAYNE SCHULTZ,‡ and JOHN D. WALKER§  
†U.S. Environmental Protection Agency, National Health and Environmental Effect

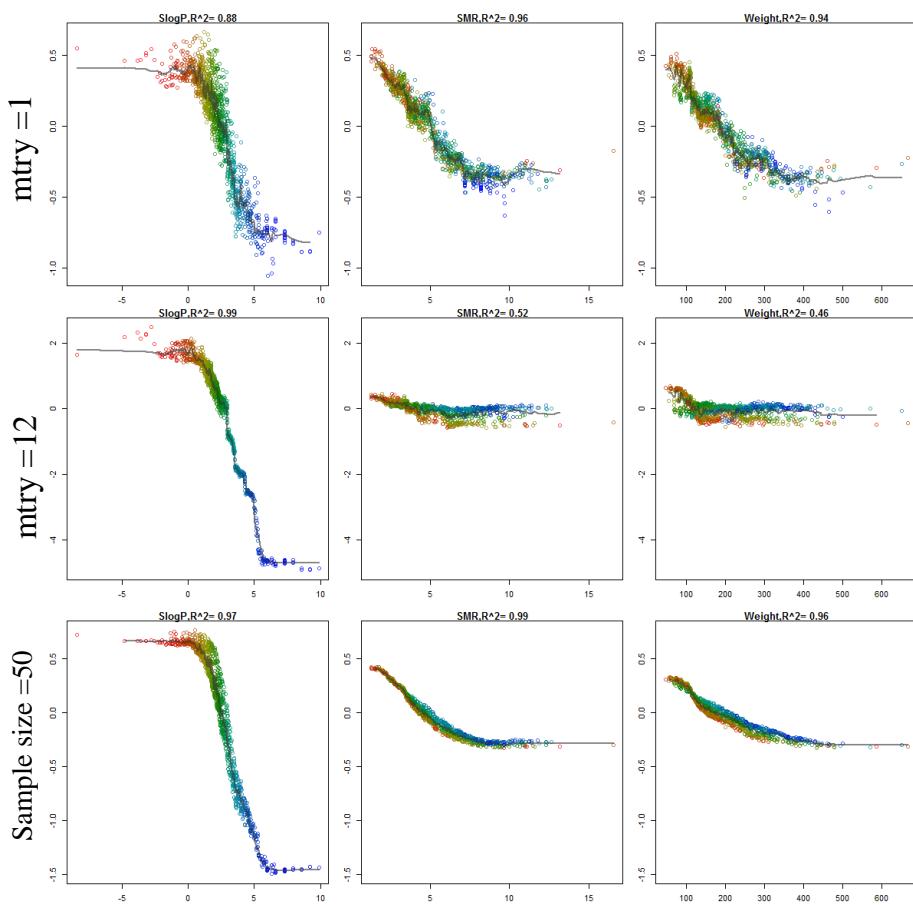
Received: ((will be filled in by the editorial staff))

Accepted: ((will be filled in by the editorial staff))

Published online: ((will be filled in by the editorial staff))

t showed that all molecules were polychlorinated biphenyl compounds. As both SlogP and SMR are computed by atomic contributions, all PCB with the same amount of substituted chloride atoms will have same SlogP and SMR values. In fact only two features chi1v and PEOE\_RPC..1 produced different feature values for PCB's with same amount of chloride atoms. But these differences in values were minute, and they were more likely to arise from a non-deterministic convergence algorithm defining the partial charges. Also there appeared to no relationship between these features, chloirde substitution configuration and actual logS. More over the OOB cross-validated predictions for these PCB varied only 0.12 units where the average variation with PCB with equal many chlorides was 1.3. And this OOB cross validated variation within PCB with equal amounts of chloride did not correlate with actual solubility. Thus the random forest model was unable with the 12 selected features to predict the relationship between PCB substitution configuration and solubility. 90 Percent of the variance of PCB solubility can be attributed the number chloride atoms or any other Ghavami et al. [6] showed how counting ortho meta and para configuration contribute to explain upto 97% cross-validated variance.

---



# CHAPTER 7

# Heading on Level 0 (chapter)

---

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

## 7.1 Heading on Level 1 (section)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

### 7.1.1 Heading on Level 2 (subsection)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

### 7.1.1.1 Heading on Level 3 (subsubsection)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

**Heading on Level 4 (paragraph)** Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

## 7.2 Lists

### 7.2.1 Example for list (itemize)

- First item in a list
- Second item in a list
- Third item in a list
- Fourth item in a list
- Fifth item in a list

### 7.2.1.1 Example for list (4\*itemize)

- First item in a list
  - First item in a list
    - \* First item in a list
      - First item in a list
      - Second item in a list
    - \* Second item in a list

- Second item in a list
- Second item in a list

## 7.2.2 Example for list (enumerate)

1. First item in a list
2. Second item in a list
3. Third item in a list
4. Fourth item in a list
5. Fifth item in a list

### 7.2.2.1 Example for list (4\*enumerate)

1. First item in a list
    - a) First item in a list
      - i. First item in a list
      - A. First item in a list
      - B. Second item in a list
    - ii. Second item in a list
  - b) Second item in a list
2. Second item in a list

## 7.2.3 Example for list (description)

**First** item in a list

**Second** item in a list

**Third** item in a list

**Fourth** item in a list

**Fifth** item in a list

### 7.2.3.1 Example for list (4\*description)

**First** item in a list

**First** item in a list

**First** item in a list

**Second** item in a list

# CHAPTER 8

## Conclusion

---

Morbi pharetra ligula integer mollis mi nec neque ultrices vitae volutpat leo ullamcorper. In at tellus magna. Curabitur quis posuere purus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Suspendisse tristique placerat feugiat. Aliquam vitae est at enim auctor ultrices eleifend a urna. Donec non tincidunt felis. Maecenas at suscipit orci.



# APPENDIX A

## An Appendix

---

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

# Package ‘forestFloor’

June 1, 2016

**Type** Package

**Title** Visualizes Random Forests with Feature Contributions

**Version** 1.9.5

**Date** 2016-06-01

**Author** Soeren Havelund Welling

**Maintainer** Soeren Havelund Welling <SOWE@DTU.DK>

**Depends**

**Suggests** randomForest, utils, devtools, tools

**Description** Form visualizations of high dimensional mapping structures of random forests and feature contributions.

**SystemRequirements** OpenGL, GLU Library, zlib

**License** GPL-2

**URL** <http://forestFloor.dk>

**Imports** Rcpp (>= 0.11.3), rgl, kknn

**LinkingTo** Rcpp

**NeedsCompilation** yes

**Repository** CRAN

**Date/Publication** 2016-06-01 14:11:11

## R topics documented:

forestFloor-package . . . . .	2
append.overwrite.alists . . . . .	2
as.numeric.factor . . . . .	3
box.outliers . . . . .	4
convolute_ff . . . . .	5
convolute_ff2 . . . . .	7
convolute_grid . . . . .	9
fcol . . . . .	12
forestFloor . . . . .	16

2

*append.overwrite.alists*

plot.forestFloor . . . . .	21
plot_simplex3 . . . . .	25
print.forestFloor . . . . .	28
recTree . . . . .	29
show3d . . . . .	30
vec.plot . . . . .	34
Xtestmerger . . . . .	36

**Index****38**

---

**forestFloor-package**      *forestFloor: visualize the random forest model structure*

---

**Description**

forestFloor visualizes randomForests models(RF). Package enables users to understand a non-linear, regression problem or a binary classification problem through RF. Any model can be separated into a series of main effect and interactions with the concept of feature contributions.

**Details**

Package: forestFloor  
Type: Package  
Version: 1.9  
Date: 2015-12-25  
License: GPL-2

**Author(s)**

Soren Havelund Welling

**References**

Interpretation of QSAR Models Based on Random Forest Methods, <http://dx.doi.org/10.1002/minf.201000173>  
Interpreting random forest classification models using a feature contribution method, <http://arxiv.org/abs/1312.1121>

---

*append.overwrite.alists*

*Combine two argument lists*

---

`as.numeric.factor`

3

**Description**

First argument list is master, second list slave

**Usage**

```
append.overwrite.alists(masterArgs,slaveArgs)
```

**Arguments**

- |                         |  |
|-------------------------|--|
| <code>masterArgs</code> | List of arguments, of which will stay unchanged  |
| <code>slaveArgs</code>  | List of arguments, conflicts with <code>masterArgs</code> will be deleted. Additional args will be appended. |
|                         | <code>s</code>   |

**Details**

This function combines to lists of arguments. Conflicts will be resolved by `masterArgs`.

**Value**

List of arguments, being `masterArgs` appended by `slaveArgs`

**Author(s)**

Soren Havelund Welling

**Examples**

```
arglist1 = alist(monkey="happy",telephone.no=53)
arglist2 = alist(monkey="sad",house.no=12)

#this should yield a alist(monkey="happy", telephone.no=53, house.no=12)
forestFloor:::append.overwrite.alists(arglist1,arglist2)
```

`as.numeric.factor`      *Convert a factor to numeric.vector.*

**Description**

Internal function which will drop unused levels and convert remaining to a number from 1 to `n.levels`.

**Usage**

```
as.numeric.factor(x,drop.levels=TRUE)
```

4

*box.outliers***Arguments**

- x**              Normally a factor, can be a numeric vector(will be output unchanged)  
**drop.levels**    Boolean, should unused levels be dropped?

**Details**

Simple internal function, used to direct categorical variables to a 1 dimensional scale.

**Value**

A vector of same length, where each category/level is replaced with number from 1 to n

**Author(s)**

Soren Havelund Welling

**Examples**

```
as.numeric.factor = forestFloor:::as.numeric.factor #import to environment
some.factor = factor(c("dog","cat","monkey")[c(1,3,2,1,3,2,1,1)]) #make factor
a.numeric.vector = as.numeric.factor(some.factor) #convert factor representation.
```

box.outliers

*Box Outliers***Description**

Squeeze all outliers onto standard.dev-limits and/or normalize to [0;1] scale

**Usage**

```
box.outliers(x, limit = 1.5, normalize = TRUE)
```

**Arguments**

- x**              numeric vector, matrix, array, data.frame  
**limit**            limit(SD,standard deviation) any number deviating more than limit from mean  
                      is an outlier  
**normalize**        TRUE/FALSE should output range be normalized to [0;1]?

**Details**

Can be used to squeeze high dimensional data into a box, hence the name box.outliers. Box.outliers is used internally in forestFloor-package to compute colour gradients without assigning unique colours to few outliers. It's a box because the borders uni-variate/non-interacting.

**convolute\_ff**

5

**Value**

matrix(n x p) of normalized values

**Author(s)**

Soren Havelund Welling, 2014

**See Also**

scale()

**Examples**

```

box.outliers = function (x, limit = 1.5) {
  x = scale(x)
  x[x > limit] = limit
  x[-x > limit] = -limit
  x = x - min(x)
  x = x/(limit * 2)
  return(x)
}
n=1000 #some observations
p = 5 #some dimensions
X = data.frame(replicate(p,rnorm(n))) # a dataset
Xboxed =box.outliers(X,limit=1.5) #applying normalization
plot(Xboxed[,1],Xboxed[,2],col="#00000088") #plot output for first two dimensions

```

**convolute\_ff***Cross-validated main effects interpretation for all feature contributions.***Description**

convolute\_ff estimates feature contributions of each feature separately as a function of the corresponding variable/feature. The estimator is a k-nearest neighbor function with Gaussian distance weighting and LOO cross-validation see [train.kknn](#).

**Usage**

```

convolute_ff(ff,
             these.vars=NULL,
             k.fun=function() round(sqrt(n.obs)/2),
             userArgs.kknn = alist(kernel="gaussian"))

```

6

*convolute\_ff***Arguments**

<code>ff</code>	forestFloor object "forestFloor_regression" or "forestFloor_multiClass" consisting of at least <code>ff\$X</code> and <code>ff\$FCmatrix</code> with two matrices of equal size
<code>these.vars</code>	vector of col.indices to <code>ff\$X</code> . Convolution can be limited to these.vars
<code>k.fun</code>	function to define k-neighbors to consider. <code>n.obs</code> is a constant as number of observations in <code>ff\$X</code> . Hereby k neighbors is defined as a function <code>k.fun</code> of <code>n.obs</code> . To set k to a constant use e.g. <code>k.fun = function() 10</code> . <code>k</code> can also be overridden with <code>userArgs.kknn = alist(kernel="Gaussian",kmax=10)</code> .
<code>userArgs.kknn</code>	argument list to pass to <code>train.kknn</code> function for each convolution. See ( <a href="#">link</a> ) <code>kknn.args</code> . Conflicting arguments to this list will be overridden e.g. <code>k.fun</code> .

**Details**

`convolute_ff` uses `train.kknn` from `kknn` package to estimate feature contributions by their corresponding variables. The output inside a `ff$FCfit` will have same dimensions as `ff$FCmatrix` and the values will match quite well if the learned model structure is relative smooth and main effects are dominant. This function is e.g. used to estimate fitted lines in `plot.forestFloor` function "`plot(ff,...)`". LOO cross validation is used to quantify how much of feature contribution variation can be explained as a main effect.

**Value**

`ff$FCfit` a matrix of predicted feature contributions has same dimension as `ff$FCmatrix`. The output is appended to the input "forestFloor" object as `$FCfit`.

**Author(s)**

Soren Havelund Welling

**Examples**

```
## Not run:
library(forestFloor)
library(randomForest)

#simulate data
obs=1000
vars = 6
X = data.frame(replicate(vars,rnorm(obs)))
Y = with(X, X1^2 + 2*sin(X2*pi) + 8 * X3 * X4)
Yerror = 5 * rnorm(obs)
cor(Y,Y+Yerror)^2
Y= Y+Yerror

#grow a forest, remeber to include inbag
rfo=randomForest(X,Y,keep.inbag=TRUE)

ff = forestFloor(rfo,X)
```

`convolute_ff2`

7

```
ff = convolute_ff2(ff) #return input object with ff$FCfit included

#the convolutions correlation to the feature contribution
for(i in 1:6) print(cor(ff$FCmatrix[,i],ff$FCfit[,i])^2)

#plotting the feature contributions
pars=par(no.readonly=TRUE) #save graphicals
par(mfrow=c(3,2),mar=c(2,2,2,2))
for(i in 1:6) {
  plot(ff$X[,i],ff$FCmatrix[,i],col="#00000030",ylim=range(ff$FCmatrix))
  points(ff$X[,i],ff$FCfit[,i],col="red",cex=0.2)
}
par(pars) #restore graphicals

## End(Not run)
```

`convolute_ff2`

*Low-level function to estimate a specific set of feature contributions by corresponding features with kknn-package. Used to estimate goodness-of-fit of surface in show3d.*

### Description

Low-level function to estimate a selected combination feature contributions as function of selected features with leave-one-out k-nearest neighbor.

### Usage

```
convolute_ff2(
  ff,
  Xi,
  FCi = NULL,
  k.fun=function() round(sqrt(n.obs)/2),
  userArgs.kknn = alist(kernel="gaussian") )
```

### Arguments

<code>ff</code>	forestFloor object class "forestFloor_regression" or "forestFloor_multiClass" consisting of at least ff\$X and ff\$FCmatrix with two matrices of equal size
<code>Xi</code>	integer vector, of column indices of ff\$X to estimate by.
<code>FCi</code>	integer vector, column indices of features contributions in ff\$FCmatrix to estimate. If more than one , these columns will be summed by samples/rows. If NULL then FCi will match Xi.
<code>k.fun</code>	function to define k-neighbors to consider. n.obs is a constant as number of observations in ff\$X. Hereby k neighbors is defined as a function k.fun of n.obs. To set k to a constant use e.g. k.fun = function() 10. k can also be overridden with userArgs.kknn = alist(kernel="Gaussian",kmax=10).
<code>userArgs.kknn</code>	argument list passed to train.kknn function for each convolution, see <a href="#">train.kknn</a> . Arguments in this list have priority of any arguments passed by default by this wrapper function. See argument merger <a href="#">train.kknn</a>

8

*convolute\_ff2***Details**

*convolute\_ff2* is a wrapper of `train.kknn` to estimate feature contributions by a set of features. This function is e.g. used to estimate the visualized surface layer in `show3d` function. LOO CV is used to quantify how much of a feature contribution variation can be explained by a given surface. Can in theory also be used to quantify higher dimensional interaction effects, but randomForest do not learn much 3rd order (or higher) interactions. Do not support orderByImportance, thus  $X_i$  and  $FC_i$  points to column order of training matrix  $X$ .

**Value**

an numeric vector with one estimated feature contribution for any observation

**Author(s)**

Soren Havelund Welling

**Examples**

```
## Not run:
library(forestFloor)
library(randomForest)
library(rgl)
#simulate data
obs=2500
vars = 6
X = data.frame(replicate(vars,rnorm(obs)))
Y = with(X, X1^2 + 2*sin(X2*pi) + 8 * X3 * X4)
Yerror = 15 * rnorm(obs)
cor(Y,Yerror)^2 #relatively noisy system
Y= Y+Yerror

#grow a forest, remeber to include inbag
rfo=randomForest(X,Y,keep.inbag=TRUE,ntree=1000,sampsize=800)

#obtain
ff = forestFloor(rfo,X)

#convolute the interacting feature contributions by their feature to understand relationship
fc34_convolved = convolute_ff2(ff,Xi=3:4,FCi=3:4, #arguments for the wrapper
                                userArgs.kknn = alist(kernel="gaussian",k=25)) #arguments for train.kknn

#plot the joined convolution
plot3d(ff$X[,3],ff$X[,4],fc34_convolved,
       main="convolution of two feature contributions by their own variables",
       #add some colour gradients to ease visualization
       #box.outliers squeeze all observations in a 2 std.dev box
       #univariately for a vector or matrix and normalize to [0;1]
       col=rgb(.7*box.outliers(fc34_convolved),
              .7*box.outliers(ff$X[,3]),
              .7*box.outliers(ff$X[,4]))
)
```

*convolute\_grid*

9

```
## End(Not run)
```

---

*convolute\_grid*

*Model structure grid estimated by feature contributions*

---

### Description

Low-level n-dimensional grid wrapper of [kknn](#) (not [train.kknn](#)). Predicts a grid structure on the basis of estimated feature contributions. Is used to draw one 2D surface in a 3D plot ([show3d](#)) on basis of feature contributions.

### Usage

```
convolute_grid      (ff,
                     Xi,
                     FCi = NULL,
                     grid = 30,
                     limit = 3,
                     zoom = 3,
                     k.fun=function() round(sqrt(n.obs)/2),
                     userArgs.kknn = alist(kernel="gaussian") )
```

### Arguments

<b>ff</b>	the forestFloor object of class "forestFloor_regression" or "forestFloor_multiClass" at least containing ff\$X and ff\$FCmatrix with two matrices of equal size
<b>Xi</b>	the integer vector, of col indices of ff\$X to estimate by, often of length 2 or 3. Note total number of predictions is equal grid^"length of this vector".
<b>FCi</b>	the integer vector, of col indices of ff\$FCmatrix. Those feature contributions to combine(sum) and estimate. If FCi=NULL, will copy Xi vector, which is the trivial choice.
<b>grid</b>	Either, an integer describing the number of grid.lines in each dimension(trivial choice) or, a full defined matrix of any grid position as defined by this function.
<b>limit</b>	a numeric scalar, number of standard deviations away from mean by any dimension to disregard outliers when spanning observations with grid. Set to limit=Inf outliers never should be disregarded.
<b>zoom</b>	numeric scalar, the size of the grid compared to the uni-variate range of data. If zoom=2 the grid will by any dimension span the double range of the observations. Outliers are disregarded with limit argument.
<b>k.fun</b>	function to define k-neighbors to consider. n.obs is a constant as number of observations in ff\$X. Hereby k neighbors is defined as a function k.fun of n.obs. To set k to a constant use e.g. k.fun = function() 10. k can also be overridden with userArgs.kknn = alist(kernel="Gaussian",kmax=10).
<b>userArgs.kknn</b>	argument list to pass to train.kknn function for each convolution, see <a href="#">kknn</a> for possible args. Arguments in this list will have priority of any passed by default by this wrapper function, see argument merger <a href="#">append.overwrite.alists</a>

10

*convolute\_grid***Details**

This low-level function predicts feature contributions in a grid with [train.kknn](#) which is k-nearest neighbor + Gaussian weighting. This wrapper is used to construct the transparent grey surface in [show3d](#).

**Value**

a data frame, 1 + X variable columns. First column is the predicted summed feature contributions as a function of the following columns feature coordinates.

**Author(s)**

Soren Havelund Welling

**Examples**

```
## Not run:
## avoid testing of rgl 3D plot on headless non-windows OS
## users can disregard this sentence.
if(!interactive() && Sys.info()["sysname"]!="Windows") skip=TRUE

library(rgl)
library(randomForest)
library(forestFloor)

#simulate data
obs=1500
vars = 6
X = data.frame(replicate(vars,runif(obs)))*2-1
Y = with(X, X1*2 + 2*sin(X2*pi) + 3* (X3+X2)^2 )
Yerror = 1 * rnorm(obs)
var(Y)/var(Y+Yerror)
Y= Y+Yerror

#grow a forest, remember to include inbag
rfo=randomForest::randomForest(X,Y,
                                keep.inbag=TRUE,
                                ntree=1000,
                                replace=TRUE,
                                sampsize=500,
                                importance=TRUE)

#compute ff
ff = forestFloor(rfo,X)

#print forestFloor
print(ff)

#plot partial functions of most important variables first
Col=fcol(ff,1)
plot(ff,col=Col,orderByImportance=TRUE)
```

*convolute\_grid*

11

```

#the pure feature contributions
rgl::plot3d(ff$X[,2],ff$X[,3],apply(ff$FCmatrix[,2:3],1,sum),
            #add some colour gradients to ease visualization
            #box.outliers squeeze all observations in a 2 std.dev box
            #univariately for a vector or matrix and normalize to [0;1]
            col=fcol(ff,2,orderByImportance=FALSE))

#add grid convolution/interpolation
#make grid with current function
grid23 = convolute_grid(ff,Xi=2:3,userArgs.kknn= alist(k=25,kernel="gaus"),grid=50,zoom=1.2)
#apply grid on 3d-plot
rgl::persp3d(unique(grid23[,2]),unique(grid23[,3]),grid23[,1],alpha=0.3,
             col=c("black","grey"),add=TRUE)
#anchor points of grid could be plotted also
rgl::plot3d(grid23[,2],grid23[,3],grid23[,1],alpha=0.3,col=c("black"),add=TRUE)

## and we see that there is almost no variance out of the surface, thus is FC2 and FC3
## well explained by the feature context of both X3 and X4

### next example show how to plot a 3D grid + feature contribution
### this 4D application is very experimental

#Make grid of three effects, 25^3 = 15625 anchor points
grid123 = convolute_grid(ff,
                         Xi=c(1:3),
                         FCi=c(1:3),
                         userArgs.kknn = alist(
                           k= 100,
                           kernel = "gaussian",
                           distance = 1),
                         grid=25,
                         zoom=1.2)

#Select a dimension to place in layers
uni2 = unique(grid123[,2]) #2 points to X1 and FC1
uni2=uni2[c(7,9,11,13,14,16,18)] #select some layers to visualize

## plotting any combination of X2 X3 in each layer(from red to green) having different value of X1
count = 0
add=FALSE
for(i in uni2) {
  count = count +1
  this34.plane = grid123[grid123[,2]==i,]
  if (count==2) add=TRUE

  # plot3d(ff$X[,1],ff$X[,2]
  persp3d(unique(this34.plane[,3]),
          unique(this34.plane[,4]),
          this34.plane[,1], add=add,
          col=rgb(count/length(uni2),1-count/length(uni2),0),alpha=0.1)
}

```

12

*fcol*

```

## plotting any combination of X1 X3 in each layer(from red to green) having different value of X2
uni3 = unique(grid123[,4]) #2 points to X1 and FC1
uni3=uni3[c(7,9,11,13,14,16,18)] #select some layers to visualize
count = 0
add=FALSE
for(i in uni3) {
  count = count +1
  this34.plane = grid123[grid123[,4]==i,]
  if (count==2) add=TRUE

  #plot3d(ff$X[,1],ff$X[,2])
  persp3d(unique(this34.plane[,2]),
          unique(this34.plane[,3]),
          this34.plane[,1], add=add,
          col=rgb(count/length(uni3),1-count/length(uni3),0),alpha=0.1)

}
## End(Not run)

```

*fcol**Generic colour module for forestFloor objects*

### Description

This colour module colour observations by selected variables. PCA decomposes a selection more than three variables. Space can be inflated by random forest variable importance, to focus coloring on influential variables. Outliers(>3std.dev) are automatically suppressed. Any colouring can be modified.

### Usage

```
fcol(ff, cols = NULL, orderByImportance = NULL, plotTest=NULL, X.matrix = TRUE,
     hue = NULL, saturation = NULL, brightness = NULL,
     hue.range = NULL, sat.range = NULL, bri.range = NULL,
     alpha = NULL, RGB = NULL, byResiduals=FALSE, max.df=3,
     imp.weight = NULL, imp.exp = 1,outlier.lim = 3,RGB.exp=NULL)
```

### Arguments

**ff** a object of class "forestFloor\_regression" or "forestFloor\_multiClass" or a matrix or a data.frame. No missing values. X.matrix must be set TRUE for "forestFloor\_multiClass" as colouring by multiClass feature contributions is not supported.

fcol

13

<b>cols</b>	vector of indices of columns to colour by, will refer to ff\$X if X.matrix=T and else ff\$FCmatrix. If ff itself is a matrix or data.frame, indices will refer to these columns
<b>orderByImportance</b>	logical, should cols refer to X column order or columns sorted by variable importance. Input must be of forestFloor -class to use this. Set to FALSE if no importance sorting is wanted. Otherwise leave as is.
<b>plotTest</b>	NULL(plot by test set if available), TRUE(plot by test set), FALSE(plot by train), "andTrain"(plot by both test and train)
<b>X.matrix</b>	logical, true will use feature matrix false will use feature contribution matrix. Only relevant if input is forestFloor object.
<b>hue</b>	value within [0,1], hue=1 will be exactly as hue = 0 colour wheel settings, will skew the colour of all observations without changing the contrast between any two given observations.
<b>saturation</b>	value within [0,1], mean saturation of colours, 0 is grey tone and 1 is maximal colourful.
<b>brightness</b>	value within [0,1], mean brightness of colours, 0 is black and 1 is lightly colours.
<b>hue.range</b>	value within [0,1], ratio of colour wheel, small value is small slice of colour wheel those little variation in colours. 1 is any possible colour except for RGB colour system.
<b>sat.range</b>	value within [0,1], for colouring of 2 or more variables, a range of saturation is needed to obtain more degrees of freedom in the colour system. But as saturation of is preferred to be >.75 the range of saturation cannot here exceed .5. If NULL sat.range will set widest possible without exceeding range.
<b>bri.range</b>	value within [0,1], for colouring of 3 or more variables, a range of brightness is needed to obtain more degrees of freedom in the colour system. But as brightness of is preferred to be >.75 the range of saturation cannot here exceed .5. If NULL bri.range will set widest possible without exceeding range.
<b>alpha</b>	value within [0;1] transparency of colours.
<b>RGB</b>	logical TRUE/FALSE, RGB=NULL: will turn TRUE if one variable selected RGB=TRUE: Red-Green-Blue colour: a system with fewer colours(~3) but more contrast. Can still be altered by hue, saturation, brightness etc. RGB=FALSE: True-colour-system: Maximum colour detail. Sometimes more confusing.
<b>byResiduals</b>	logical, should coloring be residuals of main effect fit(overrides X.matrix=). If no fit has been computed "is.null(ff\$FCfit)", a temporarily main effect fit will be computed. Use ff = convolute_ff(ff) to only compute once and/or to modify fit parameters.
<b>max.df</b>	integer 1, 2, or 3 only. Only for true-colour-system, the maximal allowed degrees of freedom in a colour scale. If more variables selected than max.df, PCA decompose to request degrees of freedom. max.df = 1 will give more simple colour gradients
<b>imp.weight</b>	Logical?, Should importance from a forestFloor object be used to weight selected variables? obviously not possible if input ff is a matrix or data.frame. If

14

*fcol*

	randomForest(importance=TRUE) during training, variable importance will be used. Otherwise the more unreliable gini_importance coefficient.
imp.exp	exponent to modify influence of imp.weight. 0 is not influence. -1 is counter influence. 1 is linear influence. .5 is square root influence etc..
outlier.lim	number from 0 to Inf. Any observation which univariately exceed this limit will be suppressed, as if it actually were on this limit. Normal limit is 3 standard deviations. Extreme outliers can otherwise reserve alone a very large part of a given linear colour gradient. This leads to visualization where outlier have one colour and any other observation another but same colour.
RGB.exp	value between ]1;>1]. Defines steepness of the gradient of the RGB colour system Close to one green middle area is missing. For values higher than 2, green area is dominating

### Details

*fcol* produces colours for any observation. These are used plotting.

### Value

a character vector specifying the colour of any observations. Each element is something like "#F1A24340", where F1 is the hexadecimal of the red colour, then A2 is the green, then 43 is blue and 40 is transparency.

### Author(s)

Soren Havelund Welling

### Examples

```
## Not run:
#example 1 - fcol used on data.frame or matrix
library(forestFloor)
X = data.frame(matrix(rnorm(1000),nrow=1000,ncol=4))
X[] = lapply(X,jitter,amount = 1.5)

#single variable gradient by X1 (Unique colour system)
plot(X,col=fcol(X,1))
#double variable gradient by X1 and X2 (linear colour system)
plot(X,col=fcol(X,1:2))
#triple variable gradient (PCA-decomposed, linear colour system)
plot(X,col=fcol(X,1:3))
#higher based gradient (PCA-decomposed, linear colour system)
plot(X,col=fcol(X,1:4))

#force linear col + modify colour wheel
plot(X,col=fcol(X,
                 cols=1, #colouring by one variable
                 RGB=FALSE,
                 hue.range = 4, #cannot exceed 1, if colouring by more than one var
```

fcol

15

```

#except if max.df=1 (limits to 1D gradient)
saturation=1,
brightness = 0.6))

#colour by one dimensional gradient first PC of multiple variables
plot(X,col=fcol(X,
                 cols=1:2, #colouring by multiple
                 RGB=TRUE, #possible because max.df=1
                 max.df = 1, #only 1D gradient (only first principal component)
                 hue.range = 2, #can exceed 1, because max.df=1
                 saturation=.95,
                 brightness = 0.8))

##example 2 - fcol used with forestFloor objects
library(forestFloor)
library(randomForest)

X = data.frame(replicate(6,rnorm(1000)))
y = with(X,.3*X1^2+sin(X2*pi)+X3*X4)
rf = randomForest(X,y,keep.inbag = TRUE,sampsize = 400)
ff = forestFloor(rf,X)

#colour by most important variable
plot(ff,col=fcol(ff,1))

#colour by first variable in data set
plot(ff,col=fcol(ff,1,orderByImportance = FALSE),orderByImportance = FALSE)

#colour by feature contributions
plot(ff,col=fcol(ff,1:2,order=FALSE,X.matrix = FALSE,saturation=.95))

#colour by residuals
plot(ff,col=fcol(ff,3,orderByImportance = FALSE,byResiduals = TRUE))

#colour by all features (most useful for colinear variables)
plot(ff,col=fcol(ff,1:6))

#disable importance weighting of colour
#(important colours get to define gradients more)
plot(ff,col=fcol(ff,1:6,imp.weight = FALSE)) #useless X5 and X6 appear more colourful

#insert outlier in data set in X1 and X2
ff$X[1,1] = 10; ff$X[1,2] = 10

plot(ff,col=fcol(ff,1)) #colour not distorted, default: outlier.lim=3
plot(ff,col=fcol(ff,1,outlier.lim = Inf)) #colour gradient distorted by outlier
plot(ff,col=fcol(ff,1,outlier.lim = 0.5)) #too little outlier.lim

## End(Not run)

```

---

16	<i>forestFloor</i>
----	--------------------

---

<b>forestFloor</b>	<i>Compute out-of-bag cross-validated feature contributions to visualize model structures of randomForest models.</i>
--------------------	---

---

### Description

Computes a cross validated feature contribution matrix from a randomForest model-fit and outputs a forestFloor S3 class object (a list), including unscaled importance and the original training set. The output object is the basis for all visualizations.

### Usage

```
forestFloor(rf.fit, X, Xtest=NULL, calc_np = FALSE, binary_reg = FALSE,
           bootstrapFC = FALSE, ...)
```

### Arguments

<b>rf.fit</b>	rf.fit, a random forest object as the output from randomForest::randomForest
<b>X</b>	data.frame of input variables, numeric(continuous), discrete(treated as continuous) or factors(categorical). n_rows observations and n_columns features X MUST be the same data.frame as used to train the random forest, see above item.
<b>Xtest</b>	data.frame of input variables, numeric(continuous), discrete(treated as continuous) or factors(categorical). n_rows test_examples and n_columns features Xtest MUST have same number and order of columns(variables) as X. Number of rows can vary.
<b>calc_np</b>	TRUE/FALSE. Calculate Node Predictions(TRUE) or reuse information from rf.fit(FALSE)? Slightly faster when FALSE for regression. calc_np=TRUE will only take effect for rf.fit of class "randomForest" and type="regression". This option, is only for developmental purposes. Just set =FALSE always, as function will override this choice if not appropriate.
<b>binary_reg</b>	boolean, if TRUE binary classification can be changed to "percentage votes" of class 1, and thus be treated as regression.
<b>bootstrapFC</b>	boolean, if TRUE an extra column is added to FCmatrix or one extra matrix to FCarray accounting for the minor feature contributions attributed to random bootstraps or stratifications. Mainly useful to check FC row sums actually are equal to OOB-CV predictions, or to tweak randomForest into a "probability forest"-like model.
<b>...</b>	For classification it is possible to manually set majorityTerminal=FALSE. For the randomForest classification implementation majorityTerminal is by default set to TRUE, as each tree uses majority vote within terminal nodes. In other implementations terminal nodes are not necessarily reduced by majority voting before aggregation on ensemble level. majorityTerminal, does not apply to random forest regressions.

*forestFloor*

17

### Details

*forestFloor* computes out-of-bag cross validated feature contributions for a "randomForest" class object. Other packages will be supported in future, mail me a request. *forestFloor* guides you to discover the structure of a randomForest model fit. Check examples of how latent interactions can be identified with colour gradients.

What is FC?: Feature contributions are the sums over all local increments for each observation for each feature divided by the number of trees. A local increment is the change of node prediction from parent to daughter node split by a given feature. Thus a feature contribution summarizes the average outcome for all those times a given sample was split by a given feature. *forestFloor* use inbag samples to calculate local increments, but only sum local increments over out-of-bag samples divided with OOBtimes. OOBtimes is the number of times a given observation have been out-of-bag. which is roundly  $n_{trees} / 3$ . In practice this removes a substantial self-leverage of samples to the corresponding feature contributions. Hereby visualizations becomes less noisy.

What is FC used for?: Feature contributions is smart way to decompose a RF mapping structure into additive components. Plotting FC's against variables values yields at first glance plots similar to marginal-effect plots, partial dependence plots and vector effect characteristic plots. This package for *forestFloor*, make use of feature contributions to separate main effects and identify plus quantify latent interactions. The advantages of *forestFloor* over typical partial.dependence plots are: (1) Easier to identify interactions. (2) Training samples is a part of plot, such that extrapolated model structure can be disregarded. (3) The "goodness of visualization" (how exactly the plot represent the higher dimensional model structure) can be quantified. (4) Cheerful colours and 3D graphics thanks to the *rgl* package.

RF regression takes input features and outputs a target value. RF classification can output a pseudo probability vector with predicted class probability for each sample. The RF mapping topology of classification is different than for regression as the output is no longer a scalar, the output is a vector with predicted class probability for each class. For binary classification this topology can be simplified to a regression-like scalar as the probability of class\_1 = 1 - class\_2. Set `binary_reg=TRUE` for a binary RF classification to get regression like visualizations. For multi-class the output space is probability space where any point is a probability prediction of each target class.

To plot *forestFloor* objects use plot-method `plot.forestFloor` and function `show3d`. Input parameters for classification or regression are not entirely the same. Check help-file `plot.forestFloor` and `show3d`. For 3-class problems the special function `plot_simplex3` can plot the probability predictions in a 2D phase diagram (K-1 simplex).

### Value

the *forestFloor* function outputs(depending on type `rf.fit`) an object of either class "forestFloor\_regression" or "forestFloor\_multiClass" with following elements:

<code>X</code>	a copy of the training data or feature space matrix/data.frame, <code>X</code> . The copy is passed unchanged from the input of this function. <code>X</code> is used in all visualization to expand the feature contributions over the features of which they were recorded.
<code>Y</code>	a copy of the target vector, <code>Y</code> .
<code>importance</code>	The gini-importance or permutation-importance a.k.a variable importance of the random forest object (unscaled). If <code>rfo=randomForest(X,Y,importance=FALSE)</code> , gini-importance is used. Gini-importance is less reproducible and more biased. The extra time used to compute permutation-importance is negligible.

18

*forestFloor*

<code>imp_ind</code>	the importance indices is the order to sort the features by descending importance. <code>imp_ind</code> is used by plotting functions to present most relevant feature contributions first. If using gini-importance, the order of plots is more random and will favor continuous variables. The plots themselves will not differ.
<code>FC_matrix</code>	[ONLY <code>forestFloor_regression</code> .] feature contributions in a matrix. <code>n_row</code> observations and <code>n_column</code> features - same dimensions as <code>X</code> .
<code>FC_array</code>	[ONLY <code>forestFloor_multiClass</code> .] feature contributions in a array. <code>n_row</code> observations and <code>n_column</code> features and <code>n_layer</code> classes. First two dimensions will match dimensions of <code>X</code> .

**Note**

check out more guides at [forestFloor.dk](#)

**Author(s)**

Soren Havelund Welling

**References**

Interpretation of QSAR Models Based on Random Forest Methods, <http://dx.doi.org/10.1002/minf.201000173>  
Interpreting random forest classification models using a feature contribution method, <http://arxiv.org/abs/1312.1121>

**See Also**

[plot.forestFloor](#), [show3d](#),

**Examples**

```
## Not run:
## avoid testing of rgl 3D plot on headless non-windows OS
## users can disregard this sentence.
if(!interactive() && Sys.info()["sysname"]!="Windows") skipRGL=TRUE

#1 - Regression example:
set.seed(1234)
library(forestFloor)
library(randomForest)

#simulate data y = x1^2+sin(x2*pi)+x3*x4 + noise
obs = 5000 #how many observations/samples
vars = 6   #how many variables/features
#create 6 normal distr. uncorr. variables
X = data.frame(replicate(vars,rnorm(obs)))
#create target by hidden function
Y = with(X, X1^2 + sin(X2*pi) + 2 * X3 * X4 + 0.5 * rnorm(obs))

#grow a forest
rfo = randomForest(
  X, #features, data.frame or matrix. Recommended to name columns.
```

forestFloor

19

```

Y, #targets, vector of integers or floats
keep.inbag = TRUE, # mandatory,
importance = TRUE, # recommended, else ordering by giniImpurity (unstable)
sampszie = 1500 , # optional, reduce tree sizes to compute faster
ntree = if(interactive()) 500 else 50 #speedup CRAN testing
)

#compute forestFloor object, often only 5-10% time of growing forest
ff = forestFloor(
  rfo.fit = rfo,      # mandatory
  X = X,             # mandatory
  calc_np = FALSE,   # TRUE or FALSE both works, makes no difference
  binary_reg = FALSE # takes no effect here when rfo$type="regression"
)

#print forestFloor
print(ff) #prints a text of what an 'forestFloor_regression' object is
plot(ff)

#plot partial functions of most important variables first
plot(ff,                  # forestFloor object
     plot_seq = 1:6,       # optional sequence of features to plot
     orderByImportance=TRUE # if TRUE index sequence by importance, else by X column
)

#Non interacting features are well displayed, whereas X3 and X4 are not
#by applying color gradient, interactions reveal themselves
#also a k-nearest neighbor fit is applied to evaluate goodness-of-fit
Col=fcol(ff,3,orderByImportance=FALSE) #create color gradient see help(fcol)
plot(ff,col=Col,plot_GOF=TRUE)

#feature contributions of X3 and X4 are well explained in the context of X3 and X4
# as GOF R^2>.8

show3d(ff,3:4,col=Col,plot_GOF=TRUE,orderByImportance=FALSE)

#if needed, k-nearest neighbor parameters for goodness-of-fit can be accessed through convolute_ff
#a new fit will be calculated and saved to forstFloor object as ff$FCfit
ff = convolute_ff(ff,userArgs.kknn=alist(kernel="epanechnikov",kmax=5))
plot(ff,col=Col,plot_GOF=TRUE) #this computed fit is now used in any 2D plotting.

###  

#2 - Multi classification example: (multi is more than two classes)
set.seed(1234)
library(forestFloor)
library(randomForest)

data(iris)
X = iris[,-names(iris) %in% "Species"]
Y = iris[, "Species"]

rf = randomForest(

```

20

*forestFloor*

```

X,Y,
keep.forest=TRUE, # mandatory
keep.inbag=TRUE, # mandatory
samp=20,          # reduce complexity of mapping structure, with same OOB%-explained
importance = TRUE # recommended, else ordering by giniImpurity (unstable)
)

ff = forestFloor(rf,X)

plot(ff,plot_GOF=TRUE,cex=.7,
      colLists=list(c("#FF0000A5"),
                    c("#00FF0050"),
                    c("#0000FF35")))

#...and 3D plot, see show3d
show3d(ff,1:2,1:2,plot_GOF=TRUE)

#...and simplex plot (only for three class problems)
plot_simplex3(ff)
plot_simplex3(ff,zoom.fit = TRUE)

#...and 3d simplex plots (rough look, Z-axis is feature)
plot_simplex3(ff,fig3d = TRUE)

#####
#3 - binary regression example
#classification of two classes can be seen as regression in 0 to 1 scale
set.seed(1234)
library(forestFloor)
library(randomForest)
data(iris)
X = iris[-1:-50,!names(iris) %in% "Species"] #drop third class virginica
Y = iris[-1:-50,"Species"]
Y = droplevels((Y)) #drop unused level virginica

rf = randomForest(
  X,Y,
  keep.forest=TRUE, # mandatory
  keep.inbag=TRUE, # mandatory
  samp=20,          # reduce complexity of mapping structure, with same OOB%-explained
  importance = TRUE # recommended, else giniImpurity
)

ff = forestFloor(rf,X,
                 calc_np=TRUE,    #mandatory to recalculate
                 binary_reg=TRUE) #binary regression, scale direction is printed
Col = fcol(ff,1) #color by most important feature
plot(ff,col=Col) #plot features

#interfacing with rgl:::plot3d
show3d(ff,1:2,col=Col,plot.rgl.args = list(size=2,type="s",alpha=.5))

## End(Not run)

```

*plot.forestFloor*

21

---

 plot.forestFloor      *plot.forestFloor\_regression*


---

### Description

A method to plot an object of forestFloor-class. Plot partial feature contributions of the most important variables. Colour gradients can be applied to show possible interactions. Fitted function(plot\_GOF) describe FC only as a main effect and quantifies 'Goodness Of Fit'.

### Usage

```
## S3 method for class 'forestFloor_regression'
plot(
  x,
  plot_seq=NULL,
  plotTest = NULL,
  limitY=TRUE,
  orderByImportance=TRUE,
  cropXaxes=NULL,
  crop_limit=4,
  plot_GOF = TRUE,
  GOF_args = list(col="#33333399"),
  speedup_GOF = TRUE,
  ...)

## S3 method for class 'forestFloor_multiClass'
plot(
  x,
  plot_seq = NULL,
  label.seq = NULL,
  plotTest = NULL,
  limitY = TRUE,
  col = NULL,
  collists = NULL,
  orderByImportance = TRUE,
  fig.columns = NULL,
  plot_GOF = TRUE,
  GOF_args = list(),
  speedup_GOF = TRUE,
  jitter_these_cols = NULL,
  jitter.factor = NULL,
  ...)
```

### Arguments

- x      forestFloor-object, also abbreviated ff. Basically a list of class="forestFloor" containing feature contributions, features, targets and variable importance.

22

*plot.forestFloor*

<b>plot_seq</b>	a numeric vector describing which variables and in what sequence to plot. Ordered by importance as default. If <code>orderByImportance = F</code> , then by feature/column order of training data.
<b>label.seq</b>	[only classification] a numeric vector describing which classes and in what sequence to plot. NULL is all classes ordered in levels in <code>x\$Y</code> of <code>forestFloor_multiclass</code> object <code>x</code> .
<b>plotTest</b>	NULL(plot by test set if available), TRUE(plot by test set), FALSE(plot by train), "andTrain"(plot by both test and train)
<b>fig.columns</b>	[only for multiple plotting], how many columns per page. default(NULL) is 1 for one plot, 2 for 2, 3 for 3, 2 for 4 and 3 for more.
<b>limitY</b>	TRUE/FLASE, constrain all Yaxis to same limits to ensure relevance of low importance features is not over interpreted
<b>col</b>	Either a color vector with one colour per plotted class label or a list of colour vectors. Each element is a colour vector one class. Colour vectors in list are normally either of length 1 with or of length equal to number of training observations. NULL will choose standard one colour per class.
<b>colLists</b>	Depreciated, will be replaced by col input
<b>jitter_these_cols</b>	vector to apply jitter to x-axis in plots. Will refer to variables. Useful to for categorical variables. Default=NULL is no jitter.
<b>jitter.factor</b>	value to decide how much jitter to apply. often between .5 and 3
<b>orderByImportance</b>	TRUE / FALSE should plotting and <code>plot_seq</code> be ordered after importance. Most important feature plot first(TRUE)
<b>cropXaxes</b>	a vector of indices of which zooming of x.axis should look away from outliers
<b>crop_limit</b>	a number often between 1.5 and 5, referring limit in sigmas from the mean defining outliers if <code>limit = 2</code> , above selected plots will zoom to +/- 2 std.dev of the respective features.
<b>plot_GOF</b>	Boolean TRUE/FALSE. Should the goodness of fit be plotted as a line?
<b>GOF_args</b>	Graphical arguments fitted lines, see <code>points</code> for parameter names.
<b>speedup_GOF</b>	Should GOF only computed on reasonable sub sample of data set to speedup computation. GOF estimation leave-one-out-kNN becomes increasingly slow for +1500 samples.
<b>...</b>	... other arguments passed to <code>par</code> or <code>plot</code> . e.g. <code>mar=</code> , <code>mfrow=</code> , is passed to <code>par</code> , and <code>cex=</code> is passed to <code>plot</code> . <code>par()</code> arguments are reset immediately as <code>plot</code> function returns.

### Details

The method `plot.forestFloor` visualizes partial plots of the most important variables first. Partial dependence plots are available in the `randomForest` package. But such plots are single lines(1d-slices) and do not answer the question: Is this partial function(PF) a fair generalization or subject to global or local interactions.

*plot.forestFloor*

23

**Author(s)**

Soren Havelund Welling

**Examples**

```

## Not run:
## avoid testing of rgl 3D plot on headless non-windows OS
## users can disregard this sentence.
if(!interactive() && Sys.info()["sysname"]!="Windows") skipRGL=TRUE

### 
#1 - Regression example:
set.seed(1234)
library(forestFloor)
library(randomForest)

#simulate data y = x1^2+sin(x2*pi)+x3*x4 + noise
obs = 5000 #how many observations/samples
vars = 6   #how many variables/features
#create 6 normal distr. uncorr. variables
X = data.frame(replicate(vars,rnorm(obs)))
#create target by hidden function
Y = with(X, X1^2 + sin(X2*pi) + 2 * X3 * X4 + 0.5 * rnorm(obs))

#grow a forest
rfo = randomForest(
  X, #features, data.frame or matrix. Recommended to name columns.
  Y, #targets, vector of integers or floats
  keep.inbag = TRUE, # mandatory,
  importance = TRUE, # recommended, else ordering by giniImpurity (unstable)
  sampsize = 1500 , # optional, reduce tree sizes to compute faster
  ntree = if(interactive()) 1000 else 25 #speedup CRAN testing
)

#compute forestFloor object, often only 5-10% time of growing forest
ff = forestFloor(
  rf.fit = rfo,      # mandatory
  X = X,            # mandatory
  calc_np = FALSE,  # TRUE or FALSE both works, makes no difference
  binary_reg = FALSE # takes no effect here when rfo$type="regression"
)

#print forestFloor
print(ff) #prints a text of what an 'forestFloor_regression' object is
plot(ff)

#plot partial functions of most important variables first
plot(ff,                  # forestFloor object
     plot_seq = 1:6,       # optional sequence of features to plot
     orderByImportance=TRUE # if TRUE index sequence by importance, else by X column
)

```

24

*plot.forestFloor*

```

#Non interacting features are well displayed, whereas X3 and X4 are not
#by applying color gradient, interactions reveal themself
#also a k-nearest neighbor fit is applied to evaluate goodness-of-fit
Col=fcol(ff,3,orderByImportance=FALSE) #create color gradient see help(fcol)
plot(ff,col=Col,plot_GOF=TRUE)

#feature contributions of X3 and X4 are well explained in the context of X3 and X4
# as GOF R^2>.8

show3d(ff,3:4,col=Col,plot_GOF=TRUE,orderByImportance=FALSE)

#if needed, k-nearest neighbor parameters for goodness-of-fit can be accessed through convolute_ff
#a new fit will be calculated and saved to forstFloor object as ff$FCfit
ff = convolute_ff(ff,userArgs.kknn=alist(kernel="epanechnikov",kmax=5))
plot(ff,col=Col,plot_GOF=TRUE) #this computed fit is now used in any 2D plotting.

###  

#2 - Multi classification example: (multi is more than two classes)
set.seed(1234)
library(forestFloor)
library(randomForest)

data(iris)
X = iris[,!names(iris) %in% "Species"]
Y = iris[,"Species"]

rf = randomForest(
  X,Y,
  keep.forest=TRUE, # mandatory
  keep.inbag=TRUE, # mandatory
  samp=20,           # reduce complexity of mapping structure, with same OOB%-explained
  importance = TRUE, # recommended, else ordering by giniImpurity (unstable)
  ntree = if(interactive()) 1000 else 25 #speedup CRAN testing
)

ff = forestFloor(rf,X)

plot(ff,plot_GOF=TRUE,cex=.7,
  col=c("#FF0000A5","#00FF0050","#0000FF35") #one col per plotted class
)

#...and 3D plot, see show3d
show3d(ff,1:2,1:2,plot_GOF=TRUE)

#...and simplex plot (only for three class problems)
plot_simplex3(ff)
plot_simplex3(ff,zoom.fit = TRUE)

#...and 3d simplex plots (rough look, Z-axis is feature)
plot_simplex3(ff,fig3d = TRUE)

###

```

plot\_simplex3

25

```

#3 - binary regression example
#classification of two classes can be seen as regression in 0 to 1 scale
set.seed(1234)
library(forestFloor)
library(randomForest)
data(iris)
X = iris[-1:-50,!names(iris) %in% "Species"] #drop third class virginica
Y = iris[-1:-50,"Species"]
Y = droplevels((Y)) #drop unused level virginica

rf = randomForest(
  X,Y,
  keep.forest=TRUE,  # mandatory
  keep.inbag=TRUE,   # mandatory
  samp=20,           # reduce complexity of mapping structure, with same OOB%-explained
  importance = TRUE, # recommended, else giniImpurity
  ntree = if(interactive()) 1000 else 25 #speedup CRAN testing
)

ff = forestFloor(rf,X,
                 calc_np=TRUE,      #mandatory to recalculate
                 binary_reg=TRUE) #binary regression, scale direction is printed
Col = fcol(ff,1) #color by most important feature
plot(ff,col=Col)  #plot features

#interfacing with rgl:::plot3d
show3d(ff,1:2,col=Col,plot.rgl.args = list(size=2,type="s",alpha=.5))

## End(Not run)

```

plot\_simplex3        3-class simplex forestFloor plot

**Description**

3-class forestFloor plotted in a 2D simplex. The plot describes with feature contributions the change of predicted class probability for each sample due a single variable given all other variables. This plot is better than regular multiclass plots (plot.forestFloor\_multiClass) to show the change of class probabilities, but the feature values can only be depicted as a colour gradient. But (fig3d=TRUE) allows the feature value to be depicted by the Z-axis as a extra pop-up 3D plot.

**Usage**

```

plot_simplex3(
  ff,
  Xi          = NULL,
  includeTotal = TRUE,
  label.col    = NULL,
  fig.cols     = 3,
  fig.rows     = NULL,

```

26

*plot\_simplex3*

```

auto.alpha    = 0.25,
fig3d        = FALSE,
restore_par   = TRUE,
set_pars     = TRUE,
zoom.fit     = NULL,
var.col      = NULL,
plot.sep.centroid = TRUE)

```

### Arguments

<code>ff</code>	x also abbreviated ff, forestFloor_multiclass the output from the forestFloor function. Must have 3 classes exactly.
<code>Xi</code>	vector of integer indices (referring to column order of trainingset) to what feature contributions should be plotted in individual plots.
<code>includeTotal</code>	TRUE / FALSE. Combined separation of all feature contributions, which is equal to the separation of the entire model can be included.
<code>label.col</code>	a colour vector of K classes length defining the colour of each class for plotting. NULL is auto.
<code>fig.cols</code>	How many columns should be plotted sideways, is passed to par(mfrow=c(fig.rows,fig.cols))
<code>fig.rows</code>	How many rows should be plotted, is passed to par(mfrow=c(fig.rows,fig.cols)) NULL is auto
<code>auto.alpha</code>	a scalar between 0.5 to 1 most often. Low values increase transparency of points used to avoid overplotting. auto.alpha is alpha corrected of samplesize such that less adjustment is needed.
<code>fig3d</code>	TRUE/FALSE, a 3D plot including the variable as an axis can be co-plotted with rgl.
<code>restore_par</code>	TRUE/FALSE, calls to graphics par() will be reset
<code>set_pars</code>	TRUE/FALSE, if FALSE plot function will rather inherit plot settings global pars. Useful for multi plotting loops.
<code>zoom.fit</code>	NULL/TRUE, if TRUE zooming on samples will be applied. Do not set to FALSE.
<code>var.col</code>	a single colour or a colour vector of N samples length. Samples will be coloured accordingly. use function fcol to make colour gradient e.g. by the variable values themselves. See example <code>fcol</code> .
<code>plot.sep.centroid</code>	TRUE/FALSE. Should the average bootstrap prediction be plotted? If no bootstrap stratification, the average bootstrap prediction is equal to class distribution training set. RF model probabilistic predictions is equal to average bootstrap prediction plus all feature contributions.

### Details

Random forest 3 class maps from a feature space to a 3 dimensional ( $K-1$ ) probability simplex space, which can be plotted in 2D because class probabilities sum to one, and class feature contributions sum to zero. The centroid these plots is the prior of the random forest model. The prior, unless modified with stratification is the target class distribution. Default majority voting lines would run from middle to the corners.

*plot\_simplex3*

27

**Author(s)**

Soren Havelund Welling

**Examples**

```

## Not run:
library(randomForest)
library(forestFloor)
require(utils)

data(iris)

X = iris[, !names(iris) %in% "Species"]
Y = iris[, "Species"]
as.numeric(Y)
rf.test42 = randomForest(X,Y,keep.forest=TRUE,
    replace=FALSE,keep.inbag=TRUE,samp=15,ntrree=100)
ff.test42 = forestFloor(rf.test42,X,calc_np=FALSE,binary_reg=FALSE)

plot(ff.test42,plot_GOF=TRUE,cex=.7,
    colLists=list(c("#FF0000A5"),
        c("#00FF0050"),
        c("#0000FF35")))

show3d(ff.test42,1:2,3:4,plot_GOF=TRUE)

#plot all effect 2D only
pars = plot_simplex3(ff.test42,Xi=c(1:3),restore_par=FALSE,zoom.fit=NULL,
    var.col=NULL,fig.cols=2,fig.rows=1,fig3d=FALSE,includeTotal=TRUE,auto.alpha=.4
    ,set_pars=TRUE)

pars = plot_simplex3(ff.test42,Xi=0,restore_par=FALSE,zoom.fit=NULL,
    var.col=alist(alpha=.3,cols=1:4),fig3d=FALSE,includeTotal=TRUE,
    auto.alpha=.8,set_pars=FALSE)

for (I in ff.test42$imp_ind[1:4]) {
    #plotting partial OOB-CV separation(including interactions effects)
    #coloured by true class
    pars = plot_simplex3(ff.test42,Xi=I,restore_par=FALSE,zoom.fit=NULL,
        var.col=NULL,fig.cols=4,fig.rows=2,fig3d=TRUE,includeTotal=FALSE,label.col=1:3,
        auto.alpha=.3,set_pars = (I==ff.test42$imp_ind[1]))

    #coloured by variable value
    pars = plot_simplex3(ff.test42,Xi=I,restore_par=FALSE,zoom.fit=TRUE,
        var.col=alist(order=FALSE,alpha=.8),fig3d=FALSE,includeTotal=(I==4),
        auto.alpha=.3,set_pars=FALSE)
}

## End(Not run)

```

28

*print.forestFloor*

---

**print.forestFloor**      *print summary of forestFloor.Object*

---

### Description

This function simply states the obvious and returns the elements inside the object list.

### Usage

```
## S3 method for class 'forestFloor_regression'  
  print(x,...)  
## S3 method for class 'forestFloor_multiClass'  
  print(x,...)
```

### Arguments

x	x also abbreviated ff, forestFloor_Object the output from the forestFloor function
...	... other arguments passed to generic print function

### Details

prints short help text for usage of a forestFloor\_object

### Author(s)

Soren Havelund Welling

### Examples

```
## Not run:  
#simulate data  
obs=1000  
vars = 6  
X = data.frame(replicate(vars,rnorm(obs)))  
Y = with(X, X1^2 + sin(X2*pi) + 2 * X3 * X4 + 0.5 * rnorm(obs))  
  
#grow a forest, remeber to include inbag  
rfo=randomForest::randomForest(X,Y,keep.inbag=TRUE)  
  
#compute topology  
ff = forestFloor(rfo,X)  
  
#print forestFloor  
print(ff)  
  
## End(Not run)
```

recTree

29

---

recTree*recursiveTree: cross-validated feature contributions*

---

**Description**

internal C++ functions to compute feature contributions for a random Forest

**Usage**

```
recTree( vars, obs, ntree, calculate_node_pred, X,Y,majorityTerminal, leftDaughter,
         rightDaughter, nodestatus, xbestsplit, nodepred, bestvar,
         inbag, varLevels, OOBtimes, localIncrements)

multiTree(vars, obs, ntree, nClasses,           X,Y,majorityTerminal, leftDaughter,
          rightDaughter, nodestatus, xbestsplit, nodepred, bestvar,
          inbag, varLevels, OOBtimes, localIncrements)
```

**Arguments**

vars	number of variables in X
obs	number of observations in X
ntree	number of trees starting from 1 function should iterate, cannot be higher than columns of inbag
nClasses	number of classes in classification forest
calculate_node_pred	should the node predictions be recalculated(true) or reused from nodepred-matrix(false & regression)
X	X training matrix
Y	target vector, factor or regression
majorityTerminal	bool, majority vote in terminal nodes? Default is FALSE for regression. Set only to TRUE when binary_reg=TRUE.
leftDaughter	a matrix from the output of randomForest rf\$forest\$leftDaughter the node.number/row.number of the leftDaughter in a given tree by column
rightDaughter	a matrix from the output of randomForest rf\$forest\$rightDaughter the node.number/row.number of the rightDaughter in a given tree by column
nodestatus	a matrix from the output of randomForest rf\$forest\$nodestatus the nodestatus of a given node in a given tree
xbestsplit	a matrix from the output of randomForest rf\$forest\$xbestsplit. The split point of numeric variables or the binary split of categorical variables. See help file of randomForest::getTree for details of binary expansion for categorical splits.
nodepred	a matrix from the output of randomForest rf\$forest\$xbestsplit. The inbag target average for regression mode and the majority target class for classification

30

*show3d*

<b>bestvar</b>	a matrix from a the output of randomForest rf\$forest\$xbestsplit the inbag target average for regression mode and the majority target class for classification
<b>inbag</b>	a matrix as the output of randomForest rf\$inbag. Contain counts of how many times a sample was selected for a given tree.
<b>varLevels</b>	the number of levels of all variables, 1 for continuous or discrete, >1 for categorical variables. This is needed for categorical variables to interpret binary split from xbestsplit.
<b>OOBtimes</b>	number of times a certain observation was out-of-bag in the forest. Needed to compute cross-validated feature contributions as these are summed local increments over out-of-bag observations over features divided by this number. In previous implementation(rfFC), articles(see references) feature contributions are summed by all observations and is divived by ntrees.
<b>localIncrements</b>	an empty matrix to store localIncrements during computation. As C++ function returns, the input localIncrement matrix contains the feature contributions.

### Details

This function is excuted by the function forestFloor. This is a c++/Rcpp implementation computing feature contributions. The main differences from this implementation and the rfFC-package(Rforge), is that these feature contributions are only summed over out-of-bag samples yields a cross-validation. This implementation allows sample replacement, binary and multi-classification.

### Value

no output, the feature contributions are wrtten directly to localIncrements input

### Author(s)

Soren Havelund Welling

### References

Interpretation of QSAR Models Based on Random Forest Methods, <http://dx.doi.org/10.1002/minf.201000173>  
 Interpreting random forest classification models using a feature contribution method, <http://arxiv.org/abs/1312.1121>

show3d

*make forestFloor 3D-plot of random forest feature contributions*


---

### Description

2 features features(horizontal XY-plane) and one combined feature contribution (vertical Z-axis). Surface response layer will be estimated(kknn package) and plotted alongside the data points. 3D graphic device is rgl. Will dispatch methods show3d.forestFloor\_regression for regression and show3d\_forestFloor\_multiClass for classification.

*show3d*

31

**Usage**

```
## S3 method for class 'forestFloor_regression'
show3d(
  x,
  Xi = 1:2,
  FCi = NULL,
  col = "#12345678",
  plotTest = NULL,
  orderByImportance = TRUE,
  surface=TRUE,
  combineFC = sum,
  zoom=1.2,
  grid.lines=30,
  limit=3,
  cropPointsOutSideLimit = TRUE,
  kknnGrid.args = alist(),
  plot.rgl.args = alist(),
  surf.rgl.args = alist(),
  user.gof.args = alist(),
  plot_GOF = TRUE,
  ...)

## S3 method for class 'forestFloor_multiClass'
show3d(
  x,
  Xi,
  FCi=NULL,
  plotTest = NULL,
  label.seq=NULL,
  kknnGrid.args=list(NULL),
  plot.rgl.args=list(),
  plot_GOF=FALSE,
  user.gof.args=list(NULL),
  ...)
```

**Arguments**

<i>x</i>	forestFloor" class object
<i>Xi</i>	integer vector of length 2 indices of feature columns
<i>FCi</i>	integer vector of length 1 to p variables indices of feature contributions columns
<i>col</i>	a colour vector. One colour or colour palette(vector).
<i>plotTest</i>	NULL(plot by test set if available), TRUE(plot by test set), FALSE(plot by train), "andTrain"(plot by both test and train)
<i>orderByImportance</i>	should indices order by 'variable importance' or by matrix/data.frame order?
<i>surface</i>	should a surface be plotted also?

32	<i>show3d</i>
combineFC	a row function applied on selected columns(FCi) on \$FCmatrix or \$FCarray. How should feature contributions be combined? Default is <code>sum</code> .
zoom	grid can be expanded in all directions by a factor
grid.lines	how many grid lines should be used. Total surface anchor points in plot is grid.lines^2. May run slow above 200-500 depending on hardware.
limit	a number. Sizing of grid does not consider outliers outside this limit of e.g. 3 SD deviations univariately.
cropPointsOutSideLimit	#if points exceed standard deviation limit, they will not be plotted
kknnGrid.args	argument list, any possible arguments to kknnknn These default wrapper arguments can hereby be overwritten: wrapper = alist( formula=fc~, # do not change train=Data, # do not change k=k, # integer < n_observations. k>100 may run slow. kernel="gaussian", #distance kernel, other is e.g. kernel="triangular" test=gridX #do not change ) see kknnknn to understand parameters. k is set by default automatically to a half times the square root of observations, which often gives a reasonable balance between robustness and adeptness. k neighbors and distance kernel can be changed be passing kknnGrid.args = alist(k=5,kernel="triangular",scale=FALSE), hereby will default k and default kernel be overwritten. Moreover the scale argument was not specified by this wrapper and therefore not conflicting, the argument is simply appended.
plot.rgl.args	pass argument to rgl::plot3d, can override any argument of this wrapper, defines plotting space and plot points. See plot3d for documentation of graphical arguments. wrapper_arg = alist( x=xaxis, #do not change, x coordinates y=yaxis, #do not change, y coordinates z=zaxis, #do not change, z coordinates col=col, #colouring evaluated within this wrapper function xlab=names(X)[1], #xlab, label for x axis ylab=names(X)[2], #ylab, label for y axis zlab=paste(names(X[,FCi]),collapse=" - "), #zlab, label for z axis alpha=.4, #points transparency size=3, #point size scale=.7, #z axis scaling avoidFreeType = T, #disable freeType=T plug-in. (Postscript labels) add=FALSE #do not change, should graphics be added to other rgl-plot? )
surf.rgl.args	wrapper_arg = alist( x=unique(grid[,2]), #do not change, values of x-axis y=unique(grid[,3]), #do not change, values of y-axis z=grid[,1], #do not change, response surface values add=TRUE, #do not change, surface added to plotted points alpha=0.4 #transparency of surface, [0;1] )

*show3d*

33

see `rgl::persp3d` for other graphical arguments notice the surface is added onto plotting of points, thus can e.g. labels not be changed from here.

<code>label.seq</code>	a numeric vector describing which classes and in what sequence to plot. NULL is all classes ordered in levels in <code>x\$Y</code> of <code>forestFloor_multiclass</code> object <code>x</code> .
<code>user.gof.args</code>	argument list passed to internal function <code>ff2</code> , which can modify how goodness-of-fit is computed. Number of neighbors and kernel can be set manually with e.g. <code>list(kmax=40,kernel="gaussian")</code> . Default pars should work already in most cases. Function <code>ff2</code> computed leave-one-out CV prediction the feature contributions from the chosen context of the visualization.
<code>plot_GOF</code>	Boolean TRUE/FALSE. Should the goodness of fit be computed and plotted in main of 3D plot? If false, no GOF input pars are useful.
<code>...</code>	not used at the moment

### Details

`show3d` plot one or more combined feature contributions in the context of two features with points representing each data point. The input object must be a "forestFloor\_regression" or "forestFloor\_multiClass" S3 class object , and should at least contain `$X` the data.frame of training data, `$FCmatrix` the feature contributions matrix. Usually this object are formed with the function `forestFloor` having a random forest model fit as input. Actual visualization differs for each class.

### Value

no value

### Author(s)

Soren Havelund Welling

### Examples

```
## Not run:
## avoid testing of rgl 3D plot on headless non-windows OS
## users can disregard this sentence.
if(!interactive() && Sys.info()["sysname"]!="Windows") skipRGL=TRUE

library(forestFloor)
library(randomForest)
#simulate data
obs=2500
vars = 6

X = data.frame(replicate(vars,rnorm(obs)))
Y = with(X, X1^2 + sin(X2*pi) + 2 * X3 * X4 + 1 * rnorm(obs))

#grow a forest, remeber to include inbag
rfo=randomForest(X,Y,keep.inbag = TRUE,sampsize=1500,ntree=500)

#compute topology
```

34

*vec.plot*

```

ff = forestFloor(rfo,X)

#print forestFloor
print(ff)

#plot partial functions of most important variables first
plot(ff)

#Non interacting functions are well displayed, whereas X3 and X4 are not
#by applying different colourgradient, interactions reveal themself
Col = fcol(ff,3)
plot(ff,col=Col)

#in 3D the interaction between X3 and X reveals itself completely
show3d(ff,3:4,col=Col,plot.rgl=list(size=5))

#although no interaction, a joined additive effect of X1 and X2
Col = fcol(ff,1:2,X.m=FALSE,RGB=TRUE) #colour by FC-component FC1 and FC2 summed
plot(ff,col=Col)
show3d(ff,1:2,col=Col,plot.rgl=list(size=5))

#.or two-way gradient is formed from FC-component X1 and X2.
Col = fcol(ff,1:2,X.matrix=TRUE,alpha=0.8)
plot(ff,col=Col)
show3d(ff,1:2,col=Col,plot.rgl=list(size=5))

## End(Not run)

```

**vec.plot**

*Compute and plot vector effect characteristics for a given multivariate model*

### Description

`vec.plot` visualizes the vector effect characteristics of a given model. Geometrically it corresponds to a specific 2D or 3D slice of a higher dimensional mapping structure. One variable (2D plot) or two variables (3D plot) are screened within the range of the training data, while remaining variables are fixed at the univariate means (as default). If remaining variables do not interact strongly with plotted variable(s), `vec.plot` is a good tool to break up a high-dimensional model structure into separate components.

### Usage

```

vec.plot(model,X,i.var,grid.lines=100,VEC.function=mean,
         zoom=1,limitY=F,moreArgs=list(),...)

```

*vec.plot*

35

**Arguments**

<code>model</code>	model_object which has a defined method predict.model, which can accept arguments as showed for randomForest e.g. library(randomForest) model = randomForest(X,Y) predict(model,X)
<code>X</code>	where X is the training features and Y is the training response vector(numeric) matrix or data.frame being the same as input to model
<code>i.var</code>	vector, of column_numbers of variables to scan. No plotting is available for more than two variables.
<code>grid.lines</code>	scalar, number of values by each variable to be predicted by model. Total number of combinations = grid.lines^length(i.var).
<code>VEC.function</code>	function, establish one fixed value for any remaining variables(those not chosen by i.var). Default is to use the mean of variables.
<code>zoom</code>	scalar, number defining the size.factor of the VEC.surface compared to data range of scanned variables. Bigger number is bigger surface.
<code>limitY</code>	boolean, if TRUE Y-axis is standardized for any variable. Useful for composite plots as shown in example.
<code>moreArgs</code>	any lower level graphical args passed to rgl::surface3d or points depending on number of variables(length of i.var)
<code>...</code>	any lower level graphical args passed to rgl::plot3d or plot depending on number of variables(length of i.var)

**Details**

`vec.plot` visualizes the vector effect characteristics of a given model. One(2D plot) or two(3D plot) variables are screened within the range of the training data, while remaining variables are fixed at the univariate means of each them(as default). If remaining variables do not interact strongly with plotted variable(s), `vec.plot` is a good tool to break up a high-dimensional model topology in separate components.

**Value**

no value

**Author(s)**

Soren Havelund Welling

**Examples**

```
## Not run:
## avoid testing of rgl 3D plot on headless non-windows OS
## users can disregard this sentence.
if(!interactive() && Sys.info()["sysname"]!="Windows") skipRGL=TRUE
library(randomForest)
library(forestFloor)

#simulate data
```

36

*Xtestmerger*

```

obs=2000
vars = 6
X = data.frame(replicate(vars,rnorm(obs)))
Y = with(X, X1^2 + 2*sin(X2*pi) + 2 * X3 * (X4+.5))
Yerror = 1 * rnorm(obs)
var(Y)/var(Y+Yerror)
Y= Y+Yerror

#grow a forest, remeber to include inbag
rfo2=randomForest(X,Y,keep.inbag=TRUE,sampsize=800)

#plot partial functions of most important variables first
pars=par(no.readonly=TRUE) #save previous graphical parameters
par(mfrow=c(2,3),mar=c(2,2,1,1))
for(i in 1:vars) vec.plot(rfo2,X,i,zoom=1.5,limitY=TRUE)
par(pars) #restore

#plot partial functions of most important variables first
for(i in 1:vars) vec.plot(rfo2,X,i,zoom=1.5,limitY=TRUE)

#plotvariable X3 and X4 with vec.plot
Col = fcol(X,3:4)
vec.plot(rfo2,X,3:4,zoom=1,grid.lines=100,col=Col)

## End(Not run)

```

*Xtestmerger**merge training set (X) and (test) set***Description**

... and expand inbag matrix and training target vector to compute FC for a test set.

**Usage**

```
Xtestmerger(X,test,inbag=NULL,y=NULL)
```

**Arguments**

X	training set data.frame used to train a random forest model
test	a test set data.frame which feature contributions should be computed for
inbag	matrix of inbag sampling to expande with training set, which is set OOB for any tree
y	random forest target vector, which is set to first value for observation

**Details**

*Xtestmerger* is a low-level function to merge a test set with *X* training set. There can be no names, column class, column number mismatch. Moreover any level in any factor of test must be present in *X*, as RF/forestFloor cannot score a unknown factor level / category.

*Xtestmerger*

37

**Value**

List of merged bigX, bigInbag and bigy. The two latter may be NULL if not provided.

**Author(s)**

Soren Havelund Welling

**Examples**

```
library(randomForest)
library(forestFloor)
#X y could be a training set
X = data.frame(numeric = c(1,5,2,7,-4.3),
               factor1 = factor(c("jim","freddy","marley","marley","alfred")),
               factor2 = factor(c("jill","ann","liz","leila","vicky")))
y = factor(1:5)
set.seed(1)
rf = randomForest(X,y,keep.inbag=TRUE,ntree=7)
#should not raise any error
test = data.frame(numeric = rnorm(5),
                  factor1 = factor(c("jim","jim","jim","freddy","freddy")),
                  factor2 = factor(c("jill","jill","vicky","leila","vicky")))
out = Xtestmerger(X,test,inbag=rf$inbag,y=y)
```

# Index

\*Topic **models**  
  forestFloor, 16  
  forestFloor-package, 2  
  plot.forestFloor, 21

\*Topic **multivariate**  
  forestFloor, 16  
  forestFloor-package, 2  
  plot.forestFloor, 21

\*Topic **non-linear**  
  forestFloor-package, 2

\*Topic **nonlinear**  
  forestFloor, 16  
  plot.forestFloor, 21

\*Topic **outlier.filtration**  
  box.outliers, 4

\*Topic **robust**  
  forestFloor, 16  
  forestFloor-package, 2  
  plot.forestFloor, 21

append.overwrite.alists, 2, 9  
as.numeric.factor, 3

box.outliers, 4

convolute\_ff, 5  
convolute\_ff2, 7  
convolute\_grid, 9

fcol, 12, 26  
forestFloor, 16  
forestFloor-package, 2  
forestFloor\_randomForest\_classification  
  (forestFloor), 16  
forestFloor\_randomForest\_regression  
  (forestFloor), 16  
forestFloorPackage  
  (forestFloor-package), 2

kknn, 9

multiTree (recTree), 29  
par, 22  
plot, 22  
plot.forestFloor, 17, 18, 21  
plot.forestFloor\_multiClass  
  (plot.forestFloor), 21  
plot.forestFloor\_regression  
  (plot.forestFloor), 21  
plot\_simplex3, 17, 25  
points, 22  
print.forestFloor, 28  
print.forestFloor\_classification  
  (print.forestFloor), 28  
print.forestFloor\_multiClass  
  (print.forestFloor), 28  
print.forestFloor\_regression  
  (print.forestFloor), 28

recTree, 29

show3d, 8–10, 17, 18, 30  
sum, 32

train.kknn, 5, 7–10

vec.plot, 34

Xtestmerger, 36



# Bibliography

---

- [Ada80] Douglas Adams. *The hitchhiker's guide to the galaxy*. Hitchhiker's Guide to the Galaxy Series. Harmony Books, 1980. ISBN: 9780517542095.
- [Wai99] D. Waitzman. *IP over Avian Carriers with Quality of Service*. RFC 2549 (Informational). Internet Engineering Task Force, April 1999. URL: <http://www.ietf.org/rfc/rfc2549.txt> (visited on September 31, 2014).

