Exploring the Factory: Into the Guts of the Monster

Greg Landrum, Anna Martin, Daria Goldmann

KNIME
Outline

• Review of the whole process
• The model factory (Anna)
• Extracting the data (Greg)
• Transform (Generating fingerprints) (Greg)
• Model building and validation (Daria)
• Model report (Daria)
• REST service and prediction (Greg)
• Results view (Anna)
• Wrap up and what’s next (Greg)
Reminder 1: the problem

- I have >1500 datasets that I would like to build models for
- I want to actually use the models, so they need to be deployed
- The whole process needs to be automated and reproducible so that I can do it again when the database is updated
Reminder 2: the solution

- Created workflows for:
  - automating dataset extraction from the database
  - generating features for a dataset
  - automated model building and validation
  - deploying models that pass validation
- Automate the execution of that using the Model Factory
- Deployed models to KNIME Server as:
  - WebPortal application for model details
  - A REST service to generate predictions
  - WebPortal application for viewing predictions
Model Factory

Init  Load  Transform  Learn  Score  Evaluate  Deploy
**Model Factory**

**Abstract Model Factory**

```
Load: load the assay data from the db
Transform: WF2
CASE Switch Data (Start)
  case 1: existing
  case 2: initialisation
  case 3: learn model only
Select Model: currentModel
End IF
Score: Node 58
Evaluate: Node 59
IF Switch: WF3
End IF
Train
```

**Execute all Modelling Configuration**

```
Delete Old Models
  remove previously trained models
Extract Assay IDs from Database
Get assay_id from the database
Get model configuration table
create modeling_configuration_table
Table Row To Variable Loop Start
Loop End
Select model process definition
Cleanup and collect
Write Models
```

© 2018 KNIME AG. All Rights Reserved.
Model Factory

Abstract Model Factory

Execute all Modelling Configuration

<table>
<thead>
<tr>
<th>assay_id</th>
<th>standard_type</th>
<th>description</th>
<th>assay_chembl_id</th>
<th>target_chembl_id</th>
<th>num_active</th>
<th>num_inactive</th>
<th>active_thresh</th>
<th>inactive_thresh</th>
</tr>
</thead>
<tbody>
<tr>
<td>1520</td>
<td>Ki</td>
<td>Binding affinity for rat cortex 5-hydro...</td>
<td>CHEMBL616154</td>
<td>CHEMBL273</td>
<td>53</td>
<td>5</td>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>1700</td>
<td>IC50</td>
<td>Displacement of [3H]5-HT binding to...</td>
<td>CHEMBL616907</td>
<td>CHEMBL1983</td>
<td>60</td>
<td>0</td>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>2087</td>
<td>Ki</td>
<td>In vitro binding affinity was determin...</td>
<td>CHEMBL616730</td>
<td>CHEMBL1805</td>
<td>55</td>
<td>0</td>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>5728</td>
<td>MIC</td>
<td>Minimum inhibitory concentration aga...</td>
<td>CHEMBL621290</td>
<td>CHEMBL612545</td>
<td>54</td>
<td>0</td>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>6309</td>
<td>ED50</td>
<td>Cytotoxic concentration required to in...</td>
<td>CHEMBL839886</td>
<td>CHEMBL392</td>
<td>72</td>
<td>0</td>
<td>100</td>
<td>1000</td>
</tr>
</tbody>
</table>
Model Factory

Abstract Model Factory

MFP - Select Model
- CASE Switch Data (Start)
  - Node 58
  - CASE 1: existing
  - CASE 2: initialization
  - CASE 3: learn model only
- Select Model
  - Node 59
  - IF Switch
    - WF3
    - END IF
  - IF Switch
    - 1: re-learn model
    - 2: nothing to do
- Select Train Choice
  - END IF
- Score
- Evaluate
- END IF
- Load
  - load the assay data from the db
- Transform
  - WF2
- Learn
- Score
- Evaluate
- Deploy

Execute all Modelling Configuration

Delete Old Models
- remove previously trained models

Extract Assay IDs from Database
- Get assay ID from the database

Get model configuration table
- create modeling configuration table

Table Row To Variable Loop Start
- loop over the assay_ids

Select model process definition

<table>
<thead>
<tr>
<th>Process_Name</th>
<th>Step</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioactivity Prediction</td>
<td>LOAD</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_01_Load</td>
</tr>
<tr>
<td>Bioactivity Prediction</td>
<td>TRANSFORM</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_02_Transform</td>
</tr>
<tr>
<td>Bioactivity Prediction</td>
<td>LEARN</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_03_Learn</td>
</tr>
<tr>
<td>Bioactivity Prediction</td>
<td>SCORE</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_04_Score</td>
</tr>
<tr>
<td>Bioactivity Prediction</td>
<td>EVALUATE</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_05_Evaluate</td>
</tr>
</tbody>
</table>
Model Factory

Abstract Model Factory

Execute all Modelling Configuration

<table>
<thead>
<tr>
<th>Process_Name</th>
<th>Step</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioactivity Prediction</td>
<td>LOAD</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_01_Load</td>
</tr>
<tr>
<td>Bioactivity Prediction</td>
<td>TRANSFORM</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_02_Transform</td>
</tr>
<tr>
<td>Bioactivity Prediction</td>
<td>LEARN</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_03_Learn</td>
</tr>
<tr>
<td>Bioactivity Prediction</td>
<td>SCORE</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_04_Score</td>
</tr>
<tr>
<td>Bioactivity Prediction</td>
<td>EVALUATE</td>
<td>knime://knime.workflow/.../Process/_Process_Step_Templates/workflows/Templates/Bioactivity Prediction_05_Evaluate</td>
</tr>
</tbody>
</table>
Model Factory

Abstract Model Factory

Load
Transform
Select Model
CASE Switch Data (Start)
Select Model
Int Model
CASE Switch 1: existing 2. Initialisation 3. Learn model only
Select Train Choice
Node 58
Node 59
End IF
Score
Evaluate
End IF
IF Switch
WF3
1 (re-)learn model 2 nothing to do
End IF
Train

Execute all Modelling Configuration

Delete Old Models
Extract Assay_IDs from Database
Get model configuration table
Table Row To Variable Loop Start
Select model process definition
Cleanup and collect
Loop End
Write Models

init
load
transform
learn
score
evaluate
deploy
Model Factory
The heart of the factory...

The heart of the factory: Call Local Workflow

The heart of the factory: Call Local Workflow

- Executes another workflow in the same local repository
- Interface is via JSON

Abstract node
Abstract node
Abstract node

Load

Varible to Table Column

Table Row to Variable

Load

Row Splitter

Table Load Workflow

Concatenate

Init Load Transform Learn Score Evaluate Deploy

```
{
  "Process_Name":"Bioactivity Prediction",
  "assay_id":1585583,
  "assay_chembl_id":"CHEMBL3819006",
  "active_thresh":100,
  "inactive_thresh":1000,
  "inactive_file_name":"N/A",
  "Metric":"kappa",
  "Update_Type":"Automatic",
  "Threshold":0.5,
  "Responsible":"anna.martin@knime.com",
  "Last_Execution":"missing",
  "seed_outer":587992
}
```
Abstract node
Abstract node

Load

Transform

Init | Load | Transform | Learn | Score | Evaluate | Deploy
Abstract node

Load

Transform

Init Load Transform Learn Score Evaluate Deploy

© 2018 KNIME AG. All Rights Reserved.
Extracting the data

- **Data source:** ChEMBL 23
- **Activity types:**
  - ('GI50', 'IC50', 'Ki', 'MIC', 'EC50', 'AC50', 'ED50', 'GI', 'Kd', 'CC50', 'LC50', 'MIC90', 'MIC50', 'ID50') -> 6.5 million points
- **Define active:**
  - Standard_value < 100nM -> 1.3 million actives
- **Define an interesting assay**
  - At least 50 actives -> 1556 assays
- **Define inactive:**
  - Standard_value > 1uM
- **Final dataset size:** 2.5 million data points, 1.5 million compounds
Finding more inactives

• The ChEMBL datasets almost all have an unrealistically high ratio of actives to inactives
• “Fix” that by adding enough assumed inactives to each dataset to get a 1:10 active:inactive ratio
• Pick those assumed inactives to be roughly the same “size” as the actives: *tanimoto* similarity of at least 0.7 using RDKit ECFC0 fingerprints
Extracting the data

Init Load Transform Learn Score Evaluate Deploy
Transform

- Convert SMILES from database into chemical structures
- Cleanup the chemical structures
Transform

- Convert SMILES from database into chemical structures
- Cleanup the chemical structures
- Generate several chemical fingerprints for each structure
Learn

• Four model types
• Parameter optimization
EXAMPLES Server:
04_Analytics/11_Optimization/06_Parameter_Optimization_two_examples*
Parameter Optimization

Flow variable with objective function value

Function should be...

- maximized
- minimized
Parameter Optimization

Init  Load  Transform  Learn  Score  Evaluate  Deploy

Parameter Optimization Loop Start
Compute additional method parameters
Tree Ensemble Learner

WrappedNode Input
Partitioning
stratified 80-20

Tree Ensemble Predictor
Compute Kappa & AUC
Parameter Optimization Loop End
Collect Statistics
WrappedNode Output

Brute Force
String Manipulation (Variable)
method
Pick the Best Model

Init → Load → Transform → Learn → Score → Evaluate → Deploy

- Naive Bayes
- Logistic Regression
- Random Forest
- H2O Gradient Boosting

Partitioning

80/20

Pick the Best Model

Build Model with Selected Parameters

Score Model with External Data
Pick the Best Model

- Naive Bayes
- Logistic Regression
- Random Forest
- H2O Gradient Boosting

Pick the Best Model

Build Model with Selected Parameters

Score Model with External Data

Init > Load > Transform > Learn > Score > Evaluate > Deploy
Pick the Best Model
Pick the Best Model
On to deploy...
Model Report for Assay: CHEMBL1212560

Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>active</td>
</tr>
<tr>
<td>active</td>
<td>11</td>
</tr>
<tr>
<td>inactive</td>
<td>3</td>
</tr>
</tbody>
</table>

Performance:

- **AUC:** 0.953
- **Enrichment factor at 1%, 5%, 10%:** 8.333, 6.667, 7.500
- **Balanced Accuracy:** 0.945
- **Cohen’s kappa:** 0.829
- **F-Score:** 0.846

Assay Details

- **Description:** Antimicrobial activity against Pseudomonas aeruginosa ATCC 13525 by MTT assay
- **standard_type:** MIC
- **actives:** 58; **measured inactives:** 0; **synthetic inactives:** 116

Model Details

- **Fingerprints:** ECFP4
- **Method:** Random Forest
- **trees:** 200; **depth:** 15; **min_leaf_size:** 3
Model Report for Assay: CHEMBL1212560

Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>active</td>
<td>inactive</td>
</tr>
<tr>
<td>Real</td>
<td></td>
<td></td>
</tr>
<tr>
<td>active</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>inactive</td>
<td>3</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>0.79</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Performance:

- **AUC**: 0.953
- **Enrichment factor at 1%, 5%, 10%**: 8.333, 6.667, 7.500
- **Balanced Accuracy**: 0.945
- **Cohen's kappa**: 0.829
- **F-Score**: 0.846

Assay Details

- **Description**: Antimicrobial activity against Pseudomonas aeruginosa ATCC 13525 by MTT assay
- **standard_type**: MIC
- **actives**: 58; **measured inactives**: 0; **synthetic inactives**: 116

Model Details

- **Fingerprints**: ECFP4
- **Method**: Random Forest
- **trees**: 200; **depth**: 15; **min_leaf_size**: 3
## Model Report

<table>
<thead>
<tr>
<th>assay_id</th>
<th>description</th>
<th>PortObject</th>
<th>FPs</th>
<th>method</th>
<th>forest...</th>
<th>AUC_...</th>
<th>Cohen...</th>
</tr>
</thead>
<tbody>
<tr>
<td>643696</td>
<td>Antimicrobial act... 200 classifiers... ECFP4</td>
<td></td>
<td></td>
<td>RF</td>
<td>200</td>
<td>0.953</td>
<td>0.829</td>
</tr>
</tbody>
</table>

![Diagram of the model](image-url)
Model Report

<table>
<thead>
<tr>
<th>assay_id</th>
<th>description</th>
<th>PortObject</th>
<th>FPs</th>
<th>method</th>
<th>forest...</th>
<th>AUC...</th>
<th>Cohen...</th>
</tr>
</thead>
<tbody>
<tr>
<td>643696</td>
<td>Antimicrobial act... 200 classifiers... ECFP4</td>
<td></td>
<td></td>
<td>RF</td>
<td>200</td>
<td>0.953</td>
<td>0.829</td>
</tr>
</tbody>
</table>
### Model Report

<table>
<thead>
<tr>
<th>assay_id</th>
<th>description</th>
<th>PortObject</th>
<th>FPs</th>
<th>method</th>
<th>forest...</th>
<th>AUC...</th>
<th>Cohen...</th>
</tr>
</thead>
<tbody>
<tr>
<td>643696</td>
<td>Antimicrobial act... 200 classifiers... ECFP4</td>
<td>RF</td>
<td>200</td>
<td>0.953</td>
<td>0.829</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model Report

<table>
<thead>
<tr>
<th>assay_id</th>
<th>description</th>
<th>PortObject</th>
<th>FPs</th>
<th>method</th>
<th>forest...</th>
<th>AUC_...</th>
<th>Cohen...</th>
</tr>
</thead>
<tbody>
<tr>
<td>643696</td>
<td>Antimicrobial act... 200 classifiers... ECFP4</td>
<td>RF</td>
<td>200</td>
<td>0.953</td>
<td>0.829</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Generic JavaScript View

WrappedNode Input

Confusion Matrix

Collect Variables

CASE Switch Variable (Start)

CASE Switch Variable (End)

Text Output

Text Output

Text Output

Title

Assay Info

H2O

WrappedNode Output
Deploy

- RESTful web service:

```json
"json-output-102:2": [
  {
    "canonical_smiles": "c1ccc(O)cc1N",
    "ids": "mol1",
    "Salt Stripped Molecule": "Nc1cccc(O)c1",
    "assay_104094-method_NB-P(activity=active)": 1,
    "assay_104094-method_NB-P(activity=inactive)": 0,
    "assay_104094-method_NB-Prediction (activity)": "inactive",
    "assay_105108-method_NB-P(activity=inactive)": 0,
    "assay_105108-method_NB-P(activity=active)": 1,
    "assay_105108-method_NB-Prediction (activity)": "inactive",
    "assay_107296-method_RF-Prediction (activity)": "inactive",
    "assay_107296-method_RF-P(activity=active)": 0,
    "assay_107296-method_RF-P(activity=inactive)": 1,
    "assay_110580-method_NB-P(activity=active)": 1,
    "assay_110580-method_NB-P(activity=inactive)": 0,
    "assay_110580-method_NB-Prediction (activity)": "inactive",
    ...
  }
]```
Results: the Heatmap
Results
Results
Results: the Heatmap
Results: the Heatmap

Compound molecule
Results: the Heatmap

<table>
<thead>
<tr>
<th>ChEMBL assay ID</th>
<th>Compound ID</th>
<th>Activity prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>170355</td>
<td>170584</td>
<td></td>
</tr>
<tr>
<td>170646</td>
<td>170563</td>
<td></td>
</tr>
<tr>
<td>170498</td>
<td>170730</td>
<td></td>
</tr>
<tr>
<td>170899</td>
<td>170782</td>
<td></td>
</tr>
<tr>
<td>170354</td>
<td>170475</td>
<td></td>
</tr>
</tbody>
</table>

P (active) = 0.125
Results: the Heatmap

Assay info

Escherichia coli DNA cleavage endpoint value which is the minimum amount required (in micro g/mL) to induce detectable cleavage of supercoiled pBR322 substrate to linear form (gymase)

AUC: 0.949
Number of actives: 52

more statistics...
Results: the Heatmap

Assay info

Escherichia coli DNA cleavage endpoint value which is the minimum amount required (in micro g/mL) to induce detectable cleavage of supercoiled pBR322 substrate to linear form (gyrase)

AUC: 0.988
Number of actives: 52

[Link to more statistics]
Model Report for Assay: CHEMBL1212560

Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>active</td>
<td>11</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>inactive</td>
<td>3</td>
<td>113</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performance:

- **AUC**: 0.953
- **Enrichment factor at 1%, 5%, 10%**: 8.333, 6.667, 7.500
- **Balanced Accuracy**: 0.945
- **Cohen’s kappa**: 0.829
- **F-Score**: 0.846

Assay Details

- **Description**: Antimicrobial activity against Pseudomonas aeruginosa ATCC 13525 by MTT assay
- **standard_type**: MIC
- **actives**: 58; **measured inactives**: 0; **synthetic inactives**: 116

Model Details

- **Fingerprints**: ECFP4
- **Method**: Random Forest
- **trees**: 200; **depth**: 15; **min_leaf_size**: 3
Wrapping up

- Identified >1500 datasets from ChEMBL
- Created workflows for:
  - automating dataset extraction from the database
  - generating features for a dataset
  - automated model building and validation
  - deploying models that pass validation
- Automate the execution of that using the Model Factory
- Deployed models to KNIME Server as:
  - WebPortal application for model details
  - A REST service to generate predictions
  - WebPortal application for viewing predictions
Next steps

• Last bits of testing and tweaking of the workflows and views
• Do the big model building run using the new KNIME distributed executors
• Put workflows on the EXAMPLES server
• Analysis of the results
• Blog post(s)/white paper
• Repeating the whole process when the next version of ChEMBL is released.