

KNIMEを使った 材料探索 基本操作(1)

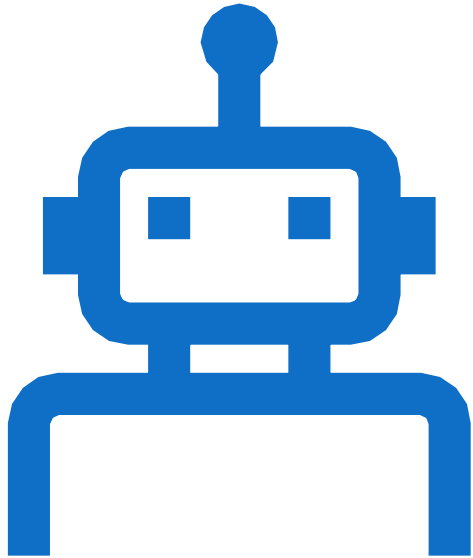
早稲田大学 応用化学科

講師(任期付) 畠山 歓

<https://github.com/KanHatakeyama/>

satokan@toki.早稲田.jp

マテリアルズ・ インフォマティクス (MI)



- 機械学習を使った材料探索
- 人間の代わりにAIが新規材料を見つけられるかもしれない
- 材料情報をどのようにAIに認識させるか、が課題

書籍

機械学習 (色々あります)

- 東京大学のデータサイエンティスト育成講座

マテリアル・インフォマティクス

(未だ種類があまりありません)

- マテリアルズ・インフォマティクス-材料開発のための機械学習超入門-
- 化学のための Pythonによるデータ解析・機械学習入門
- 実践 マテリアルズインフォマティクス など

どう
実践する
か？

Python (プログラミング言語)

- 機械学習のデファクトスタンダード
- 最先端のコードを利用出来る
- 入門者にとっては敷居が高い (Python を使えるようになるまで数十時間...?)

専用のソフトウェア

- 自由度は下がるが、敷居も格段に低い
- 概観を知るには丁度良い
- **KNIME(無料)**、SPSS(有料)、...

KNIME

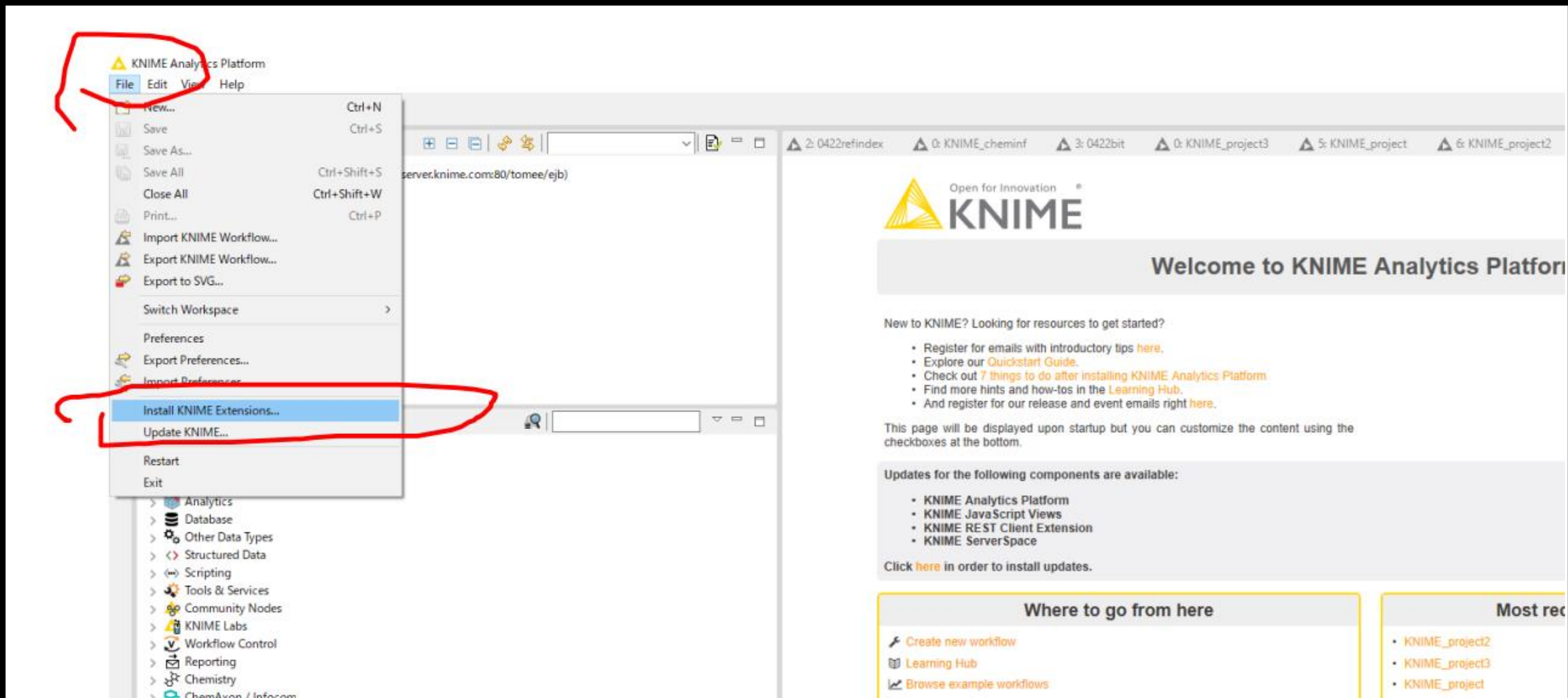
- あらゆるデータの連携・統合・分析を自動化するオープンソースプラットフォーム ([出典](#))
- 機械学習や統計処理に特化した無料ツール
- 化合物を扱う為のExtensionも無料で利用可能
- 必要に応じ、Pythonのコードも追加出来るので便利

ダウンロードはこちら

- <https://www.knime.com/downloads/download-knime>
- Win, Mac, Linuxに対応
- 以後、スライドはWindows10で動作確認

化合物を扱う Extensionの インストール

RDKitと呼ばれるモジュール

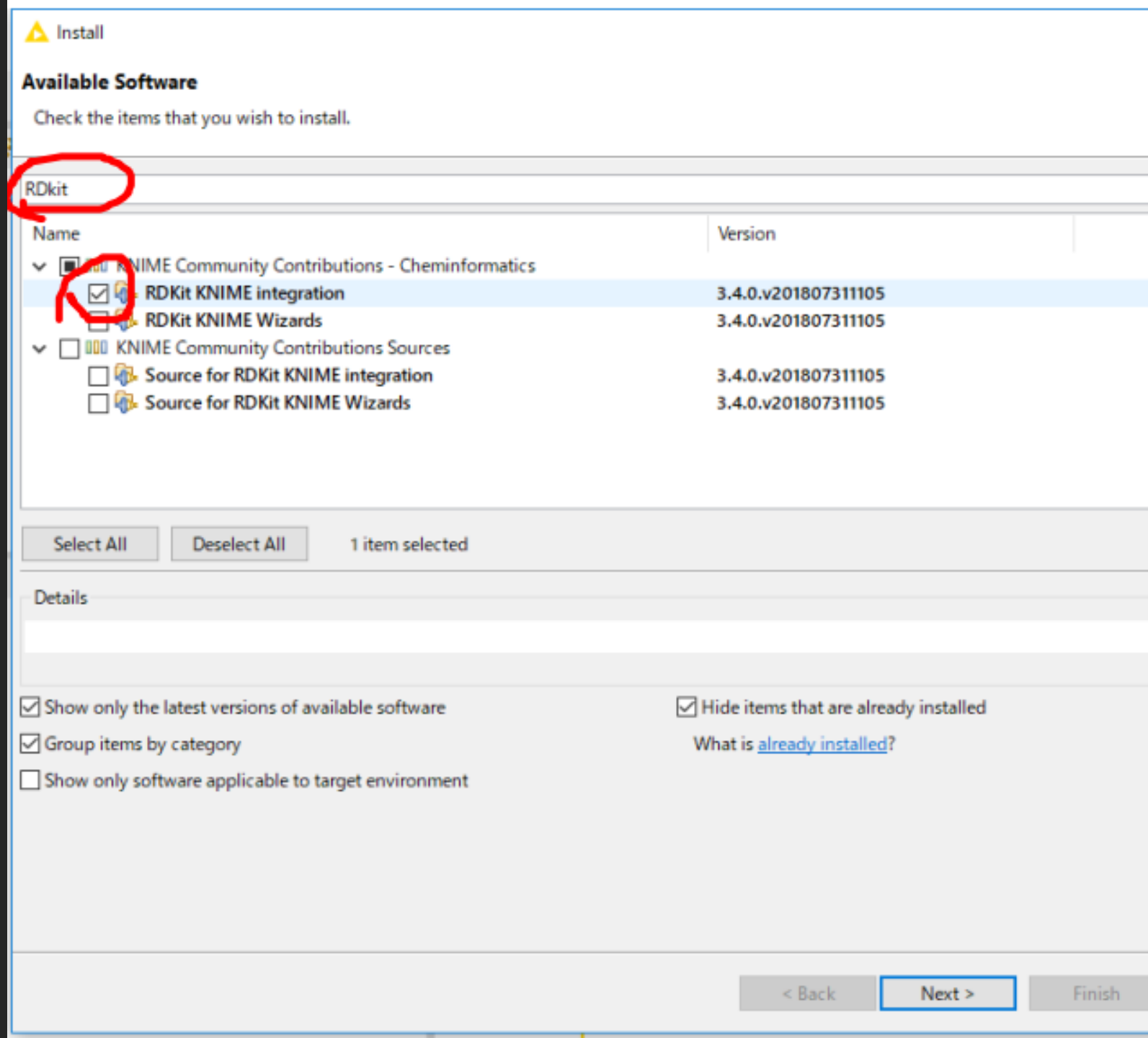


KNIME起動→File→Install KNIME Extensions

RDKitと入力

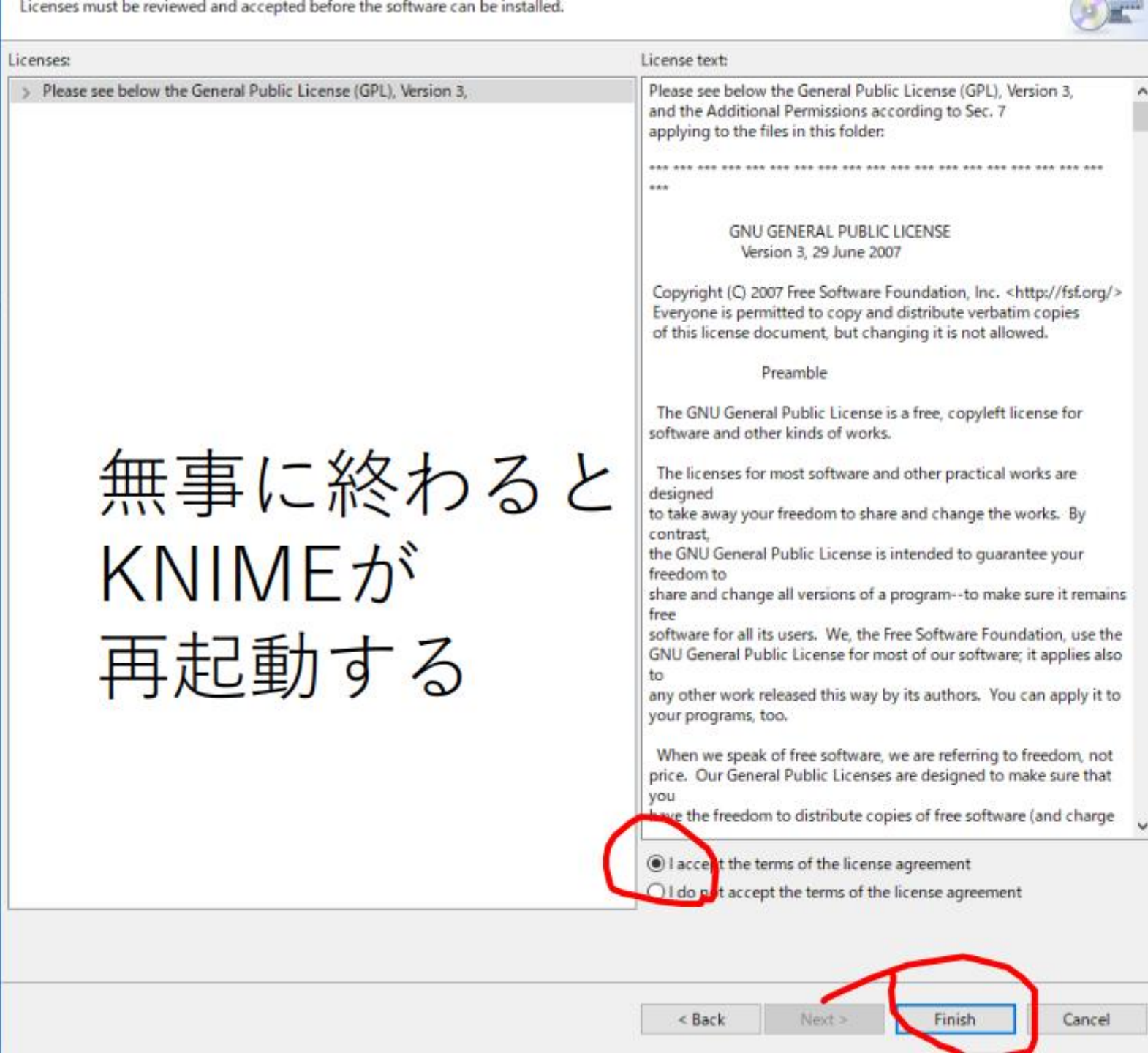
→

RDKit KNIME integratioを チェック

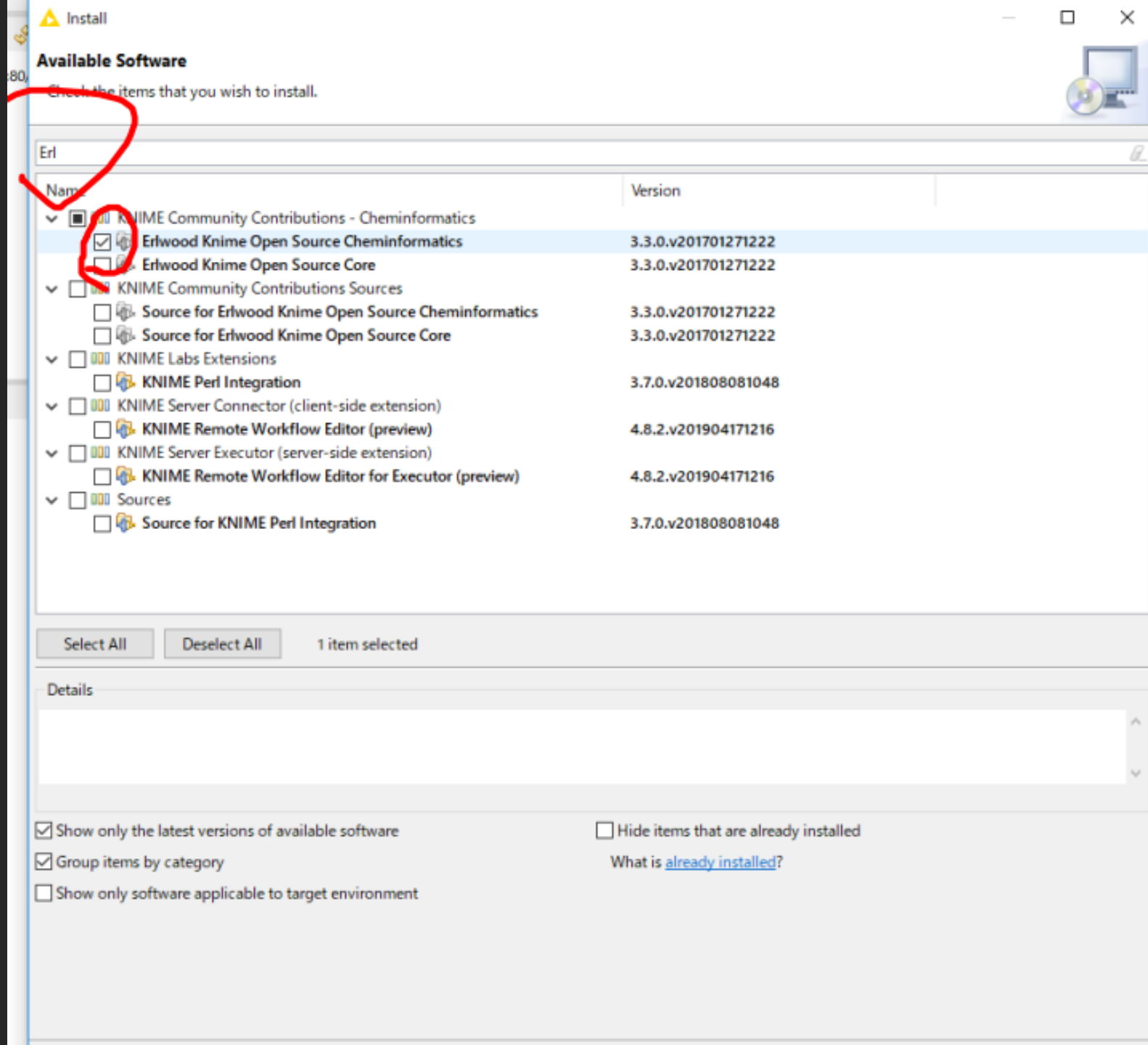


利用規約に 同意

無事に終わると
KNIMEが
再起動する



再起動後、
ついでに
こちらも
Install
すると便利

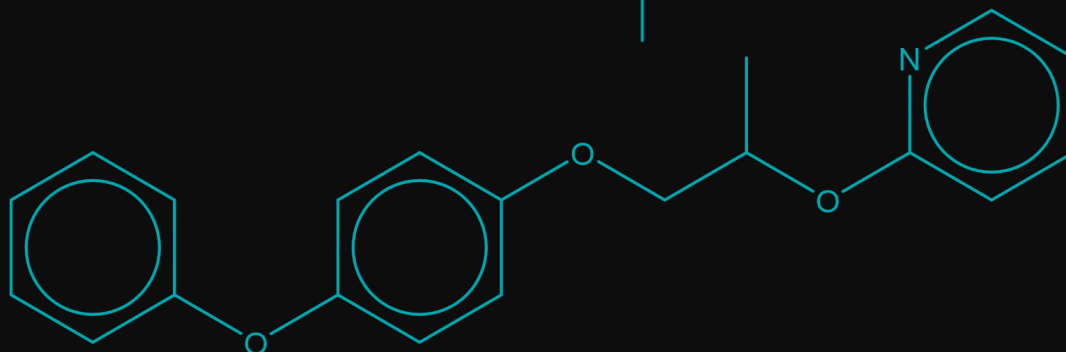
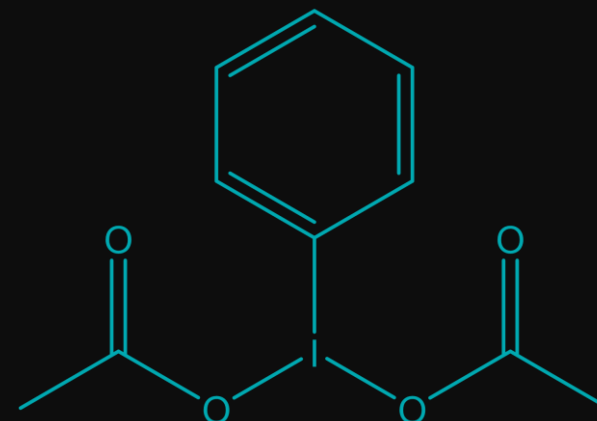
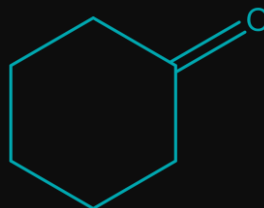


はじめてのMI

KNIMEで 化合物の融点を 予測

今回のタスク

- Wikipediaに掲載されている1000種類程度のデータを機械学習する
- 以下の化合物の融点を予測する

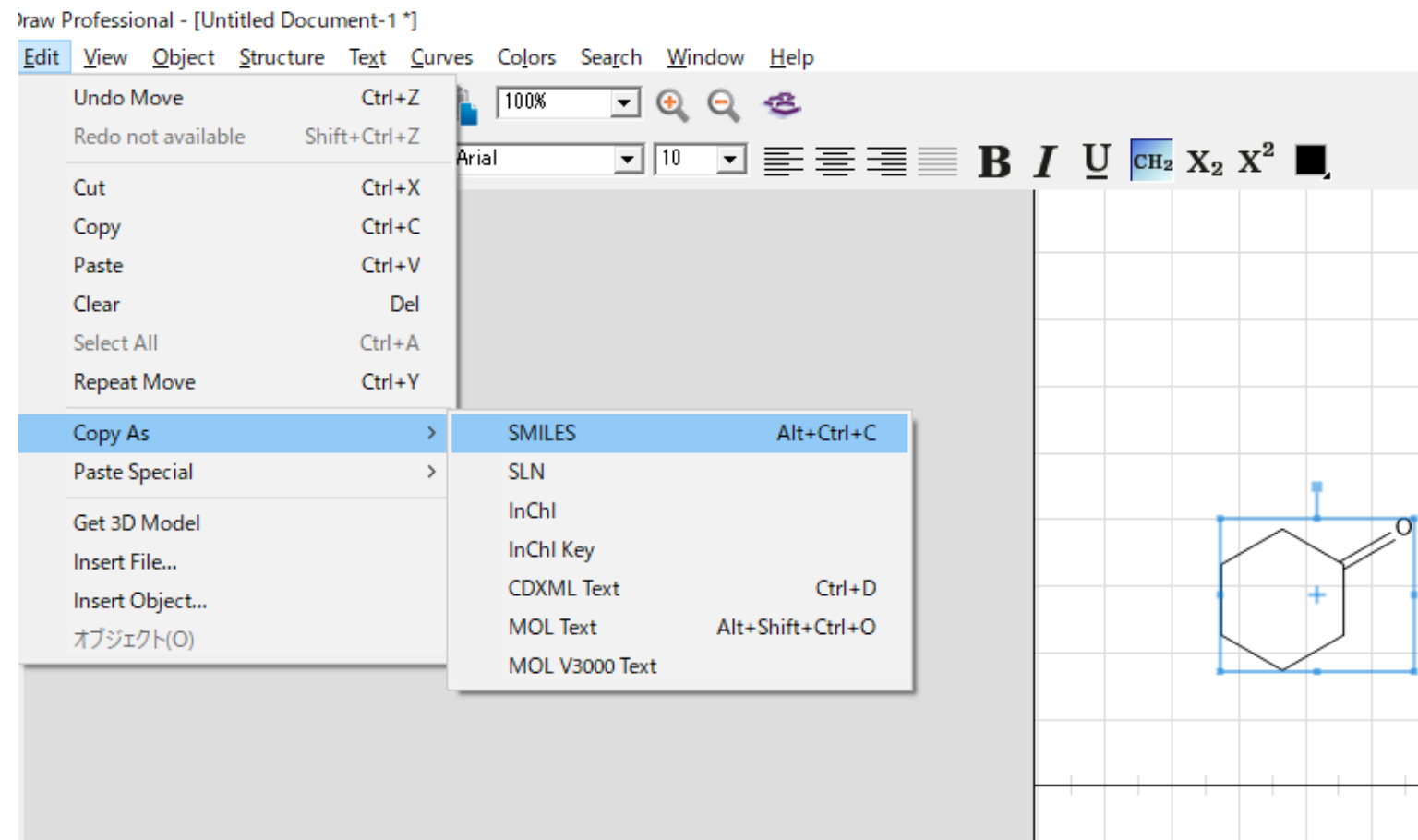


これらの構造を見ただけで
融点が予測出来る方には
MIは不要かもしれません

データベース

- [ここ](#)からダウンロード出来ます
- Excelに纏まっています
- 化学構造はSMILESという表記法で格納されています
- 融点が記録されています

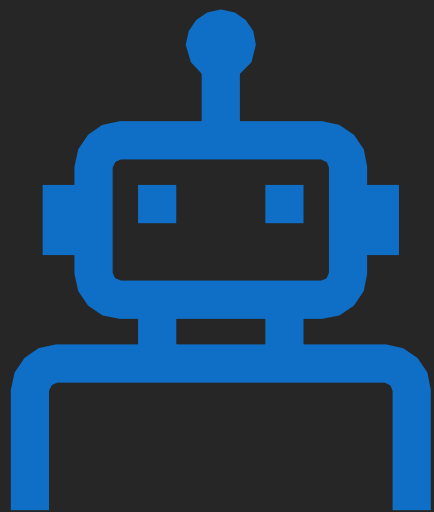
A	B	C	D
ID	SMILES	Melting temperature	
1	[Cu]=S	500	
2	c1cc2ccc3c	117	
3	O1[Fe]2O	1539	
4	O=C1NC(=	245	
5	P#[Y]	200.78	
6	C1=CC=C(c	290	
7	ClC(Cl)C(=	98	
8	FC(F)F	-155.2	
9	O=[N+]([C	108	
10	CCC[C@@	86	
11	c1ccc2c/c1	-30	



- 化学構造のままだとExcelに記録するのが難しいので、SMILESと呼ばれるアルゴリズムで文字列に変換します
- ChemDraw等のソフトウェアで可逆に変換出来ます

SMILES とは?

simplified molecular input line entry system



スキーム

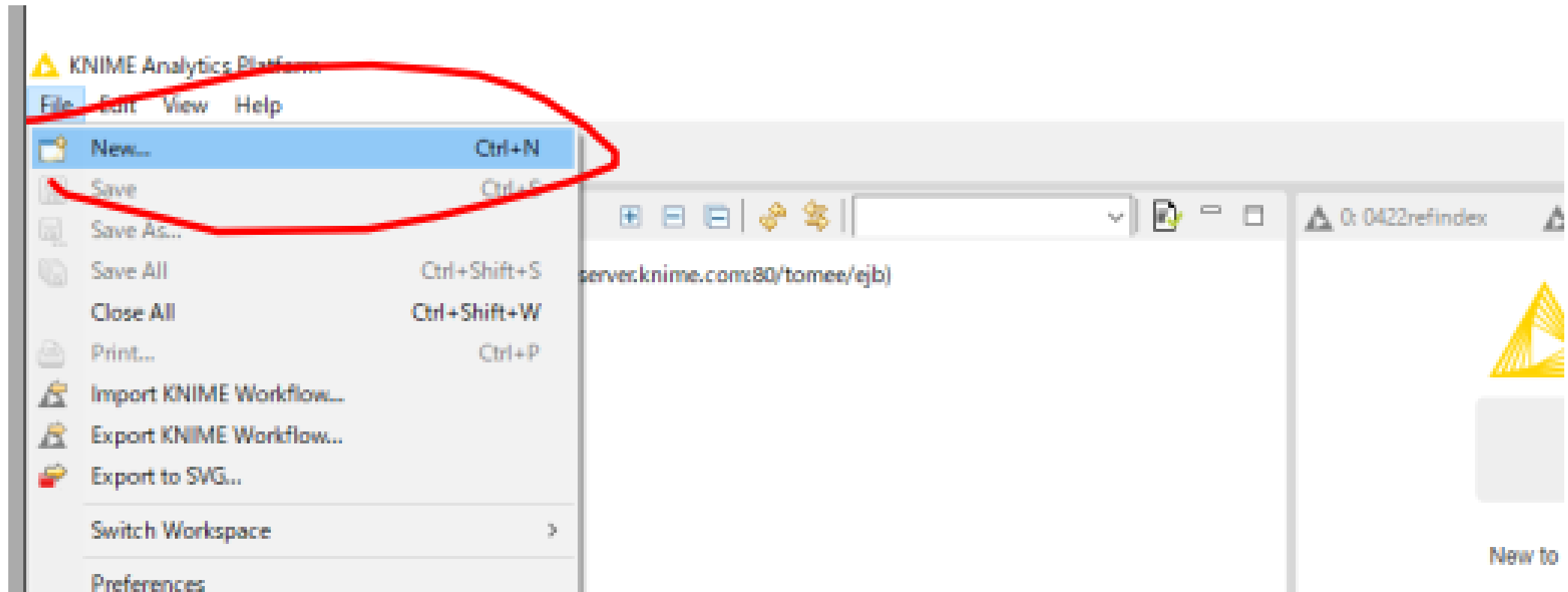
—

1. データベースの読み込み
2. 化学構造の数値化
3. 構造と融点の関係性を機械学習
4. 未知化合物の融点を予測

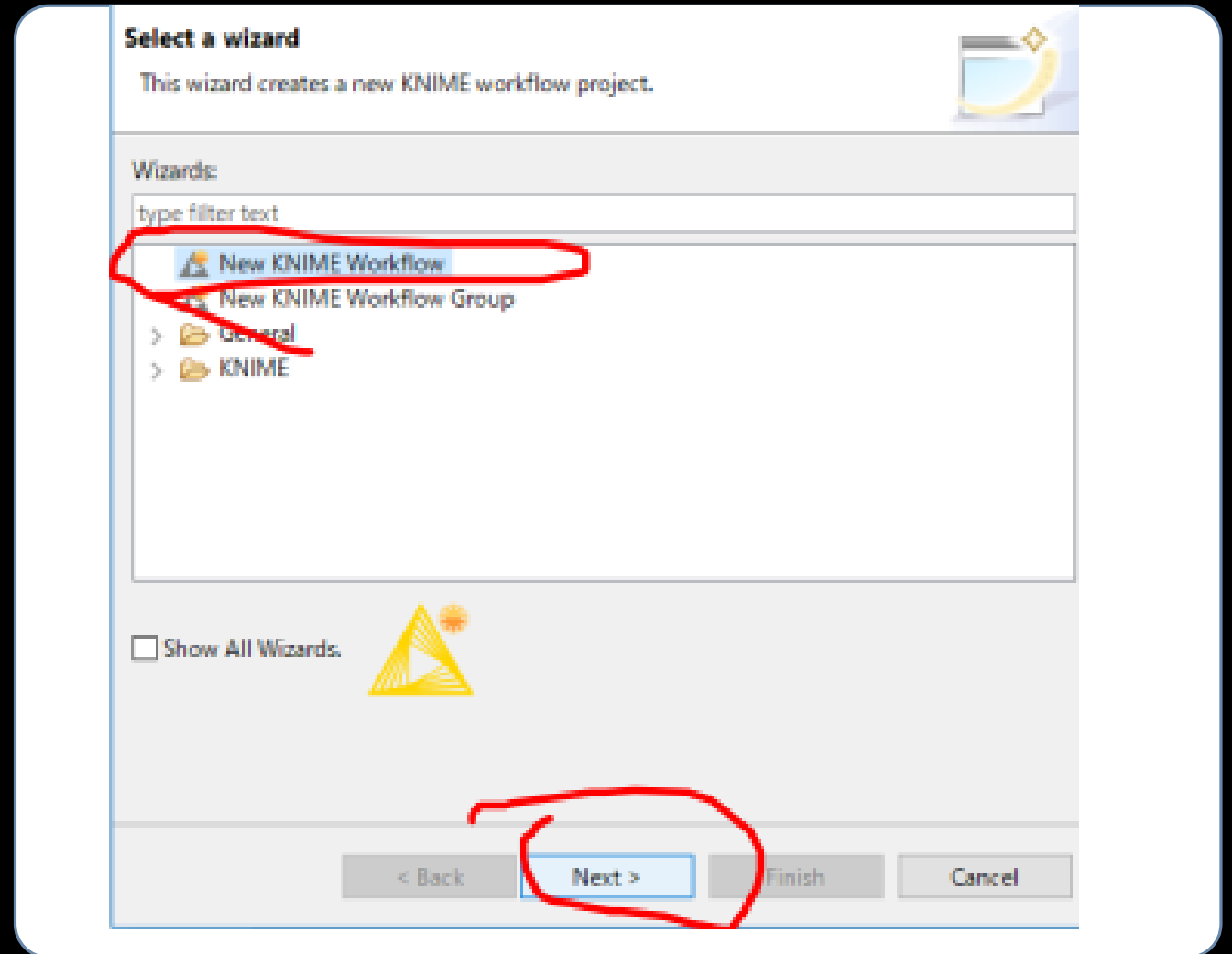
KNIMEの操作

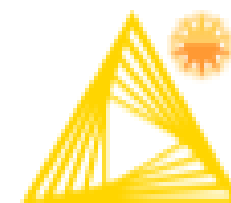
まずは真似してみよう

新規プロジェクト



新規Workflow





New KNIME Workflow Wizard

Create a new KNIME workflow.

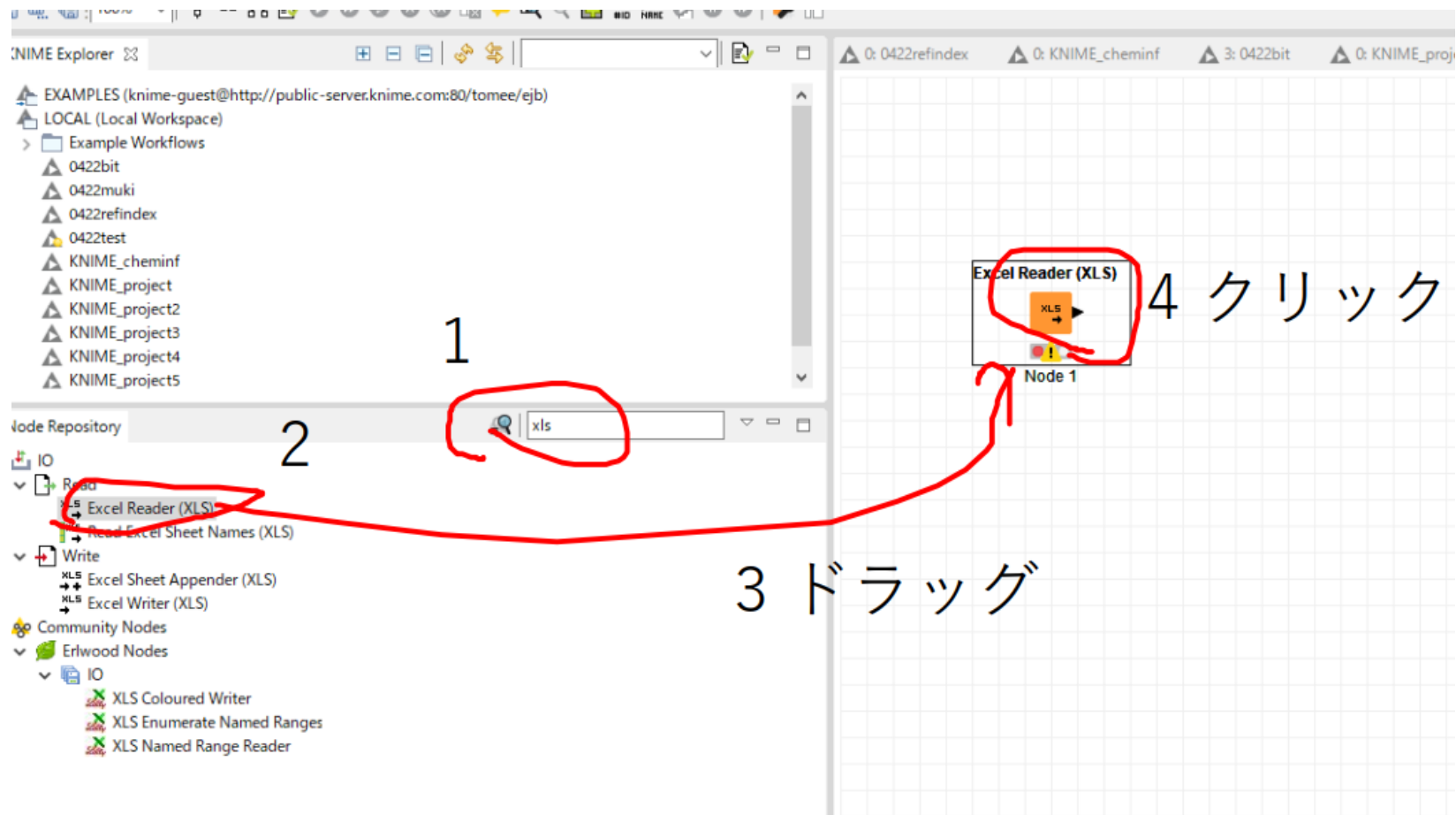
Name of the workflow to create: 0422test

Destination of new workflow : LOCAL:/

Browse...

好きなファイル名

ここにノードと呼ばれるオブジェクトを配置
していく



Excel readerの配置

XLS Reader Settings | Flow Variables | Memory Policy

Select file to read:

Browse...

Adjust Settings:

Select the sheet to read: Connect timeout [s]:

Column Names:

☐ Table contains column names in row number: (Row numbers start with 1. Mouse over header to see row number.)

Row IDs:

☒ Generate RowIDs (index incrementing, starting with 'Row0') ☐ Generate RowIDs (index as per sheet content, skipped rows will increment index)

☐ Table contains row IDs in column: ☐ Make row IDs unique

Select the columns and rows to read:

☒ Read entire data sheet, or ...

read columns from: to:

and read rows from: to:

Tip: Mouse over the column and row headers in the "File Content" tab to identify cell coordinates

On evaluation error:

☒ Insert an error pattern:

☐ Insert a missing cell

More Options:

☒ Skip empty columns ☐ Reevaluate formulas (leave unchecked if uncertain; see node description for details)

☒ Skip hidden columns ☐ Disable Preview (does not compute the output table structure)

☒ Skip empty rows

Preview | File Content

Preview with current settings

Set a filename.

読み込むexcelを選択

先ほどダウンロードしたファイル
(wikipedia_db.xlsx)を選択

データベースの読み込み

※2の意味

Excelの1行目は実際の生データではなく
ID, SMILES等の文字列(カラム名)が記録されているので
1行目はデータとしては読み込まない、という指示

1. ファイルが選択されていることを確認

The screenshot shows the 'Select file to read' dialog box. The file path 'wikipedia_db.xlsx' is selected and circled in red. Below, the 'Adjust Settings' section has 'Table contains column names in row number: 1' checked and circled in red. The 'Row IDs' section has 'Generate RowIDs (index incrementing, starting with 'Row0')' selected. The 'Select the columns and rows to read' section has 'Read entire data sheet, or ...' checked. The 'On evaluation error' section has 'Insert an error pattern: #XL_EVAL_ERROR#' selected. The 'More Options' section has 'Skip empty columns', 'Skip hidden columns', and 'Skip empty rows' checked. The 'Preview' tab is active, showing a table of data. The 'refresh' button is circled in red. The 'OK' button at the bottom is also circled in red.

2. チェックする

3. 押す

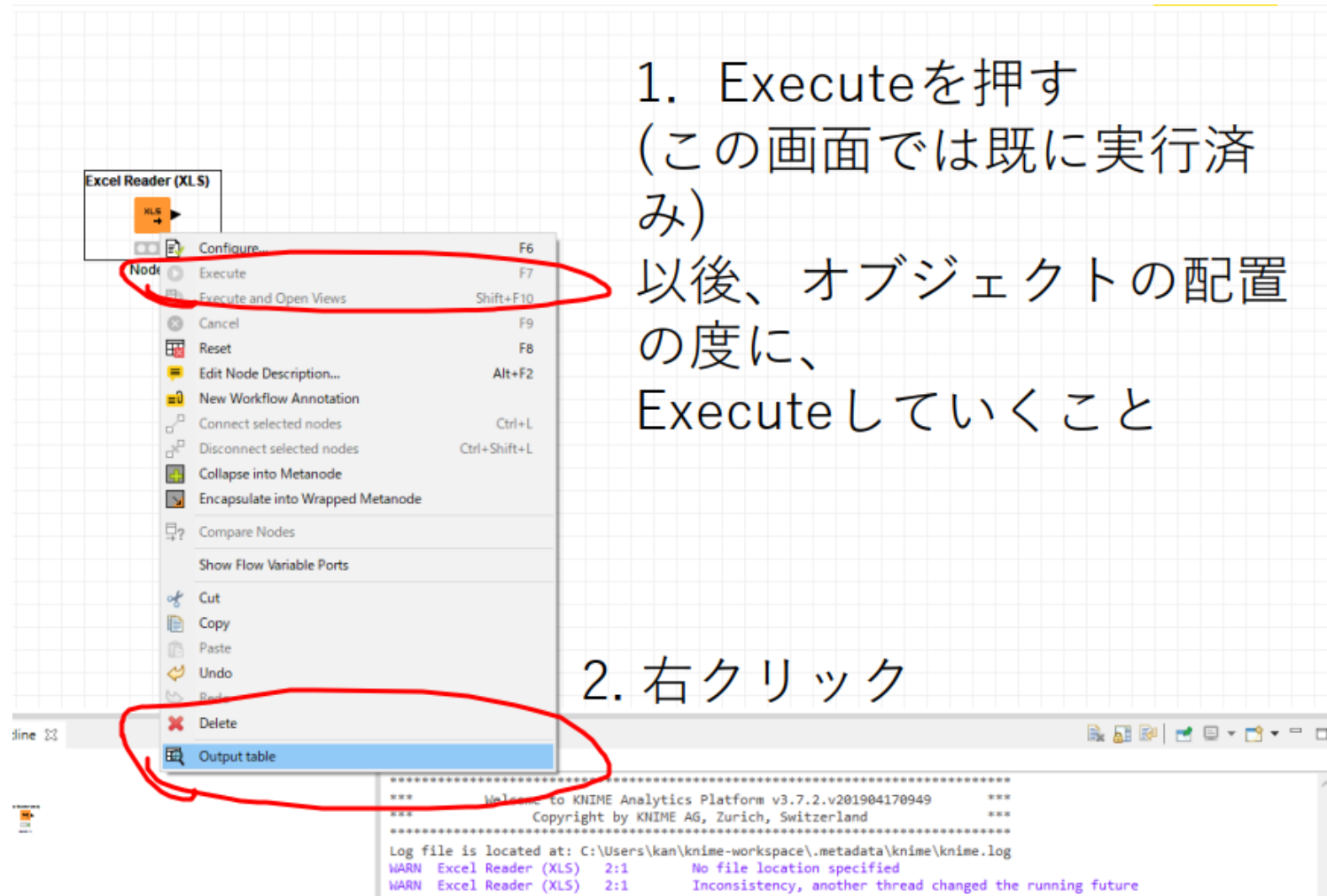
4. 押す

Row ID	ID	SMILES	Melting ...
Row0	1	[Cu]=S	500
Row1	2	c1cc2ccc3cc...	117
Row2	3	O1[Fe]2O[F...	1,539
Row3	4	O=C1NC(=O...	245
Row4	5	P#[Y]	200.78
Row5	6	C1=CC=C(C...	290
Row6	7	ClC(C1)C(=O...	98
Row7	8	FC(F)F	-155.2
Row8	9	O=[N+](O-)	108
Row9	10	CCC[C@@H]...	86
Row10	11	c1ccc2c(c1)...	-30
Row11	12	O=C(O)[C@...	285
Row12	13	F[Co](F)F	927
Row13	14	[Cs+].[I-]	632
Row14	15	C1CCC(CC1...	4
Row15	16	O=C2c3c(O...	251
Row16	17	BrC(F)(F)F	-167.78

ノードの 実行

1. Executeを押す
(この画面では既に実行済み)
以後、オブジェクトの配置の度に、
Executeしていくこと

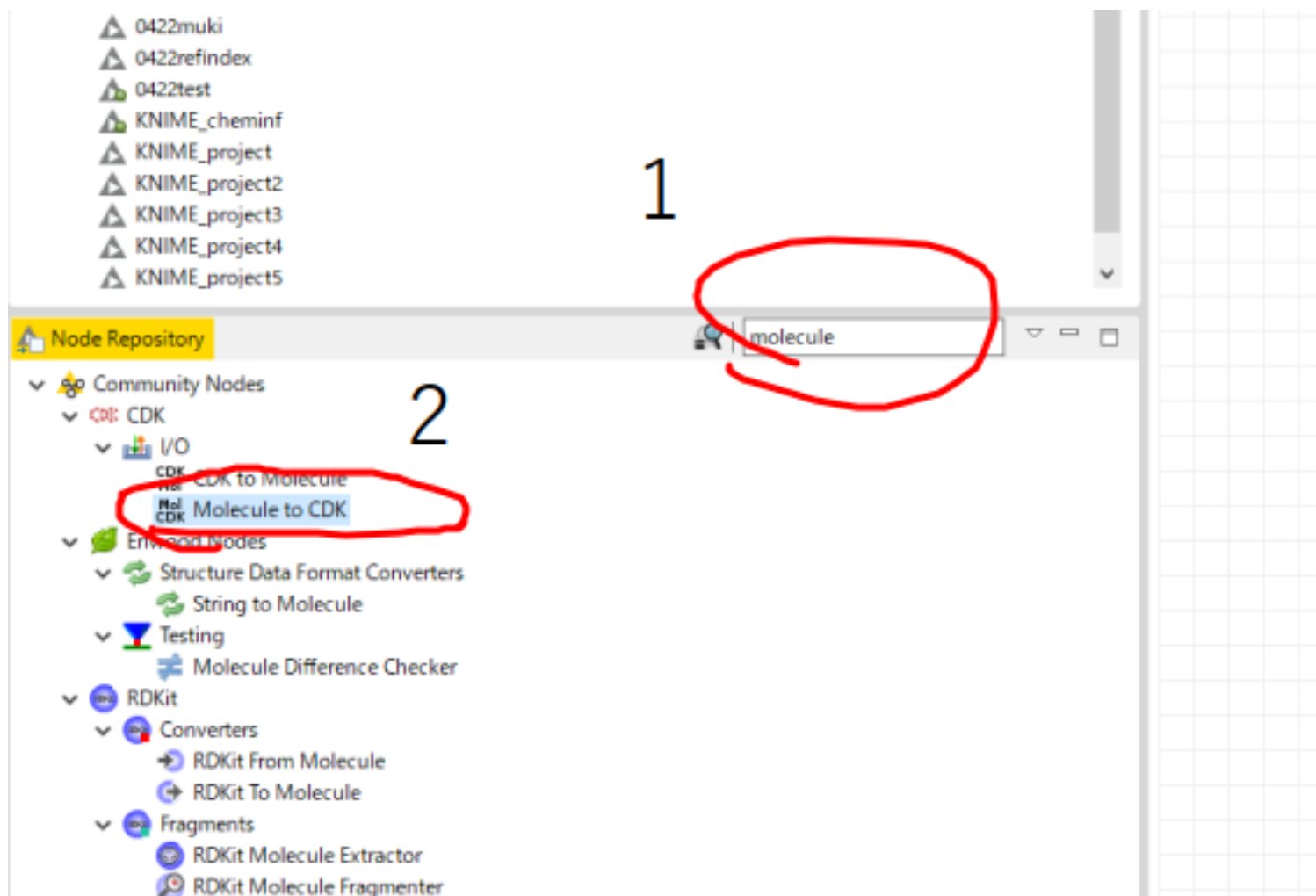
2. 右クリック

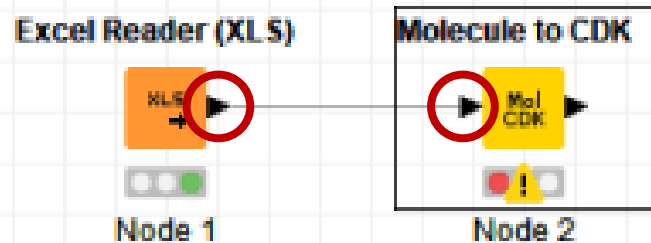


Row ID	I ID	S SMILES	D Melting ...
Row0	1	[Cu]=S	500
Row1	2	c1cc2ccc3cc...	117
Row2	3	O1[Fe]2O[F...	1,539
Row3	4	O=C1NC(=O...	245
Row4	5	P#[Y]	200.78
Row5	6	C1=CC=C(C...	290
Row6	7	ClC(Cl)C(=O...	98
Row7	8	FC(F)F	-155.2
Row8	9	O=[N+](O-)...	108
Row9	10	CCC[C@@H]...	86
Row10	11	c1ccc2c(c1)...	-30
Row11	12	O=C(O)[C@...	285
Row12	13	F[Co](F)F	927
Row13	14	[Cs+].[I-]	632
Row14	15	C1CCC(CC1...	4
Row15	16	O=C2c3c(O[...	251
Row16	17	BrC(F)(F)F	-167.78
Row17	18	FCC(F)(F)F	-103.3
Row18	19	[H]1[BH]2[H...	-46.8
Row19	20	C(Nc1ncnc2...	269
Row20	21	C1CC2C(O2...	-108.9
Row21	22	OC[C@H](O)...	145
Row22	23	c2(=C=O)c1...	122

データが
読み込ま
れている

分子を 処理する ノード の設置

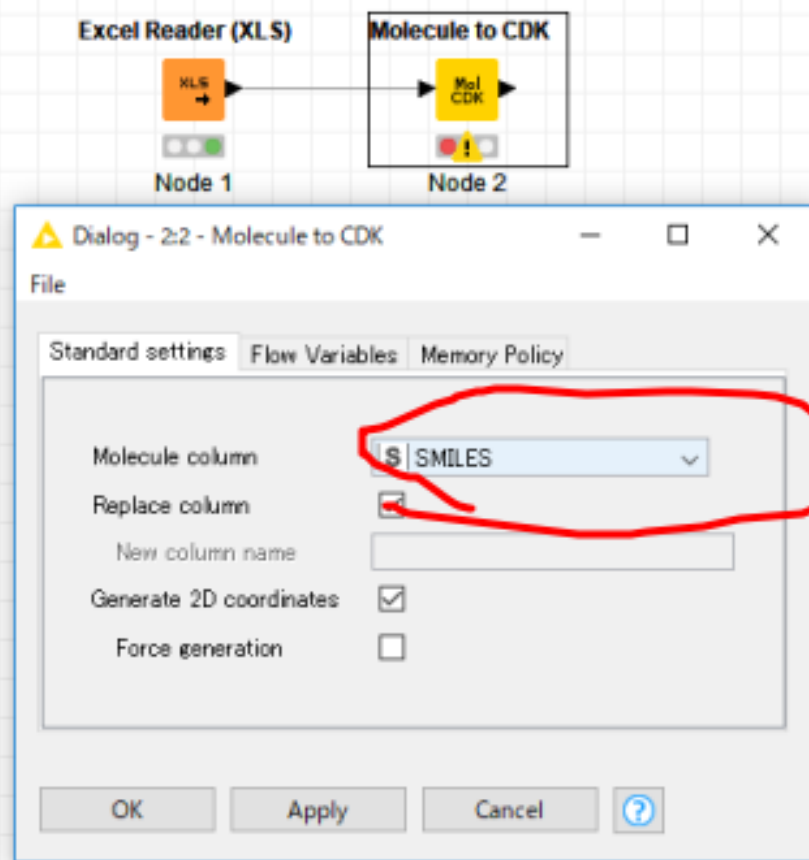




三角マークと三角マークをドラッグして繋げます

1. 配置して結合
2. 左クリックして設定画面を開く

ノードの結合



SMILESカラムを読み込む
ように設定されていることを
確認

設定確認


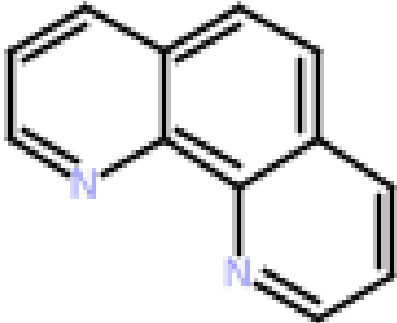
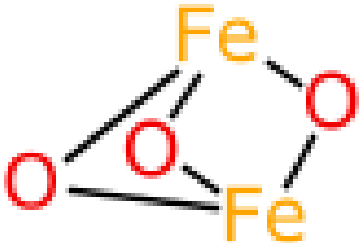
Node 1

- Configure... F6
- Execute F7
- Execute and Open Views Shift+F10
- Cancel F9
- Reset F8
- Edit Node Description... Alt+F2
- New Workflow Annotation
- Connect selected nodes Ctrl+L
- Disconnect selected nodes Ctrl+Shift+L
- Collapse into Metanode
- Encapsulate into Wrapped Metanode
- Compare Nodes
- Show Flow Variable Ports
- Cut
- Copy
- Paste
- Undo
- Redo
- Delete
- Parsed molecules**

1. execute
2. 右クリック
3. Parsed molecules
をクリック

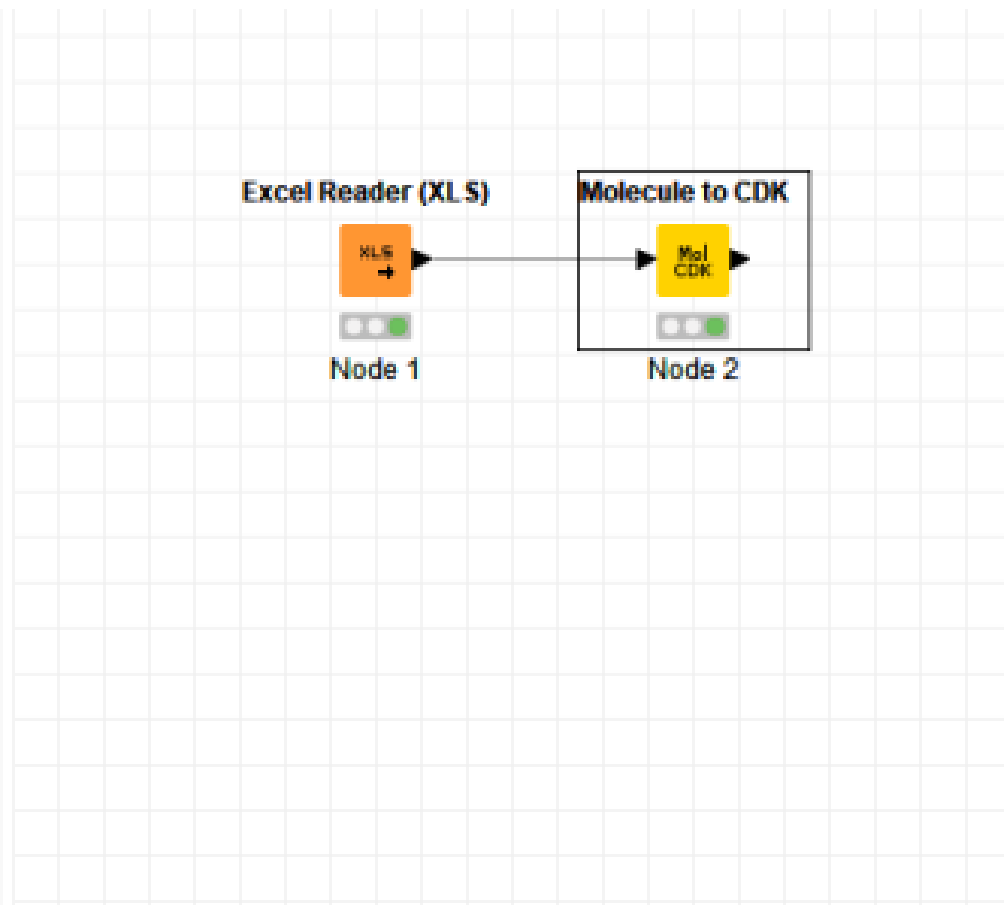
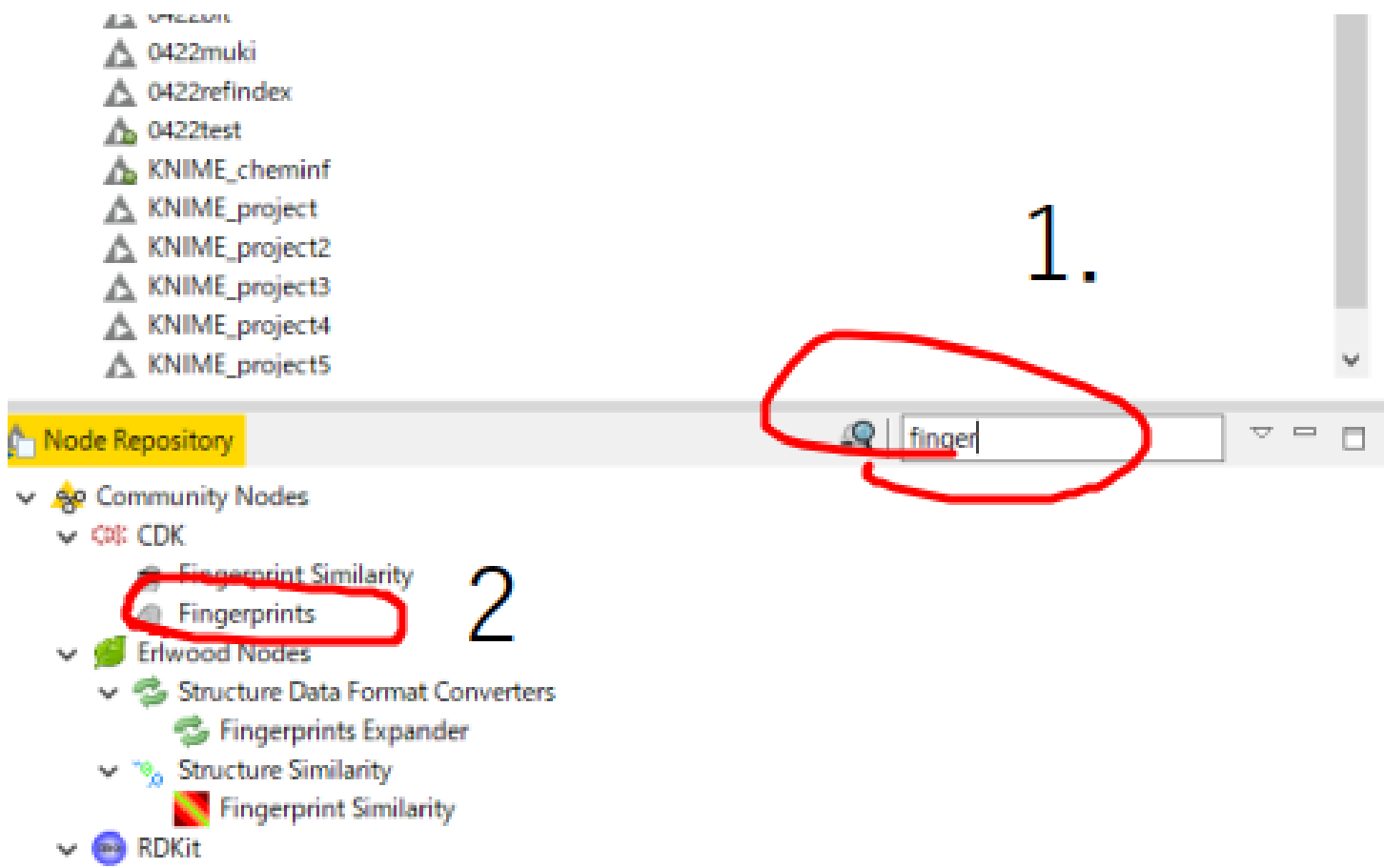
Outline

KNIMEが
SMILESから
分子構造を
認識!

Row ID	ID	SMILES	Melting ...
Row0	1	 <chem>[Cu]=S</chem>	500
Row1	2	 <chem>c1ccc2c(c1)cnc3ccccc23</chem>	117
Row2	3	 <chem>[Fe]([O])([O])[Fe]([O])</chem>	1,539

化学構造を数値化する ノードの設置

- Fingerprintと呼ばれる考え方を 사용합니다 (後述)



Excel Reader (XLS)



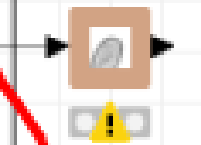
Node 1

Molecule to CDK



Node 2

Fingerprints

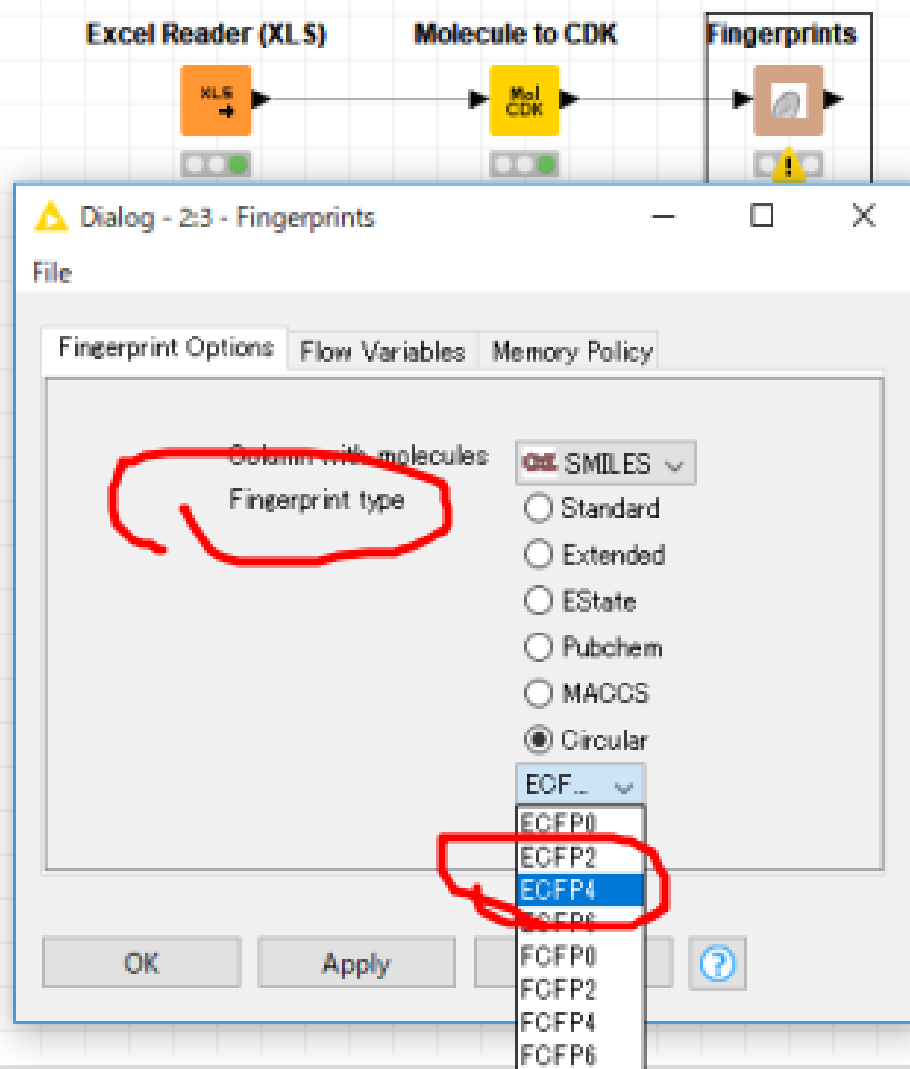


Node 3

クリック

ノード設置→ 接続 → 設定→実行
が KNIMEの基本操作です

繋いでクリック



自分の好みの
Fingerprintを読み込む

よく分からない方は適当に選んでも大丈夫です

Excel Reader (XLS)

Molecule to CDK

Fingerprints



Node 1



Node 2



Node 3

- Configure... F6
- Execute F7
- Execute and Open Views Shift+F10
- Cancel F9
- Reset F8
- Edit Node Description... Alt+F2
- New Workflow Annotation
- Connect selected nodes Ctrl+L
- Disconnect selected nodes Ctrl+Shift+L
- Collapse into Metanode
- Encapsulate into Wrapped Metanode
- Compare Nodes
- Show Flow Variable Ports
- Cut
- Copy
- Paste
- Undo
- Redo
- Delete
- Input with fingerprints

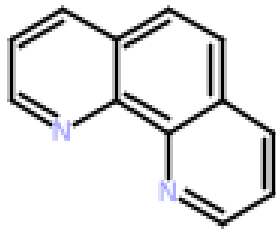
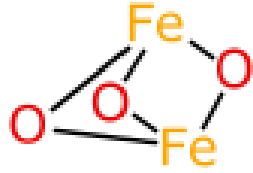
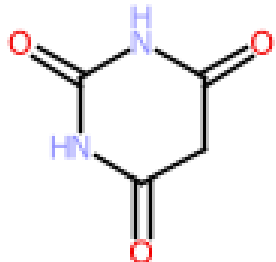
1. execute
2. 右クリック

ne

Console

KNIME Console

Fingerprint の生成 を確認

Row ID	ID	SMILES	Melting ...	Circular fingerprints fo...
Row0	1	<chem>Cu=S</chem>	500	0000000000000000000010...
Row1	2		117	0000000000000000000000...
Row2	3		1,539	0000000000000000000000...
Row3	4		245	0000000000000000000000...

Fingerprintと は？



機械学習モデルは基本的に数値しか理解出来ない



化学構造も認識出来ない



何らかのアルゴリズムで数値に変換する必要有り

→ **Fingerprint**、記述子、...

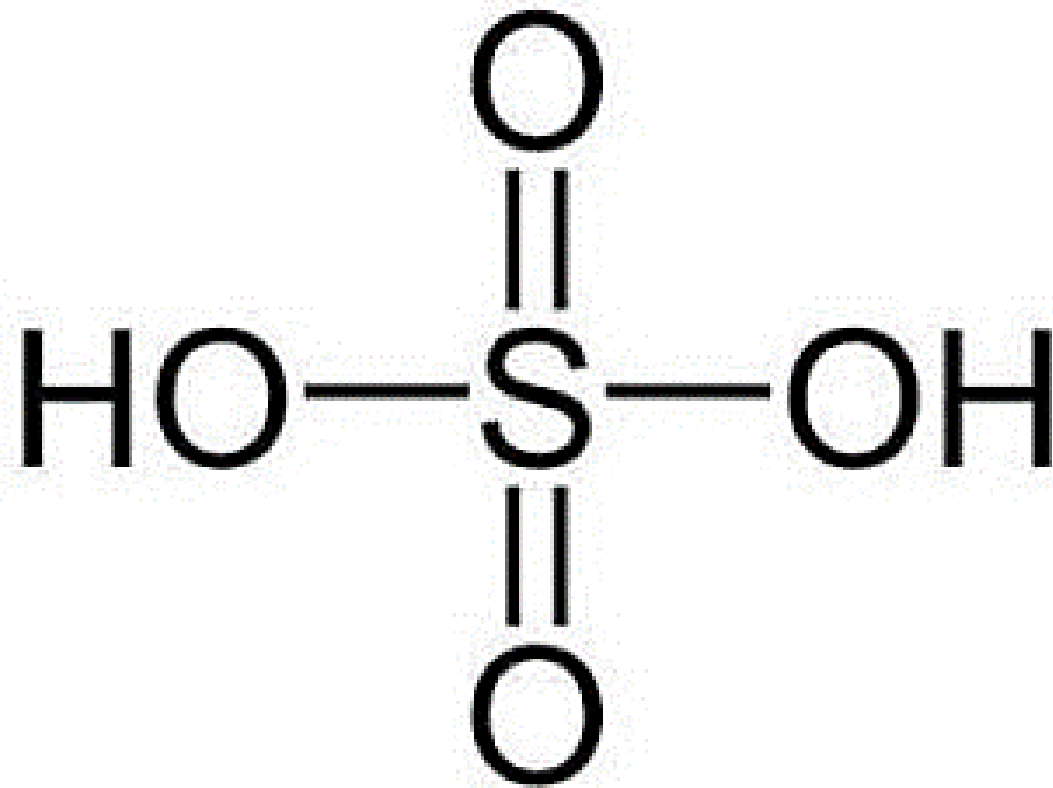
Fingerprintの イメージ

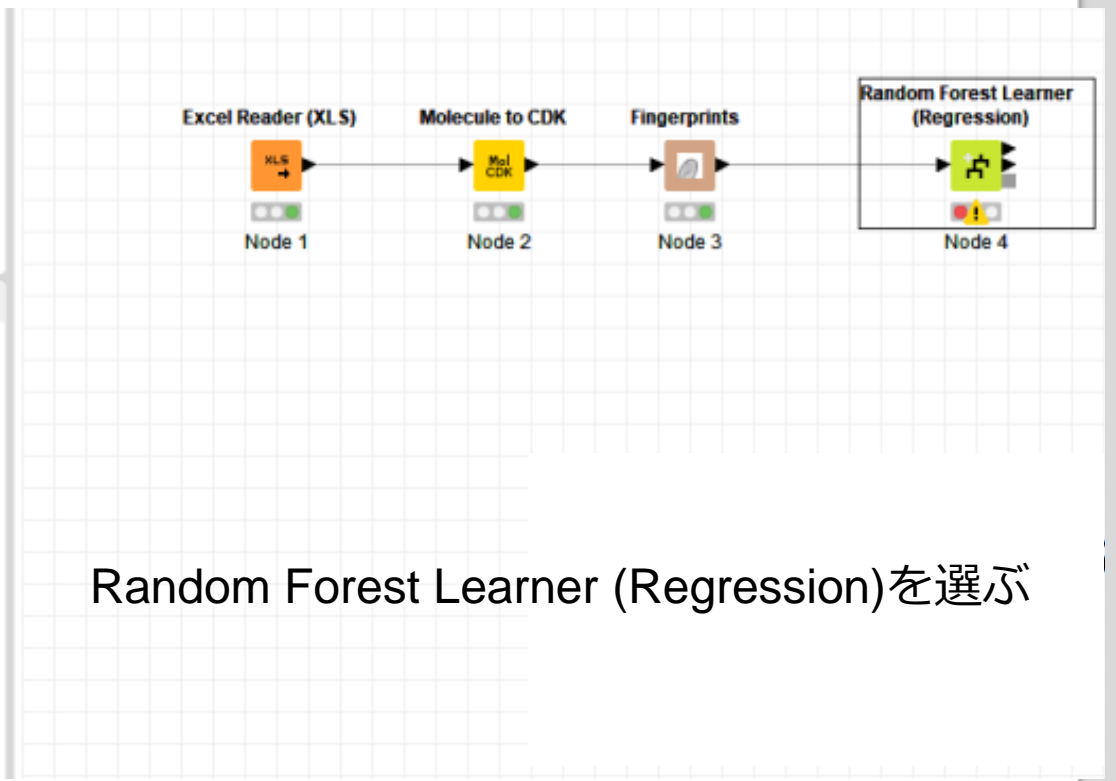
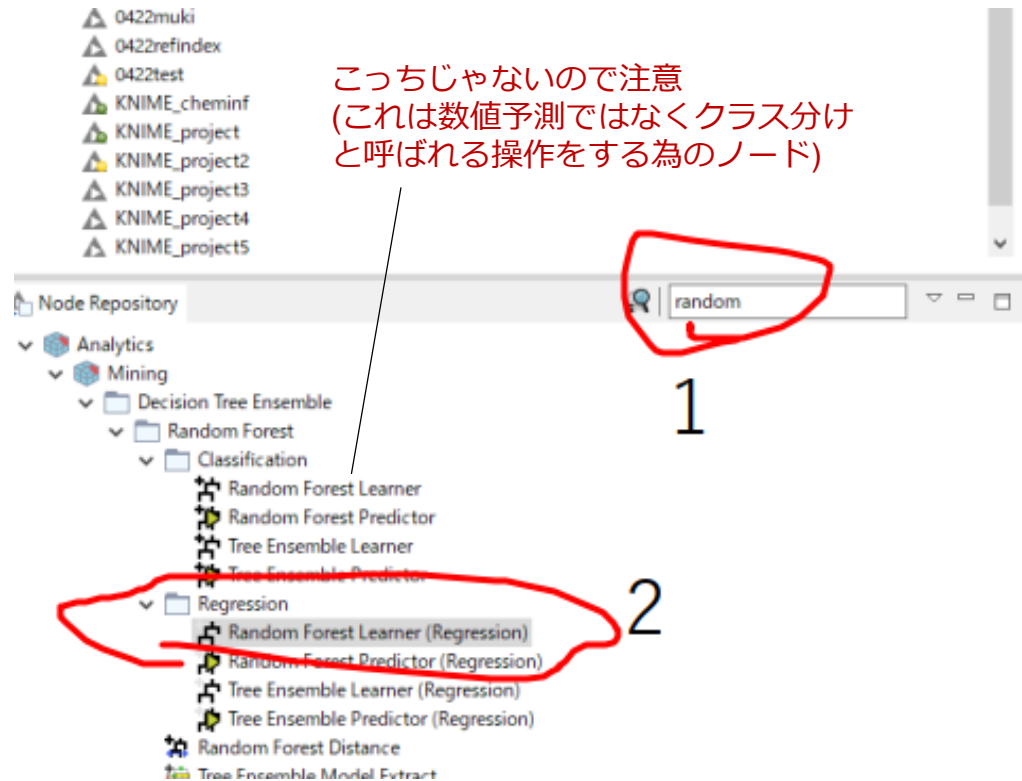
100-2000 dim.

1011010100001110001110
100000000001010000111
000111010010100001110
001110100000000000000
0000101000011100011110
000101101010001110001
110100000000000000000
1000101000011100011111
101000000000010100010
1110001110100101100001
110001110100000001000
000000000000101001001
1110000000111010000110
1011000001110001110100

...

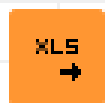
000001101100000000101





機械学習モデルの設置

Excel Reader (XLS)



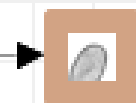
Node 1

Molecule to CDK



Node 2

Fingerprints



Node 3

Random Forest Learner
(Regression)



Node 4

繋いだらクリックして
設定画面を開く

繋ぐ

設定確認

(何もいじらなくてOK)

Options | Flow Variables | Memory Policy

Target Column yは融点 D Melting temperature

Attribute Selection xは化学構造 (Fingerprint)

☒ Use fingerprint attribute

☐ Use column attributes

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

☒ Enforce exclusion

Include

Filter

ID

☐ Enforce inclusion

> >> < <<

Misc Options

☐ Enable Hilighting (#patterns to store) 2,000

Tree Options

☐ Limit number of levels (tree depth) 10

☐ Minimum node size 5

Forest Options

Number of models 100

☒ Use static random seed 1595321365108 New

0422refindex
0422test
KNIME_cheminf
KNIME_project
KNIME_project2
KNIME_project3
KNIME_project4
KNIME_project5

Repository
Analytics
Mining
Decision Tree Ensemble
Random Forest
Classification
Random Forest Learner
Random Forest Predictor
Tree Ensemble Learner
Tree Ensemble Predictor
Regression
Random Forest Learner (Regression)
Random Forest Predictor (Regression)
Tree Ensemble Learner (Regression)
Tree Ensemble Predictor (Regression)
Random Forest Distance
Tree Ensemble Model Extract
Tree Ensemble Statistics

random

1

2

Excel Reader (XLS)
Node 1

Molecule to CDK
Node 2

Fingerprints
Node 3

Random Forest Learner (Regression)
Node 4

Random Forest Predictor (Regression)
Node 5

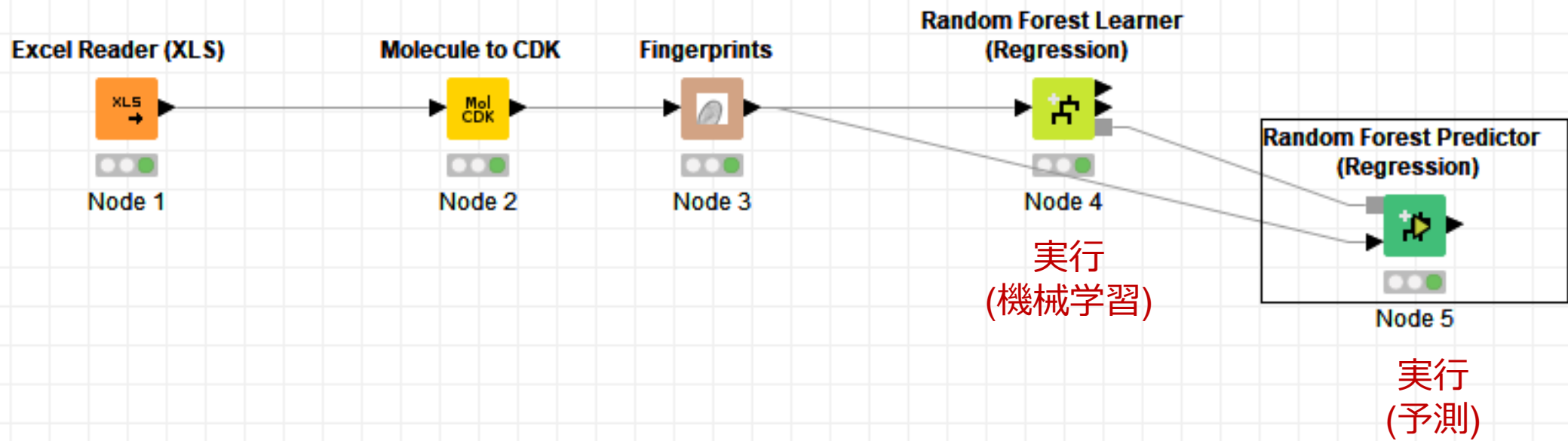
4. 四角同士を繋ぐ

5. 三角同士を繋ぐ

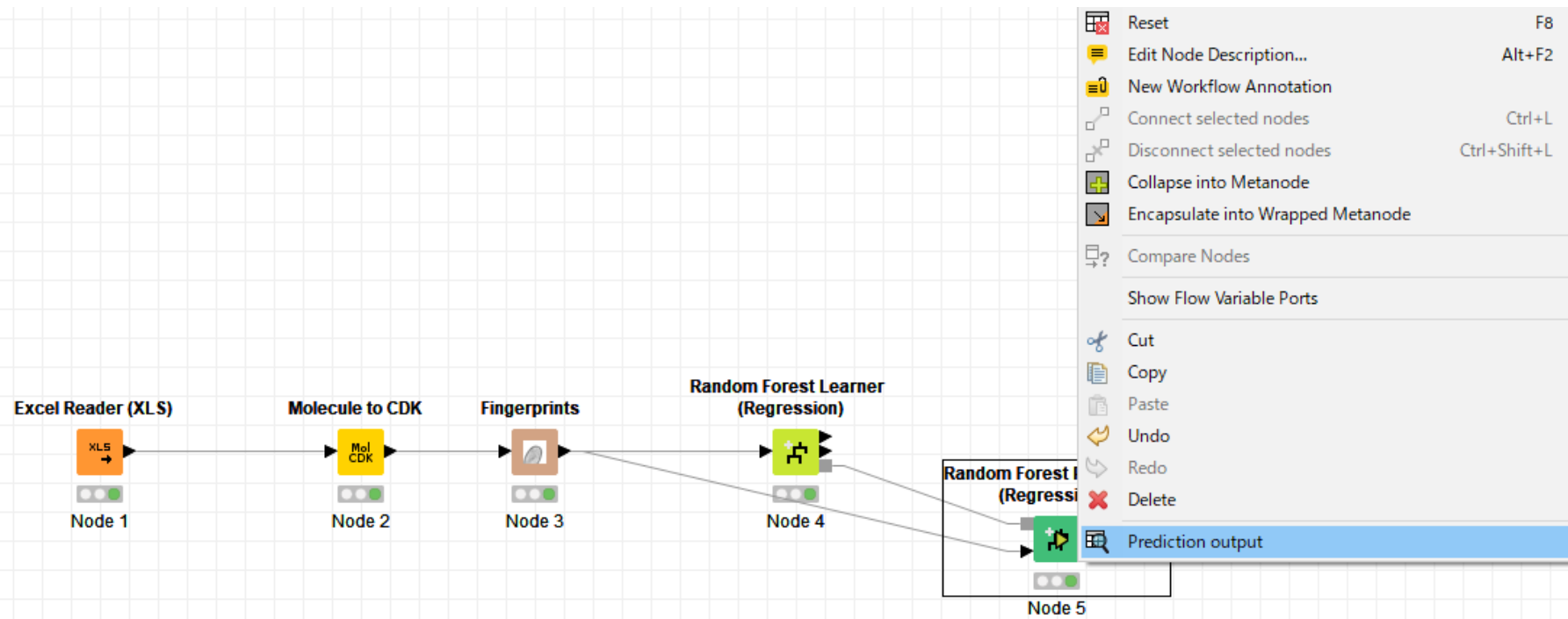
3 設置

学習したモデル(learner)を用い、
訓練データのfingerprintから融点を予測する、という処理をしています

予測用ノードの設置



実行

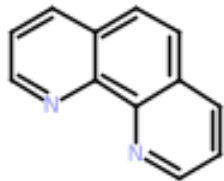
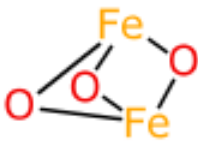
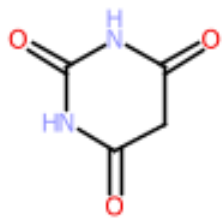


予測結果を見してみる

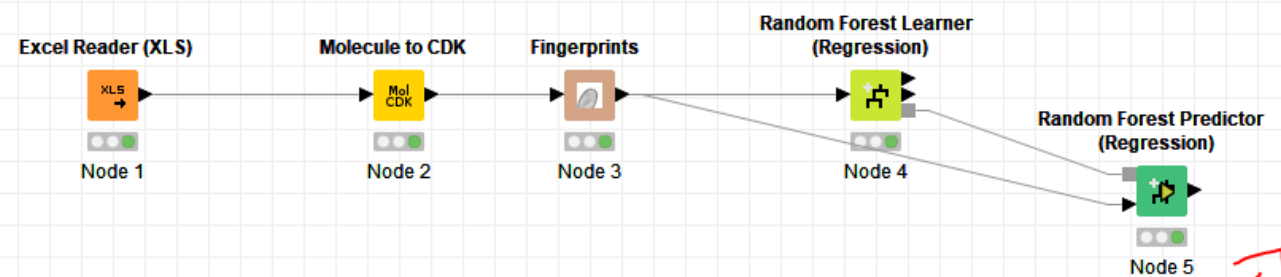
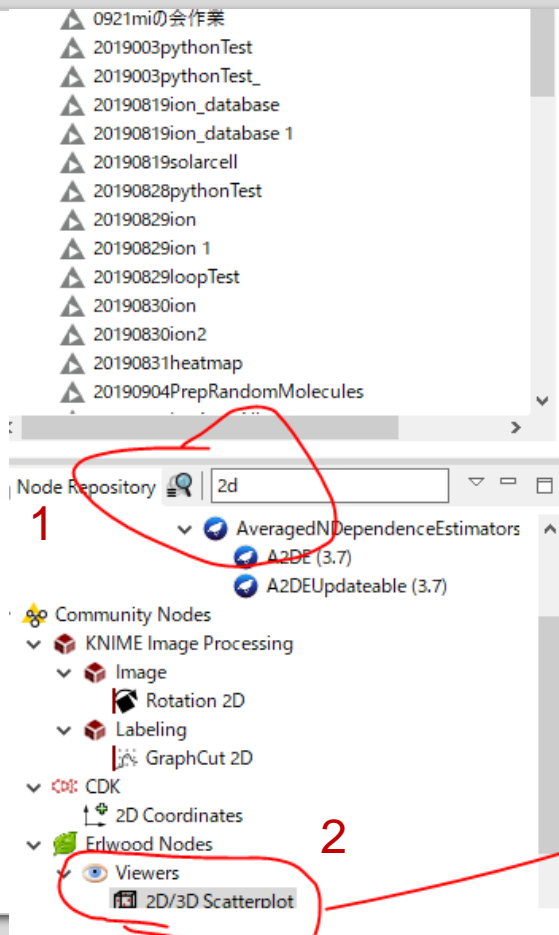
実測値

予測値 (無視してOK)



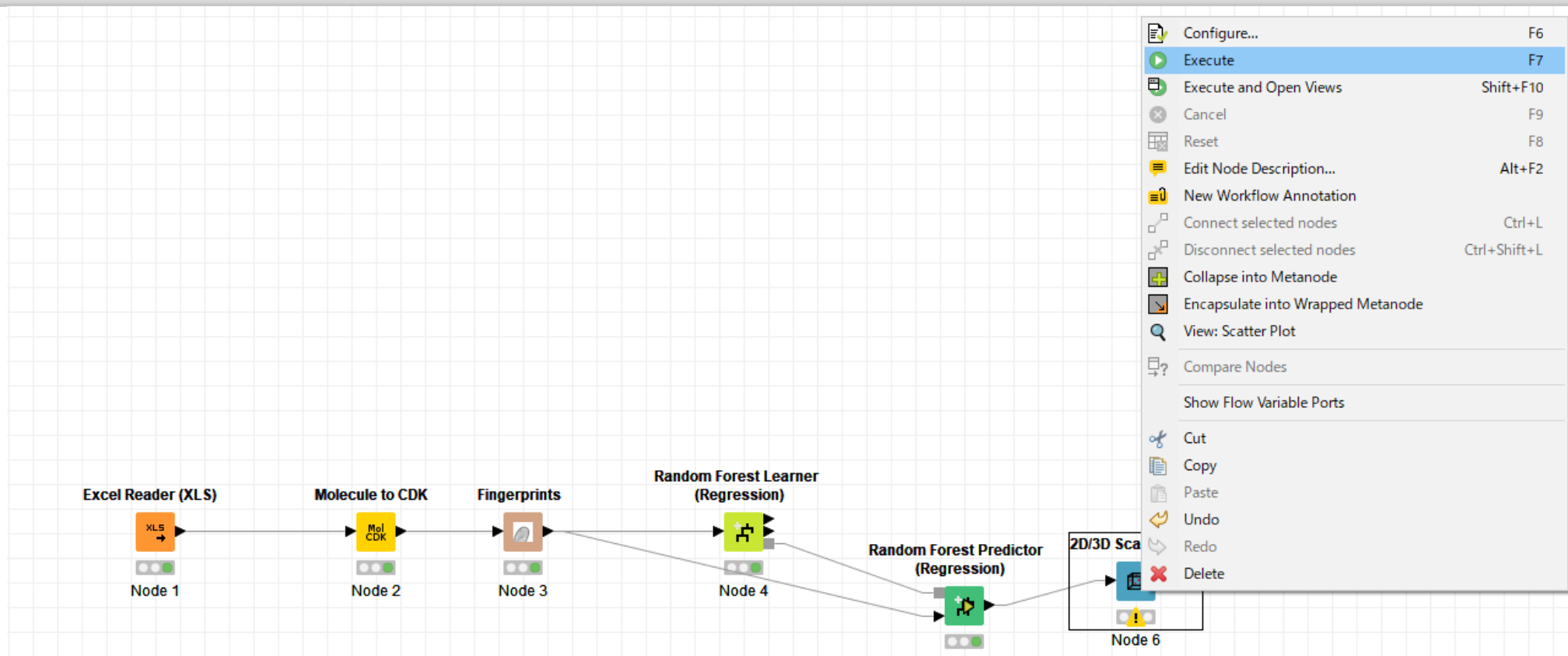
Row ID	ID	SMILES	Melting ...	Circular fingerprints fo...	Predicti...	Predicti...
Row0	1	<chem>Cu=S</chem>	500	0000000000000000000010...	533.276	109,853.248
Row1	2		117	0000000000000000000000...	111.938	21,720.777
Row2	3		1,539	0000000000000000000000...	1,073.579	367,562.537
Row3	4		245	0000000000000000000000...	259.956	41,448.713

結果の
確認

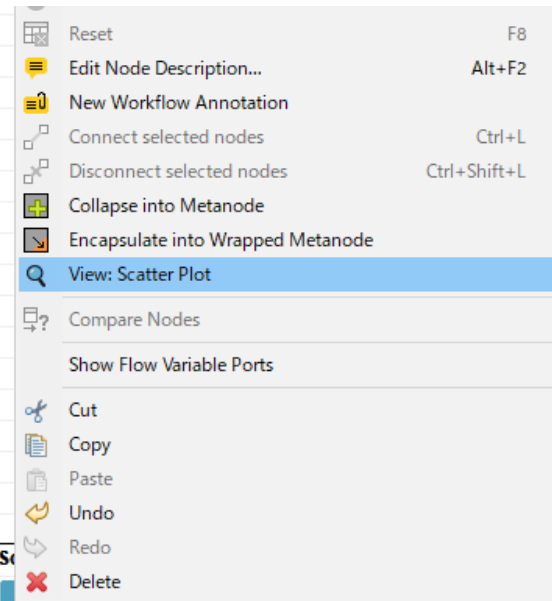
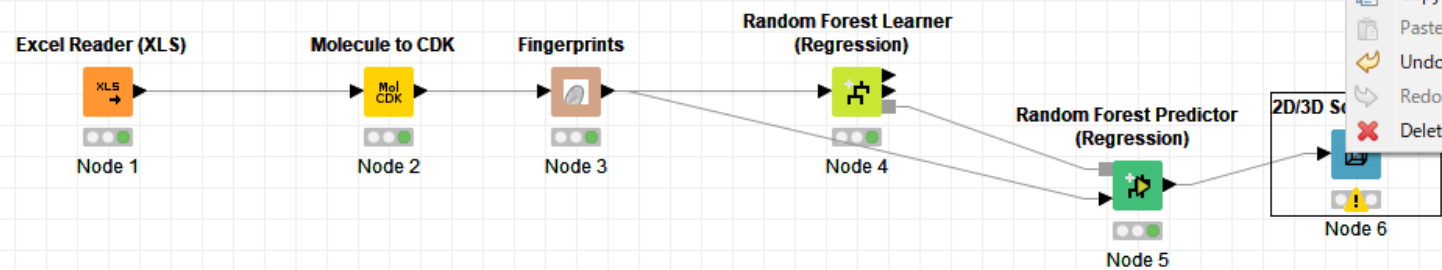


3. 設置

結果をグラフ化する



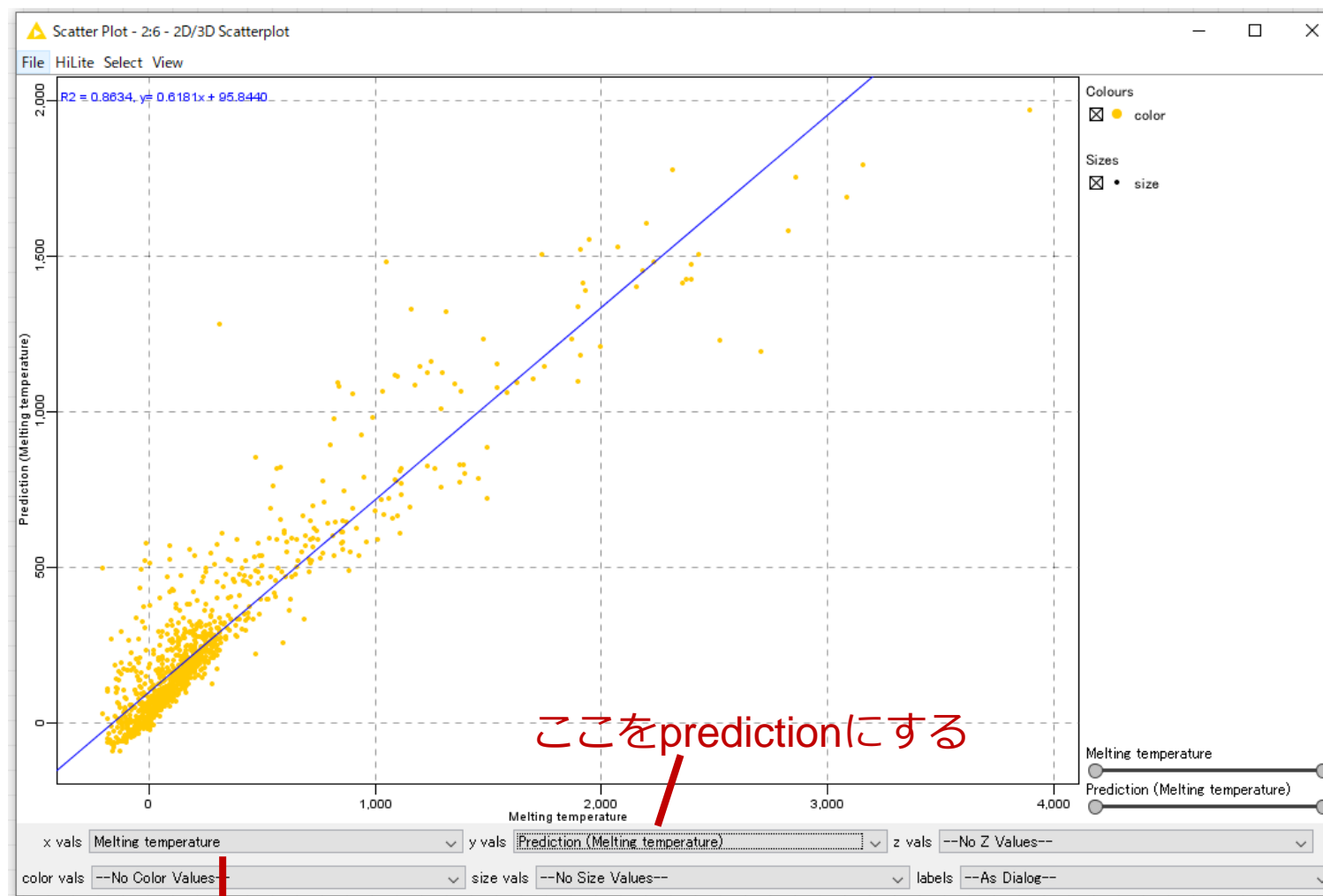
繋いで実行



グラフ表示

可視化

対角線上にプロット
が載っているほど、
高精度です



ここをpredictionにする

ここをmelting temperatreにする

最後のタスク

未知化合物
の性能予測

データベースの
questionシート中
の化合物は融点
が分からない!
→予測したい

	A	B	C	D	E
	ID	SMILES			
1	1	O=C=S			
2	2	CC(=O)OI(C1=CC=CC=C1)OC(=O)C			
3	3	C1CCC(=O)CC1			
4	4	Oc1cccnc1			
5	5	SC(C)(C)C			
6	6	O(c1ncccc1)C(COc3ccc(Oc2ccccc2)cc3)C			
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					
33					
34					
35					
36					
37					
38					
39					
40					
41					
42					
43					
44					
45					
46					
47					
48					
49					
50					
51					
52					
53					
54					
55					
56					
57					
58					
59					
60					
61					
62					
63					
64					
65					
66					
67					
68					
69					
70					
71					
72					
73					
74					
75					
76					
77					
78					
79					
80					
81					
82					
83					
84					
85					
86					
87					
88					
89					
90					
91					
92					
93					
94					
95					
96					
97					
98					
99					
100					
101					
102					
103					
104					
105					
106					
107					
108					
109					
110					
111					
112					
113					
114					
115					
116					
117					
118					
119					
120					
121					
122					
123					
124					
125					
126					
127					
128					
129					
130					
131					
132					
133					
134					
135					
136					
137					
138					
139					
140					
141					
142					
143					
144					
145					
146					
147					
148					
149					
150					
151					
152					
153					
154					
155					
156					
157					
158					
159					
160					
161					
162					
163					
164					
165					
166					
167					
168					
169					
170					
171					
172					
173					
174					
175					
176					
177					
178					
179					
180					
181					
182					
183					
184					
185					
186					
187					
188					
189					
190					
191					
192					
193					
194					
195					
196					
197					
198					
199					
200					
201					
202					
203					
204					
205					
206					
207					
208					
209					
210					
211					
212					
213					
214					
215					
216					
217					
218					
219					
220					
221					
222					
223					
224					
225					
226					
227					
228					
229					
230					
231					
232					
233					
234					
235					
236					
237					
238					
239					
240					
241					
242					
243					
244					
245					
246					
247					
248					
249					
250					
251					
252					
253					
254					
255					
256					
257					
258					
259					
260					
261					
262					
263					
264					
265					
266					
267					
268					
269					
270					
271					
272					
273					
274					
275					
276					
277					
278					
279					
280					
281					
282					
283					
284					
285					
286					
287					
288					
289					
290					
291					
292					
293					
294					
295					
296					
297					
298					
299					
300					
301					
302					
303					
304					
305					
306					
307					
308					
309					
310					
311					
312					
313					
314					
315					
316					
317					
318					
319					
320					
321					
322					
323					
324					
325					
326					
327					
328					
329					
330					
331					
332					
333					
334					
335					
336					
337					
338					
339					
340					
341					
342					
343					
344					
345					
346					
347					
348					
349					
350					
351					
352					
353					
354					
355					
356					
357					
358					
359					
360					
361					
362					
363					
364					
365					
366					
367					
368					
369					
370					
371					
372					
373					
374					
375					
376					
377					
378					
379					
380					
381					
382					
383					
38					

1. Excel readerを コピーで複製

Excel Reader (XLS)



Excel Reader (XLS)



Adjust Settings:

Select the sheet to read: <first sheet with data>

Column Names: <first sheet with data>

☒ Table contains column names: database, question, answer

Row IDs: ☒ Generate RowIDs (index incrementing, starting from 0) ☐ Generate RowIDs (incrementing, starting from 1) ☐ Make row IDs unique

Select the columns and rows to read:

☒ Read entire data sheet, or ... read columns from: A to: and read rows from: 1 to:

Tip: Mouse over the column and row headers in the "File Content"

On evaluation error:

☒ Insert an error pattern: #XL_EVAL_ERROR#

☐ Insert a missing cell

More Options:

☒ Skip empty columns ☐ Reevaluate formulas (leave unchecked if uncertain)

☒ Skip hidden columns ☐ Disable Preview (does not compute the output to the preview)

☒ Skip empty rows

Preview File Content

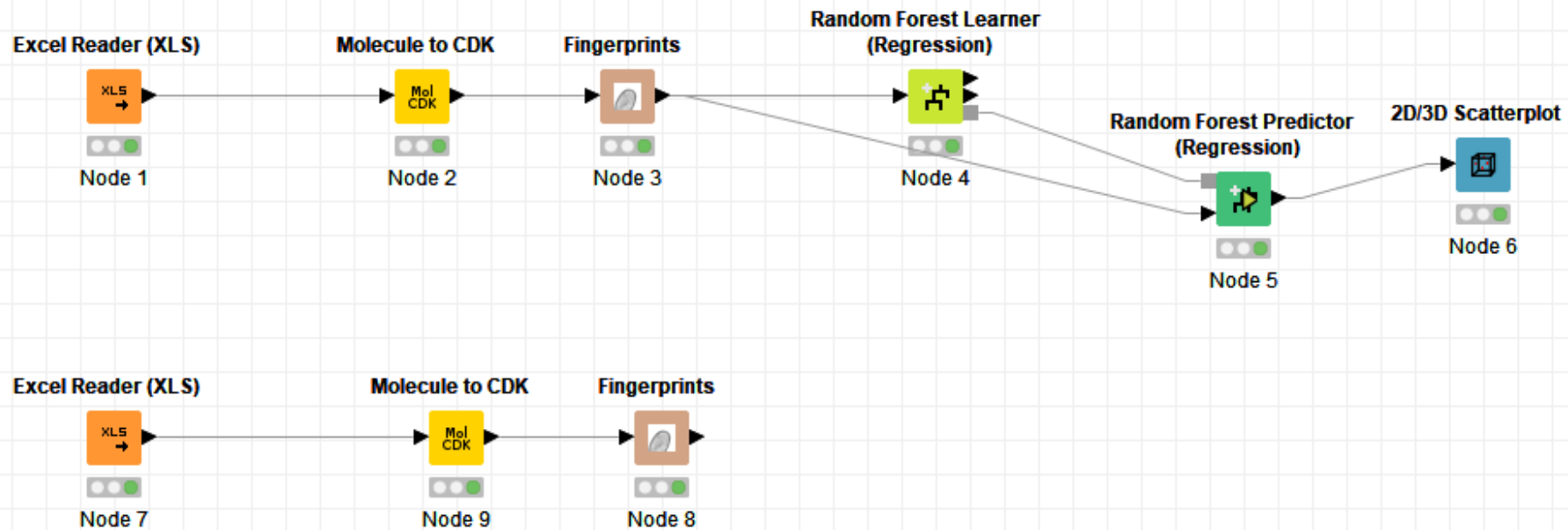
Preview with current settings: wikipedia_db.xlsx [database]

refresh

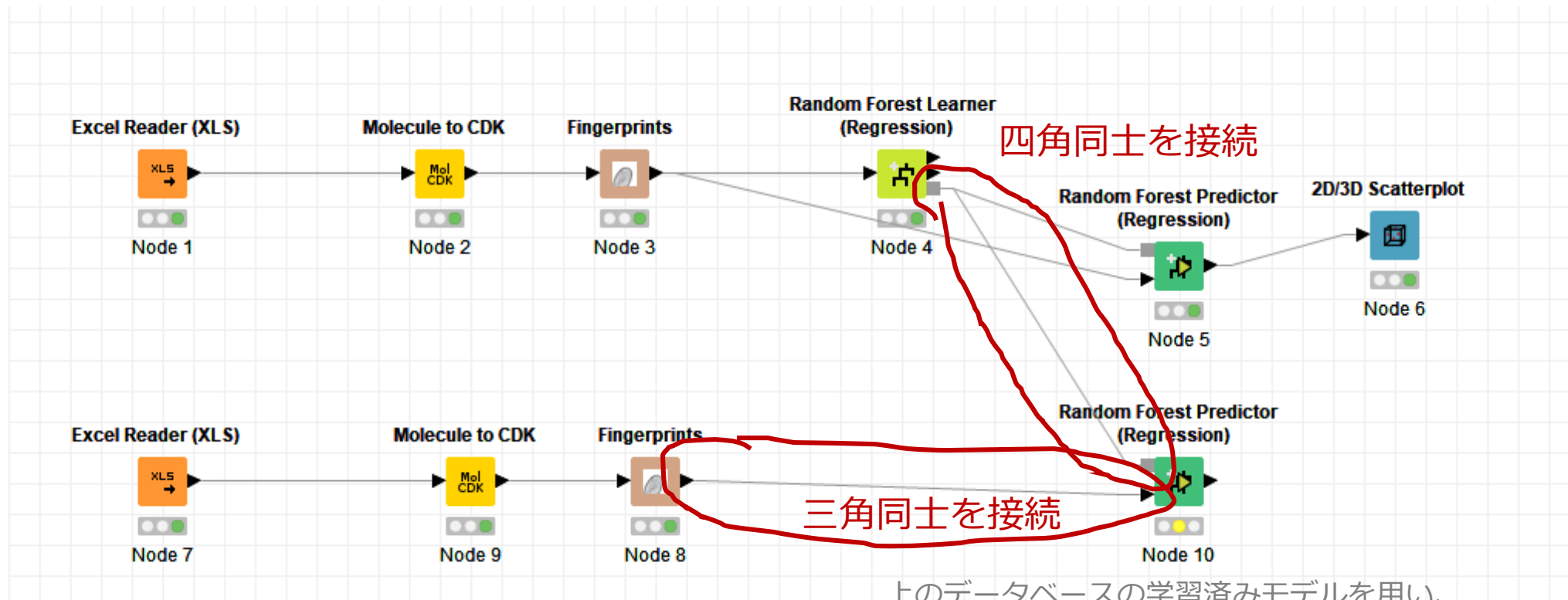
Row ID	I ID	S SMILES	D Melting ...
Row0	1	[Cu]=S	500
Row1	2	c1cc2ccc3cc...	117
Row2	3	O1[Fe]2O[F...	1,539
Row3	4	O=C1NC(=O...	245
Row4	5	P#[Y]	200.78

2. Question シート を読み込む

データベース の 読み込み

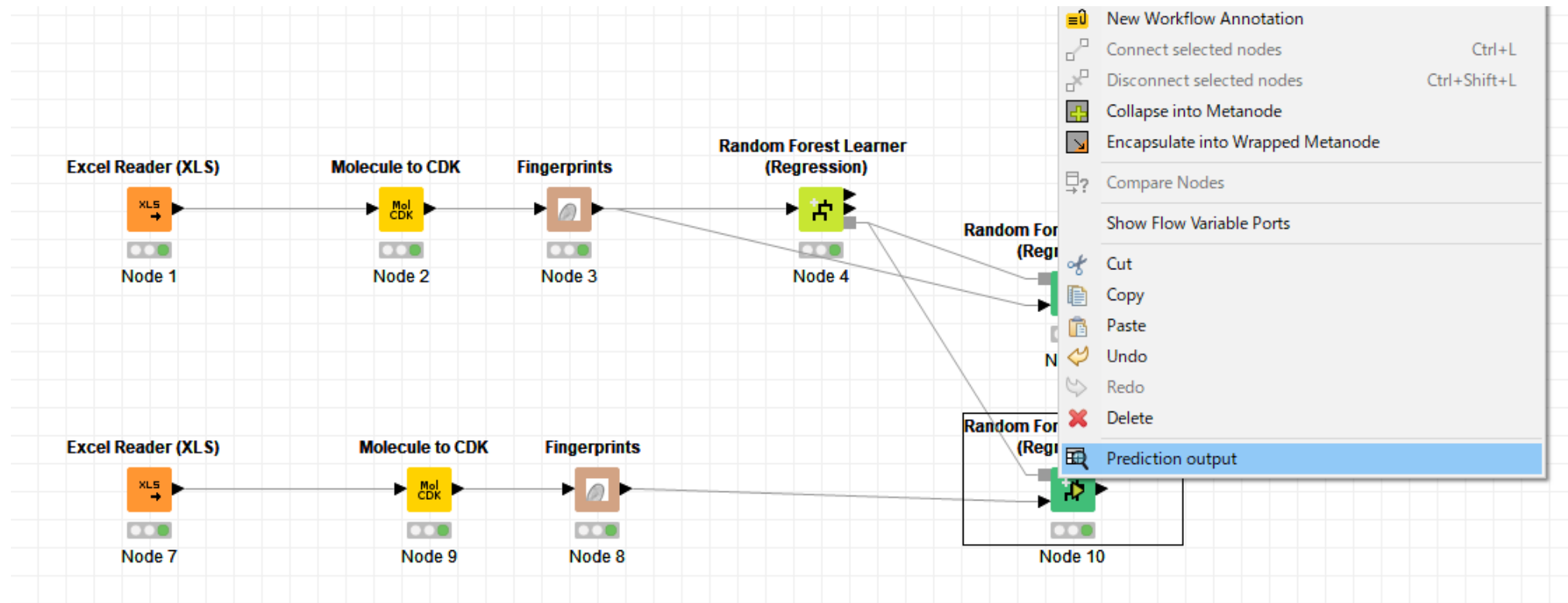


Molecule to CDK, Fingerprintsもコピーして実行



上のデータベースの学習済みモデルを用い、
下のデータの化合物の融点を予測する、という処理

予測ノードをコピーして設置・接続



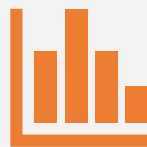
実行し、予測結果を確認

今回のworkflowファイル

sample_workflow.knwf
をダウンロード出来ます



まとめ



KNIMEと呼ばれるソフトウェアを使うと、比較的簡単に有機化合物の物性予測を実現出来る



MI操作の8割くらいはKNIMEでカバー出来る印象(?)



逆に、残りの2割(いわゆる最先端)を追求するには、多くの労力と技能が必要 (Pythonなど)

今回は 省略した 内容

Train/Testデータの準備

- 普通はデータベースをTrain/Testに分けてモデル精度を調べたりしますが、今回は割愛しました

各種モデルの検討

- 今回はRandom forestと呼ばれる、お手軽ながら強力なモデルを使用しました
- KNIMEだとGradient boosting, 線形モデル等を使えます
- それ以外のモデルの場合はPythonで専用ノードを作ったりする必要があります (配布予定...?)

yの正規化

- 多くのモデルではyをそのまま使うのではなく、標準得点等に正規化する必要があります