分子設計

山本 典史　著

Version 0.2.9（2026 年 11 月 3 日）

はじめに

研究室に入ると、次に何を合成するか、どの条件で測るかを、自分で決める場面が増えます。候補はいくつも思いつくのに、全部を試す時間はありません。手元には先輩の記録や文献の値がありますが、条件がそろっていなかったり、知りたい分子の値がなかったりします。そのとき、どの記録を使い、何を予測し、どの実験を先にするかを、根拠を添えて決められるでしょうか。

この授業ノートは、その決め方を、一つの依頼を最後まで進めながら練習するための本です。題材は、部品を洗う溶媒の 1-ペンタノールを、水に溶けやすい分子へ改良するという架空の依頼です。分子は炭素数 5〜6 のアルコールに限り、使うデータも十数個の分子の分だけです。小さな題材ですが、データを整える、構造を数で表す、予測する、予測を確かめる、次の実験を選ぶ、という分子設計の一通りの段を、すべて自分の手で通ります。

この授業ノートのねらい

序章で依頼を受け取り、第 1 章から第 6 章で一段ずつ進めます。第 1 章 で比べられるデータを整え、第 2 章 で分子の構造を 3 つの数（記述子）で表します。第 3 章 では似た分子の値から、第 4 章 では式から沸点を予測し、第 5 章 でその予測を、まだ使っていない資料値と照らし合わせます。第 6 章 では候補を選び、次に何を確かめるかを依頼人に提案します。各章の結果は次の章の出発点になるので、章を進めるごとに、依頼についての判断が具体的になります。

ここで扱う方法は、ケモインフォマティクス（化学の情報をデータとして扱い、分子や実験の設計に使う分野）の、いちばん小さな形です。実際の研究では、分子の数も記述子の数もずっと多く、計算はプログラムに任せます。それでも、比べられるデータを用意する、予測に使っていないデータで確かめる、言えることと言えないことを分ける、という考え方は変わりません。各章の「実際のケモインフォでは」の箱で、この授業の最小版が実際の研究のどこにあたるかを示しました。

ねらいは、アルコールの沸点の式を覚えることではありません。データの整え方、比べ方、確かめ方、次の実験の選び方を、自分の研究の記録（合成の収率、材料の特性など）に移せるようになることです。そのため、各章に実験系への接続の項を置き、研究室の場面に当てはめた例を示しました。

対象読者と前提

化学専攻の学部 3 年生、とくに研究室に配属されたばかりの人を読者に想定しています。前提にするのは、有機化合物の構造式の読み方、一級・二級アルコールの区別、分子間相互作用（水素結合や分散力）の基本です。数学は、四則演算、平均、一次式、散布図の読み方で足ります。使う計算は付録 A にまとめました。

プログラムを書くことと、機械学習の知識は前提にしません。繰り返しの計算は、授業のスライドに組み込んだアプリが受け持ちます。そのぶん、予想する、比べる、理由を述べる、判断を改める、という考える部分に時間を使ってください。

この授業ノートの使い方

各章の進み方、箱の見方、値の種類と色、演習アプリの使い方は、序章の「授業ノートと演習アプリの使い方」にまとめました。ここでは、一冊を通して守ってほしい約束を 2 つだけ挙げます。

一つ目は、演習の答えを手元に残すことです。答えは提出しません。第 6 章 で、第 1 章のときの自分の候補分子と根拠を見比べ、何が分かって判断がどう変わったかを確かめます。

二つ目は、資料値を見る順番です。構造を変えた分子の沸点は 第 5 章 で、水溶性は 第 6 章 で初めて示します。一冊の本なので後ろを先に読むこともできますが、予測を記録してから資料値を見ると、予測がどこまで当たったかを正直に確かめられます。

前付けには「記号の約束」と「授業回との対応表」を置きました。記号に迷ったら、記号の約束に戻ってください。巻末には、計算とグラフの読み方（付録 A）、用語集（付録 B）、理解確認問題と総合問題（付録 C）、演習の参考解答（付録 D）、記録の例（付録 E）、資料と出典（付録 F）があります。この本で使う実在の分子の値には、すべて付録 F に出典を示しました。

記号の約束

読み始めで記号に迷わないよう、本書で使う主な記号の約束をまとめます。

表 1　記号の約束

記号	意味
n	炭素数
s	OH が結合する炭素が二級なら 1、一級なら 0
b	炭素骨格で 3 個以上の炭素と結合する炭素（分岐点）の個数
(n, s, b)	記述子の組
T	常圧の沸点（℃）
w_{25}	25 ℃ の水に富む飽和溶液 100 g 中のアルコールの質量（g）
d	記述子の違いから決める距離
α	距離で s の違いに掛ける重み（既定は 2。第 3 章で変える）
M0	炭素数だけの式 $T = 20n + 38$（M はモデル、0 は補正なし）
M1	補正式 $T = 20n + 38 - 17s - 10b$（M0 に二級と分岐の補正の項を入れた式）
c_s, c_b	二級アルコール・分岐の補正係数（℃）
残差	資料値 − 予測値
SSE	残差平方和（係数の選択に使う）
MAE	平均絶対誤差（予測の成績の要約に使う）

授業回との対応表

表 1　授業回と本書の章の対応

授業回	本書の章	範囲・備考
第 1 回	序章・第 1 章	データ駆動とケモインフォマティクスの考え方、依頼の提示、比較できるデータの整理
第 2 回	第 2 章	記述子 (n, s, b) と、表せない違い
第 3 回	第 3 章	距離・重み・最近傍による予測
第 4 回	第 4 章	基準式 M0、残差、補正式 M1
第 5 回	第 5 章	未使用データでの検証と MAE。確認用の資料値はこの回で初めて示す
第 6 回	第 6 章	候補の選択、次の確認実験、自分の研究への移し替え
第 7 回	付録 C	定期試験（60 分・マークシート）

目次

はじめに i
この授業ノートのねらい i
対象読者と前提 ii
この授業ノートの使い方 ii
記号の約束 iii
授業回との対応表 v
序章　データから分子を設計する 2
今回の依頼：1-ペンタノールを改良する 2
目標、制約、最終的に提案するもの 4
分子・材料を設計する 3 つの進め方 5
データ駆動の循環：予測して、次の実験を選ぶ 7
ケモインフォマティクスとは：分子をデータとして扱う 9
この授業の最小版と、実際の研究の規模 11
授業ノートと演習アプリの使い方 11
1 設計に使うデータを整える 14
1.1 データ表を作る：行・列・目的変数 15
1.2 構造と物性の関係をデータから見る 16
1.3 比較できるデータとは：単位・温度・相・測定条件 18
1.4 欠測・出典・転載・誤記 20
1.5 例題・演習：データを並べ、使える記録を選ぶ 22
1.6 設計判断：出発候補の不足と、次に必要な情報 24
2 分子構造を数値で表す 30
2.1 分子を原子と結合のつながりとして見る 31
2.2 記述子とは何か 32
2.3 炭素数、OH 結合炭素の級数、分岐を数える 33
2.4 同じ記述子になる異性体：表せない違い 36
2.5 例題・演習：構造を表にし、追加する特徴を選ぶ 37

2.6 設計判断：今回の表現で何が分かり、何が分からないか 39
3 似た分子から物性を予測する 46
3.1 似た分子の値を使う：直感と最近傍 47
3.2 記述子の違いを距離にする 48
3.3 尺度と重み：何の違いを重く見るか 50
3.4 重みを変えると、最近傍と予測が変わる 51
3.5 例題・演習：参照分子と重みを変えて予測する 53
3.6 設計判断：どの予測を採用し、何を未確認とするか 54
4 データから予測式を作る 60
4.1 構造と物性の関係を式にする：QSPR の考え方 61
4.2 炭素数から沸点を予測する基準式 62
4.3 残差を見る：式が説明できていない違い 63
4.4 構造による補正と、誤差を小さくする係数の選択 65
4.5 例題・演習：補正式を作り、未知分子の予測を記録する 68
4.6 設計判断：経験式の化学的な解釈と限界 69
5 予測を確かめ、信用できる範囲を考える 78
5.1 当てはまりと予測を分ける：記録してから確かめる 79
5.2 絶対誤差と平均絶対誤差：結果をまとめ、何が言えるか 80
5.3 過学習とデータ漏洩 83
5.4 記述子だけでは区別できない分子 85
5.5 適用範囲：どこまで使ってよいか 86
5.6 例題・演習：予測と資料値を照合し、言える結論を選ぶ 87
5.7 設計判断：予測をどこまで根拠として使うか 89
6 分子候補と次の実験を提案する 94
6.1 複数の物性を同時に満たす候補を探す 95
6.2 要求値からの余裕と、予測値・資料値・欠測の区別 98
6.3 目標から構造を考える：逆設計 100
6.4 限られた測定機会で、次に何を確かめるか 102
6.5 例題・演習：候補を選び、確認結果から判断を更新する 104
6.6 最終提案：候補、根拠、未確認事項、次の確認 105
6.7 自分の研究で考える：構造・組成・調製条件・測定条件 108
付録 A　計算とグラフの読み方 114

平均 114
一次式：傾きと切片 114
残差 115
平方誤差と残差平方和 115
絶対値と平均絶対誤差 116
付録 B　用語集 118
付録 C　理解確認問題と総合問題 122
章末の理解確認問題 122
総合問題 122
付録 D　演習の参考解答 128
第 1 章 128
第 2 章 130
第 3 章 133
第 4 章 135
第 5 章 137
第 6 章 140
付録 C の総合問題 142
付録 E　記録の例 146
データを集めて表にする 146
構造や条件を表の列にする 147
予測を記録する（封をする） 148
予測を確かめる 148
候補の状態と、次の確認を計画する 149
自分の研究で考える 150
付録 F　資料と出典 152
この本の資料値 152
本文で引用した文献 153
資料値の示し方について 154
改訂履歴 156

序章　データから分子を設計する

研究室で実験をしていると、「次に何を合成するか」「どの条件で測るか」を決める場面がたびたびあります。候補はたくさん思いつくのに、試せる数は限られています。そのとき、これまでのデータを手がかりに、次の一手を選べないでしょうか。

この授業ノートでは、一つの依頼を 6 章かけて進めます。ある溶媒を、水に溶けやすい溶媒へ改良したいという依頼です。データを整え、分子の構造を数で表し、物性を予測し、その予測を確かめ、最後に候補と次に確かめる実験を提案します。序章では、まずその依頼を受け取ります。そのあとで、依頼に取り組むための進め方と、データを使う考え方を見ておきます。

今回の依頼：1-ペンタノールを改良する

まず、6 章かけて取り組む依頼を見ていきます。場面は授業のための架空の設定です。

小さな部品メーカーが、機械部品についた油を溶媒で洗い落としています。いまの溶媒は 1-ペンタノール（$CH_3(CH_2)_4OH$）です。工程は 図 0-1 のように、溶媒で洗い、水ですすぎ、乾かす、の 3 段です。困っているのは 2 段目のすすぎで、1-ペンタノールが水に溶けにくいため、部品に残ってしまいます。そこで、工程の担当者から次の依頼が届きました。

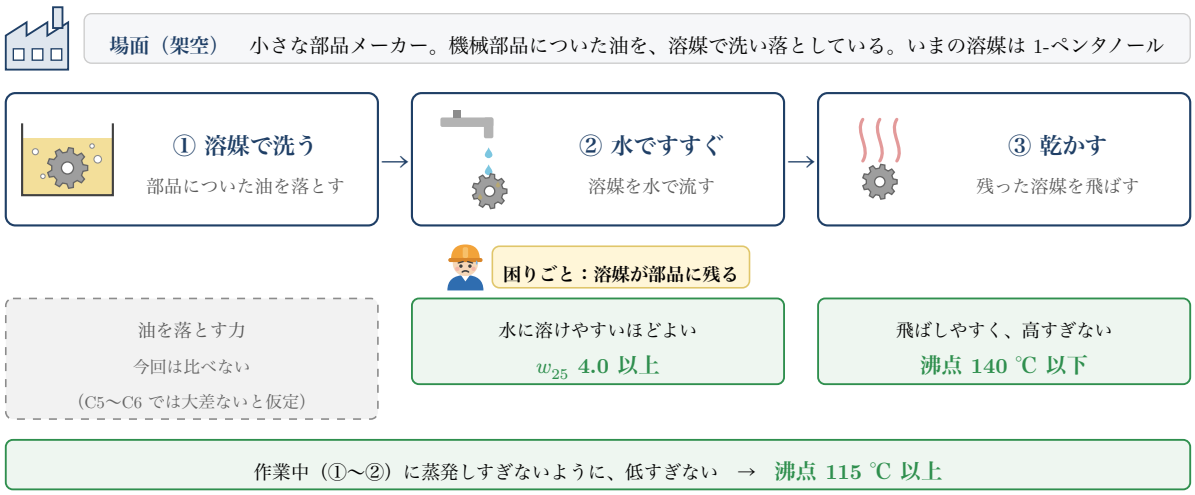


図 0-1　依頼の背景（架空）です。溶媒で洗い、水ですすぎ、乾かす工程の各段に、どの物性が影響するかを緑の札で示しました。困っているのは ② のすすぎです。すすぎには水に溶けやすいことが、乾かす段と作業中の蒸発には沸点が影響します。油を落とす力は、この授業では比べません。



依頼メモ　工程担当者より

材料加工の工程で使う溶媒の候補を検討してほしい。出発の候補の 1-ペンタノールは、沸点 138 ℃、25 ℃ の水への溶解度 2.19 である。沸点を 115〜140 ℃ に保ちながら、水への溶解度を 4.0 以上にできる候補を考えてほしい。限られたデータと測定の機会を使い、既知の候補 1 つと、追加で調べる候補 1 つ、次に確かめる分子と物性を提案してほしい。

依頼の「水への溶解度」は、この授業ノートでは次の量で表します。

定義：水溶性 w_{25}

25 ℃ でアルコールと水を混ぜ、2 層に分かれたときの水相（水に富む相）100 g 中に含まれるアルコールの質量（g）です。分母は飽和溶液 100 g で、水 100 g でも 100 mL でもありません。単位は g / 飽和溶液 100 g です。この授業ノートでは、w_{25} の数値には単位を付けず、表の見出しやグラフの軸に書きます。

1-ペンタノールの w_{25} は 2.19 です。沸点 138 ℃ と w_{25} 2.19 は、どちらも文献に載った測定値（資料値）です（沸点は国際化学物質安全性カード、w_{25} は IUPAC の溶解

度データ集に再録された測定値)。分母や温度の違う値と混ぜないことは、第 1 章で詳しく扱います。

図 0-1 では、油を落とす力を比べないことにしました。炭素数 5〜6 のアルコールどうしなら大きく変わらないと仮定しています。これは仮定であって、確かめた事実ではありません。また、純物質の沸点と水溶性だけでは、実際の工程での性能（混ざった液のふるまい、洗い上がり、安全性など）までは保証できません。この点は第 6 章で、もう一度考えます。

目標、制約、最終的に提案するもの

依頼の要求は 2 つで、型が違います（図 0-2)。

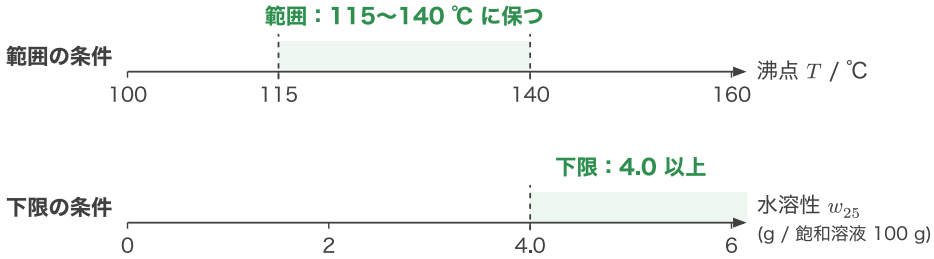


図 0-2　要求値の 2 つの条件です。沸点は範囲の条件で、両側の境界と比べます。w_{25} は下限の条件で、下限とだけ比べます。

- **沸点 115〜140 ℃**：範囲の条件です。高すぎても低すぎても満たしません。乾かす段で飛ばしやすいように上限があり、作業中に蒸発しすぎないように下限があります。
- **w_{25} 4.0 以上**：下限の条件です。4.0 以上なら、どれだけ大きくても満たします。すすぎで水に流しやすいほどよいからです。

依頼には、次の制約も付けます。候補は炭素数 5〜6 の、一級または二級の一価アルコールに限ります。これは授業のための設定です。炭素数 4 以下が実際に使えないという意味ではなく、炭素数の同じ分子どうし（異性体）を比べる練習のための制約です。

図 0-3 に、2 つの要求を同時に満たす領域を示しました。1-ペンタノールは沸点の範囲には入っていますが、w_{25} が 4.0 に 1.81 足りず、領域の左にあります。

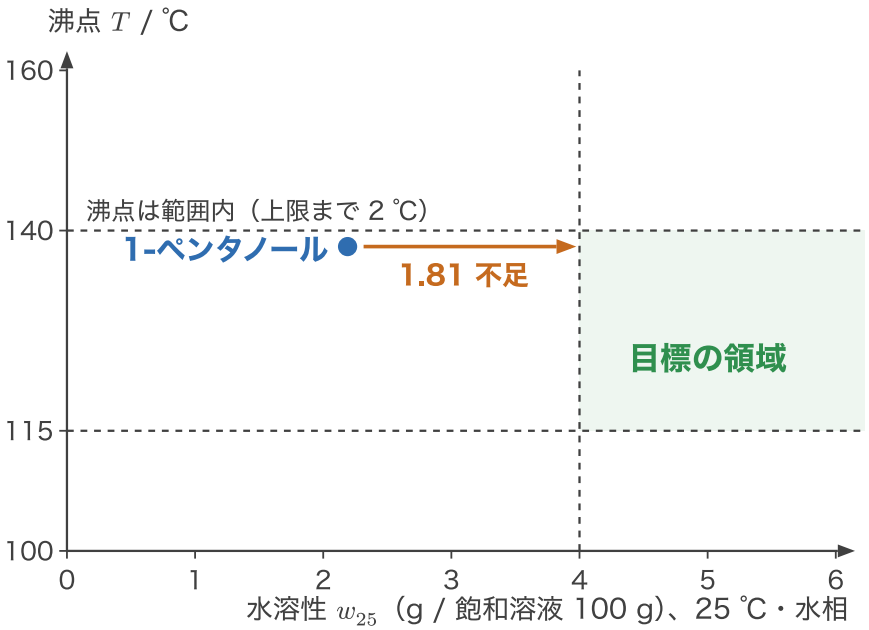


図 0-3　依頼の目標の領域です。横軸が w_{25}、縦軸が沸点で、破線が要求値の境界、緑の四角が 2 つの要求を同時に満たす領域です。青い点は 1-ペンタノールの資料値（沸点 138 ℃、w_{25} 2.19）で、沸点は範囲に入っていますが（上限まで 2 ℃）、w_{25} が 1.81 足りません。

最終的に提案するものは 3 つです。

1. **既知の候補 1 つ**：資料値で、2 つの要求を満たすと確かめられる分子
2. **追加で調べる候補 1 つ**：資料値はまだないが、予測から有望と考える新しい分子
3. **次に確かめる分子と物性**：限られた測定の機会を、どれに使うか。その理由と、結果ごとの判断も添えます

ここで、最初の考えを書き留めておきましょう。1-ペンタノールの構造をどう変えれば、水に溶けやすくなりそうでしょうか。根拠は問いません。直感の候補分子を 1 つ、ノートに残してください。第 1 章 では、その候補分子を判断するのにどんなデータが必要かを考えます。第 6 章 では、最終的な提案とこの最初の候補分子を比べ、何を根拠に判断を変えたかを振り返ります。

分子・材料を設計する 3 つの進め方

この依頼に、どう取り組めばよいでしょうか。分子を設計するとは、ほしい性質をもつ構造を選ぶことです。今回の依頼なら、沸点を 115〜140 ℃ に保ったまま、w_{25} を

4.0 以上にできそうな構造を考え、確かめることになります。そのための進め方は、大きく 3 つに分けられます（図 0-4）。

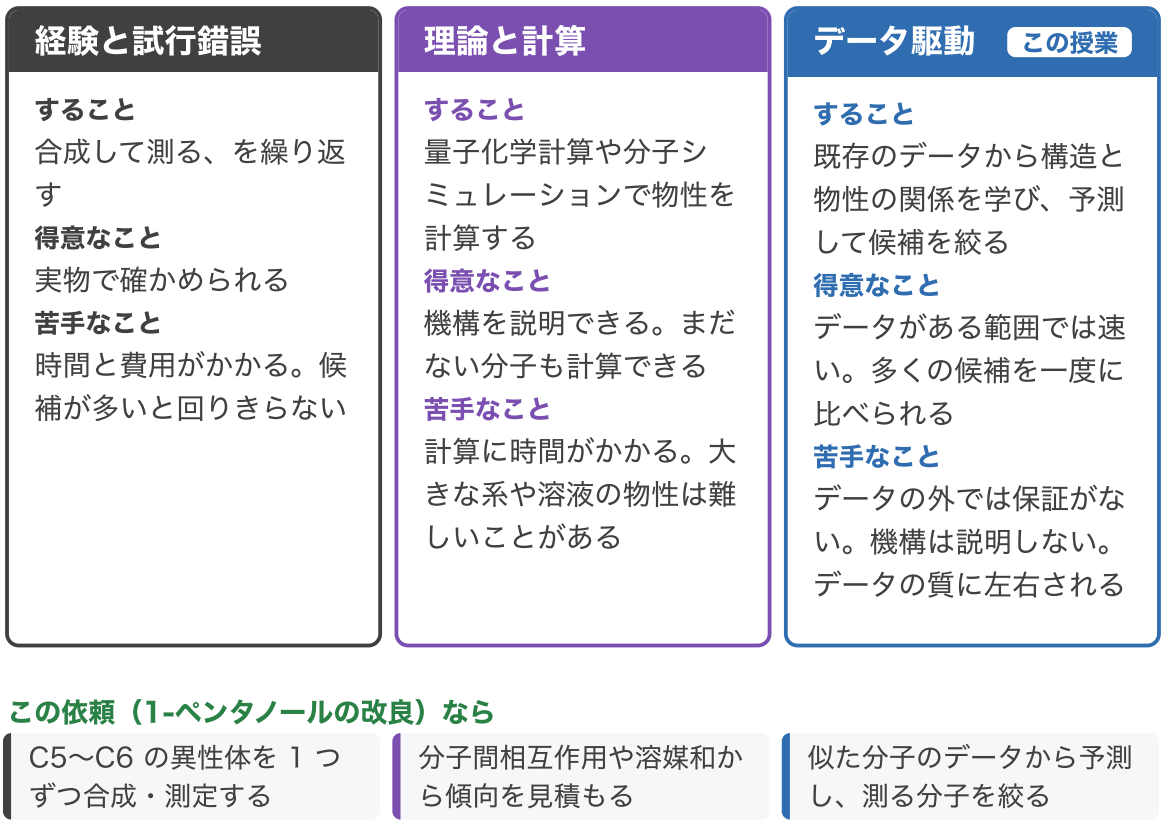


図 0-4　分子・材料を設計する 3 つの進め方です。どれか 1 つが優れているわけではなく、得意なことと苦手なことが違います。下の帯のように、組み合わせて使います。

3 つは、どれか 1 つを選ぶものではありません。実際の研究では組み合わせて使います。たとえば、図 0-4 の下の帯のように、計算でデータを作り、そのデータから学んだ予測で候補を絞り、残った候補を実験で確かめる、という使い方です。

この授業で手を動かすのは、3 つ目のデータ駆動の最小版です。ただし、予測の値が出たら終わりではありません。予測がなぜそうなるのかを考えるときには、分子間の力などの化学の知識を使います。データ駆動は、化学の理屈や実験に代わるものではなく、それらと組み合わせて使う進め方です。

データ駆動の循環：予測して、次の実験を選ぶ

データ駆動の進め方は、一度予測して終わる一本道ではなく、図 0-5 のような循環になります。データを集めて整え、分子の構造を数にし、それを使って物性を予測します。予測がどれだけ当たるかを確かめ、それを踏まえて次に測る分子を選びます。測った結果はデータに加わり、次の予測に使えます。

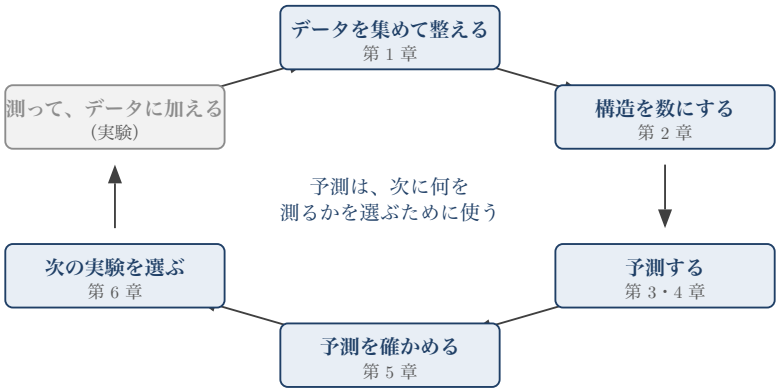


図 0-5　データ駆動の循環です。データを整え、構造を数にして予測し、予測を確かめてから次の実験を選びます。測った結果をデータに加えて、また予測に戻ります。各段の下は、この授業ノートでその段を扱う章です。

ここで大事なのは、予測の役目です。予測がすべての分子で当たる必要はありません。予測は、限られた実験の機会を、どの分子に使うかを選ぶために使います。当たりそうな順に並べられれば、手当たりしだいに試すより、少ない実験で目的に近づけます。

今回の依頼も、この循環に沿って進めます。この授業ノートの第 1〜6 章は、図 0-5 の段を 1 つずつたどります。第 1 章でデータを整え、第 2 章で構造を数にし、第 3・4 章で予測し、第 5 章で予測を確かめ、第 6 章で次の実験を選びます。測る段だけは、授業の中では資料を開くことで代えます。

> **実際のケモインフォでは　予測で候補を絞った 2 つの研究**　試験の範囲外
>
> データ駆動で候補を絞った研究を 2 つ紹介します。数値は覚えなくてかまいません。

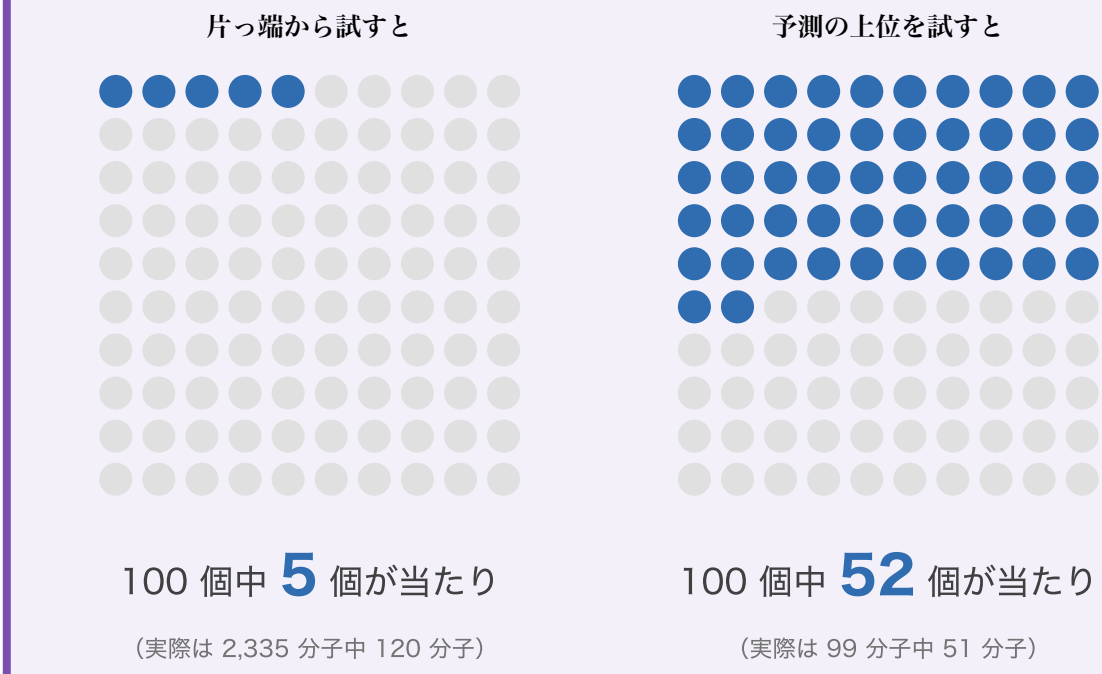


図 0-6　抗菌薬の探索で、100 個あたりに直した当たり（大腸菌の増殖を止めた分子）の数です。左は学習に使った 2,335 分子の実験結果、右は別の 6,111 分子のうち予測の点数が高い順に試した 99 分子です。左と右は別の分子の集まりです（Stokes ほか, 2020）。

抗菌薬の探索：2,335 分子の実験結果から、大腸菌の増殖を止めるかどうかを予測するモデルを作り、別の 6,111 分子に点数を付けました（図 0-6）。予測の上位 99 分子を試すと、51 分子が増殖を止めました。同じ集まりの下位 63 分子では、止めたのは 2 分子でした。上位から見つかった分子の 1 つは、既存の抗菌薬と構造が似ていない新しい候補でした（Stokes ほか, 2020）。

有機 EL の発光分子の探索：部品の組み合わせで 160 万個を超える候補をコンピューターで作り、そのうち 40 万個以上を量子化学計算で調べました。計算する順番は、計算済みの結果から学んだ予測で決めています。計算の結果で絞った約 2,500 個を化学者が評価し、素子にして確かめたのは 4 個でした（Gómez-Bombarelli ほか, 2016）。

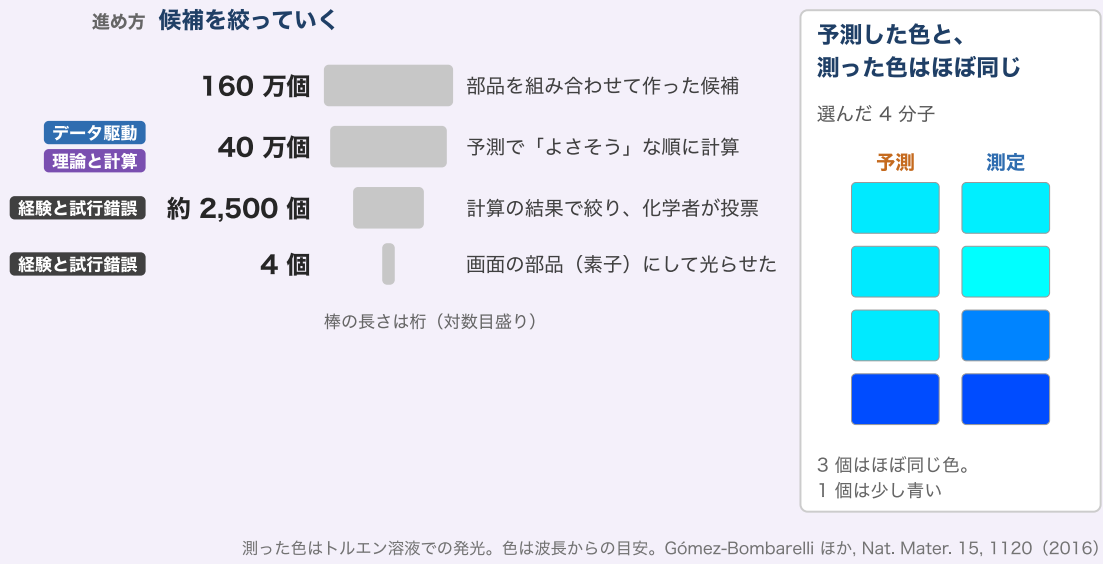


図 0-7　有機 EL の発光分子の探索です。左は候補を絞っていく流れで、各段の札は 3 つの進め方のどれにあたるかを示します。右は、素子にした 4 分子について、予測した光の色と測った色を並べたものです（色は波長からの目安）。

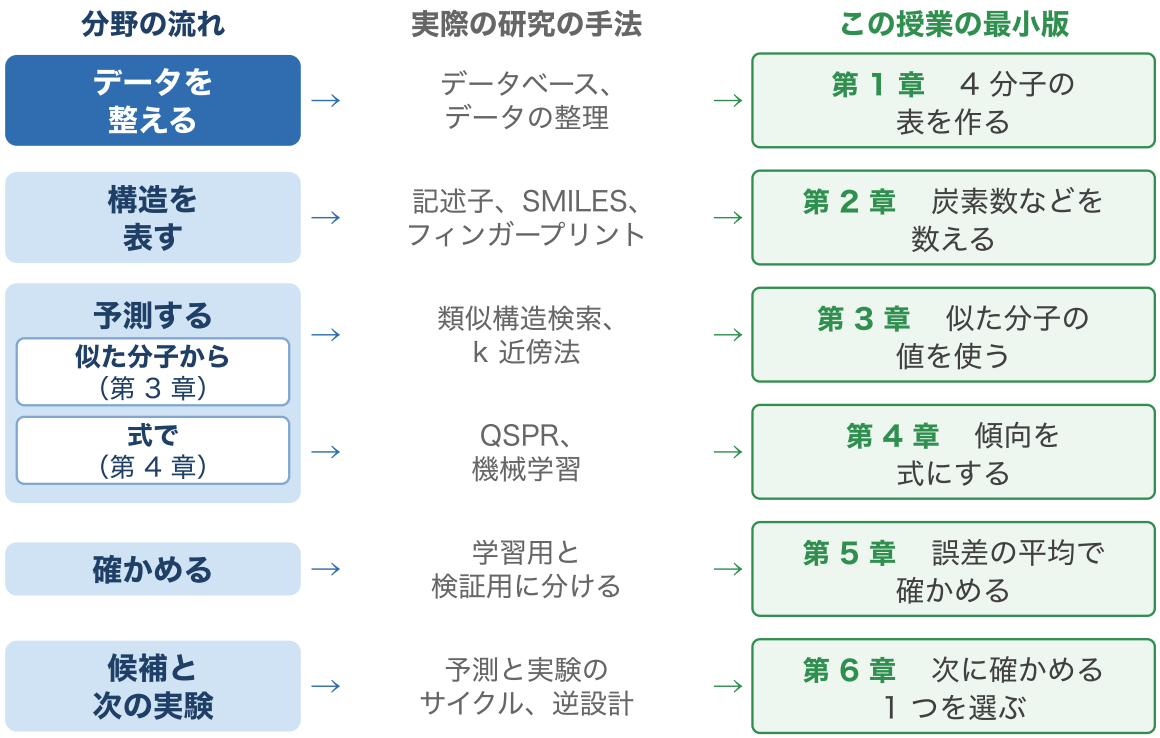
どちらの研究でも、予測は答えを決めるためではなく、実験や計算を使う分子を選ぶために使われています。この授業で次に確かめる分子を選ぶ（第 6 章）ことの、規模の大きい版にあたります。

ケモインフォマティクスとは：分子をデータとして扱う

図 0-5 の各段を、分子について行う分野があります。

> **定義：ケモインフォマティクス（chemoinformatics）**
>
> 分子の構造や物性などの化学の情報をデータとして集め、計算機で表し、比べ、予測して、分子や実験の設計に使う分野です。

ケモインフォマティクスで扱うことを、この授業ノートの章に合わせて並べると、図 0-8 と次のようになります。



灰の字は名前を知る程度でよい（試験の範囲外）。上から下へ、データ駆動の流れの順

図 0-8　ケモインフォマティクスの流れです。左の列が分野の流れ、中の灰の字が実際の研究で使う手法の名前（覚えなくてかまいません）、右の列がこの授業で各章に手を動かす最小版です。

- **データを整える**：文献やデータベースから値を集め、測定の条件と出典をそろえます（第 1 章）。
- **構造を表す**：構造を、計算機で比べられる数や文字列にします。構造の特徴を数にしたものを記述子と呼びます（第 2 章）。
- **予測する**：構造の違いを距離にして、似た分子の値を使います（第 3 章）。構造と物性の関係を式にまとめ、データから係数を決めて予測します（第 4 章）。
- **確かめる**：式や係数を決めるのに使っていないデータで、予測の成績を測ります（第 5 章）。
- **候補と次の実験**：予測と、まだ分かっていないことを踏まえて、候補を選び、次に確かめるものを決めます（第 6 章）。

この授業の最小版と、実際の研究の規模

この授業では、各段をいちばん小さな形で体験します。分子は十数個、記述子は 3 つ、計算は手とアプリでできる範囲です。実際の研究では、分子の数も記述子の数もずっと多く、計算はプログラムに任せます（表 0-1）。

表 0-1　この授業で手を動かす最小版と、実際の研究で使う手法の対応です。右の列の手法は、名前と考え方を知る程度でかまいません（試験の範囲外）。

章	この授業の最小版	実際の研究では
第 1 章	4 分子の表。条件と出典をそろえる	データベースから集めたデータの整理
第 2 章	(n, s, b) の 3 列	SMILES・分子グラフ、記述子・フィンガープリント
第 3 章	3 分子との距離と最近傍	類似構造検索、k 近傍法
第 4 章	係数 2 つの補正式	QSPR・QSAR の回帰、機械学習
第 5 章	5 分子で平均絶対誤差を求める	学習用と検証用のデータの分割、交差検証、過学習、適用範囲
第 6 章	次に確かめる 1 つを選ぶ	予測と実験を繰り返すサイクル、ベイズ最適化・能動学習、逆設計

規模は大きく違っても、考え方は共通しています。比べる前に条件をそろえる。構造の何を数にしたかを意識する。予測は、決めるのに使っていないデータで確かめる。限られた実験を、判断を変えそうな所に使う。各章の終わりの「実際のケモインフォでは」の箱で、その章の最小版が実際の研究のどれにあたるかを示します。

授業ノートと演習アプリの使い方

各章は、だいたい次の順に進みます。章の初めに、その章で考える問い（今回の問い）とねらいを示します。続いて基礎の解説、解き方の例（例題）、自分で取り組む演習、結果の解釈と設計の判断と進み、最後に依頼について分かったことと、まだ分からないことをまとめます。

箱の見方

本文の途中には、色の付いた箱を置いています。「定義」は重要な用語を正式に決める箱で、全巻で一度だけ置きます。「要点」は覚えておく結論、「よくある誤解」は間

違えやすい読み方です。「発展」と「実際のケモインフォでは」は、試験の範囲外の内容です。飛ばしても本文の筋は切れません。

値の種類と色

この授業ノートでは、値を 3 種類に分けて書きます。**資料値**は文献に載った測定値で、出典を付けます。**予測値**は、この授業で作る方法（似た分子の値を使う方法や、式）で出した値です。**架空値**は練習のために作った値で、必ず「架空」と書き添えます。図では、資料値を青、予測値を橙、架空の練習データを灰色で描きます。

演習アプリと答えの残し方

演習の一部は、授業のスライドに組み込んだアプリで取り組みます。アプリは、予想や予測を記録してから結果を表示するように作ってあります。演習の答えは提出しませんが、手元に残しておいてください。第 6 章で、第 1 章の自分の答えと比べます。演習の参考解答は付録 D にまとめました。

資料値を見る順番

確かめるための資料値（構造を変えた分子の沸点）は、第 5 章 で初めて示します。第 2〜4 章では予測を立て、記録してから、第 5 章で資料値と照らし合わせます。この授業ノートは一冊にまとまっているので、後ろの章や付録を先に見ることもできます。それでも、予測を記録してから資料値を見る順番を守ると、予測がどれだけ当たったかを正直に確かめられます。

1

第 1 章

設計に使うデータを整える

序章では、1-ペンタノールを水に溶けやすい溶媒へ改良する依頼を受け取りました（序章の「今回の依頼：1-ペンタノールを改良する」）。要求は、沸点を 115〜140 ℃ に保ち、w_{25} を 4.0 以上にすることでした。



工程担当者

1-ペンタノールの、どこが足りないの？ 代わりを探すには、何を調べる？

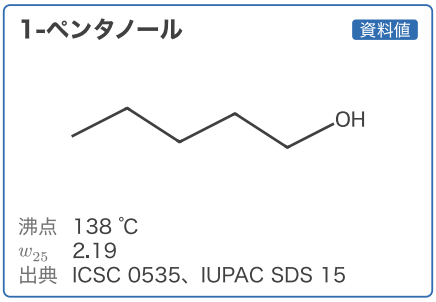


図 1-1　出発候補の 1-ペンタノールです。沸点も w_{25} も資料値で、出典を下に書きました。

この章の問いは、次のとおりです。

出発候補の何が不足し、どのデータが必要か。

答えるには、まず手元のデータを、比べられる形に整える必要があります。この章では、分子のデータを表にし、散布図で傾向を見て、1-ペンタノールを要求値と比べます。そのあと、文献で見つかる記録を、そのまま比べてよいかを確かめます。

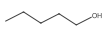
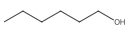
この章のねらい

- 分子のデータを、行（分子）と列（構造・物性・出典）の表に整理し、目的変数を言える。
- 散布図から読み取れること（観察）と、まだ確かめていない予想（仮説）を区別して書ける。

- 1-ペンタノールを要求値と比べ、何がどれだけ足りないかを数値で示せる。
- 記録の温度・相・溶質・分母を照らし合わせ、目標の量とそのまま比べられるかを判断できる。
- 欠測を 0 として扱わず、出典・転載・誤記に注意して記録を選べる。

1.1　データ表を作る：行・列・目的変数

データを集めたら、最初に表にします。この授業では、1 行を 1 つの分子にし、列に構造の特徴、物性、出典を並べます（図 1-2）。条件は、列の見出しに書いて固定します。沸点は常圧での値、w_{25} は 25 ℃・水相での値です。w_{25} の単位（g / 飽和溶液 100 g）も見出しに書き、表の数値には付けません。見出しで条件を決めておけば、行ごとに条件が変わることを防げます。

構造の列 → 第 2 章：構造を数の列にする		沸点：第 3〜4 章 予測する	水溶性：第 6 章 資料で確かめる	出典 → 第 1 章 そろえて残す
分子		**炭素数 n**	**沸点 T / ℃** 常圧（101.3 kPa）	**水溶性 w_{25}** （g / 飽和溶液 100 g） 25 ℃・水相	**出典**
	エタノール	2	78	混和	ICSC 0044
	1-ブタノール	4	117	—	ICSC 0111
	1-ペンタノール	5	138	2.19	ICSC 0535、IUPAC SDS 15
	1-ヘキサノール	6	157	—	ICSC 1084
	候補分子（例）	5	?	?	

青 資料値（資料集や論文に載った測定値）　— まだ集めていない（0 ではない）　混和 2 層に分かれない（w_{25} は定義されない）
? 知りたい値（予測するか、確かめる）　緑 見出しで固定した条件

図 1-2　分子のデータの表です。値はすべて資料値で、出典を右の列に書きました。条件は列の見出しに固定しています。エタノールの w_{25} の「混和」は、エタノールが水とどんな割合でも混ざって 2 層に分かれないので、w_{25} が定義されないことを表します。「—」はまだ集めていない値で、0 ではありません。「？」は、これから予測するか、確かめる値です。上の帯は、構造の列と物性の列を、どの章で扱うかを示します。

表の列には役割があります。分子の構造を表す列（ここでは炭素数）と、知りたい物性の列です。

> **定義：目的変数**
>
> 予測したい物性の列のことです。この依頼では、沸点 T と水溶性 w_{25} が目的変数です。

表の最下行は、構造を変えた候補分子です。物性の欄は「？」になっています。この欄を埋めること、つまり候補分子の沸点と w_{25} を知ることが、6 章を通しての課題です。構造の列をどう作るかは 第 2 章 で扱います。沸点の「？」をどう予測するかは第 3・4 章で扱い、w_{25} の「？」は 第 6 章 で資料により確かめます。

1.2　構造と物性の関係をデータから見る

表の数字を眺めるだけでは、傾向は見えにくいものです。構造の特徴を横軸に、物性を縦軸にとって、散布図にしてみましょう（図 1-3）。

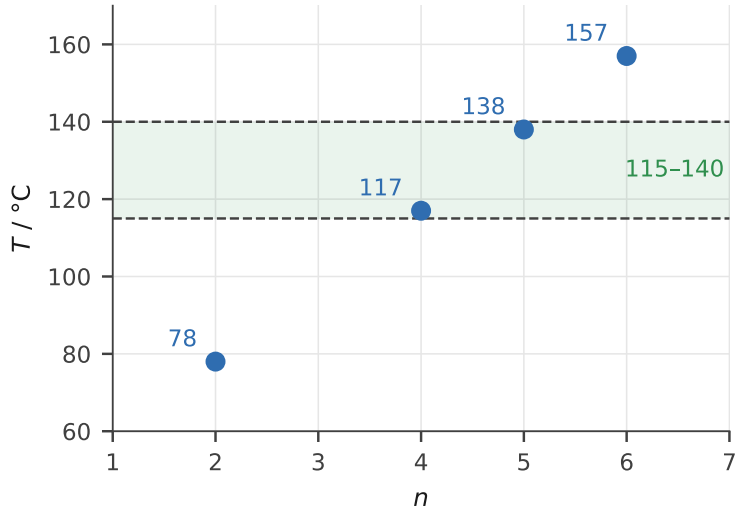


図 1-3　直鎖の一級アルコール 4 分子の炭素数 n と沸点 T です。点はすべて資料値（ICSC）です。破線と淡い緑は、沸点の要求の範囲（115〜140 ℃）です。この 4 例では、炭素数が多いほど沸点が高くなっています。

点は右上がりに並んでいます。炭素数が 2 → 4 → 5 → 6 と増えると、沸点は 78 → 117 → 138 → 157 ℃ と上がります。この 4 例では、炭素数が多いほど沸点が高い、と読み取れます。

ここで、データから読み取ったことと、まだ確かめていないことを分けておきます。

定義：観察と仮説

観察は、手元のデータで確かめられたことです。**仮説**は、データでまだ確かめていない予想です。

「この 4 例では、炭素数が多いほど沸点が高い」は観察です。一方、「OH の位置を変えれば、w_{25} も変わる」は、いまのデータには OH の位置が違う分子がないので、仮説です。仮説を立てること自体はよいことです。仮説は、次に何を測るかを決める手がかりになるからです。ただし、仮説を観察のように書かないことが大切です。

よくある誤解

4 例の観察を、ほかの分子にまで広げてしまうことです。図 1-3 の 4 分子は、どれも直鎖の一級アルコールです。枝分かれのある分子や、OH が鎖の内側にある分子が同じ並びに乗るかは、このデータからは分かりません。水溶性が炭素数でどう変わるかも、このデータには入っていません。観察は「この 4 例では」まで、と範囲を付けて書きます。

1-ペンタノールを要求値と比べる

次に、出発候補の 1-ペンタノールが、依頼の 2 つの要求を満たすかを確かめます。要求の型が違うので、判定のしかたも違います（序章の「目標、制約、最終的に提案するもの」）。

まず、沸点を確かめます。沸点の要求は 115〜140 ℃ の範囲の条件なので、下限と上限の両方と比べます。1-ペンタノールの沸点は 138 ℃ です。

$$115 \le 138 \le 140 \tag{1.1}$$

範囲に入っているので、沸点の要求は満たします。ただし、上限までは $140 - 138 = 2$ ℃ しかありません。

次に、w_{25} を確かめます。w_{25} の要求は 4.0 以上の下限の条件なので、下限とだけ比べます。1-ペンタノールの w_{25} は 2.19 です。

$$4.0 - 2.19 = 1.81 \tag{1.2}$$

2.19 は 4.0 より小さく、1.81 足りません。

つまり、1-ペンタノールに足りないのは水溶性で、その量は w_{25} にして 1.81 です。沸点は範囲に入っていますが、上限に近い位置にあります。この 2 つの結果が、問いの前半「何が不足しているか」への答えになります（序章の 図 1-3）。

1.3　比較できるデータとは：単位・温度・相・測定条件

不足が水溶性だと分かったので、構造を変えた分子の水溶性を文献で探すことになります。すると、「水への溶解度 4.2」のような記録が見つかります。この 4.2 は、そのまま w_{25} の 4.0 と比べてよいでしょうか。

まず、w_{25} が何を量った値かを思い出します。1-ペンタノールと水を混ぜて 25 ℃ に置くと、2 層に分かれます。下の層は水に富む相（水相）で、水にアルコールが少し溶けています。上の層はアルコールに富む相で、アルコールに水が少し溶けています。w_{25} は、下の水相 100 g に溶けているアルコールの質量です。

ところが、資料の「溶解度」には、上の層の組成、つまりアルコールに溶けている水の量が載っていることもあります。2 層に分かれる組み合わせでは、便覧などが両方の層の組成を並べて載せることがあるからです。「水への溶解度」と書いてあっても、中身が上の層の値のことがあります。上の層の値は、水溶性の別の量り方ではありません。w_{25} とは別の量で、取り違えやすいだけです。

そこで、比べる前に、記録の条件を目標の量の定義と照らし合わせます。照らし合わせるのは次の 4 点です。

- **温度**：25 ℃ の値か。溶解度は温度で変わります。
- **相**（どの層を取ったか）：水相の値か。どちらの層の値かで意味が変わります（図 1-4）。
- **溶質**（その層に溶けている何の量か）：アルコールの量か。アルコール相中の水の量を書いた記録もあります。
- **分母**（何あたりか）：飽和溶液 100 g あたりか。水 100 g あたりや、溶液 100 mL あたりの値とは数が違います。

目標の w_{25} を 4 点で言えば、25 ℃・水相・アルコール・飽和溶液 100 g です。相と溶質は、続けて「水相の中のアルコール」と読みます。



図 1-4　w_{25} の定義です。左は 25 ℃ で 2 層に分かれた飽和溶液の模式図で、w_{25} は下の水相（水に富む相）の値です。右は、溶解度の書き方の例です。w_{25} と同じなのは一番上だけで、ほかは分母や溶質が違います。

4 点のどれかが違えば、数字だけを比べても意味がありません。冒頭の「4.2」は、条件が書いていないので、このままでは比べられません。

例題：分母が水 100 g の値を、溶液 100 g あたりに直す

架空の物質 X について、「25 ℃・水相・4.2 g アルコール／100 g 水」という記録 R3 があります。これを w_{25} と同じ分母（飽和溶液 100 g）に直してください。

まず、何を求めたいかを確かめます。分母を「水 100 g」から「溶液 100 g」に変えた値です。

次に、使う値を決めます。この記録では、水 100 g にアルコール 4.2 g が溶けています。溶液の質量は $100 + 4.2 = 104.2$ g です。

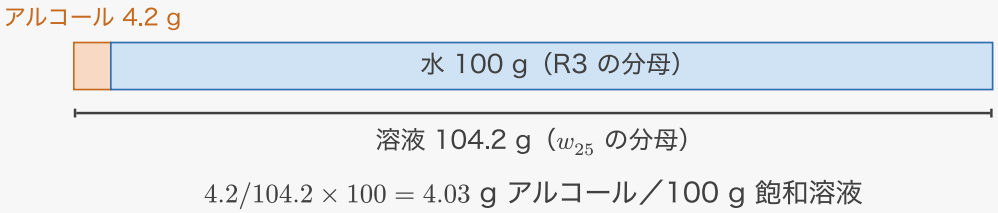


図 1-5　R3 の分母を直す様子です。水 100 g にアルコール 4.2 g が溶けているので、溶液は 104.2 g です。

計算します。溶液 104.2 g にアルコールが 4.2 g なので、溶液 100 g あたりでは

$$\frac{4.2}{104.2} \times 100 = 4.03 \tag{1.3}$$

となり、約 4.03 g アルコール／100 g 飽和溶液です。

> 結果の意味を確かめます。分母を読み替えただけで 4.2 のまま使うのは誤りです。この例では、直した後も 4.0 以上なので要求の判定は変わりませんが、いつもそうなるとは限りません。分母が違う記録は、直してから比べます。

同じ物質 X について、「25 ℃・水相・4.2 g アルコール／100 g 飽和溶液」という記録 R1 があったとします。R3 を直した 4.03 と R1 の 4.2 は一致しません。これは、どちらかが誤りだという意味でも、同じ値の別の書き方だという意味でもありません。条件をそろえることと、値が信頼できることは別の問題です。

発展　体積あたりの値を、質量あたりに直す　試験の範囲外

「溶液 100 mL あたり c g のアルコール」という記録を w_{25} に直すには、その温度の飽和溶液の密度 ρ（g/mL）が必要です。溶液 100 mL の質量は 100ρ g なので、

$$w = \frac{c}{100\rho} \times 100 = \frac{c}{\rho} \tag{1.4}$$

です。ρ には、同じ温度の飽和溶液の値を使います。水や純粋なアルコールの密度で代えることはできません。

1.4　欠測・出典・転載・誤記

条件をそろえても、記録の扱いで気をつけることがあります。

欠測は 0 にしない

表の空欄は、測っていない、または値が分からないことを表します。測って 0 だったこととは違います。たとえば、ある物質の記録が 3 つあり、1 つが空欄だったとします。空欄を 0 として平均に入れると、平均は実際より小さくなってしまいます。空欄は空欄のまま残し、平均に入れません（図 1-6）。

記録	温度	相	値	出典
a	25 ℃	水相	3.8	文献 1
b	25 ℃	水相		文献 2
c	25 ℃	水相	4.6	なし

b の空欄 → 欠測。測っていない・不明
0 を入れると、平均が下がって見える

c の出典なし → どこの値か確かめられない

架空の記録（値は練習用）

図 1-6　欠測と出典の扱いの例です（架空の記録）。b の空欄は欠測で、0 ではありません。c は出典がないので、どこの値か確かめられません。

図 1-2 の「—」も同じで、0 ではなく、まだ集めていない値です。

出典を書く

値には出典を付けます。出典があれば、後から条件を確かめ直せます。この授業の沸点は国際化学物質安全性カード（ICSC）の値です。w_{25} は、1937 年に報告された測定を、IUPAC の溶解度データ集（Solubility Data Series, Vol. 15）が再録した値を使います。同じデータ集には、評価者が丸めた推奨値も載っています。もとの測定値と丸めた推奨値は混ぜずに、どちらかにそろえます。この授業では、もとの測定値にそろえています。

転載と誤記

データベースには、ほかの資料から転載した値も多く載っています。たとえば、PubChem に ICSC の値が転載されていても、それは同じ値がもう 1 か所に書かれているだけで、確かめた測定が 2 つになったわけではありません。

誤った値が、資料から資料へ写されて広まることもあります。2-ブタノールの水溶性について、そのような誤記の広がりを扱った教育論文があります（Alger, 1991）。値に疑いがあるときは、転載をたどって、もとの測定まで確かめます。

> **要点**
>
> 比べる前に、温度・相・溶質・分母をそろえます。欠測は 0 にせず、空欄のまま残します。値には出典を付け、転載を独立した確認と数えません。

1.5　例題・演習：データを並べ、使える記録を選ぶ

ここまでの考え方を、例題と演習で使ってみます。演習で使う記録 R1〜R6 は、すべて架空の物質 X の練習用の記録です。実在の分子の値とは分けて扱います。

例題：記録を 3 つに分類する

架空の物質 X の水溶性の記録 R1・R4 を、目標の定義（25 ℃・水相・g アルコール／100 g 飽和溶液）と比べ、次の 3 つに分類してください。

- **そのまま比較できる**
- **換算や追加確認が必要**：情報を足せば、目標の量に直せる
- **現状では判断できない**：情報を足しても目標の量にならない、または値がない

記録	条件と値
R1	25 ℃・水相・4.2 g アルコール／100 g 飽和溶液
R4	25 ℃・アルコール相・8.0 g 水／100 g 飽和溶液

まず、R1 を 4 点で照らし合わせます。温度は 25 ℃、相は水相、溶質はアルコール、分母は飽和溶液 100 g です。4 点とも目標と同じなので、R1 は「そのまま比較できる」です。
次に、R4 を照らし合わせます。温度と分母は同じですが、相はアルコール相で、量っているのは水です。アルコール相の中の水の量から、水相の中のアルコールの量は出せません。情報を足しても目標の量にならないので、R4 は「現状では判断できない」です。
R4 の 8.0 を見て「4.0 より大きいから水に溶けやすい」と読むのは誤りです。量っている成分も相も違うので、4.0 と比べる意味がありません。

演習 1：何が足りないかを確かめる

授業のアプリ（データカードを並べて散布図に置くアプリ）か、手元のノートで取り組みます。

1. 図 1-2 の 4 分子を炭素数の順に並べ、散布図に点を打ってください。置く前に、炭素数と沸点の関係について仮説を 1 行書き、置いたあとに、分かったことを観察として 1 行書いてください。

2. 1-ペンタノールは、沸点と w_{25} の要求をそれぞれ満たしますか。満たさない方は、要求値との差を求めてください。

演習 2：記録を 3 つに分類する

授業のアプリ（記録を分類するアプリ）か、手元のノートで取り組みます。記録はすべて架空の物質 X のものです。



図 1-7　演習 2 の練習用の記録です（すべて架空）。例題で分けた R1・R4 は分類に入れてあります。R2・R5・R6 を分類します。

1. R5 は、どの分類に入りますか。
2. R2 と R6 を分類してください。それぞれ、目標の定義と比べるには、何の情報が必要ですか。

演習 3：最初の構造変更案と、必要なデータ

依頼の条件は、沸点 115〜140 ℃、w_{25} 4.0 以上、炭素数 5〜6 の一級または二級のアルコールです。

1. 1-ペンタノールの構造をどう変えますか。序章で書き留めた直感の候補分子を見直し、候補分子を 1 つ選んで構造式を描いてください。迷ったら、図 1-8 の構造から選んでかまいません。
2. その候補分子がよいかを判断するには、どんなデータが必要ですか。どの分子の、どの物性を、どの条件で、まで書いてください。
3. 図 1-2 の 4 分子のデータだけで、その候補分子がよいと言えますか。理由を 1 行で書いてください。

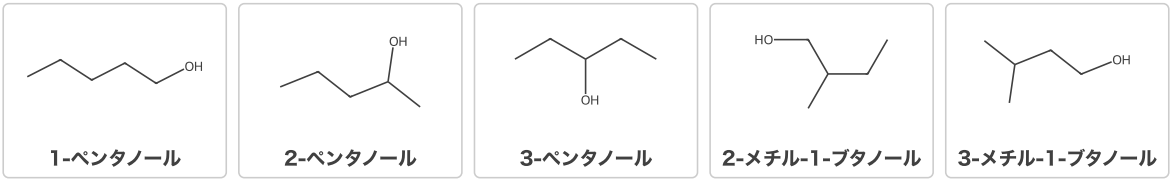


図 1-8　炭素数 5 の一級・二級アルコールの例です。左端の 1-ペンタノール（と直鎖の仲間）のほかは、まだ値を示しません。

答えは手元に残しておいてください。第 6 章 で、最終的な提案と比べます。演習の参考解答は付録 D にあります。

1.6　設計判断：出発候補の不足と、次に必要な情報

演習 3 に取り組んでから読んでください。問いに戻ります。出発候補の何が不足し、どのデータが必要でしょうか。

不足しているのは水溶性です（序章の 図 1-3）。1-ペンタノールの w_{25} は 2.19 で、要求の 4.0 に 1.81 足りません。沸点 138 ℃ は範囲に入っていますが、上限まで 2 ℃ しかありません。

構造を変えれば水溶性は変わるかもしれませんが、沸点も一緒に変わるかもしれません。ですから、候補分子を判断するのに必要なデータは、候補分子の w_{25} と沸点の 2 つです。w_{25} は、25 ℃・水相・飽和溶液 100 g あたりの値でなければ比べられません。「水溶性のデータ」と書くだけでは足りず、条件まで指定します。

いま手元にあるのは、直鎖の一級アルコール 4 分子のデータだけです。OH の位置を変えた分子や、枝分かれを付けた分子のデータはありません。ですから、どの候補分子がよいかは、いまのデータでは決められません。「OH を内側に移せば w_{25} が変わるかもしれない」と考えるのは自由ですが、それは仮説として書いておきます。変わるとしてどちらの向きかも、いまのデータでは分かりません。

候補分子は一つに決まりません。炭素数 5〜6 の一級・二級アルコールで、図 1-2 の 4 分子の資料値から要求を外れると分かっていない候補分子なら、どれも検討する価値があります。たとえば 1-ヘキサノールは、表に沸点 157 ℃ があり、上限を超えると分かっているので外れます。大事なのは、候補分子と一緒に、それを判断するのに必要なデータを、物性と条件まで書いておくことです。

物語の上では、ここで工程の担当者が、候補分子の資料（沸点と水溶性）を取り寄せることにします。資料が届くのは、沸点が 第 5 章、水溶性が 第 6 章 です。それまでの章では、届く前に、構造から沸点を見積もる方法を作ります。

実験系への接続：収率や材料特性を比べるとき

条件をそろえて比べることは、アルコールの溶解度に限りません。たとえば、同じ反応の収率の記録が 2 つあったとします（どちらも架空です）。1 つは「82 %（60 ℃・2 時間、NMR で定量）」、もう 1 つは「90 %（80 ℃・4 時間、単離して秤量）」です。温度も時間も、収率の求め方も違うので、90 % の方がよい反応条件だとは、この 2 つだけでは言えません（図 1-9）。



図 1-9　値だけでは比べられない記録の例です（すべて架空）。上は合成の収率、下は材料の電気伝導度です。どちらも、条件や求め方が違います。

材料の物性では、組成だけでなく、焼成の温度や試料の形（粉末か薄膜か）、測定の温度や方法も値に関わります。比べる前に、どの条件をそろえ、どの条件を記録しておくかを決めることは、この章で溶解度について行ったことと同じです。

> **実際のケモインフォでは　データを集めて、そろえる（キュレーション）** 試験の範囲外
>
> 実際の研究では、PubChem や NIST Chemistry WebBook のようなデータベース、資料集、論文から、ずっと多くのデータを集めます。集めたデータをそのまま使うこ

とはなく、重複と転載を除き、単位と条件をそろえ、誤記や外れ値をもとの測定までたどって確かめ、欠測を空欄のまま記録します。この作業をデータの整理（キュレーション）と呼びます（図 1-10）。

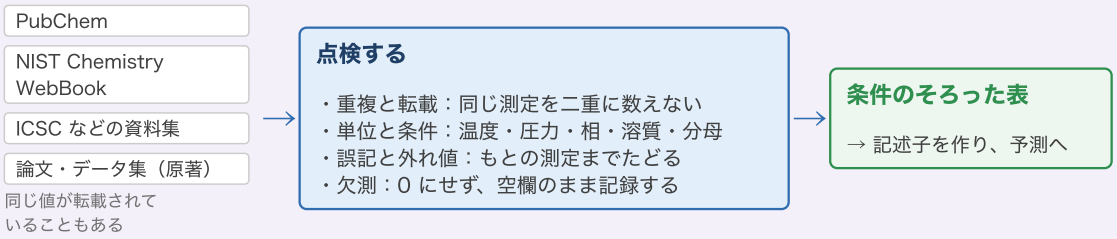


図 1-10　データの整理の流れです。集めた記録を点検し、条件のそろった表にしてから、記述子を作って予測に使います。

この章で 4 分子と練習用の記録について行ったことと共通しているのは、どんなに大きなデータでも、最初の仕事は列と条件をそろえることだ、という点です。

分かったこと・まだ分からないこと

分かったこと

- 直鎖の一級アルコール 4 例では、炭素数が多いほど沸点が高い（観察）。
- 1-ペンタノールの沸点 138 ℃ は範囲に入るが、上限まで 2 ℃。w_{25} 2.19 は要求の 4.0 に 1.81 足りない（資料値）。不足は水溶性。
- 記録は、温度・相・溶質・分母をそろえてから比べる。欠測は 0 にしない。

まだ分からないこと

- 構造の違いを、どう表して比べるか（第 2 章）。
- 構造を変えた分子の w_{25} と沸点。資料を取り寄せ中で、沸点は 第 5 章、水溶性は 第 6 章 で示す。

まとめ

- 分子のデータは、1 行を 1 分子、列を構造・物性・出典にした表にまとめ、条件は見出しで固定します。予測したい物性の列が目的変数です。

- 散布図で傾向を見るときは、データで確かめた観察と、まだ確かめていない仮説を分けて書きます。観察には「この 4 例では」のように範囲を付けます。
- 要求が範囲の条件なら両側と、下限の条件なら下限とだけ比べます。1-ペンタノールに足りないのは水溶性で、w_{25} にして 1.81 です。
- 記録は、温度・相・溶質・分母を目標の定義と照らし合わせてから比べます。分母が違うときは直してから比べ、相や溶質が違う記録は、そのままでは使えません。
- 欠測は 0 にせず、値には出典を付け、転載を独立した確認と数えません。

章末問題

手順を実行する

1. 架空の物質 E について、「25 ℃・水相・3.0 g アルコール／100 g 水」という記録があります。これを飽和溶液 100 g あたりの値に直してください。
2. 架空の物質 Z の沸点は 143 ℃、w_{25} は 4.6 です。依頼の 2 つの要求をそれぞれ満たすか判定し、満たさない方は要求値との差を求めてください。
3. 次の架空の記録 U・V・W を、例題の 3 つに分類してください。U：25 ℃・アルコール相・6.5 g 水／100 g 飽和溶液。V：25 ℃・水相・値が空欄。W：20 ℃・水相・4.4 g アルコール／100 g 飽和溶液。

理由と限界を説明する

1. 図 1-3 の 4 例から「炭素数 5 のアルコールなら、沸点は 138 ℃ 前後である」と言えますか。理由を説明してください。
2. ある沸点の値が、PubChem と ICSC の両方に載っていました。「2 つの資料で確かめられた」と言えますか。理由を説明してください。
3. 同じ物質の w_{25} の記録が 3 つあり、値は 4.0、4.4、空欄でした。空欄を 0 として平均すると、どんな問題が起きますか。

別の題材へ使う

1. 同じ反応の収率の記録が 2 つあります（架空）。75 %（室温・3 時間、NMR で定量）と、88 %（50 ℃・1 時間、単離して秤量）です。この 2 つを比べて反応条件の良し悪しを判断する前に、何をそろえる必要がありますか。

2. 自分の研究（まだなら学生実験）で、結果を比べるときにそろえるべき条件と、記録しておくべき条件を 1 つずつ挙げてください。

2

第 2 章

分子構造を数値で表す

前の章では、1-ペンタノールに足りないのは水溶性だと分かりました。候補分子を考えるには構造を変えた分子を比べる必要がありますが、その前に残った問いがありました。構造の違いを、どう表して比べるか、です（第 1 章 の「まだ分からないこと」）。



工程担当者

候補の分子がいくつもあるね。比べられる形にしてくれる？

前の章の演習 3 で見た 5 つの構造を思い出してください（図 2-5）。5 つとも炭素数 5 のアルコールで、分子式はどれも $C_5H_{12}O$ です。違いは構造にしかありません。予測に使う表にこれらを並べるには、構造を数の列にする必要があります。

この章の問いは、次のとおりです。

分子のどの特徴を、表の列にするか。

この章では、構造の何をどう数にするかを決め、その列で何が区別でき、どの違いが表せないかを確かめます。

この章のねらい

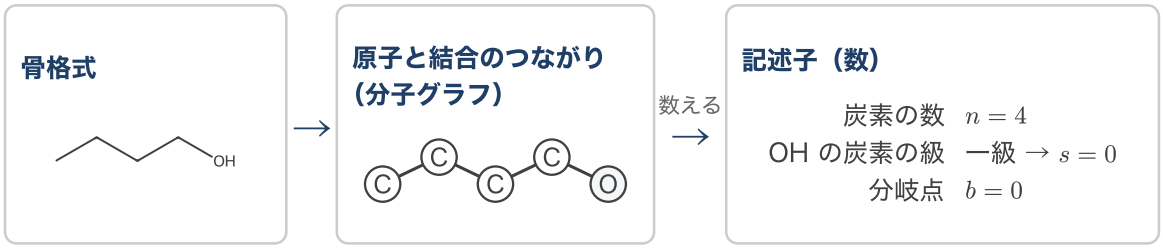
- 記述子が何かを説明し、表の構造の列として使える。
- 骨格式・縮合構造式から、炭素数 n、OH の炭素の級 s、分岐点の個数 b を数えられる。この 3 つを使ってよい分子の範囲を言える。
- 同じ記述子になる分子を見つけ、3 列だけの予測が同じ値になること、実際の物性が同じとは限らないことを説明できる。
- 列を 1 つ足したときに区別できる組とできない組を判断し、区別できることと予測がよくなることを分けて考えられる。

2.1　分子を原子と結合のつながりとして見る

構造を数にするには、まず構造式から何を読み取るかを決めます。

この本の構造式は、骨格式で描きます。骨格式では、線の端と折れ目が炭素で、炭素についた水素は省きます。酸素は O、OH は OH と書きます。たとえば 1-ブタノールの骨格式は、4 つの炭素がジグザグにつながり、端の炭素に OH がついた形です（図 2-1 の左）。縮合構造式で書けば $CH_3CH_2CH_2CH_2OH$ で、同じつながりを表しています。

骨格式から、原子を点、結合を線として取り出すと、原子と結合のつながりだけが残ります（図 2-1 の中）。このように分子を点と線のつながりとして見たものを、分子グラフと呼びます。分子グラフには、結合の長さや角度、分子の立体的な形は入っていません。入っているのは、どの原子がどの原子と結合しているか、だけです。



構造の列（記述子）→ この章　沸点：第 3・4 章 予測する　水溶性：第 6 章 資料で確かめる

分子	n	s	b	沸点 T / ℃	水溶性 w_{25} g / 飽和溶液 100 g
1-ブタノール	4	0	0	117	—

図 2-1　骨格式から原子と結合のつながりを取り出し、数えて記述子にする流れです。例は 1-ブタノールです。数えた結果は表の 1 行になり、構造の列ができます。沸点 117 ℃ は 第 1 章の表の資料値です。表の上の帯は、沸点を予測し（第 3・4 章）、水溶性は予測せずに資料で確かめる（第 6 章）ことを示します。

この章で数えるのは、炭素どうしのつながりです。何個の炭素があるか、OH のついた炭素に炭素がいくつ結合しているか、ほかの炭素 3 個以上と結合する炭素があるか、を数えます。数えるときに炭素に番号を振ることがありますが、番号は数えるための印です。どの端から番号を振っても、数えた結果は変わりません。この本の図では、番号を名前の番号（2-ペンタノールの「2」など）に合わせて振ります。

2.2　記述子とは何か

図 2-1 の右の 3 つの数は、1-ブタノールの構造を数で表したものです。このような数を記述子と呼びます。

> **定義：記述子**
>
> 分子の構造の特徴を数にしたものです。記述子（descriptor）は、データ表の構造の列になります。

記述子を列にすると、構造の違いを数の違いとして比べられるようになります。図 2-1 の下の表のように、構造の列（記述子）と物性の列（目的変数）を並べておけば、第 3・4 章では、構造の列から沸点の列を予測します（第 3 章、第 4 章）。水溶性の列は予測せず、第 6 章 で資料値によって確かめます。

どの特徴を数えるか

構造の特徴は、数えようと思えばいくらでもあります。どれを列にするかは、予測したい物性に関わりそうかどうかで選びます。

沸点を予測したいとします。前の章で見たとおり、直鎖の一級アルコール 4 例では、炭素数が多いほど沸点が高くなっていました。炭素数のほかに、沸点に関わりそうな特徴は何でしょうか。分子間にはたらく力から考えると、次のような予想が立ちます。

- **OH の位置**：OH が鎖の内側にあると、OH のまわりの立体障害が大きくなり、分子どうしの水素結合をつくりにくくなるかもしれません。
- **枝分かれ**：枝分かれがあると、分子どうしが接する面が減って、分散力が弱くなるかもしれません。

どちらも、まだデータで確かめていない仮説です（第 1 章 の観察と仮説）。ここでの化学の知識は、どの列を作るかを決める手がかりとして使います。仮説が当たっているかどうかは、列を作ったあとにデータで確かめます（第 4 章・第 5 章）。

記述子の種類

記述子には、いろいろな種類があります。代表的なものを、この章で使うものと合わせて挙げます。

- **数える記述子**：原子の数、分岐点の数、OH の数、環の数など。この章の n と b はこの種類です。
- **あるかないかを表す記述子**：ある構造があれば 1、なければ 0 とする記述子です。この章の s（二級なら 1、一級なら 0）はこの種類です。決めておいた多くの部分構造について 0 と 1 を並べたものは、フィンガープリントと呼ばれます（章末の「実際のケモインフォでは」）。
- **構造から計算する記述子**：分子量（原子の質量を足し合わせる）や、分子グラフの上で原子と原子の間の結合の数を数えた距離などです。分子の立体的な形や大きさから計算する記述子もあります。

どの種類でも、大切なのは定義がはっきりしていることです。定義に従えば、誰が数えても同じ値になる記述子でなければ、表の列として使えません。

2.3　炭素数、OH 結合炭素の級数、分岐を数える

この授業では、次の 3 つの記述子を使います。

> **定義：記述子 n・s・b**
>
> - n：炭素の数。
> - s：OH が結合する炭素が二級なら 1、一級なら 0。OH の炭素に炭素が 1 個結合していれば一級、2 個なら二級です。
> - b：炭素骨格で 3 個以上の炭素と結合する炭素（分岐点）の個数。OH との結合は数えません。

3 つの記述子を並べた (n, s, b) を、記述子の組と呼びます。4-メチル-2-ペンタノールで数えてみましょう（図 2-2）。この分子は 第 6 章 で候補の 1 つとして出てきます（値はまだ示しません）。

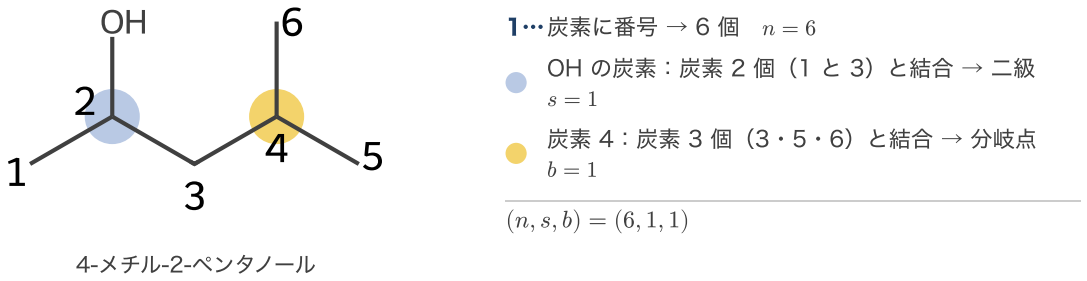


図 2-2　4-メチル-2-ペンタノールの n・s・b の数え方です。数字は炭素の番号（数えるための印）、紺の丸は OH の炭素、黄の丸は分岐点です。

まず、炭素に番号を振ります。炭素は 6 個なので $n = 6$ です。次に、OH の炭素（2 番）を見ます。炭素は 1 番と 3 番の 2 個が結合しているので二級で、$s = 1$ です。最後に、ほかの炭素 3 個以上と結合する炭素を探します。4 番の炭素は 3・5・6 番の 3 個と結合しているので分岐点で、ほかにはないので $b = 1$ です。まとめて $(n, s, b) = (6, 1, 1)$ になります。

数えるときに、間違えやすい点が 3 つあります。

- **s は OH の位置番号ではありません**。2-ペンタノールの「2」は名前の上での位置で、s は OH の炭素が一級か二級かを表します。2-ペンタノールも 3-ペンタノールも二級なので、どちらも $s = 1$ です。
- **OH との結合は、分岐に数えません**。二級アルコールの OH の炭素は、炭素 2 個と酸素 1 個に結合していますが、数えるのは炭素だけなので分岐点ではありません。
- **枝分かれと二級は別のものです**。枝分かれしていても、OH の炭素が CH_2 なら一級です。枝分かれは b で、OH の炭素の級は s で数えます。

なお、同じ炭素に枝が 2 つつく場合（炭素 4 個と結合する炭素）も、分岐点 1 個と数えます。

3 つの記述子を使ってよい範囲

n・s・b は、鎖状で（環がなく）、炭素どうしの結合がすべて単結合で、OH を 1 つもつ、一級または二級のアルコールのために決めた記述子です。依頼の条件（炭素数 5〜6 の一級または二級のアルコール。序章の「今回の依頼：1-ペンタノールを改良する」）も、この範囲に入っています。範囲の外の分子には使いません。とくに、次の 2 つに注意します。

- **三級アルコール**：OH の炭素に炭素が 3 個結合している分子です。s は一級（0）と二級（1）しか表せないので、三級を表す値がありません。そのうえ、OH の炭素自身が「炭素 3 個と結合する炭素」になるので、b の分岐点に数えられてしまいます。OH の炭素の級と、炭素骨格の分岐の区別が崩れるので、使いません。
- **エーテル**：酸素が 2 つの炭素の間にあり、OH をもたない分子です。OH の炭素がないので、s が決まりません。さらに、OH がないので、分子どうしの水素結合のしかたがアルコールと違います。アルコールのデータで作った予測を当てはめる根拠がないので、使いません。

記述子は、決めた範囲の分子どうしを比べるためのものです。数えられるかどうかだけでなく、その数が同じ意味をもつかどうかも確かめてから使います。

例題：枝分かれと二級を見分ける

2-メチル-1-プロパノールと 2-ブタノールの (n, s, b) を求めてください。
まず、炭素の数を数えます。どちらも炭素は 4 個で、$n = 4$ です。
次に、OH の炭素に結合する炭素の数を数えます。2-メチル-1-プロパノールの OH の炭素は CH_2 で、炭素 1 個と結合しているので一級（$s = 0$）です。2-ブタノールの OH の炭素は、炭素 2 個と結合しているので二級（$s = 1$）です。
最後に、炭素 3 個以上と結合する炭素を探します。2-メチル-1-プロパノールでは、2 番の炭素が炭素 3 個と結合しているので、分岐点が 1 個（$b = 1$）です。2-ブタノールの OH の炭素は炭素 2 個と酸素に結合していますが、酸素は数えないので分岐点ではなく、$b = 0$ です。

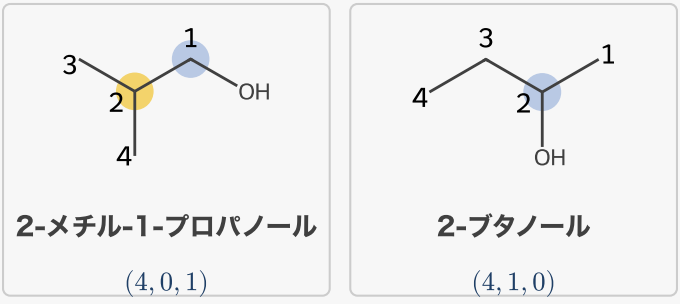


図 2-3　例題の 2 分子に印を付けたものです。紺の丸は OH の炭素、黄の丸は分岐点です。

> 答えは、2-メチル-1-プロパノールが $(4, 0, 1)$、2-ブタノールが $(4, 1, 0)$ です。2-メチル-1-プロパノールは枝分かれしていますが、一級です。枝分かれしていれば二級になる、というわけではありません。

2.4　同じ記述子になる異性体：表せない違い

構造を 3 つの数にすると、何が起きるでしょうか。2-ヘキサノールと 3-ヘキサノールで確かめます（図 2-4）。どちらも炭素は 6 個、OH の炭素は二級、分岐点はありません。記述子の組は、どちらも $(6, 1, 0)$ です。

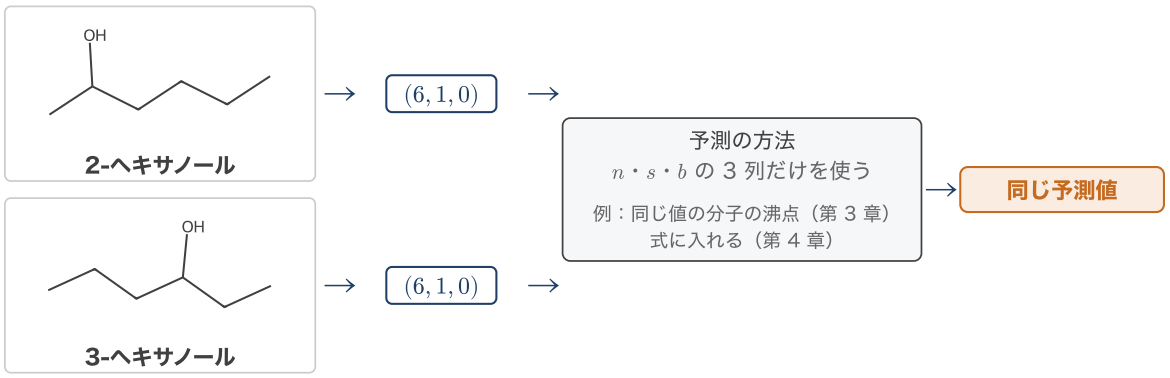


図 2-4　2-ヘキサノールと 3-ヘキサノールは、記述子の組がどちらも $(6, 1, 0)$ になります。この 3 列だけから予測すると、2 つは同じ値になります。構造だけの例で、値は使いません。

この 3 列だけから沸点を予測するとします。同じ値から同じ手順で計算すれば、予測は必ず同じ値になります。ですから、2-ヘキサノールと 3-ヘキサノールの予測値は同じになります。第 3 章 の距離で言えば、2 つの分子の距離は 0 です。第 4 章 の式で言えば、同じ予測値になります。

では、2 つの実際の沸点は同じでしょうか。それは、この表からは分かりません。2 つの分子は、OH が鎖のどこについているかが違います。その違いは、3 つの列のどこにも入っていません。構造を表にすると、表せる違いと、表せない違いがあります。ここで表せないのは、OH の位置です。

よくある誤解

記述子が同じなら物性も同じ、と読むことです。記述子が同じなのは、選んだ列では区別できない、という意味です。分子が同じという意味でも、物性が同じという意味でもありません。同じ記述子の分子の物性が違うかどうかは、資料値で確かめます（第 5 章）。

記述子が同じになる分子の組は、表の上で同じ行になります。同じ行の分子を見つけておくことは、その表で何が区別できないかを知ることです。

2.5　例題・演習：構造を表にし、追加する特徴を選ぶ

同じ行になる分子を区別したいときは、列を足すことになります。ただし、どんな列でも足せばよいわけではありません。例題で確かめてから、演習に取り組みます。

例題：列を足せば、区別できるか

2-ヘキサノールと 3-ヘキサノール（図 2-4）の分子量と OH の数を比べ、この 2 つの列を足せば 2 つを区別できるかを答えてください。

まず、分子式を確かめます。どちらも $C_6H_{14}O$ で、異性体です。分子量は分子式だけで決まるので、2 つの分子量は同じです。

次に、OH の数を数えます。どちらも 1 個です。

2 つの列のどちらも、2 つの分子で同じ値になります。ですから、分子量と OH の数を足しても、2 つは区別できません。

列は、区別したい分子の間で値が変わるものを選びます。この章の 5 つの分子（図 2-5）も分子式はどれも $C_5H_{12}O$ で、分子量も OH の数も 5 つで同じです。

演習 1：C5 の分子の記述子を数える

授業のアプリ（構造に印を付けて記述子を数えるアプリ）か、手元のノートで取り組みます。ノートでは、構造を写して、OH の炭素に丸、炭素 3 個以上と結合する炭素に三角を付けてから数えるとよいでしょう。

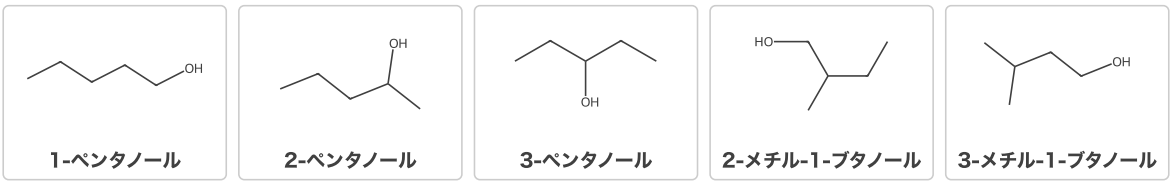


図 2-5　演習で使う、炭素数 5 の一級・二級アルコールの 5 つの構造です（図 1-8 と同じ）。

1. 1-ペンタノールと 2-ペンタノールの (n, s, b) を求めてください。違うのはどの記述子ですか。
2. 3-メチル-1-ブタノールの (n, s, b) を求めてください。s は 0 ですか、1 ですか。
3. 3-ペンタノールと 2-メチル-1-ブタノールの (n, s, b) を求めてください。

演習 2：同じ行になる分子をまとめる

授業のアプリ（同じ記述子の行をまとめるアプリ）か、手元のノートで取り組みます。先にノートに書いてから、アプリで確かめてください。

1. 演習 1 で求めた 5 つの分子の (n, s, b) を表にし、同じ行になる分子の組をすべて挙げてください。
2. この 3 列だけで沸点を予測すると、2-ペンタノールと 3-ペンタノールの予測値は同じになりますか。実際の沸点も同じと言えますか。

演習 3：区別できる列を 1 つ足す

演習 2 で見つけた同じ行の分子を区別するために、列を 1 つ足すことを考えます。足す候補として、次の記述子を使います。

> **定義：末端と、末端までの距離**
>
> **末端**は、炭素骨格で、ほかの炭素と 1 本だけ結合する炭素です。**末端までの距離**は、OH の炭素から最も近い末端までの、C–C 結合の数です。OH の炭素自身が末端なら、距離は 0 です。

1. 5 つの分子の末端までの距離を求めてください。同じ行だった 2 組は、それぞれこの列で区別できますか。
2. 2-メチル-1-ブタノールと 3-メチル-1-ブタノールで違う所を見つけ、その違いを「〔どこ〕から〔どこ〕までの〔何〕の数」の形で定義してください。

3. 通し番号（1〜5）の列でも、5 つは区別できます。区別できる列を足せば、沸点の予測は必ずよくなると言えますか。理由を 1 行で書いてください。

答えは手元に残しておいてください。演習の参考解答は付録 D にあります。

2.6　設計判断：今回の表現で何が分かり、何が分からないか

演習 3 に取り組んでから読んでください。問いに戻ります。分子のどの特徴を、表の列にすればよいでしょうか。

この授業では、n・s・b の 3 つを列にします。炭素数、OH の炭素の級、分岐点の数は、沸点に関わりそうな特徴（炭素数、OH の位置、枝分かれ）を、誰が数えても同じになる数で表したものです。5 つの分子を表にすると、3 種類の行になります（図 2-6）。

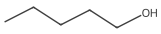
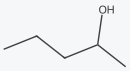
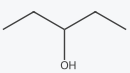
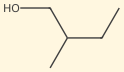
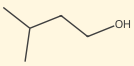
構造	分子	n	s	b	同じ行
	1-ペンタノール	5	0	0	**(5, 0, 0) 1 分子**
	2-ペンタノール	5	1	0	**(5, 1, 0) 2 分子**
	3-ペンタノール	5	1	0	
	2-メチル-1-ブタノール	5	0	1	**(5, 0, 1) 2 分子**
	3-メチル-1-ブタノール	5	0	1	

図 2-6　炭素数 5 の 5 つの分子の記述子です。同じ (n, s, b) の行をまとめると、3 種類になります。沸点などの値は、まだ載せていません。

この表で分かることと、分からないことを分けておきます。

- **分かること**：1-ペンタノール、二級の 2 つ、枝分かれのある 2 つは、表の上で違う行になります。炭素数・級・分岐の違いは、表で比べられます。
- **分からないこと**：同じ行の 2 つ（2-ペンタノールと 3-ペンタノール、2-メチル-1-ブタノールと 3-メチル-1-ブタノール）は、この表では区別できません。表せない違いは、OH の位置と、分岐の位置です。

では、列を足すべきでしょうか。末端までの距離を足せば、2-ペンタノールと 3-ペンタノールは区別できます。しかし、2-メチル-1-ブタノールと 3-メチル-1-ブタノールは、どちらも距離 0 で区別できません。末端までの距離を足しても、すべてを区別できるわけではありません。そのうえ、区別できるようになっても、その列が沸点の違いに合っているかは、資料値で確かめるまで分かりません。

この授業では、第 3〜5 章の予測を、n・s・b の 3 列のまま進めます。同じ行の分子を区別できないという限界は、そのまま残しておき、第 5 章 で資料値と比べて確かめます。末端までの距離を足して進める、という判断も成立します。その場合でも、何を区別できるようになり、何がまだ区別できないかを書いておくことは同じです。大事なのは、表に残した情報と、省いた情報を、表と一緒に書いておくことです。

物語の上では、工程の担当者に、候補を (n, s, b) の表にして渡せるようになりました。ただし、同じ行の分子はこの表では見分けられない、と一言添えて渡します。

発展　分岐点までの距離

試験の範囲外

2-メチル-1-ブタノールと 3-メチル-1-ブタノールの違いは、分岐点の位置です。そこで、OH の炭素から最も近い分岐点までの C–C 結合の数を数えると、2-メチル-1-ブタノールは 1、3-メチル-1-ブタノールは 2 で、区別できます（図 2-7）。



図 2-7　2-メチル-1-ブタノールと 3-メチル-1-ブタノールは、末端までの距離はどちらも 0 ですが、分岐点までの距離は 1 と 2 で違います。紺の丸は OH の炭素、黄の丸は分岐点です。

5 つの分子について、2 つの距離をまとめると次のようになります。

分子	(n, s, b)	末端までの距離	分岐点までの距離
1-ペンタノール	$(5, 0, 0)$	0	分岐なし
2-ペンタノール	$(5, 1, 0)$	1	分岐なし
3-ペンタノール	$(5, 1, 0)$	2	分岐なし
2-メチル-1-ブタノール	$(5, 0, 1)$	0	1
3-メチル-1-ブタノール	$(5, 0, 1)$	0	2

2 つの距離を足せば、5 つの分子はすべて違う行になります。ただし、分岐点がない分子には、分岐点までの距離が決まりません。「分岐なし」を 0 と書くと、OH の炭素自身が分岐点だという意味になってしまうので、0 とは書けません。列を足すときは、すべての分子で値が決まるかも確かめる必要があります。

実験系への接続：何を表の列にするか

表の列を決めることは、アルコールの記述子に限りません。合成なら、置換基を変えた基質の収率を表にするとき、置換基の種類と位置を列にして、反応温度・時間、溶媒、仕込みの規模を書かなければ、それらは表に残りません。材料なら、組成比と焼成条件を列にすれば、粒の大きさや試料の形は残りません（図 2-8。どちらも架空の模式で、値はありません）。

合成：置換基を変えた基質と収率

列にした	置換基の種類（Me・OMe・Cl）、位置（o・m・p）
省いた	反応温度・時間、溶媒、仕込みの規模

材料：組成を変えた試料と電気伝導度

列にした	組成比、焼成温度、焼成時間
省いた	粒の大きさ、試料の形、調製の手順の細部

図 2-8　表の列にした情報と、省いた情報の例です（架空の模式）。上は合成の収率、下は材料の電気伝導度です。

省いた情報で結果が変わるなら、表の上で同じ行の試料なのに、結果がばらつきます。そのときに原因を探せるよう、列にした情報と省いた情報を、表の横に書いておきます。これは、この章でアルコールについて行ったことと同じです。

実際のケモインフォでは　分子を、数の並びにする　試験の範囲外

実際の研究では、分子をプログラムで扱える形にします。SMILES は、分子を 1 行の文字列で書く約束です。たとえば 1-ペンタノールは CCCCCO と書きます（C が 5 つつながり、最後に O）。プログラムの中では、分子は分子グラフ（原子を点、結合を線にしたつながり）として扱われ、そこから記述子を数えます。記述子には、分

子量、環の数、水素結合をつくる原子の数など、多くの種類があり、ソフトウェアで多数の列を一度に作ります。フィンガープリントは、分子を 0 と 1 の並びにしたものです。決めておいた部分構造があれば 1 とする型（構造キー型）のほかに、分子の中の部分構造から自動で並びを作る型（ECFP など）もあります（図 2-9）。



図 2-9　分子を数の並びにする流れです。フィンガープリントは構造キー型の模式図で、実際のビットの割り当てではありません。

この章で 3 列を手で数えたことと共通しているのは、分子を比べられる数の並びにする、という点です。列がどれだけ多くても、同じ並びになる分子は区別できない、という点も同じです。

分かったこと・まだ分からないこと

分かったこと

- 構造を (n, s, b) の 3 列にすると、表の上で比べられる。炭素数 5 の 5 つの分子は、3 種類の行になる。
- 同じ行の分子（2-ペンタノールと 3-ペンタノール、2-メチル-1-ブタノールと 3-メチル-1-ブタノール）は、3 列だけから予測すると同じ値になる。物性が同じかどうかは、まだ分からない。

- 分子量や OH の数では、異性体は区別できない。末端までの距離で区別できる組もあるが、すべてではない。区別できることと、予測がよくなることは別。

まだ分からないこと

- 記述子の近い分子の沸点から、候補分子の沸点を予測できるか（第 3 章）。
- 同じ行の分子の沸点は、どれだけ違うか（第 5 章）。構造を変えた分子の資料は取り寄せ中で、沸点は 第 5 章、水溶性は 第 6 章 で示す。

まとめ

- 記述子は、分子の構造の特徴を数にしたもので、データ表の構造の列になります。定義に従えば、誰が数えても同じ値になるものを使います。
- この授業の記述子は、炭素数 n、OH の炭素が二級なら 1・一級なら 0 の s、分岐点の個数 b です。数えるのは炭素どうしの結合で、s は OH の位置番号ではなく、OH との結合は分岐に数えません。
- n・s・b は、鎖状・飽和で OH を 1 つもつ一級・二級のアルコールに使います。三級アルコールとエーテルには使いません。
- 同じ記述子の分子は、3 列だけから予測すると同じ値になります。実際の物性が同じとは限りません。OH の位置や分岐の位置のような違いは、この表では表せません。
- 列は、区別したい分子の間で値が変わるものを選びます。列を足しても、すべてを区別できるとは限らず、区別できても予測がよくなるとは限りません。

章末問題

手順を実行する

1. 次の 3 つの分子の (n, s, b) を求めてください。(a) $CH_3CH(OH)CH(CH_3)CH_2CH_3$ (b) $HOCH_2CH(CH_2CH_3)_2$　(c) $HOCH_2C(CH_3)_3$
2. 問 1 の (a) と、4-メチル-2-ペンタノール（図 2-2）の末端までの距離を求めてください。この列で 2 つを区別できますか。

3. 次の 2 つの分子に、n・s・b を使ってよいですか。それぞれ理由も答えてください。(a) $(CH_3)_2C(OH)CH_2CH_3$　(b) $CH_3OCH_2CH_2CH_2CH_3$

理由と限界を説明する

1. 問 1 の (a) と 4-メチル-2-ペンタノールは、記述子の組がどちらも $(6, 1, 1)$ です。「だから、2 つの沸点は同じである」と言えますか。理由を説明してください。
2.「分子量を列に足せば、分子式が $C_5H_{12}O$ のアルコールどうしを区別できる」は正しいですか。理由を説明してください。
3.「列を 10 個に増やせば、どの 2 つの分子も区別でき、沸点の予測もよくなる」という主張の問題点を説明してください。

別の題材へ使う

1. ある材料の試料を、組成比・焼成温度・焼成時間の 3 列で表にしました（架空）。同じ行の 2 つの試料で、電気伝導度が大きく違いました。考えられる理由と、表に足す列の候補を 1 つ挙げてください。
2. 自分の研究（まだなら学生実験）で、結果に関わりそうなのに、表の列にしていない情報を 1 つ挙げてください。それを列にするなら、どう数にしますか。

3

第 3 章

似た分子から物性を予測する

前の章では、分子の構造を (n, s, b) の 3 列にしました。表の上で分子を比べられるようになりましたが、まだ分からないことが残りました。記述子の近い分子の沸点から、候補分子の沸点を予測できるか、です（第 2 章 の「まだ分からないこと」）。



工程担当者

構造を変えると、沸点も範囲から外れないか心配。資料が届く前に、2-ペンタノールの沸点を見積もれる？

2-ペンタノールは、1-ペンタノールの OH を鎖の内側へ 1 つ移した分子で、水溶性を上げられないかと考えた候補分子の 1 つです。水溶性の資料は 第 6 章 で届きます。それまでに沸点を見積もって、要求の範囲（115〜140 ℃）を外れないかを見ておきたい、というのが依頼人の心配です。

手元には、沸点の資料値がある分子がいくつかあります。この章では、そのうち 3 つを参照分子として使います（図 3-1）。1-ペンタノールの沸点は 第 1 章 で見た値で、2-ブタノールと 2-メチル-1-プロパノールの沸点は、ここで初めて示す資料値です（どれも国際化学物質安全性カード）。

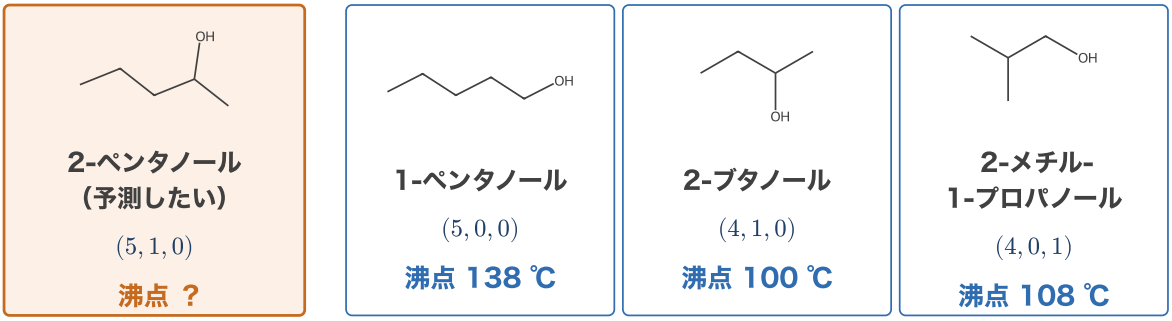


図 3-1　予測したい 2-ペンタノールと、参照する 3 つの分子です。参照分子の沸点は資料値（青）です。2-ペンタノールの沸点の資料値は、第 5 章 まで示しません。

この章の問いは、次のとおりです。

「似ている」をどう定義すると、予測がどう変わるか。

この章のねらい

- 最近傍による予測のしかた（最も似た分子の値を使う）を説明できる。
- 記述子の違いから距離を計算し、最近傍と予測値を求められる。同じ距離の最近傍が複数あるときの扱いを使える。
- 重みと尺度が何を決めているかを説明し、重みを変えると最近傍と予測が変わることを示せる。
- 最近傍が選べても予測が正しいとは限らないこと、1 例の一致から最良の決め方とは言えないことを説明できる。

3.1　似た分子の値を使う：直感と最近傍

2-ペンタノールの沸点を、図 3-1 の 3 つの参照分子から見積もるとします。最も素朴な方法は、2-ペンタノールに最も似ている分子を選び、その沸点を使うことです。

では、どれが最も似ているでしょうか。構造だけを見て、値を使いたい分子を 1 つ選び、理由を一言書いてみてください。「炭素数が同じだから 1-ペンタノール」と考える人もいれば、「OH が鎖の内側にある二級どうしだから 2-ブタノール」と考える人もいるでしょう。どちらも化学的に筋の通った考えです。人によって選ぶ分子が割れるのは、「似ている」をまだ決めていないからです。

「似ている」を数で決めれば、誰がやっても同じ分子を選べます。そのために使うのが距離です。

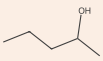
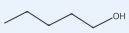
定義：距離

記述子の違いから決める、似ていなさの数です。距離が小さいほど似ていると考えます。

最も簡単なのは、炭素数だけを比べる距離です。予測したい分子の炭素数を n、参照する分子の炭素数を n' として、

$$d = |n - n'| \tag{3.1}$$

とします。2-ペンタノールは $n = 5$ なので、1-ペンタノール（$n' = 5$）までの距離は 0、2-ブタノールと 2-メチル-1-プロパノール（どちらも $n' = 4$）までの距離は 1 です（図 3-2）。

参照分子		参照分子の炭素数 n'	沸点 / ℃	距離
	**2-ペンタノール**（予測したい）	$n = 5$	?	
	1-ペンタノール	5	**138**	$\lvert 5-5\rvert = 0$ **最近傍**
	2-ブタノール	4	**100**	$\lvert 5-4\rvert = 1$
	2-メチル-1-プロパノール	4	**108**	$\lvert 5-4\rvert = 1$

→ 2-ペンタノールの予測 **138 ℃** 最近傍の値を使う

図 3-2　炭素数だけで距離を決めたときの、2-ペンタノールから 3 つの参照分子までの距離です。距離 0 の 1-ペンタノールが最近傍になり、その沸点 138 ℃ を予測値にします。予測値は橙で示します。

> **定義：最近傍による予測**
>
> 参照分子のうち、距離が最も小さいもの（最近傍）を選び、その物性の値を予測値にする方法です。

炭素数だけで決めると、最近傍は 1-ペンタノールで、2-ペンタノールの沸点は 138 ℃ と予測されます。この予測は、炭素数が同じことだけを「似ている」と見たものです。OH が鎖の端にあるか内側にあるか、つまり s の違いは、まったく見ていません。最近傍による予測は、式を作らずに似た例の値を使うだけなので、考え方は単純です。そのかわり、「似ている」の決め方が、そのまま予測を決めます。

3.2　記述子の違いを距離にする

s や b の違いも見るには、3 つの記述子の違いをまとめて 1 つの距離にします。この授業では、次の距離を使います。

$$d = |n - n'| + 2|s - s'| + |b - b'| \tag{3.2}$$

(n, s, b) が予測したい分子、(n', s', b') が参照する分子の記述子です。記述子ごとの違いを絶対値にして足します。絶対値にするのは、多いか少ないかの向きを問わず、違いの大きさだけを数えるためです。s の違いに掛けた 2 は重みで、意味は次の節で考えます。

例題：2-プロパノールの最近傍

2-プロパノール $(3, 1, 0)$ の沸点を、図 3-1 の 3 つの参照分子から、上の距離で予測してください。
まず、何を求めたいかを確かめます。3 つの参照分子までの距離を求め、最も近い参照分子の沸点を使います。
次に、各参照までの距離を計算します。記述子ごとの違いを書いてから足します。

- 1-ペンタノール $(5, 0, 0)$：$|3-5| + 2|1-0| + |0-0| = 2 + 2 + 0 = 4$
- 2-ブタノール $(4, 1, 0)$：$|3-4| + 2|1-1| + |0-0| = 1 + 0 + 0 = 1$
- 2-メチル-1-プロパノール $(4, 0, 1)$：$|3-4| + 2|1-0| + |0-1| = 1 + 2 + 1 = 4$

距離が最も小さいのは 2-ブタノール（距離 1）なので、最近傍は 2-ブタノールで、予測値は 100 ℃ です。
結果の意味を確かめます。2-プロパノールと 2-ブタノールは、どちらも二級で、炭素数が 1 つ違うだけです。この距離は、その近さを数で表しています。ただし、予測値 100 ℃ は 2-ブタノールの値を使っただけで、2-プロパノールの沸点がそうだと確かめたわけではありません。

重みを掛け忘れる（s の違いを 1 と数える）ことと、絶対値を忘れて $3 - 5 = -2$ をそのまま足すことが、よくある計算の誤りです。

発展　距離の決め方のいろいろ　試験の範囲外

この授業の距離は、違いの絶対値を足したものです。ほかに、違いを 2 乗して足し、平方根をとる距離（ユークリッド距離）もよく使われます。重みを同じように入れると、

$$d_{\mathrm{E}} = \sqrt{(n-n')^2 + (2(s-s'))^2 + (b-b')^2} \tag{3.3}$$

> です。2-ペンタノールから 3 つの参照分子までは、1-ペンタノール 2、2-ブタノール 1、2-メチル-1-プロパノール $\sqrt{6} \approx 2.45$ で、この例では最近傍は変わりません。ただし、ユークリッド距離では、1 つの記述子の大きな違いがより重く数えられます。距離の決め方が変われば最近傍が変わる例もあり、どの決め方を使うかは、決め方の一部です。

3.3　尺度と重み：何の違いを重く見るか

距離の式の 2 は、何を決めているのでしょうか。

> **定義：重み**
>
> 記述子の違い 1 つを、距離いくつと数えるかを決める数です。

この授業の距離では、炭素数の違い 1 つと分岐点の数の違い 1 つは、どちらも距離 1 と数えます。一級か二級かの違いは、距離 2 と数えます（図 3-3）。つまり、OH の炭素の級の違いを、炭素 1 個の違いの 2 倍と見ています。

違い 1 つ	例	距離への寄与	
n が 1 違う	炭素が 1 つ多い・少ない		1
s が 1 違う	一級と二級（OH の炭素の級）		2
b が 1 違う	分岐点が 1 つ多い・少ない		1

図 3-3　違い 1 つが、距離にいくつ寄与するかです。s の重みを 2 にしたので、一級と二級の違いは、炭素 1 個の違いの 2 倍に数えます。

この 2 は、仮に決めた値です。化学の理屈やデータから自然に決まる値ではありません。重みは、何の違いを重く見るかという、予測をする人の判断を表しています。

尺度：値の幅をそろえる

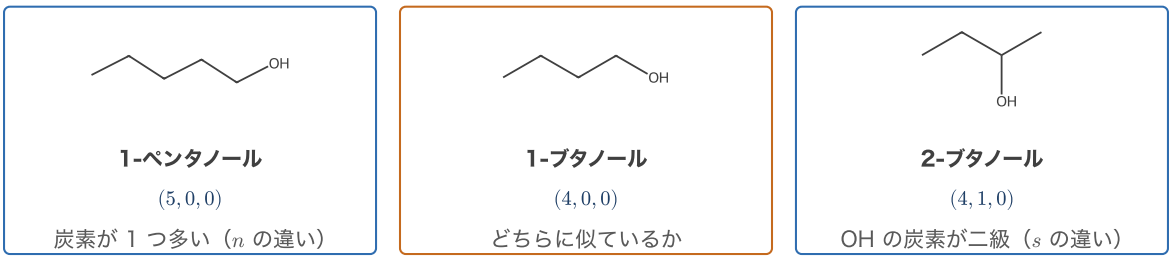
重みを考えるときには、記述子の値の幅（尺度）にも目を向けます。この授業のデータでは、炭素数 n は 2〜6 の範囲にありますが、s と b は 0 か 1 です。重みをすべ

て 1 にすると、炭素数の違いは最大で 4、s や b の違いは最大で 1 なので、距離は炭素数の違いでほぼ決まってしまいます。

値の幅の違う記述子をそろえる方法として、それぞれの記述子をその幅で割ってから距離を作る方法があります。たとえば n を幅 4 で割ると、炭素 1 個の違いは距離 0.25 になり、0 から 1 の幅の s や b と同じ尺度で比べられます。実際の研究では、このように尺度をそろえてから距離を作ることが多くあります。この授業では、尺度をそろえる代わりに、重みで直接決めています。どちらの場合も、どの違いをどれだけ重く見るかを、誰かが決めていることに変わりはありません。

3.4　重みを変えると、最近傍と予測が変わる

重みを変えると、何が起きるでしょうか。1-ブタノール $(4, 0, 0)$ に似ているのは、炭素が 1 つ多い 1-ペンタノールか、OH の炭素が二級の 2-ブタノールか、を考えてみます（図 3-4）。



1-ブタノールからの距離	1-ペンタノール	2-ブタノール	この 2 つだけで比べると
s の重み 2	1	2	1-ペンタノール
s の重み 1	1	1	**同じ距離**

図 3-4　1-ブタノールには、1-ペンタノール（炭素数の違い）と 2-ブタノール（OH の炭素の級の違い）の、どちらが似ているかです。s の重みが 2 なら 1-ペンタノールが近く、重みが 1 なら同じ距離になります。

1-ペンタノールまでの距離は、炭素数の違いだけなので、重みによらず 1 です。2-ブタノールまでの距離は、s の違いに掛ける重みそのものです。重みが 2 なら 2-ブタノールは距離 2 で、1-ペンタノールの方が似ていることになります。重みが 1 なら、2 つは同じ距離 1 になります。重みを変えるだけで、どちらが似ているかが入れ替わるのです。

3 つ目の参照の 2-メチル-1-プロパノール $(4, 0, 1)$ も比べると、重み 2 でも同じ距離が 2 つになります（1-ペンタノールと 2-メチル-1-プロパノールが、どちらも距離 1）。このように、重みを変えたときも、変えないときも、同じ距離の最近傍が 2 つ以上になることがあります。そのときの扱いを決めておきます。

要点

同じ距離の最近傍が 2 つ以上あるときは、そのすべてを使い、沸点の算術平均を予測値にします。どれか 1 つを選ぶと、選ぶ人によって予測が変わってしまうからです。

例題：同じ距離の最近傍が 2 つあるとき

1-ブタノール $(4, 0, 0)$ の沸点を、図 3-1 の 3 つの参照分子すべてと比べて、s の重み 2 の距離で予測してください。
まず、3 つの参照分子までの距離を求めます。

- 1-ペンタノール $(5, 0, 0)$：$1 + 0 + 0 = 1$
- 2-ブタノール $(4, 1, 0)$：$0 + 2 \times 1 + 0 = 2$
- 2-メチル-1-プロパノール $(4, 0, 1)$：$0 + 0 + 1 = 1$

次に、最近傍を決めます。1-ペンタノールと 2-メチル-1-プロパノールが、どちらも距離 1 で最も近くなります。
同じ距離の最近傍が 2 つあるので、2 つの沸点を平均します。

$$\frac{138 + 108}{2} = 123 \tag{3.4}$$

予測値は 123 ℃ です。1-ペンタノールの方が「似ている気がする」と片方だけを選ぶのは、決まりに合いません。
なお、1-ブタノールの資料値は 第 1 章 で見ました。予測と資料値を比べて、予測の方法のよしあしを判断する考え方は、第 5 章 で扱います。

重みを少し変えただけで最近傍が入れ替わると、予測値は、使う分子の値から別の分子の値へ一気に変わります。最近傍による予測は、決め方の小さな違いで大きく

変わることがあるのです。2-ペンタノールでこれがどう起きるかは、演習 3 で確かめます。

> **よくある誤解**
>
> 最近傍が選べたので、その予測値は正しい、と考えることです。距離は、構造がどれだけ近いかを、ある決め方で数にしたものです。距離が小さいことは、物性が同じことを保証しません。距離が 0 でも、記述子に入っていない違いで物性が違うことがあります（第 2 章 の「同じ記述子になる異性体」）。予測が合っているかは、資料値などで確かめます。

3.5　例題・演習：参照分子と重みを変えて予測する

ここまでの考え方で、2-ペンタノールの沸点を予測します。予測したい分子は 2-ペンタノール $(5, 1, 0)$ で、参照は 図 3-1 の 3 つです。

演習 1：s の重み 2 で予測する

授業のアプリ（最近傍を予想してから距離を計算するアプリ）か、手元のノートで取り組みます。

1. 1-ペンタノールまでの距離を、$d = |n - n'| + 2|s - s'| + |b - b'|$ から求めてください。
2. 2-ブタノールと 2-メチル-1-プロパノールまでの距離を求め、最近傍と予測値を出してください。計算する前に、どれが最近傍になるか予想しておきましょう。
3. 炭素数だけの予測（138 ℃）から予測が変わったのは、何の違いを見るようになったからですか。

演習 2：s の重み 1 で予測する

距離を $d = |n - n'| + 1 \cdot |s - s'| + |b - b'|$ に変えて、同じように予測します。

1. 3 つの参照分子までの距離を求め、最近傍と予測値を出してください。
2. 重み 2 の予測と比べて、予測は変わりますか。
3. 炭素数だけ、重み 2、重み 1 の 3 つの予測のうち、どれが正しいかを、いまの情報で決められますか。

演習 3：どの予測を採用するか

1. s の重みを 0.9 や 1.1 にすると、予測はいくつになりますか。重みを少しずつ変えると、予測はなめらかに変わりますか、それとも急に変わりますか。
2. 3 つの予測を表にし、それぞれ何の違いを見たかと、115〜140 ℃ に入るかを書いてください。
3. いまの時点で、どの予測を採用しますか。「〔どの予測〕を採用する。〔何の違い〕を見た。未確認は〔何か〕」の形で書いてください。正しさではなく、何の違いを見るかで選びます。「決めない」でもかまいません。

答えは手元に残しておいてください。演習 3 の答えは、第 4 章 で作る式の予測と並べて記録し、第 5 章 で資料値と比べます。演習の参考解答は付録 D にあります。

3.6　設計判断：どの予測を採用し、何を未確認とするか

演習 3 に取り組んでから読んでください。問いに戻ります。「似ている」をどう定義すると、予測がどう変わるのでしょうか。

2-ペンタノールの予測は、「似ている」の決め方によって、図 3-5 の 3 つになりました。

「似ている」の決め方	最近傍（距離）	予測	予測は 115〜140 ℃ に入るか
炭素数だけ	1-ペンタノール（0）	**138 ℃**	入る
s の重み 2	2-ブタノール（1）	**100 ℃**	**下回る**
s の重み 1	1-ペンタノールと 2-ブタノール（1） 同じ距離 → 平均	**119 ℃**	入る

図 3-5　2-ペンタノールの沸点の 3 つの予測です。どれも、決め方から出た予測値（橙）です。右の列は、予測が沸点の要求の範囲（115〜140 ℃）に入るかどうかです。

炭素数だけで決めると、炭素数の同じ 1-ペンタノールが最近傍になり、138 ℃ です。s の重みを 2 にすると、OH の炭素の級が同じ 2-ブタノールが最近傍になり、100 ℃ です。重みを 1 にすると、2 つが同じ距離になり、平均の 119 ℃ です。記述子と重みの決め方で、最近傍も予測値も変わります。

予測の違いは、依頼人の心配への答えまで変えます。138 ℃ と 119 ℃ なら沸点の範囲に入りますが、100 ℃ なら下限の 115 ℃ を下回ります。どの決め方を使うかで、「2-ペンタノールは沸点の点で使えそうか」の判断が逆になるのです。

重みを少しずつ変えると

s の重みを α として、n と b の重みは 1 のままにすると、2-ペンタノールから各参照までの距離は、1-ペンタノールが α、2-ブタノールが 1、2-メチル-1-プロパノールが $2+\alpha$ です。2-メチル-1-プロパノールは、いつも 2-ブタノールより遠くにあります。ですから、最近傍は α と 1 の大小で決まります（図 3-6）。

- $\alpha < 1$ のとき：1-ペンタノールが最近傍で、予測は 138 ℃。
- $\alpha = 1$ のとき：1-ペンタノールと 2-ブタノールが同じ距離で、予測は平均の 119 ℃。
- $\alpha > 1$ のとき：2-ブタノールが最近傍で、予測は 100 ℃。

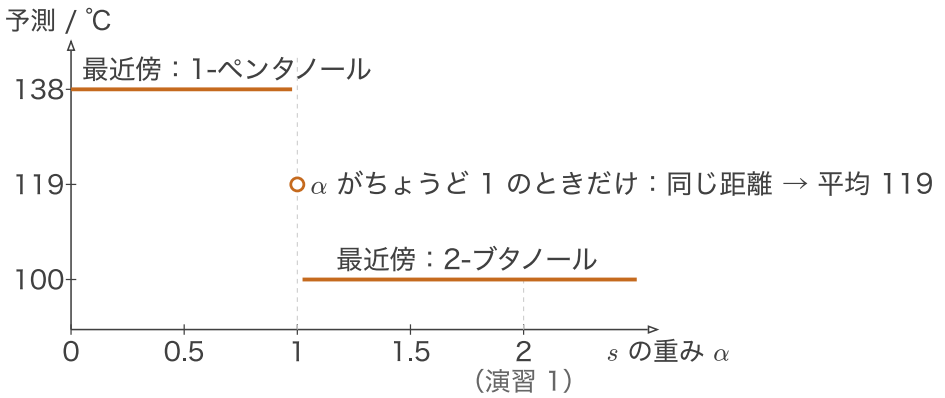


図 3-6　s の重み α を変えたときの、2-ペンタノールの予測です。最近傍が入れ替わる $\alpha = 1$ で、予測は 138 ℃ から 100 ℃ へ急に変わります。白抜きの丸は、α がちょうど 1 のときだけ出る予測値（119 ℃）です。横軸の $\alpha = 2$ は、演習 1 で使った重みです。

$\alpha = 0.9$ なら 138 ℃、$\alpha = 1.1$ なら 100 ℃ です。予測はなめらかには変わらず、最近傍が入れ替わる所で急に変わります。119 ℃ は、α がちょうど 1 で 2 つの参照が同じ距離になったときにだけ出る値です。

どの予測を採用するか

いまの時点で、どの予測を採用するかには、複数の答えが成立します。

- 138 ℃ を採用する。炭素数が沸点に最も影響すると考えた。
- 100 ℃ を採用する。OH の炭素の級の違いを、炭素 1 個の違いより重く見た。

- 119 ℃ を採用する。炭素数の違いと級の違いを、同じ重みで見た。
- まだ決めない。3 つとも決め方から出た予測なので、資料値で確かめてから選ぶ。

成立しないのは、「119 ℃ は真ん中だから正しい」「$\alpha = 1$ が最良の重み」のように、確かめていないことを結論にする答えです。真ん中にあることは、正しさの理由になりません。仮に 第 5 章 で、どれかの予測が資料値に近いと分かったとしても、それは 1 つの分子での一致です。1 例の一致から、その重みや最近傍の方法が最良だとは言えません。

ここで大切なのは、予測と一緒に、その決め方（記述子、重み、参照分子）と、まだ確かめていないことを書き残しておくことです。資料が届くまでは、どの予測がよいかは言いません。物語の上では、依頼人には「決め方によって 100〜138 ℃ と見積もりが割れる。下回るかどうかは、まだ言えない」と伝えることになります。

実験系への接続：何を似ているとするか

似た例の値を使うことは、研究室でもよく行います。新しい基質の反応の収率を、骨格の似た既知の基質の収率から見積もる。新しい組成の材料の性質を、組成の近い既知の材料から見積もる（図 3-7。どちらも架空の模式で、値はありません）。

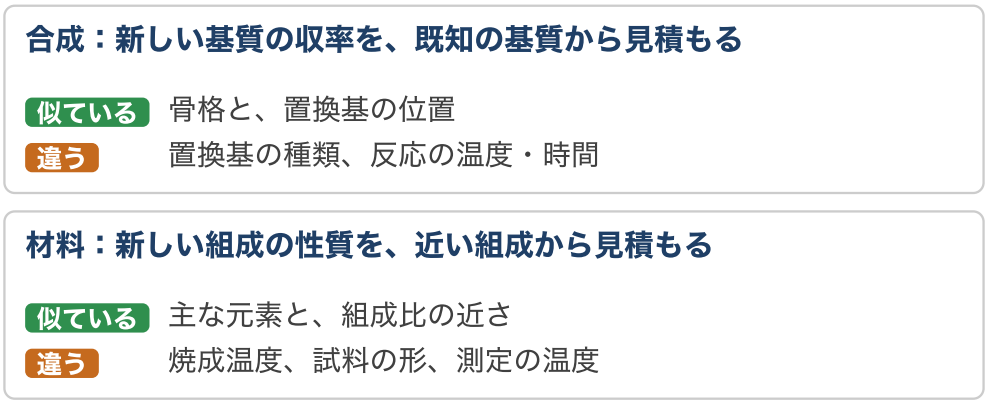


図 3-7　既知の結果を参照するときの、似ている点と違う点の例です（架空の模式）。上は合成の収率、下は材料の性質です。

このとき、似ている点だけを見て値を使うと、違う点（反応の温度や時間、焼成の条件など）の影響を見落とします。既知の結果を参照するときは、何を似ているとしたか、何が違うかを書いておきます。これは、この章で記述子と重みを書き残したことと同じです。

実際のケモインフォでは　似た分子を探し、値を使う　試験の範囲外

実際の研究では、データベースの多数の分子から、調べたい分子に似たものを探します（類似構造検索）。分子をフィンガープリント（第 2 章 の「実際のケモインフォでは」）にして、2 つの分子の並びを比べ、似ている度合いを数にします。よく使われる Tanimoto 係数は、2 つの分子で共通して 1 の部分構造の数を、どちらかで 1 の部分構造の数で割ったもので、0 から 1 の値をとり、1 に近いほど似ています（距離とは向きが逆です）。似ている順に並べた上位 k 個の値の平均で予測する方法は、k 近傍法と呼ばれます（図 3-8）。

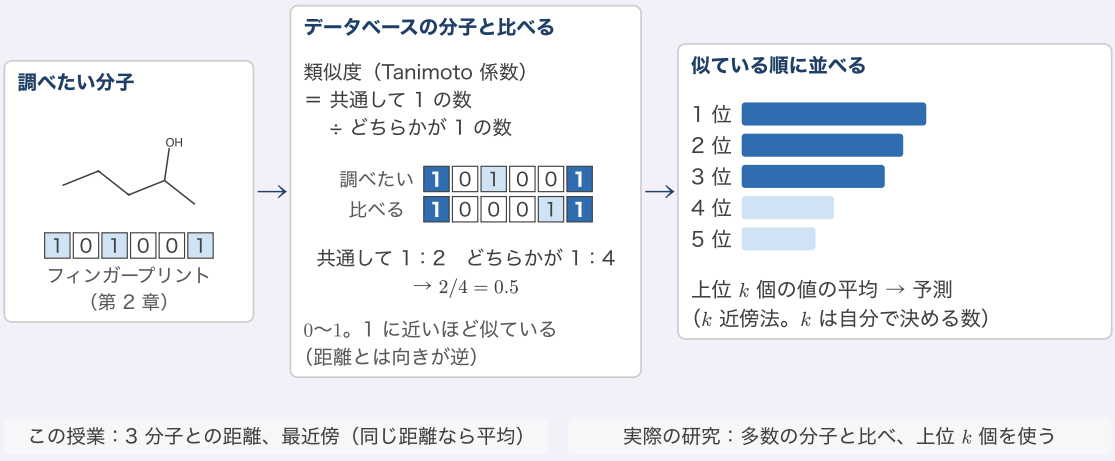


図 3-8　類似構造検索と k 近傍法の流れです。並びと棒は模式図です。

この章の最近傍は、$k = 1$ の k 近傍法（同じ距離なら平均）にあたります。実際の研究でも、どの記述子を使い、「似ている」をどう決めるかで、選ばれる分子と予測が変わる点は同じです。

分かったこと・まだ分からないこと

分かったこと

- 「似ている」は、記述子の違いから決めた距離で数にできる。最近傍の値を使って予測し、同じ距離の最近傍はすべて使って平均する。
- 2-ペンタノールの沸点の予測は、炭素数だけで 138 ℃、s の重み 2 で 100 ℃、重み 1 で 119 ℃（どれも予測値）。重みを少し変えると、予測は急に変わる。

- 最近傍が選べても、予測が正しいとは限らない。決め方しだいで、沸点の下回るかどうかの判断まで変わる。

まだ分からないこと

- 似た分子を 1 つ選んで値を使う代わりに、構造と沸点の傾向を式にまとめられるか（第 4 章）。
- 3 つの予測のどれが資料値に近いか（第 5 章）。2-ペンタノールの沸点の資料は取り寄せ中。

まとめ

- 最近傍による予測は、参照分子のうち距離が最も小さいものを選び、その値を使う方法です。同じ距離の最近傍が 2 つ以上あれば、すべて使って算術平均します。
- 距離は、記述子の違いから決める似ていなさの数です。この授業では $d = |n - n'| + 2|s - s'| + |b - b'|$ を使い、小さいほど似ていると考えます。
- 重みは、違い 1 つを距離いくつと数えるかの決め方で、自然に決まる値ではありません。尺度（値の幅）の違う記述子は、幅をそろえるか、重みで調整します。
- 2-ペンタノールの予測は、炭素数だけ 138 ℃、s の重み 2 で 100 ℃、重み 1 で 119 ℃ でした。119 ℃ は重みがちょうど 1 のときだけ出る値で、重みを少し変えると予測は急に変わります。
- 最近傍が選べても予測が正しいとは限りません。どの予測がよいかは資料値で確かめるまで言わず、1 例の一致から最良の決め方とは言いません。

章末問題

手順を実行する

1. 2-ヘキサノール $(6, 1, 0)$ の沸点を、1-ヘキサノール $(6, 0, 0)$・157 ℃、2-ブタノール $(4, 1, 0)$・100 ℃、1-ペンタノール $(5, 0, 0)$・138 ℃ の 3 つを参照として、s の重み 2 の距離で予測してください（参照分子の沸点は資料値です）。
2. 問 1 で、s の重みを 1 にすると、最近傍と予測値はどうなりますか。
3. 問 1 で、s の重みを α とします（n と b の重みは 1）。1-ヘキサノールだけが最近傍になるのは、α がどの範囲のときですか。

理由と限界を説明する

1. 問 2 の予測が、あとで 2-ヘキサノールの資料値に近いと分かったとします。「だから s の重みは 1 が最良である」と言えますか。理由を説明してください。
2. 「最近傍までの距離が 0 なら、予測は必ず正しい」は正しいですか。第 2 章 の内容を使って説明してください。
3. 炭素数の代わりに分子量（g/mol）を列にして、重みを付けずに同じ形の距離を作ると、どんな問題が起きますか。

別の題材へ使う

1. 新しい基質の反応の収率を、既知の基質の収率から見積もりたいとします。参照する基質を選ぶときに書き出しておくべきことを 2 つ挙げてください。
2. 自分の研究（まだなら学生実験）で、参照にしている既知の例を 1 つ挙げ、似ている点と違う点を書いてください。

4

第 4 章

データから予測式を作る

前の章では、似た分子の値を使って 2-ペンタノールの沸点を予測しました。ところが、「似ている」の決め方によって、予測は 138・100・119 ℃ に割れました。残った問いは、似た分子を 1 つ選ぶ代わりに、構造と沸点の傾向を式にまとめられるか、です（第 3 章 の「まだ分からないこと」）。



工程担当者

使う分子しだいで答えが変わるのは困る。式で言える？

図 4-1 は、これまでに資料値を見た 7 つの分子の炭素数と沸点です。丸の 4 点は直鎖の一級アルコール、四角の 3 点は二級か分岐のあるアルコールです。丸の 4 点に直線を引くなら、炭素 1 個あたり何 ℃ 上がりそうでしょうか。目で読んで、手元に書いておいてください。

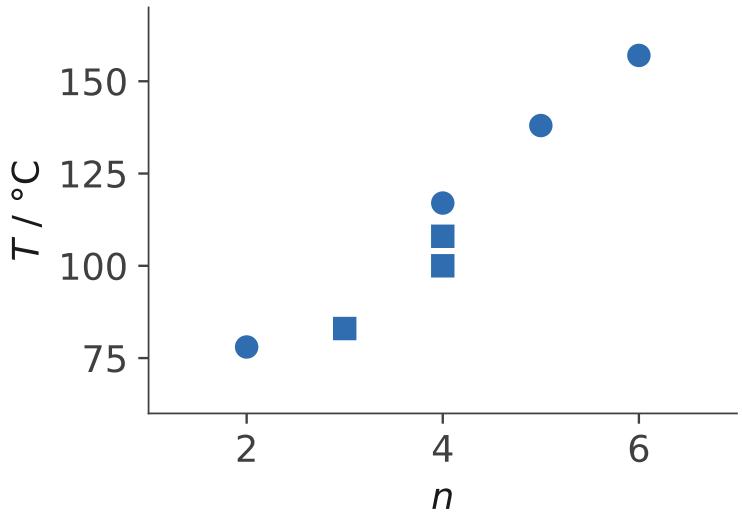


図 4-1　これまでに資料値を見た 7 分子の炭素数と沸点です（すべて資料値）。丸は直鎖の一級、四角は二級か分岐のある分子です。

この章の問いは、次のとおりです。

構造と沸点の傾向を、簡単な式にまとめられるか。

この章のねらい

- 説明変数・目的変数・係数の言葉を使って、構造から物性を出す式（QSPR）の考え方を説明できる。
- 炭素数だけの式 M0 で沸点を予測し、残差（資料値 − 予測値）を求めて、その符号の意味を読める。
- 残差平方和（SSE）を計算し、候補の中から係数を選べる。SSE と絶対誤差の違いを説明できる。
- 係数を決めた分子への当てはまりと、まだ使っていない分子への予測を区別し、係数を法則や機構の証明と読まない。

4.1　構造と物性の関係を式にする：QSPR の考え方

前の章の最近傍による予測は、似た分子の値をそのまま使う方法でした。この章では、構造の列から物性の列を計算する式を作ります（図 4-2）。

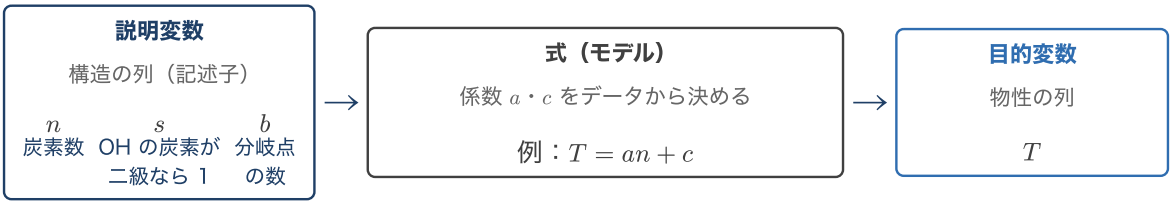


図 4-2　構造の列（説明変数）を式に入れて、物性の列（目的変数）を出します。式の係数は、データから決めます。

式に入れる列と、式で出す列を、次のように呼びます。

定義：説明変数と係数

説明変数は、式に入れる列です。この授業では構造の列（記述子 n・s・b）です。式で出す列は目的変数（第 1 章）で、ここでは沸点 T です。**係数**は、説明変数に掛ける数で、データから決めます。

構造と物性の関係を式で表したものを、定量的構造物性相関（quantitative structure–property relationship, QSPR）と呼びます。物性の代わりに薬の効き目などの生理活性を扱うときは、QSAR（activity）と呼びます。

式で予測すると、(n, s, b) を入れれば、誰が計算しても同じ予測値が出ます。似た分子を選ぶ必要はありません。ただし、式の形（どの説明変数を、どう組み合わせるか）と係数の値は、作る人が決めることです。決め方によって予測が変わる点は、前の章と同じです。この章では、式の形と係数を、どう決めたかを書き残しながら作っていきます。

4.2　炭素数から沸点を予測する基準式

まず、炭素数だけで沸点を表す式を作ります。直鎖の一級アルコール 4 分子（図 4-1 の丸）は、ほぼ一直線に並んでいます。この並びを直線の式 $T = an + c$ で表します。a が傾き、c が切片です（一次式の読み方は付録 A の「一次式：傾きと切片」）。

直線を決めるには、両端の 2 点を使います。エタノール $(n = 2, 78\,℃)$ と 1-ヘキサノール $(n = 6, 157\,℃)$ です。

まず、傾きを求めます。炭素数が 4 増えると、沸点は $157 - 78 = 79\,℃$ 上がります。

$$a = \frac{157 - 78}{6 - 2} = \frac{79}{4} = 19.75 \tag{4.1}$$

扱いやすいように、傾きを 20 に丸めます。

次に、切片を求めます。傾き 20 の直線がエタノールの点を通るように、$78 = 20 \times 2 + c$ から $c = 38$ と決めます。こうしてできた式を M0 と呼びます。

$$T = 20n + 38 \quad \text{(M0)} \tag{4.2}$$

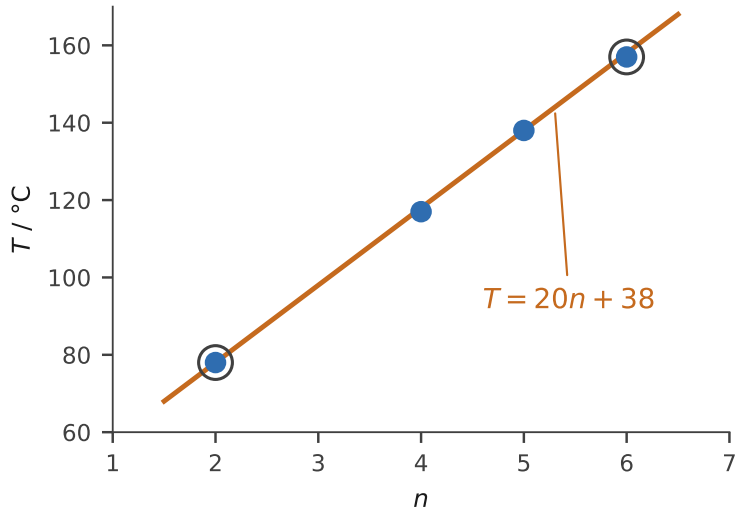


図 4-3　直鎖の一級アルコール 4 分子と、M0 の直線（予測＝橙）です。丸で囲んだ両端の 2 点から直線を決めました。

M0 の M はモデル（予測のしかた）、0 は補正なしを表します。この本では「炭素数だけの式」とも呼びます。傾き 20 は、炭素が 1 個増えると沸点が約 20 ℃ 上がる、という意味です。図 4-1 を見て書いた値と比べてみてください。

M0 は、両端の 2 点から決めた教育用の式です。4 点すべてにできるだけ合うように決めた式ではありません。また、傾き 19.75 をそのまま使うと切片は 38.5 になりますが、M0 の切片 38 は、38.5 を四捨五入したものではなく、傾きを 20 にしてからエタノールに合わせ直した値です。

4.3　残差を見る：式が説明できていない違い

式の予測は、資料値とどれだけずれるでしょうか。ずれを、次の量で表します。

定義：残差

資料値から予測値を引いた値です。

$$\text{残差} = \text{資料値} - \text{予測値} \tag{4.3}$$

引く向きは、この本を通して固定します。残差が負なら、資料値が予測より低い（式が高すぎる）ことを表します。

例題：1-ブタノールの残差

1-ブタノール $(4, 0, 0)$ の沸点の資料値は 117 ℃ です。M0 の予測と残差を求めてください。

まず、M0 に $n = 4$ を入れて予測値を求めます。

$$T = 20 \times 4 + 38 = 118 \tag{4.4}$$

予測値は 118 ℃ です。

次に、残差を求めます。資料値から予測値を引きます。

$$117 - 118 = -1 \tag{4.5}$$

残差は -1 ℃ です。

結果の意味を確かめます。残差が負なので、資料値は予測より 1 ℃ 低く、図 4-3 では点が直線のわずかに下にあります。1-ブタノールは M0 を決めた両端の 2 点には入っていませんが、直線の近くにあります。

引く向きを逆にして、予測値 − 資料値 $= +1$ としてしまう誤りがよくあります。向きを逆にすると、符号の読み方がすべて逆になります。

では、四角の 3 分子（二級か分岐のある分子）はどうでしょうか。M0 の予測と残差は、次のとおりです（図 4-4）。

表 4-1　二級・分岐の 3 分子の M0 の予測と残差です。資料値は国際化学物質安全性カードの値です。

分子	(n, s, b)	M0 の予測 / ℃	資料値 / ℃	残差 / ℃
2-プロパノール	$(3, 1, 0)$	98	83	−15
2-ブタノール	$(4, 1, 0)$	118	100	−18
2-メチル-1-プロパノール	$(4, 0, 1)$	118	108	−10

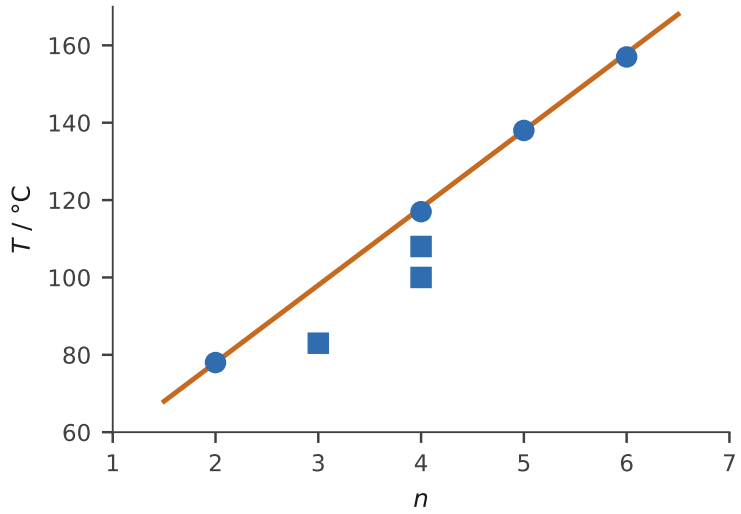


図 4-4　7 分子と M0 の直線です。四角の 3 点（二級・分岐）は、どれも直線より下にあります。点と直線の縦のずれが残差です。

3 つの残差は、どれも負です。直鎖の一級の分子では残差が 0 か −1 程度だったのに、二級・分岐の 3 分子では −10〜−18 ℃ と大きく外れ、しかも同じ向き（式が高すぎる向き）にそろっています。

2-ブタノールと 2-メチル-1-プロパノールは、どちらも炭素数 4 なので、M0 はどちらにも同じ 118 ℃ を出します。M0 は炭素数しか見ないので、炭素数の同じ分子を区別できません（第 2 章 の「同じ記述子になる異性体」と同じことです）。残差がそろって負になったのは、M0 が二級と分岐という構造の違いを表していないからだ、と考えられます。

> **要点**
>
> 残差は、資料値 − 予測値です。ある構造の分子で残差が大きく同じ向きにそろうなら、式がその構造を表していない手がかりになります。

4.4　構造による補正と、誤差を小さくする係数の選択

残差がそろって負になった分を、式から引けばよさそうです（図 4-5）。

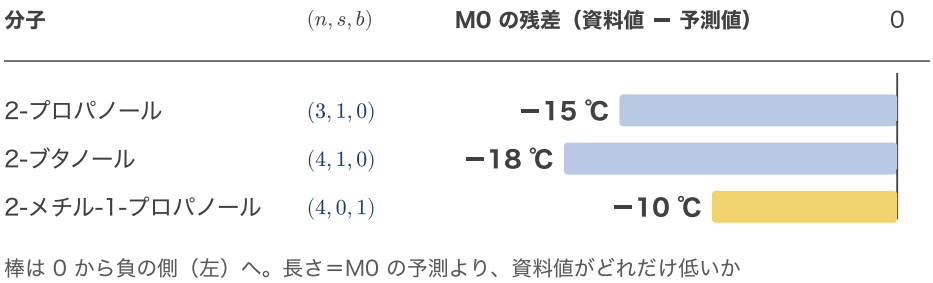


図 4-5　二級・分岐の 3 分子の M0 の残差です。棒の長さは、資料値が M0 の予測よりどれだけ低いかを表します。二級の 2 つ（−15、−18）は同じくらいで、分岐の 1 つ（−10）は少し小さくなっています。

そこで、二級なら c_s を、分岐点 1 個ごとに c_b を、M0 の予測から引く式にします。

$$T = 20n + 38 - c_s s - c_b b \tag{4.6}$$

c_s は二級の補正、c_b は分岐の補正で、どちらも正の数（℃）として決め、式ではマイナスを付けて引きます。直鎖の一級（$s = b = 0$）では、M0 と同じ予測になります。

係数を 1 つに決めるには

c_s は 1 つの数ですが、二級の分子は 2 つあり、残差は −15 と −18 で少し違います。2-プロパノールにぴったり合わせるなら $c_s = 15$、2-ブタノールにぴったり合わせるなら $c_s = 18$ です。両方にできるだけ合うように 1 つの値を選ぶには、合う度合いを 1 つの数にまとめる必要があります。

> **定義：残差平方和（SSE）**
>
> 各分子の残差を 2 乗して足した値です。
>
> $$\mathrm{SSE} = e_1^2 + e_2^2 + \cdots \tag{4.7}$$
>
> $e_1, e_2, \ldots$ は各分子の残差で、単位は ℃² です。SSE が小さいほど、式がデータに合っています（付録 A の「平方誤差と残差平方和」）。

残差平方和は sum of squared errors の略で、SSE と書きます。2 乗するのは、正と負の残差が打ち消し合わないようにするためです（図 4-6）。たとえば残差が +1 と −3 なら、そのまま足すと −2 になり、ずれが小さく見えてしまいます。2 乗して足せば $1 + 9 = 10$ です。また、2 乗すると、大きな残差ほど重く数えられます（残差 3 は残差 1 の 9 倍に数える）。

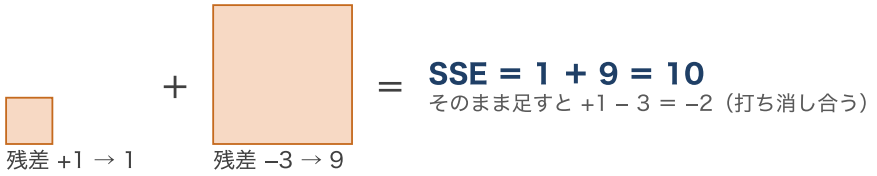


図 4-6　残差を 2 乗して足す理由です。正方形の面積が残差の 2 乗です。そのまま足すと、正と負の残差が打ち消し合います。

例題：$c_s = 20$ の残差平方和

二級の 2 分子（2-プロパノール 83 ℃、2-ブタノール 100 ℃）について、$c_s = 20$ の式 $T = 20n + 38 - 20s$ の SSE を求めてください。

まず、予測値を求めます。M0 の予測（98 ℃ と 118 ℃）から 20 を引きます。2-プロパノールは $98 - 20 = 78$ ℃、2-ブタノールは $118 - 20 = 98$ ℃ です。

次に、残差を求めます。$83 - 78 = +5$、$100 - 98 = +2$ です。

最後に、2 乗して足します。

$$\mathrm{SSE} = 5^2 + 2^2 = 25 + 4 = 29 \tag{4.8}$$

SSE は 29 ℃² です。

結果の意味を確かめます。補正をしない M0 のまま（$c_s = 0$）なら、残差は -15 と -18 で、SSE は $225 + 324 = 549$ ℃² でした。$c_s = 20$ にすると SSE は大きく下がりましたが、今度は残差が 2 つとも正になりました。引きすぎたのです。もっとよく合う c_s があるかもしれません。

この授業では、試す値の候補（15・17・20）を決めておき、それぞれの SSE を比べて最も小さいものを選びます（演習 2）。M0 の予測から c_s を引くのを忘れること、2 乗する前に足してしまうこと（$(5+2)^2$）が、よくある計算の誤りです。

分岐の補正 c_b は、分岐のある分子が 2-メチル-1-プロパノール 1 つしかないので、その残差を打ち消す値に決めます（演習 3）。1 分子なら、残差が 0 のとき SSE が最小になります。

平方誤差と絶対誤差の違い

打ち消し合いを防ぐだけなら、残差の絶対値を足す方法もあります。絶対値を使うと、残差 3 は残差 1 の 3 倍に数え、大きな残差を特に重くは数えません。どちらの基準を使うかで、選ばれる係数が変わることがあります。たとえば二級の 2 分子で

は、c_s が 15 から 18 の間なら、2 つの残差の絶対値の和はどれも 3 で、絶対値の和では 1 つに決まりません。

この授業では、係数を選ぶときは SSE を使います。第 5 章 では、予測の成績を ℃ の単位でまとめるために、絶対誤差の平均（平均絶対誤差）を使います。係数を選ぶ基準と、成績を要約する基準は別のものです。

4.5　例題・演習：補正式を作り、未知分子の予測を記録する

ここまでの考え方で、補正式を作り、まだ資料値を見ていない分子の沸点を予測します。

演習 1：M0 の残差を読む

手元のノートで取り組みます（授業のアプリでは、表 4-1 の 3 分子の残差を、符号を予想してから確かめられます）。

1. 直鎖の一級アルコール 4 分子（エタノール 78 ℃、1-ブタノール 117 ℃、1-ペンタノール 138 ℃、1-ヘキサノール 157 ℃）の M0 の残差を求めてください。
2. 問 1 の残差と、表 4-1 の二級・分岐の 3 分子の残差を比べてください。大きく同じ向きにずれる 3 分子は、s・b が直鎖の一級（どちらも 0）と何が違いますか。
3. 2-ブタノールと 2-メチル-1-プロパノールは、M0 の予測値が同じなのに、残差が違います。予測値が同じになる理由と、残差が違う理由を書いてください。

演習 2：残差平方和で c_s を選ぶ

授業のアプリ（試す値ごとに残差平方和を比べるアプリ）か、手元のノートで取り組みます。二級の 2 分子（2-プロパノール 83 ℃、2-ブタノール 100 ℃）を使います。

1. $c_s = 17$ の SSE を、例題と同じ手順で求めてください。
2. $c_s = 15$ の SSE を求め、試す値 15・17・20 のうち SSE が最も小さいものを選んでください。計算する前に、どれが最小になるか予想しておきましょう。
3. $c_s = 16$ の SSE も手で求めてください（アプリに 16 はありません）。問 2 で選んだ値だけが、最もよい係数だと言えますか。

演習 3：M1 を作る

2-メチル-1-プロパノール $(4,0,1)$ を使います。M0 の予測は 118 ℃、資料値は 108 ℃です（残差 -10 ℃）。

1. 残差が 0 になる c_b を、$118 - c_b = 108$ から求めてください。
2. $c_s = 17$ とあわせて式を書いてください。この式を M1（補正式）と呼びます。

演習 4：予測を記録する

M1 で予測し、第 3 章 の最近傍による予測と並べて記録します。

1. 表を手元に写し、M1 で 2 つの分子の沸点を予測して書いてください。

分子	最近傍（炭素数だけ）	最近傍（$\alpha = 2$）	最近傍（$\alpha = 1$）	M1
2-ペンタノール $(5,1,0)$	138 ℃	100 ℃	119 ℃	
3-メチル-1-ブタノール $(5,0,1)$	—	—	—	

   「—」は 第 3 章 で求めていない値です。
2. 「補正の係数を決めた 3 分子（2-プロパノール、2-ブタノール、2-メチル-1-プロパノール）で残差が小さい。だから、ほかの分子でもよく当たる」と言えますか。理由を 1 行で書いてください。

問 1 の記録は、資料値を見るまで書き換えないでください。これを「封をする」と呼び、第 5 章 で封を開けて資料値と比べます。演習の参考解答は付録 D にあります。

4.6　設計判断：経験式の化学的な解釈と限界

演習 4 に取り組んでから読んでください。問いに戻ります。構造と沸点の傾向を、簡単な式にまとめられるでしょうか。

まとめられます。炭素数の直線に、二級と分岐の補正の項を入れた式です。

$$T = 20n + 38 - 17s - 10b \quad \text{(M1)} \tag{4.9}$$

M1 の係数は、次のように決めました。どれも、ほかの決め方がありえた値です。

- **傾きと切片**：直鎖の一級の両端の 2 点から決めた（M0）。

- $c_s = 17$：二級の 2 分子について、候補 15・17・20 の SSE（9・5・29 ℃²）を比べ、最小の 17 を選んだ。ただし、16 でも SSE は 5 ℃² で同じなので、17 は試した 3 つの中で最小というだけで、唯一の最もよい値ではない。この本では、以後 17 を使う。
- $c_b = 10$：分岐のある分子 1 つ（2-メチル-1-プロパノール）の残差 −10 を打ち消す値にした。分岐の効果が一般に 10 ℃ だと確かめたわけではない。

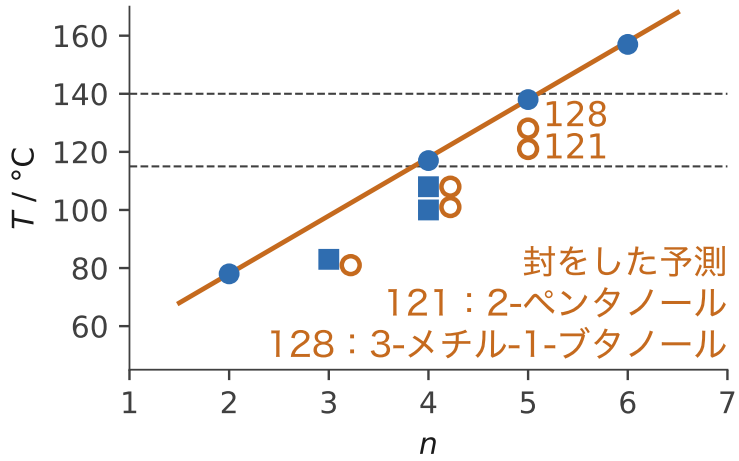


図 4-7　M0 の直線と、資料値（丸：直鎖の一級、四角：二級・分岐）、M1 の予測（白抜きの橙の丸）です。破線は沸点の要求の範囲（115〜140 ℃）です。M1 の予測が四角の 3 点に近いのは、この 3 点に合わせて係数を決めたからです。$n = 5$ の 2 つの白抜きの丸（札の 121・128）は、封をした 2-ペンタノール（121 ℃）と 3-メチル-1-ブタノール（128 ℃）の予測で、2 つとも要求の範囲の中です（当たるかは 第 5 章 で確かめます）。

係数を決めた 3 分子の M1 の残差は、2-プロパノール +2、2-ブタノール −1、2-メチル-1-プロパノール 0 ℃ で、どれも小さくなりました（図 4-7）。しかし、これは 3 分子に合うように係数を選んだのですから、合うのは当然です。係数を決めるのに使った分子への当てはまりは、まだ使っていない分子で当たることの証拠になりません。

M1 での予測は、2-ペンタノールが $20 \times 5 + 38 - 17 = 121$ ℃、3-メチル-1-ブタノールが $20 \times 5 + 38 - 10 = 128$ ℃ です。2-ペンタノールについては、最近傍の 3 つの予測（138・100・119 ℃）と並べて記録し、封をしました。c_s に 16 を選ぶ判断も成立し、その場合の 2-ペンタノールの予測は 122 ℃ です。どちらを選んでも、選び方を書いておくことが大切です。

発展　連続の値で最もよい係数と、最小二乗法　試験の範囲外

二級の 2 分子の残差は、c_s を使うと $83-(98-c_s)=c_s-15$ と $100-(118-c_s)=c_s-18$ です。したがって

$$\mathrm{SSE}=(c_s-15)^2+(c_s-18)^2 \tag{4.10}$$

で、これは c_s が 15 と 18 のちょうど真ん中、つまり平均の 16.5 のとき最小（4.5 ℃²）になります（図 4-8）。16 と 17 は、16.5 から同じだけ離れているので、SSE が同じ 5 ℃² になります。

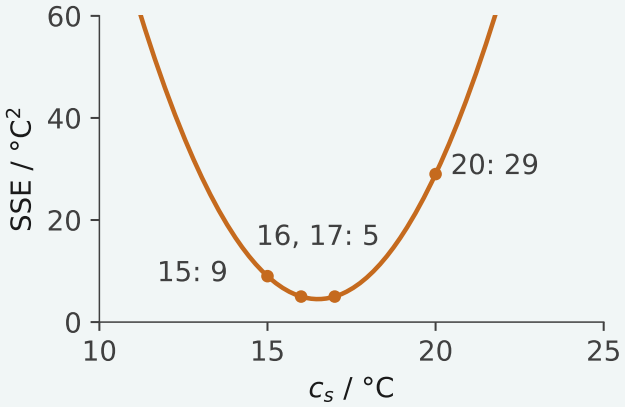


図 4-8　二級の 2 分子の SSE を、c_s の関数として描いた曲線です。SSE が最小になるのは $c_s=16.5$ で、16 と 17 は同じ高さです。

このように、SSE が最も小さくなる係数を求める方法を最小二乗法と呼びます。直鎖の一級 4 分子すべてに最小二乗法で直線を当てはめると、傾き約 19.8、切片約 38.2 となり、M0 とほぼ同じ直線になります。実際の研究では、すべての係数を最小二乗法でプログラムに同時に求めさせます。この授業の M1 は、係数を 1 つずつ、候補から手で選んだ式です。

係数の向きを化学で説明する

M1 の傾き 20 は、炭素数が多いほど沸点が上がる向きです。補正は、どちらも沸点を下げる向きです。これらの向きは、化学で説明できるでしょうか。図 4-9 は、1-ブタノールと、OH の炭素だけが二級になった 2-ブタノール、分岐だけがある 2-メチル-1-プロパノールの沸点の資料値を比べたものです。

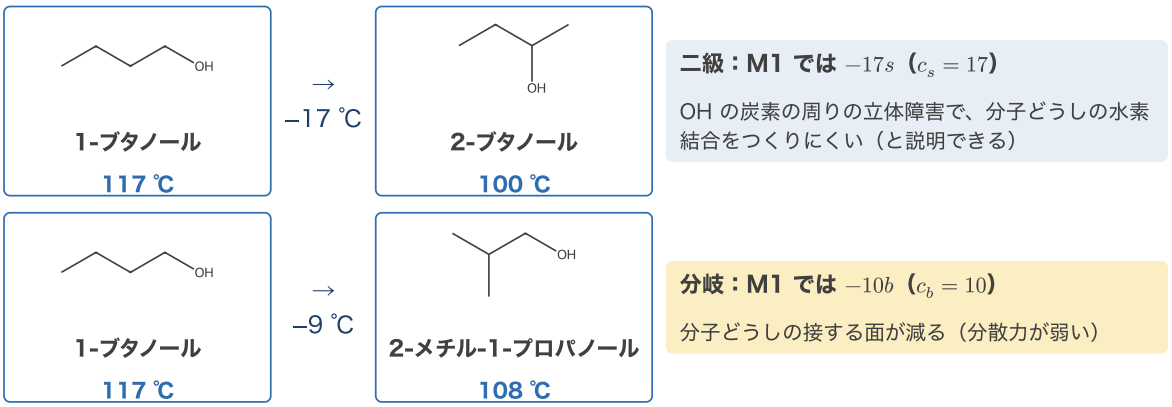


図 4-9　1-ブタノールと比べた、2-ブタノール（二級）と 2-メチル-1-プロパノール（分岐）の沸点の差です（資料値）。右は、補正の向きの化学的な説明です。

- **炭素数**：鎖が長いほど分子が大きくなり、分子どうしの分散力が強くなります。沸点が上がる（傾きが正）、と説明できます。
- **二級**：OH のついた炭素に炭素が 2 つつき、その周りに立体障害が生じます。分子どうしの水素結合をつくりにくくなり、沸点が下がる、と説明できます。
- **分岐**：枝分かれがあると、分子どうしが接する面が減ります。分散力が弱くなり、沸点が下がる、と説明できます。

どれも、データで見た向きを納得するための仮説です。確かめたのは向き（データの観察）までで、理由そのものを確かめたわけではありません。

資料値の差は 17 ℃ と 9 ℃ で、係数の 17 と 10 に近い値です。分岐の差 9 ℃ と係数 10 がずれるのは、1-ブタノール自身が M0 より 1 ℃ 低い（残差 −1）からです。ただし、差と係数は別の数です。差の 17 ℃ は 1-ブタノールと 2-ブタノールの資料値の差で、係数の 17 は二級の 2 分子（2-プロパノールと 2-ブタノール）の SSE から候補の中で選んだ値です。17 がそろったのは、たまたまです。

ここで、役割を分けておきます。式（データ駆動）は「どれだけ下がるか」を出し、化学の理屈は「なぜその向きか」を説明します（序章の「分子・材料を設計する 3 つの進め方」）。説明が係数の向きと合うことは、式を化学的に納得できる手がかりになります。けれども、説明は係数の値（17 や 10）を決めるものではなく、式が正しいことの証明でもありません。

よくある誤解

経験式の係数を、自然の法則のように読むことです。「二級にすれば、どの分子でも沸点が必ず 17 ℃ 下がる」とは言えません。$c_s = 17$ は、二級の 2 分子のデータから、試した候補の中で選んだ平均的な補正です。ほかの構造の分子や、二級と分岐を両方もつ分子で同じだけ下がるかは、確かめていません。

物語の上では、依頼人に「式で予測すると、2-ペンタノールは 121 ℃。ただし、まだ確かめていない」と伝え、予測の記録に封をしました。資料が届くのは 第 5 章 です。

実験系への接続：式から何が言えるか

研究室でも、焼成温度と電気伝導度、反応温度と収率のように、条件と結果の傾向を直線などの式に当てはめることがあります（図 4-10。架空の模式で、値はありません）。

例（架空）：焼成温度を変えた試料で、温度が高いほど電気伝導度が高い傾向。直線の式に当てはめた

式から言えること

- 調べた温度の範囲での、平均的な傾向
- その範囲の中の、別の温度での見積もり

式だけでは言えないこと

- なぜ上がるのか（機構）
- 範囲の外の温度でも同じ傾きか
- 温度を上げれば、必ず上がるか

図 4-10　条件と結果の傾向を式に当てはめたときに、言えることと言えないことの例です（架空の模式）。

式から言えるのは、調べた範囲での平均的な傾向と、その範囲の中での見積もりです。なぜそうなるのか（機構）、範囲の外でも同じ傾きか、条件を変えれば必ず結果が変わるか、は式だけでは言えません。傾向を式にしたら、調べた範囲と、言えないことを書いておきます。これは、この章で M1 の決め方と限界を書き残したことと同じです。

実際のケモインフォでは　データから式を決める　試験の範囲外

この章の M1 と同じ形（記述子に係数を掛けて足す）の式は、実際の研究でも使われています。たとえば水溶性を予測する ESOL という式は、記述子 4 つ（水と油へ

の分かれやすさの計算値、分子量、回転できる結合の数、芳香族の原子の割合）の一次式で、係数を測定値から、回帰（SSE が最小になる係数をプログラムで求める方法）で決めています（Delaney, 2004。図 4-11）。

この授業（第 4 章の M1）：沸点

$$T = 20n + 38 - 17s - 10b$$

記述子　3 つ（炭素数・OH の炭素の級・分岐点の個数）
決め方　直鎖の 2 点と二級・分岐の 3 分子から、手で決めた

実際の研究：水溶性の予測式 ESOL（Delaney 2004）

$$\log S = c_0 + c_1\,\mathrm{clogP} + c_2\,\mathrm{MW} + c_3\,\mathrm{RB} + c_4\,\mathrm{AP}$$

記述子　4 つ（油と水への分かれやすさ・分子量など）
S　水への溶解度（mol/L）
決め方　測定値から、回帰（SSE が最小になる係数をプログラムで求める）で決めた

図 4-11　この授業の M1 と、実際の研究の水溶性の予測式 ESOL の比較です。ESOL の式は形だけを示しました（c_0〜c_4 はデータから決めた係数）。

構造から物性や活性を出す式をデータから作ることを QSPR・QSAR と呼び、式には一次式のほか、機械学習の方法（決定木、ニューラルネットワークなど）も使います。どの方法でも、係数や規則は、誤差が小さくなるようにプログラムが決めます。この章で係数を候補から選んだことは、機械学習でいう「学習」の簡単な形にあたります。列や係数が多いほど、係数を決めた分子にはよく合うようになるので、使っていない分子で確かめることが、いっそう大切になります（第 5 章）。

分かったこと・まだ分からないこと

分かったこと

- 直鎖の一級の傾向は、M0（$T = 20n + 38$）にまとめられる。二級・分岐の分子では、M0 の残差がそろって負になる（−15、−18、−10 ℃）。

- 補正の項を入れた M1（$T = 20n + 38 - 17s - 10b$）を作った。c_s は候補から SSE で選び（16 も同じ SSE）、c_b は 1 分子から決めた。
- M1 の予測は、2-ペンタノール 121 ℃、3-メチル-1-ブタノール 128 ℃（予測値）。最近傍の 3 つの予測と並べて記録し、封をした。
- 係数の向きは化学で説明できるが、説明は係数の値も式の正しさも保証しない。

まだ分からないこと

- M1 の予測は、係数を決めるのに使っていない分子でも合うか（第 5 章）。
- 記録した予測（最近傍の 3 つと M1）のどれが、資料値に近いか（第 5 章）。2-ペンタノールなどの資料は取り寄せ中。

まとめ

- 構造と物性の関係を式で表したものを QSPR と呼びます。式に入れる列が説明変数、式で出す列が目的変数で、説明変数に掛ける係数はデータから決めます。
- M0（$T = 20n + 38$）は、直鎖の一級の両端の 2 点から決めた教育用の式です。すべての点に合わせた式ではありません。
- 残差は資料値 − 予測値です。二級・分岐の分子で残差がそろって負になったことから、その構造の補正を式に足しました。
- 係数は、試す値の候補ごとに残差平方和（SSE）を求め、最小のものを選びました。SSE は大きな残差を重く数え、絶対誤差とは別の基準です。
- M1（$T = 20n + 38 - 17s - 10b$）が係数を決めた分子に合うのは当然で、未知の分子で当たる証拠にはなりません。係数は法則ではなく、化学の説明も式の正しさを証明しません。

章末問題

手順を実行する

1. 架空の物質 Y1 は、記述子の組が $(6, 1, 0)$、沸点（架空）が 143 ℃ です。M0 の予測と残差を求めてください。

2. 架空の物質 Y1（$(6,1,0)$、143 ℃）と Y2（$(7,1,0)$、159 ℃）の 2 つを使って、二級の補正 c_s を候補 14・16・18 から SSE で選んでください。どの候補が選ばれますか。
3. M1 で、3-メチル-2-ブタノール $(5,1,1)$ と 2-エチル-1-ブタノール $(6,0,1)$ の沸点を予測し、それぞれ沸点の要求の範囲（115〜140 ℃）に入るかを答えてください。

理由と限界を説明する

1. 「$c_b = 10$ なので、分岐点を 1 つ増やせば、どの分子でも沸点はちょうど 10 ℃ 下がる」は正しいですか。理由を説明してください。
2. 「M0 は直鎖の一級 4 分子で残差が 0 か -1 と小さい。だから、炭素数 3 の二級アルコールにもよく合うはずだ」という主張の問題点を説明してください。
3. 係数を選ぶときに、残差を 2 乗せずにそのまま足し、その和が 0 に最も近い係数を選ぶと、どんな問題が起きますか。

別の題材へ使う

1. ある反応で、反応温度 40〜80 ℃ の範囲の収率に直線を当てはめたところ、「温度が 10 ℃ 上がると収率が約 5 % 上がる」という式になりました（架空）。この式から言えることと、言えないことを 1 つずつ挙げてください。
2. 自分の研究（まだなら学生実験）で、傾向を式や直線で表しているものを 1 つ挙げ、その式からは言えないことを書いてください。

5

第 5 章

予測を確かめ、信用できる範囲を考える

前の章では、補正式 M1 を作り、2-ペンタノールと 3-メチル-1-ブタノールの沸点を予測して、記録に封をしました。残った問いは、M1 の予測は、係数を決めるのに使っていない分子でも合うか、です（第 4 章 の「まだ分からないこと」）。

物語の上では、取り寄せていた資料が届きました。



工程担当者

資料が届いたよ。その予測、本当に信じていいの？

届いたのは、構造を変えた 5 つの分子の沸点の資料です（図 5-1）。どの分子も、M0・M1 の式を作るのにも、第 3 章 の参照分子にも使っていません。この 5 分子で、式の予測を確かめます。

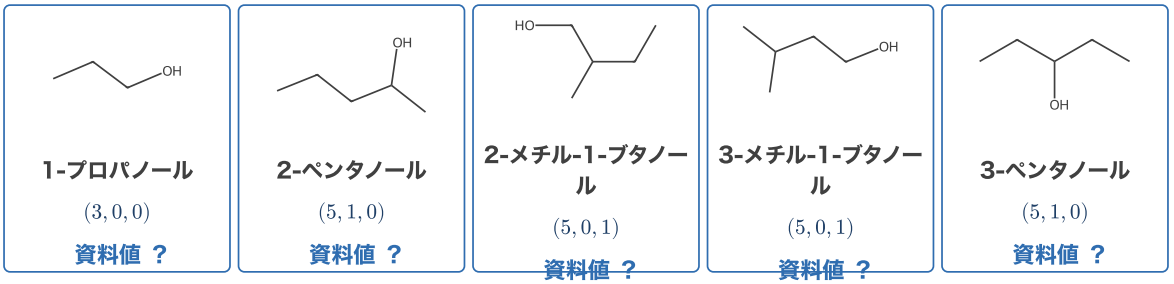


図 5-1　確かめる 5 分子です。どれも、式の係数を決めるのにも、最近傍の参照分子にも使っていません。沸点の資料値は、予測を記録してから見ます。

この章の問いは、次のとおりです。

予測がよくなったことと、信用できることはどう確かめるか。

この章のねらい

- 当てはまり（式を決めるのに使ったデータに合うこと）と、予測の確認（使っていないデータで確かめること）を区別し、予測を記録してから資料値と比べる理由を説明できる。
- 絶対誤差と平均絶対誤差（MAE）を計算し、MAE から言えることと言えないことを区別できる。
- 過学習とデータ漏洩が、予測のよさの判断をどう誤らせるかを説明できる。
- 同じ記述子の分子に残る誤差と、確かめた範囲（適用範囲）を踏まえて、予測を使ってよい範囲を言える。

5.1　当てはまりと予測を分ける：記録してから確かめる

これまでに使った分子は、役割で 3 つに分けられます（図 5-2）。

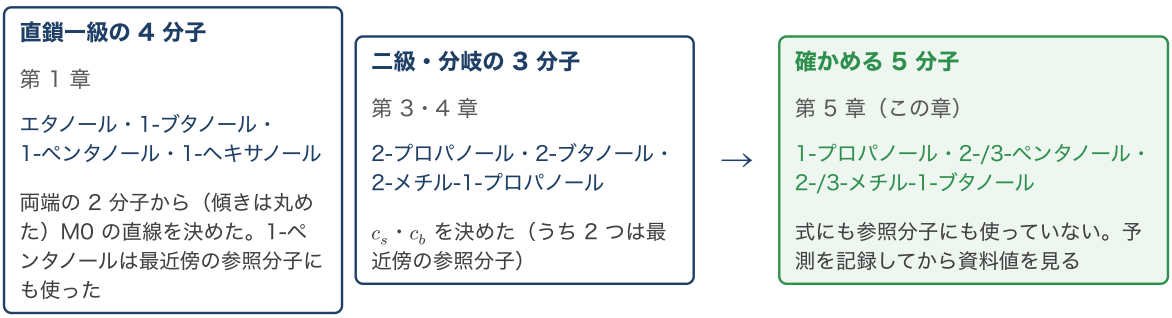


図 5-2　これまでに使った分子の役割です。直鎖の一級 4 分子と二級・分岐の 3 分子は、式の係数や最近傍の参照分子を決めるのに使いました。確かめる 5 分子は、どちらにも使っていません。

直鎖の一級 4 分子のうち両端の 2 つは M0 の直線を決めるのに、二級・分岐の 3 分子は c_s・c_b を決めるのに使いました。これらの分子に式が合うのは、合うように決めたからです。一方、確かめる 5 分子は、式を決めるときに見ていません。この違いを、言葉で分けておきます。

定義：当てはまりと予測の確認

当てはまりは、式や参照分子を決めるのに使った（見ていた）データに、予測が合うことです。**予測の確認**は、使っていないデータで、予測を先に記録してから資料値と比べることです。

予測のよさを知りたいなら、確かめるのは予測の確認の方です。当てはまりは、合うように決めたのですから、よく見えて当然です。

予測を先に記録するのは、資料値を見てから予測や式を直せないようにするためです。資料値を見たあとで「やはり重みは $\alpha = 1$ がよい」「係数を少し変えよう」と直すと、その資料値は式を決めるのに使ったデータになってしまい、予測の確認ではなくなります。前の章で予測の記録に「封をした」のは、このためです。

なお、教員は教材を作るときに、5 分子の資料値を見ています。分子の選び方も教育のためのものです。ですから、これは研究として完全に独立した検証ではなく、読者にとって使っていないデータで確かめる、教育用の確認です。

ここで、予測を記録する

資料値を見る前に、5 分子の予測を記録しましょう。M0（$T = 20n + 38$）と M1（$T = 20n + 38 - 17s - 10b$）に、図 5-1 の記述子の組を入れて、5 分子それぞれの予測値をノートに書いてください。2-ペンタノールと 3-メチル-1-ブタノールの M1 の予測は、前の章で封をした記録（121 ℃、128 ℃）と同じになるはずです。s と b がどちらも 0 の分子では、M1 の予測は M0 と同じです。M1 の予測が同じになる分子の組は、(n, s, b) がどうなっているかも見ておいてください。

書き終えてから、次の節に進んでください。

5.2　絶対誤差と平均絶対誤差：結果をまとめ、何が言えるか

予測と資料値のずれを、分子ごとに、また全体として数にします。

定義：絶対誤差と平均絶対誤差（MAE）

絶対誤差は、残差（資料値 − 予測値）の絶対値です。外れの大きさだけを表します。**平均絶対誤差**（mean absolute error, MAE）は、分子ごとの絶対誤差の平均です。

$$\mathrm{MAE} = \frac{1}{N}(|e_1| + |e_2| + \cdots + |e_N|) \tag{5.1}$$

N は分子の数、$e_1, e_2, \ldots$ は分子ごとの残差です。単位は ℃ です。

絶対値と平均の計算は付録 A の「絶対値と平均絶対誤差」 にまとめました。まず、式を決めた分子で MAE を求めてみます。

例題：係数を決めた 3 分子の MAE

M1 の予測は、2-プロパノール 81 ℃、2-ブタノール 101 ℃、2-メチル-1-プロパノール 108 ℃ です。資料値は 83・100・108 ℃ です。MAE を求めてください。
まず、分子ごとの絶対誤差を求めます。$|83 - 81| = 2$、$|100 - 101| = 1$、$|108 - 108| = 0$ です。
次に、平均します。

$$\mathrm{MAE} = \frac{2 + 1 + 0}{3} = 1.0 \tag{5.2}$$

MAE は 1.0 ℃ です。
結果の意味を確かめます。この 3 分子は、M1 の係数を決めるのに使った分子です。ですから 1.0 ℃ は当てはまりの MAE で、予測のよさを示すものではありません。

残差を符号のまま足す（$+2 - 1 + 0$）のは誤りです。絶対値をとってから足します。

封を開ける

記録した予測を、5 分子の資料値と比べます（表 5-1）。資料値は、ここで初めて示す国際化学物質安全性カードの値です。

表 5-1　確かめる 5 分子の資料値と、M0・M1 の予測、絶対誤差です（単位はすべて ℃）。予測は資料値を見る前に記録したものです。

分子	(n, s, b)	資料値	M0	M1	M0 の絶対誤差	M1 の絶対誤差
1-プロパノール	$(3, 0, 0)$	97	98	98	1	1
2-ペンタノール	$(5, 1, 0)$	119	138	121	19	2
2-メチル-1-ブタノール	$(5, 0, 1)$	128	138	128	10	0
3-メチル-1-ブタノール	$(5, 0, 1)$	132	138	128	6	4
3-ペンタノール	$(5, 1, 0)$	116	138	121	22	5
MAE					11.6	2.4

代表として、2-ペンタノールを計算します。資料値は 119 ℃ です。M0 の予測 138 ℃ との絶対誤差は $|119 - 138| = 19$ ℃、M1 の予測 121 ℃ との絶対誤差は $|119 - 121| = 2$ ℃ です。5 分子の絶対誤差の和は M0 が 58 ℃、M1 が 12 ℃ なので、MAE は

$$\mathrm{M0}\colon \frac{58}{5} = 11.6, \quad \mathrm{M1}\colon \frac{12}{5} = 2.4 \tag{5.3}$$

で、M0 が 11.6 ℃、M1 が 2.4 ℃ です。使っていない 5 分子で、予測を記録してから確かめた結果、この 5 分子では M1 の方が M0 より予測がよかった、と言えます。

MAE 2.4 ℃ から言えること、言えないこと

MAE は平均です。M1 の分子ごとの絶対誤差を見ると（図 5-3）、2.4 ℃ より大きく外れた分子が 2 つあります（3-メチル-1-ブタノール 4 ℃、3-ペンタノール 5 ℃）。

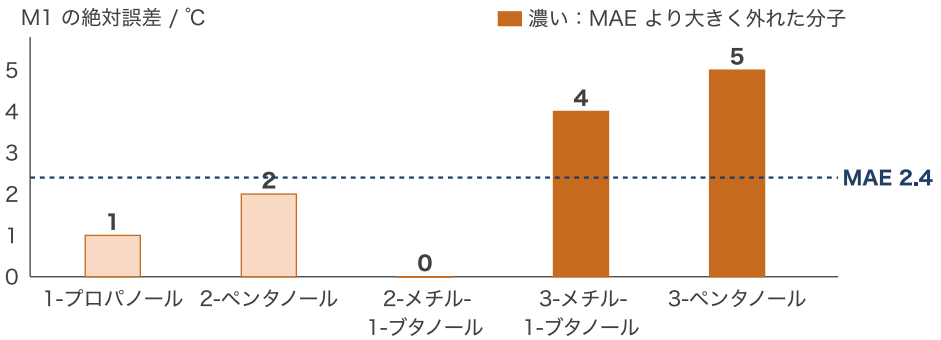


図 5-3　確かめる 5 分子での、M1 の絶対誤差です。破線が MAE（2.4 ℃）で、それより大きく外れた分子（濃い棒）が 2 つあります。

よくある誤解

MAE を、個々の分子の誤差の上限と読むことです。「M1 なら、どの分子でも誤差は 2.4 ℃ 以内」とは言えません。MAE は平均なので、それより大きく外れる分子があってかまいません。この 5 分子でも、4 ℃ と 5 ℃ 外れた分子があります。

もう 1 つ、比べておくことがあります。係数を決めた 3 分子での当てはまりの MAE は 1.0 ℃（例題）でしたが、使っていない 5 分子での MAE は 2.4 ℃ で、大きくなりました。係数を決めたデータだけで評価すると、予測はよく見えすぎます。ただし、これはこの例でそうなったということで、いつも何倍になると決まっているわけではありません。

資料値の不確かさと、予測の誤差

MAE は、資料値の測定の不確かさとも別のものです。資料値にも測定の不確かさがあり、資料によって値が少し違うこともあります（第 1 章 の「出典を書く」）。MAE 2.4 ℃ は、この 5 分子で、予測が資料値から平均してどれだけずれたかを表す数で、測定そのものがどれだけ確かかを表す数ではありません。

また、第 4 章 では係数を選ぶのに残差平方和（SSE）を使いましたが、ここでは成績のまとめに MAE を使っています。SSE は大きな残差を重く数える基準で、単位は ℃² です。MAE は外れの大きさをそのまま平均し、単位が ℃ なので、「平均して何 ℃ ずれたか」と読めます。係数を選ぶ基準と、成績を要約する基準は別のものです（付録 A の「絶対値と平均絶対誤差」）。

5.3　過学習とデータ漏洩

当てはまりの MAE (1.0 ℃) が、使っていない分子での MAE (2.4 ℃) より小さかったことから、1 つの疑問が出てきます。係数をもっと細かく変えれば、当てはまりはいくらでもよくできるのではないか、という疑問です。

できます。極端な例として、分子ごとに別々の補正係数を付けることを考えます (図 5-4)。第 4 章 までに使った 7 分子に、それぞれ専用の係数 $c_1, c_2, \ldots, c_7$ を付ければ、どの分子の資料値もそのまま再現でき、SSE は 0 になります。

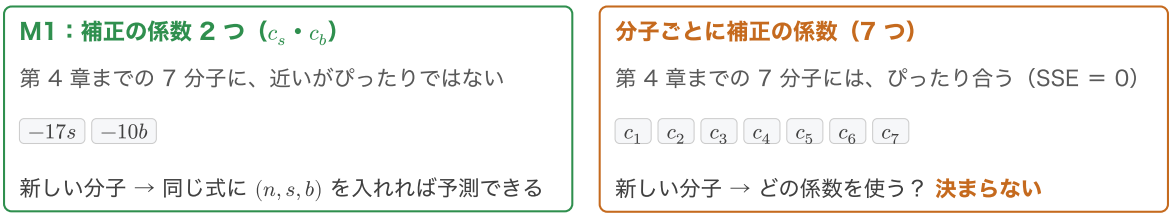


図 5-4　補正の係数が 2 つの M1（左）と、分子ごとに係数を付けた式（右）です。右は 7 分子にはぴったり合いますが、新しい分子にどの係数を使えばよいかが決まりません。

しかし、この式で予測がよくなったとは言えません。新しい分子に、7 つの係数のどれを使えばよいかが決まらないからです。また、係数を分子の数だけもてば、資料値に含まれる測定のばらつきまで写し取ってしまいます。このように、決めるのに使ったデータに合わせすぎて、使っていないデータで外れるようになることを、過学習と呼びます。係数を増やすほど当てはまりはよくなりますが、それは予測のよさの証拠になりません。

データ漏洩

もう 1 つの落とし穴は、確かめ用のデータを見て、式や重みを直してしまうことです。

たとえば、表 5-1 を見て「3-メチル-1-ブタノールと 3-ペンタノールの誤差が大きいから、係数を少し変えよう」と直したとします。直したあとの式で同じ 5 分子の MAE を求めると、小さくなるかもしれません。けれども、その 5 分子は、もう係数を直すのに使ったデータです。直した式の予測のよさは、同じ 5 分子では確かめられません。最近傍の重みを、結果を見て選び直すのも同じです。

> **定義：データ漏洩**
>
> 確かめるためのデータを見て、式や重み、係数を決めたり直したりすることです。データ漏洩が起きると、確かめたつもりの成績が、実際より良く見えます。

データを見て式を直すこと自体は、悪いことではありません。直したなら、その式は、さらに別の使っていないデータで確かめ直す必要がある、ということです。

5.4　記述子だけでは区別できない分子

第 2 章 で、同じ記述子の組になる分子には、3 列だけを入力にする予測は同じ値を出す、と約束しました。確かめる 5 分子には、その組が 2 つあります（図 5-5）。

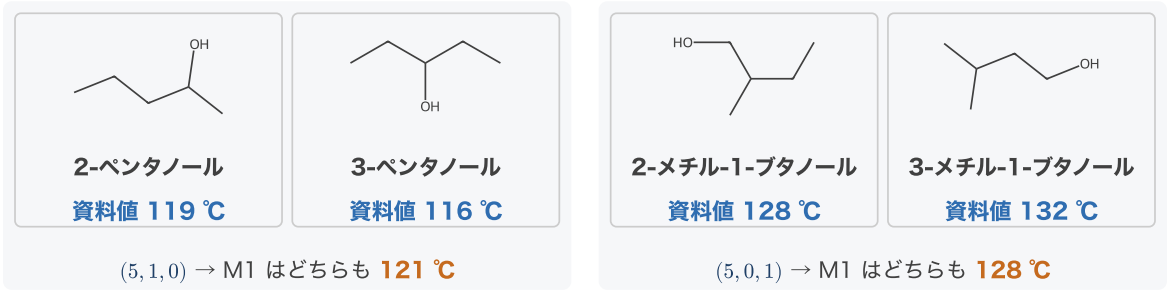


図 5-5　確かめる 5 分子のうち、同じ記述子の組になる 2 組です。M1 はどちらの組にも同じ予測値を出しますが、資料値は違います。

2-ペンタノールと 3-ペンタノールはどちらも $(5, 1, 0)$ で、M1 の予測はどちらも 121 ℃ ですが、資料値は 119 ℃ と 116 ℃ で、3 ℃ 違います。2-メチル-1-ブタノールと 3-メチル-1-ブタノールはどちらも $(5, 0, 1)$ で、予測はどちらも 128 ℃ ですが、資料値は 128 ℃ と 132 ℃ で、4 ℃ 違います。

同じ入力には同じ予測値しか出せないので、M1 の係数をどう変えても、組の 2 分子をどちらも誤差 0 にすることはできません。この違いは、第 2 章 で「表せない違い」とした OH の位置や分岐の位置によるものと考えられ、(n, s, b) の 3 列には入っていません。この誤差を小さくするには、係数ではなく、表現（列）を変える必要があります。たとえば 第 2 章 の演習 3 で考えた末端までの距離を足せば、2-ペンタノールと 3-ペンタノールは区別できます。ただし、区別できるようになっても、予測がよくなるかは、また使っていないデータで確かめるまで分かりません。

発展　この 5 分子で、(n, s, b) の予測が届く MAE の下限　試験の範囲外

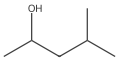
同じ記述子の組には同じ予測値しか出せないので、(n, s, b) だけを入力にする予測は、どう工夫しても、この 5 分子で MAE を 0 にはできません。どこまで小さくできるかを確かめておきます。

2-ペンタノール（119 ℃）と 3-ペンタノール（116 ℃）に共通の予測値を 1 つ選ぶとき、その値を 116〜119 ℃ の間のどこにとっても、2 つの絶対誤差の和は 3 ℃ で、これより小さくはできません。同じように、2-メチル-1-ブタノールと 3-メチル-1-ブ

タノールでは 4 ℃ です。1-プロパノールにぴったり合わせられたとしても、5 分子の絶対誤差の和は最小で $3+4=7$ ℃ で、MAE の下限は $\frac{7}{5}=1.4$ ℃ です。
これは、この表現と、この 5 分子の資料値についての計算です。未知の分子に対する予測の精度の下限を表すものではありません。

5.5　適用範囲：どこまで使ってよいか

最後に、M1 をどんな分子に使ってよいかを考えます。図 5-6 は、これまでに資料値を見た分子（式を決めるのに使った分子と、使わずに確かめた分子）の、記述子の組み合わせです。

(s,b) ＼ n	2	3	4	5	6	…	$n=9$（代入はできる）
$(0,0)$ 直鎖一級	●	●	●	●	●		範囲外
$(1,0)$ 二級		●	●	●			範囲外
$(0,1)$ 分岐			●	●			範囲外
$(1,1)$ 二級＋分岐			4-メチル-2-ペンタノール →		?		範囲外

● 式を決めるのに使った分子　● 確かめた分子（使っていない 5 分子）　? まだ資料値がない

図 5-6　資料値がある分子の記述子の組み合わせです。紺の丸は式を決めるのに使った分子（直鎖の一級 4 分子と二級・分岐の 3 分子）、緑の丸は使わずに確かめた 5 分子、? は 4-メチル-2-ペンタノール $(6,1,1)$ です。$n=9$ の列は、代入はできても範囲の外です（灰）。n は 2〜6、s と b はそれぞれ 0 か 1 の分子がありますが、$s=b=1$（二級で分岐あり）の分子は 1 つもありません。

n は 2〜6、s も b も 0 と 1 の両方があります。列ごとに見れば、どの値にも資料値のある分子があります。しかし、組み合わせで見ると、二級で分岐もある $s=b=1$ の分子は、資料値のある分子に 1 つもありません。

定義：適用範囲

式や予測の方法を使ってよいと考える範囲です。この本では、資料値のある分子（式を決めるのに使った分子と、確かめた分子）が覆う範囲とします。列ごとの値の範囲と、列の組み合わせの両方で考えます。

この本で「確かめた範囲」と言うときは、この範囲のことです。組み合わせは、主に s と b について見ます。炭素数まで含めて見れば、たとえば $n = 6$ の資料値は 1-ヘキサノール $(6, 0, 0)$ だけなので、$n = 6$ の二級や分岐も確かめていないことになります。これは授業用に簡単にした決め方で、実際の研究では、学習用のデータからの近さで決めます（この章の「実際のケモインフォでは」）。

たとえば 4-メチル-2-ペンタノール $(6, 1, 1)$ は、式に入れれば M1 の予測値が出ます。けれども、二級と分岐の両方の補正を同時に引く分子で、M1 が合うかを確かめたことはありません。同じように、$n = 9$ のような炭素数の分子にも代入はできますが、資料値のある炭素数（2〜6）の範囲の外です。式に入れて値が出ることと、その値を信用できることは別です。適用範囲の外の予測は、使うとしても「確かめていない範囲の予測」と明記します。

要点

予測から言えるのは、確かめた範囲までです。MAE は平均で上限ではなく、見て直したデータも、合わせすぎた式も、予測のよさを示しません。適用範囲は、列ごとの範囲と組み合わせの両方で見ます。

5.6　例題・演習：予測と資料値を照合し、言える結論を選ぶ

ここまでの考え方で、予測についての主張を判定します。まず、判定の手順を例題で確かめます。

例題：主張の言える範囲を確かめる

「5 分子のどれでも、M1 は M0 より誤差が小さかった」と言えるかを、表 5-1 から判定してください。

まず、主張がどの範囲の話かを確かめます。「どれでも」と言っているので、5 分子すべての行を見る必要があります。

次に、反例がないかを確かめます。1-プロパノールは、M0 も M1 も絶対誤差 1 ℃ で同じです。M1 の方が小さいとは言えません。

最後に、言える範囲に直します。「この 5 分子のどれでも、M1 の誤差は M0 の誤差以下だった」なら、表から言えます。

> 答えは、そのままでは言えない（1 分子は同じ）が、「以下」に直せば言える、です。主張を判定するときは、主張が何のデータ、どの範囲の話かを確かめ、根拠を確かめ、超えていたら言える範囲に直す、の順に考えます。

演習 1：最近傍と M1 を、5 分子で比べる

授業のアプリ（最近傍の 5 分子の予測を示すアプリ）か、手元のノートで取り組みます。最近傍の参照分子と重みは 第 3 章 で決めたもので、資料値を見て選び直していません。

1. 2-ペンタノールについて、第 3 章 の最近傍による 3 つの予測（炭素数だけ 138 ℃、s の重み $\alpha = 2$ で 100 ℃、$\alpha = 1$ で 119 ℃）の絶対誤差を求めてください。
2. 同じ 3 つの方法で、確かめる 5 分子すべての予測を求め、方法ごとの MAE を出してください（参照分子は 1-ペンタノール 138 ℃、2-ブタノール 100 ℃、2-メチル-1-プロパノール 108 ℃。同じ距離の最近傍は平均）。5 分子の MAE が最も小さいのは、最近傍の 3 つと M0・M1 のどれですか。

演習 2：結果から言えること

例題と同じ手順で、次の主張を判定します。主張が何のデータ、どの範囲の話かに印をつけ、根拠を確かめて「言える」か「言えない」かを選び、理由を 1 行で書いてください。問 2 は、言える文に直してください。

1.「この 5 分子では、M1 は M0 より予測がよかった」
2.「新しい分子でも、M1 なら必ず ±2.4 ℃ 以内に入る」
3.「この資料値を見て係数を直したあとも、同じ 5 分子で予測のよさを確かめられる」

演習 3：組み合わせと係数の数

同じ手順で判定します。問 1 は、言える文に直してください。

1.「n・s・b がそれぞれ確かめた範囲なら、$(6, 1, 1)$ の予測も確かめたことになる」
2.「分子ごとに別の補正係数を付ければ、第 4 章 までの 7 分子の SSE は 0 にできる。だから予測もよくなる」

答えは手元に残しておいてください。演習の参考解答は付録 D にあります。

5.7　設計判断：予測をどこまで根拠として使うか

演習 3 に取り組んでから読んでください。問いに戻ります。予測がよくなったことと、信用できることは、どう確かめればよいでしょうか。

予測がよくなったことは、使っていないデータで、予測を記録してから資料値と比べて確かめます。この 5 分子では、MAE は M0 が 11.6 ℃、M1 が 2.4 ℃ でした。最近傍による 3 つの予測も同じ 5 分子で比べると、MAE は炭素数だけ 12.8 ℃、$\alpha = 2$ で 15.0 ℃、$\alpha = 1$ で 7.1 ℃ で、この 5 分子では M1 が最も小さくなりました（図 5-7）。

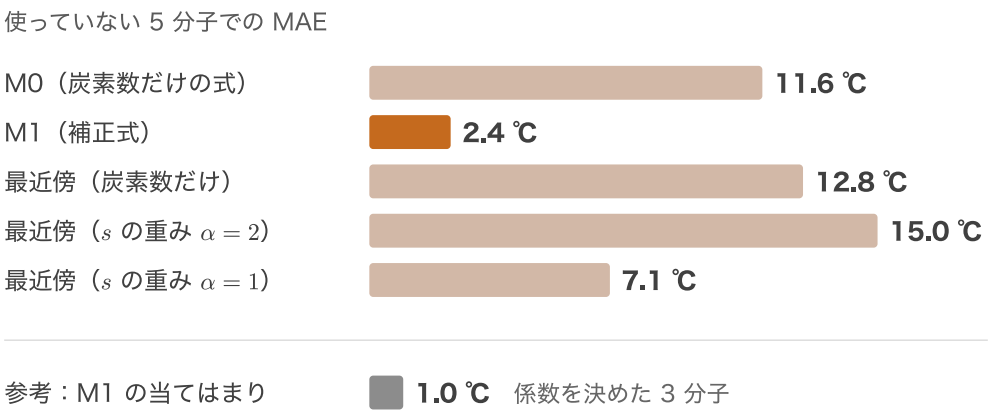


図 5-7　使っていない 5 分子での、予測の方法ごとの MAE です。灰色の棒は、係数を決めた 3 分子での M1 の当てはまりの MAE（例題）です。

ここで、第 3 章 と 第 4 章 で記録した予測を振り返ります（図 5-8）。2-ペンタノールの資料値 119 ℃ に、$\alpha = 1$ の最近傍の予測 119 ℃ はぴったり一致しました。

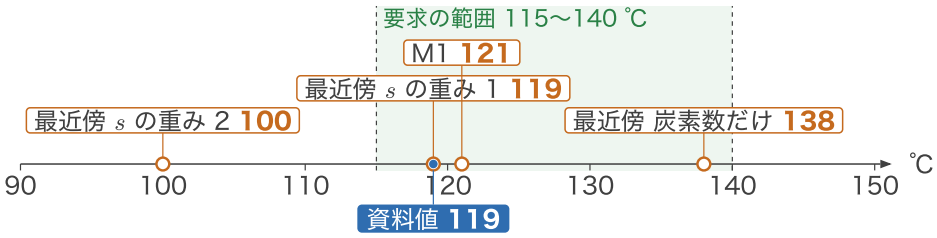


図 5-8　2-ペンタノールの 4 つの予測（白抜き）と資料値（青）です。帯は沸点の要求の範囲です。

けれども、同じ方法の 5 分子での MAE は 7.1 ℃ で、M1 の 2.4 ℃ より大きくなりました。最近傍の参照分子が 3 つしかなく、炭素数 3 の 1-プロパノールは参照分子の炭素数（4〜5）の範囲の外にあるため、大きく外れたのです（$\alpha = 2$ で 26 ℃）。1 例

の一致から、その方法や重みが最良だとは言えない理由が、ここで数字で見えます。ただし、最近傍と式の一般的な優劣も、この 5 分子からは言えません。参照分子を増やせば、最近傍の予測は変わります。

予測を信用できる範囲は、次の 3 つで限ります。

- **言える範囲**：MAE 2.4 ℃ は、この 5 分子での成績で、平均です。個々の分子の誤差の上限ではありません。また、5 分子は記述子の組み合わせとしては 3 種類しかなく、広い範囲の分子を網羅した検証ではありません。
- **表現の限界**：同じ記述子の組の分子の違い（3 ℃、4 ℃）は、M1 では予測できません。
- **適用範囲**：確かめた組み合わせの中で使います。$s = b = 1$ の分子は、その外です。

では、M1 の予測を、次の章の候補選びの根拠として使ってよいでしょうか。この章の結果からは、確かめた範囲の中なら、M1 の予測は、資料値のない分子の沸点を見積もる目安として使えます。ただし、予測値と要求の境界との差が数 ℃ しかないときは、MAE 2.4 ℃ やこの 5 分子での最大の外れ（5 ℃）と比べて、境界を外れるおそれがある、と書いておくべきです。適用範囲の外の分子（$s = b = 1$ など）の予測は、確かめていない予測として扱い、資料値で確かめる候補にします。予測をどこまで使うかは、成立する判断が複数ありますが、どの判断でも、予測と資料値と未確認のことを区別して書きます。

物語の上では、依頼人に「この 5 分子で確かめると、M1 の予測は平均 2.4 ℃ のずれだった。ただし 5 ℃ 外れた分子もあり、二級で分岐もある分子は確かめていない」と答えることになります。

実験系への接続：何で確かめたか

研究室でも、条件を決めた実験と、確かめた実験が同じになっていることがあります。ロット 1 の原料で反応の温度と時間を決め、同じロット 1 で収率を確かめたなら、見ているのは当てはまりです。別の日に別のロットの原料で同じ収率が出るかを確かめて、はじめて予測の確認になります（図 5-9。架空の模式で、値はありません）。



図 5-9　条件を決めた実験と、確かめた実験を分ける例です（架空の模式）。上は合成の反応条件、下は材料の組成と性質です。

「うまくいった」と書く前に、それが決めるのに使ったデータでの結果か、別のデータでの結果かを書いておきます。これは、この章で当てはまりと予測の確認を分けたことと同じです。

実際のケモインフォでは　使っていないデータで確かめる　試験の範囲外

実際の研究では、データを学習用と検証用に分け、式やモデルは学習用だけで決めます。検証用を見ながら式を直すと、検証用の成績が良く見えすぎる（データ漏洩）からです。データが少ないときは、検証用にする部分を入れ替えて何度も確かめ、成績を平均します（交差検証）。列や係数を増やすほど学習用にはよく合いますが、検証用で外れることがあります（過学習）。また、学習用のデータに近い分子かどうかを数で示し、遠い分子の予測には注意を付けます（適用範囲。図 5-10）。



図 5-10　学習用と検証用の分割、交差検証（5 分割）、過学習と適用範囲の模式図です。適用範囲は、学習用のデータから遠い分子では予測を信用しにくい、という考え方です。

序章で紹介した抗菌薬の探索でも、学習に使っていない分子を実験で確かめ、予測の上位 99 分子のうち 51 分子が大腸菌の増殖を止めました（Stokes ほか, 2020）。この章で、使っていない 5 分子で 1 回確かめたことと共通しているのは、決めるのに使っていないデータで確かめる、という点です。

分かったこと・まだ分からないこと

分かったこと

- 予測を記録してから、使っていない 5 分子の資料値と比べた。MAE は M0 で 11.6 ℃、M1 で 2.4 ℃。この 5 分子では、M1 の予測が最もよかった（最近傍の 3 つの方法より小さい）。
- MAE は平均で、上限ではない（4 ℃・5 ℃ 外れた分子がある）。当てはまりの MAE（1.0 ℃）より大きかった。
- 同じ記述子の組の分子の違い（3 ℃、4 ℃）は、M1 では予測できない。$s = b = 1$ の分子は、確かめた範囲の外。
- $\alpha = 1$ の最近傍は 2-ペンタノールで一致したが、5 分子では MAE 7.1 ℃。1 例の一致から最良とは言えない。

まだ分からないこと

- 沸点と水溶性の両方を満たす候補はどれか。確かめていない組み合わせ（$s = b = 1$）の候補をどう扱うか（第 6 章）。
- 構造を変えた分子の水溶性 w_{25}。資料は取り寄せ中で、第 6 章 で示す。

まとめ

- 当てはまりは、式を決めるのに使ったデータに合うことで、予測の確認は、使っていないデータで、予測を記録してから資料値と比べることです。予測のよさは、予測の確認で判断します。
- 絶対誤差は残差の絶対値、MAE はその平均です。使っていない 5 分子での MAE は、M0 で 11.6 ℃、M1 で 2.4 ℃ でした。MAE は平均で、個々の分子の誤差の上限でも、資料値の測定の不確かさでもありません。
- 分子ごとに係数を付けるような合わせすぎ（過学習）は、当てはまりを完全にしても予測をよくしません。確かめ用のデータを見て式を直したら（データ漏洩）、別の使っていないデータで確かめ直します。
- 同じ記述子の組の分子の違いは、係数ではなく表現を変えないと予測できません。
- 適用範囲は、列ごとの範囲と組み合わせの両方で見ます。式に入れて値が出ることと、その値を信用できることは別です。

章末問題

手順を実行する

1. 架空の物質 K1・K2・K3 について、ある方法の予測は 120・131・104 ℃、資料値（架空）は 125・129・104 ℃ でした。絶対誤差と MAE を求めてください。
2. 問 1 の 3 物質について、別の方法の予測は 118・126・110 ℃ でした。MAE を求め、問 1 の方法と比べてください。この 3 物質での比較から、どこまで言えますか。
3. 資料値で確かめた分子の記述子の組が、n は 3〜6、(s, b) は $(0,0)$・$(1,0)$・$(0,1)$ の 3 種類だったとします。次の分子のうち、確かめた範囲の中にあるものはどれですか。$(4,0,0)$、$(8,0,0)$、$(6,1,1)$、$(5,0,2)$

理由と限界を説明する

1. ある予測の方法の MAE が 3 ℃ でした。「だから、この方法で予測すれば、どの分子の誤差も 3 ℃ 以下になる」は正しいですか。理由を説明してください。
2. 表 5-1 を見て c_s を 16 に直したところ、5 分子の MAE が小さくなりました。「だから M1 は改善した」と言えますか。理由を説明してください。
3.「$\alpha = 1$ の最近傍は、2-ペンタノールで誤差 0 だった。だから $\alpha = 1$ の最近傍が最良の予測の方法である」という主張の問題点を、この章の数値を使って説明してください。

別の題材へ使う

1. ある合成で、原料のロット 1 を使って反応の温度と時間を決め、同じロット 1 で収率 85 % を確かめました（架空）。「この条件なら収率 85 % が出る」と報告する前に、何をすべきですか。
2. 自分の研究（まだなら学生実験）で、条件を決めたデータと、確かめたデータが同じになっている例を探してください。別のデータで確かめるには、何をすればよいですか。

6

第 6 章

分子候補と次の実験を提案する

前の章では、記録しておいた予測を、使っていない 5 分子の資料値と比べました。M1 の予測は、この 5 分子では平均 2.4 ℃ のずれでした。一方で、二級で分岐もある分子（$s = b = 1$）は、まだ確かめていないことも分かりました。残った問いは、沸点と水溶性の両方を満たす候補はどれか、確かめていない組み合わせの候補をどう扱うか、です（第 5 章 の「まだ分からないこと」）。

物語の上では、水溶性の資料も届きました。



工程担当者

水溶性の資料も届いたよ。で、どれを使えばいい？ 次に何を測る？

序章で受け取った依頼は、次の 3 つを報告することでした（序章の「目標、制約、最終的に提案するもの」）。

1. **既知の候補 1 つ**：資料値で、2 つの要求を満たすと確かめられる分子
2. **追加で調べる候補 1 つ**：資料値はまだないが、予測から有望と考える新しい分子
3. **次に確かめる分子と物性**：限られた測定の機会を、どれに使うか。その理由と、結果ごとの判断

この章で、その報告をまとめます。この章の問いは、次のとおりです。

複数の要求と未確認事項を踏まえ、次に何を確かめるか。

この章のねらい

- 沸点と水溶性の資料値から、2 つの要求を同時に満たす候補を選べる。
- 要求値の境界からの余裕を物性ごとに求め、資料値・予測値・欠測・未確認を区別して候補の状態を書ける。
- ほしい物性から記述子の組と構造を探し（逆設計）、確かめた範囲の外の候補を見分けられる。

- 次に確かめる 1 つを、目的と結果ごとの行動を先に決めて選び、最終提案（候補・根拠・未確認事項・次の確認）にまとめられる。

6.1　複数の物性を同時に満たす候補を探す

まず、資料値がそろった 5 分子で、2 つの要求を同時に見ます。届いた水溶性の資料値は、次のとおりです（表 6-1）。w_{25} は、1-ペンタノールと同じく、25 ℃ の水相（水に富む相）の飽和溶液 100 g あたりの値で、IUPAC の溶解度データ集に再録された同じ測定の値です。沸点は 第 5 章 で示した値です。

表 6-1　炭素数 5 の 5 分子の、沸点と w_{25} の資料値です（沸点の高い順）。

分子	(n, s, b)	沸点 / ℃	w_{25}
1-ペンタノール	$(5, 0, 0)$	138	2.19
3-メチル-1-ブタノール	$(5, 0, 1)$	132	2.67
2-メチル-1-ブタノール	$(5, 0, 1)$	128	2.97
2-ペンタノール	$(5, 1, 0)$	119	4.46
3-ペンタノール	$(5, 1, 0)$	116	5.15

5 分子を、横軸に w_{25}、縦軸に沸点をとった平面に置きます（図 6-1）。序章の 図 6-3 と同じ平面です。

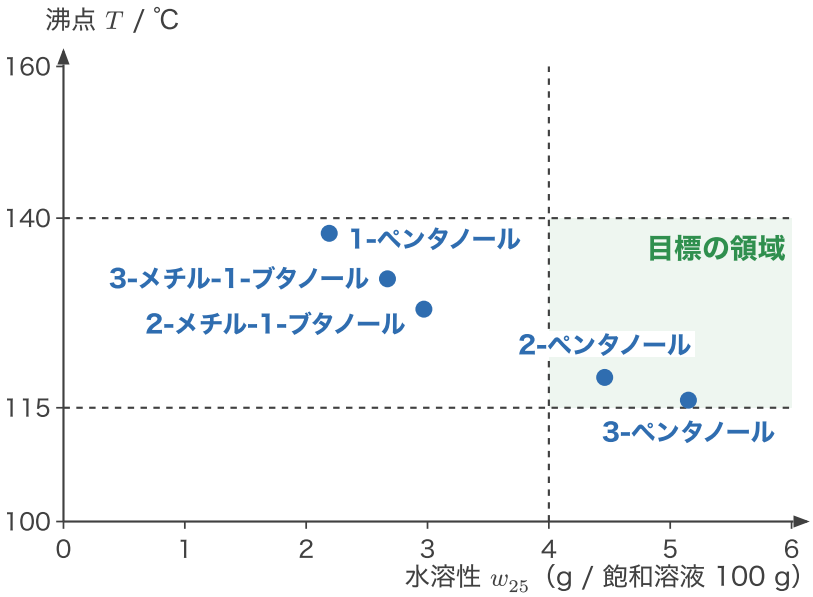


図 6-1　5 分子の沸点と w_{25} の資料値です。破線が要求値の境界、緑の四角が 2 つの要求を同時に満たす目標の領域です。

目標の領域に入るのは、2-ペンタノールと 3-ペンタノールの 2 つです。ほかの 3 つは、沸点は範囲に入っていますが、w_{25} が 4.0 に足りません。この判定は予測ではなく、資料値に基づくものです。

沸点と水溶性の兼ね合い

表 6-1 を沸点の高い順に見ると、w_{25} は 2.19 → 2.67 → 2.97 → 4.46 → 5.15 と、順に大きくなっています。この 5 分子では、沸点が低い分子ほど、水に溶けやすくなっています（図 6-2）。これは、この 5 分子での観察です。

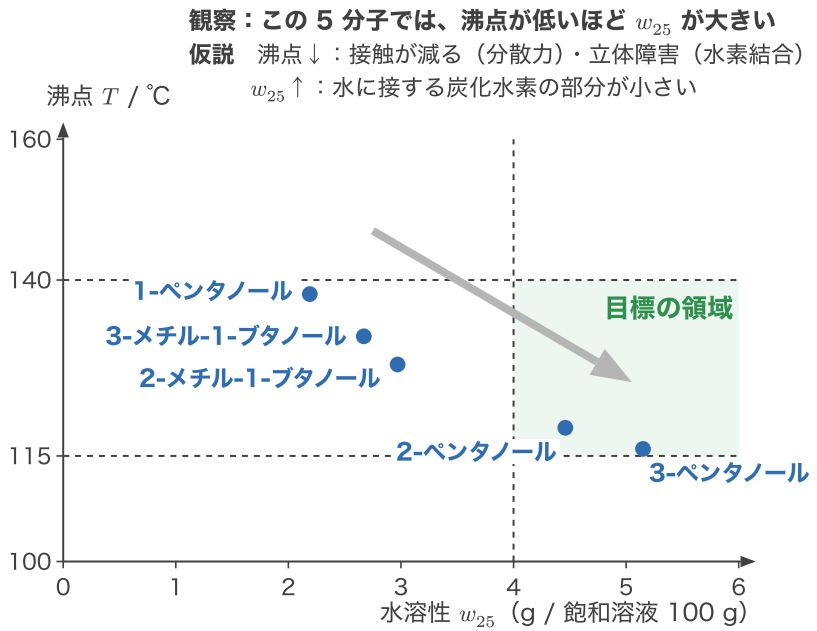


図 6-2　この 5 分子では、沸点が低いほど w_{25} が大きくなっています。灰色の矢印は説明のための模式で、データに当てはめた直線ではありません。

なぜ、沸点と水溶性が逆向きに変化するのでしょうか。構造の変化と分子間の力から、次のような仮説が立てられます。枝分かれを付けたり、OH を鎖の内側に移したりすると、分子どうしが接する面が減って分散力が弱まり、沸点が下がります。同時に、水に接する炭化水素の部分が小さくなるので、水に溶けやすくなります。同じ構造の変化が、沸点と水溶性の両方を変える、という仮説です。OH を鎖の内側に移す変化では、OH の炭素の周りの立体障害で水素結合をつくりにくくなることも、沸点を下げる側にはたらくと考えられます（第 4 章 の二級の説明）。

この仮説が正しければ、水溶性を上げようとすると、沸点も下がります。沸点には下限の 115 ℃ があるので、水溶性だけを上げ続けることはできません。ここが、この依頼の兼ね合いです。実際、w_{25} が最も大きい 3-ペンタノールは、沸点が下限まで 1 ℃ しかありません。

ただし、これは 5 分子での観察と、その説明の仮説です。炭素数 5 のアルコール一般の法則とは言えません。仮説を確かめるには、ほかの分子でも測る必要があります。第 1 章 の演習 3 で「OH を内側に移す」「枝分かれを付ける」と書いていたなら、この 5 分子ではその向きだった、と言ってよいでしょう。

6.2　要求値からの余裕と、予測値・資料値・欠測の区別

2 つの要求を満たすかどうかだけでなく、境界からどれだけ離れているかも見ておきます。

> **定義：余裕**
>
> 物性の値と、要求値の境界との差です。範囲の条件の要求（沸点）では下限からと上限までの両方を、下限の条件の要求（w_{25}）では下限からの差を求めます。境界の内側なら正、外側なら負（不足）で表します。

序章の 1-ペンタノールなら、沸点は下限から $138 - 115 = 23$ ℃、上限まで $140 - 138 = 2$ ℃ で、上限に近い位置にあります。w_{25} は $2.19 - 4.0 = -1.81$ で、下限に 1.81 足りません（図 6-3）。

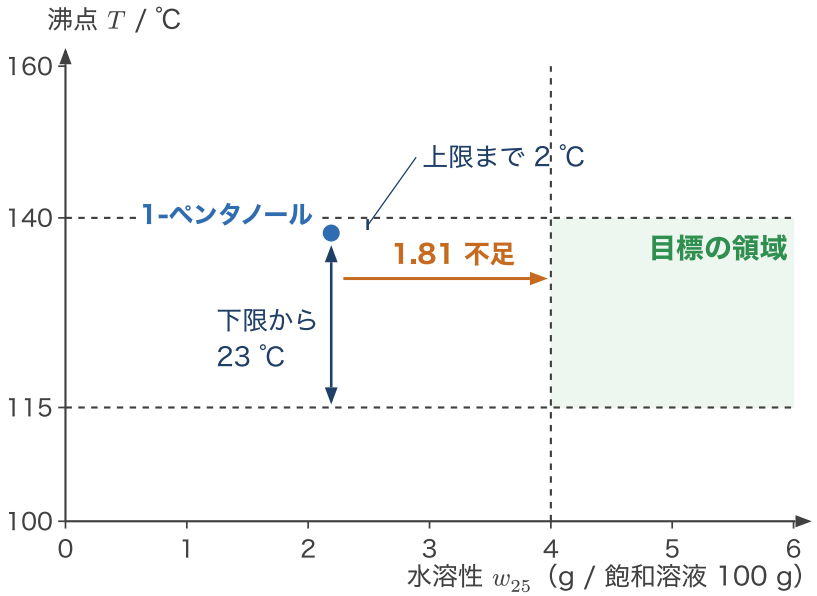


図 6-3　1-ペンタノールの余裕です。沸点は下限から 23 ℃、上限まで 2 ℃、w_{25} は 4.0 に 1.81 足りません。

余裕が大きいほど、測り直しや条件の小さな違いで要求を外れにくい、という見方ができます。余裕を使うときは、次の 3 つに気をつけます。

- **都合のよい側だけを見ない**。沸点は範囲の条件なので、下限からと上限までの両方を見ます。

- **単位の違う差をそのまま足さない**。沸点の余裕（℃）と w_{25} の余裕は単位が違うので、足して 1 つの点数にしても意味がありません。
- **差から、外れる確率は出せない**。余裕が 1 ℃ だから外れる確率が何 %、とは計算できません。

> **例題：3-メチル-1-ブタノールの余裕**
>
> 3-メチル-1-ブタノール（沸点 132 ℃、w_{25} 2.67）の余裕を求め、2 つの要求を満たすか判定してください。
> まず、沸点の余裕を両側で求めます。下限から $132 - 115 = 17$ ℃、上限まで $140 - 132 = 8$ ℃ で、範囲に入っています。
> 次に、w_{25} の余裕を求めます。$2.67 - 4.0 = -1.33$ で、下限に 1.33 足りません。
> 最後に、2 つを合わせて判定します。1 つでも満たさない要求があれば、目標の領域の外です。
> 沸点は満たしますが、w_{25} が 1.33 足りないので、既知の候補にはなりません。沸点の片側（下限から）だけを見て「17 ℃ も余裕がある」と書くのは、上限まで 8 ℃ であることを見落としています。

値の種類を区別して書く

候補の状態を書くときは、物性ごとに、値がどの種類かを区別します。

- **資料値**：条件のそろった測定の値がある。
- **予測値**：M1 などで出した値で、測っていない。
- **欠測**：条件のそろった値が手元にない（文献にまったく値がない、という意味ではない）。
- **未確認**：予測の前提（たとえば $s = b = 1$ の組み合わせで M1 が合うこと）を確かめていない。

予測値に余裕を考えるときは、その余裕を、予測の誤差の大きさと比べます。M1 の誤差は、第 5 章 で確かめた 5 分子で 0〜5 ℃（MAE 2.4 ℃）でした。予測値の余裕が数 ℃ しかなければ、予測が範囲に入っていても、実際には外れているかもしれません。余裕がそれより十分に大きければ、範囲に入りそうだと言えます。ただし、MAE は上限ではなく、確かめた範囲の外の分子では誤差がもっと大きいかもしれないので、「入りそうだが、未確認」と書きます。

6.3　目標から構造を考える：逆設計

2-ペンタノールと 3-ペンタノールは、資料値で要求を満たす既知の候補です。次に、追加で調べる新しい候補を探します。

これまでの章では、構造から物性を予測してきました。ここでは逆に、ほしい物性（沸点 115〜140 ℃）から、それを満たしそうな記述子の組と構造を探します。このように、目標の物性から構造を探すことを逆設計と呼びます。

M1 を使って、炭素数 5〜6 の記述子の組ごとに沸点を予測すると、図 6-4 のようになります。水溶性は、条件をそろえたデータが少ないので予測しません。

n	(s, b)	M1 の予測	予測が 115〜140 ℃ か
5	$(0, 0)$	138 ℃	入る △
5	$(0, 1)$	128 ℃	入る
5	$(1, 0)$	121 ℃	入る
5	$(1, 1)$	111 ℃	入らない △ ＊
6	$(0, 0)$	158 ℃	入らない
6	$(0, 1)$	148 ℃	入らない
6	$(1, 0)$	141 ℃	入らない △
6	$(1, 1)$	131 ℃	入る ＊
6	$(0, 2)$	138 ℃	入る △ ＊

△ 境界から 5 ℃ 以内（確かめた 5 分子での M1 の誤差の大きさ）
＊ 確かめた範囲の外

図 6-4　M1 で予測した、記述子の組ごとの沸点です（予測値）。△は境界から 5 ℃ 以内（確かめた 5 分子での M1 の誤差の大きさ）、＊は確かめた範囲の外です。

表を読むときは、次の点に注意します。

- **炭素数 5 の 3 行**には、これまでの分子が入っています。$(5, 0, 1)$ には、四級炭素（炭素 4 個と結合する炭素）をもつ 2,2-ジメチル-1-プロパノールも入りますが、四級炭素の分岐は、確かめた範囲の外です（c_b を決めた分子も、確かめた 5 分子も、分岐点はどれも炭素 3 個と結合する炭素でした）。
- **境界に近い行**（141・138・138・111 ℃）は、予測が境界から 5 ℃ 以内で、確かめた 5 分子での M1 の誤差の大きさ（0〜5 ℃）と変わりません。範囲に入る・入らないを言い切りません。

- **炭素数 6** で範囲に入るのは $(6,1,1)$（131 ℃）と $(6,0,2)$（138 ℃）で、どちらにも＊が付きます。ただし、確かめていない中身が違います。$(6,0,2)$ は $b=2$ の値そのものを確かめていません（確かめたデータの b は 0 と 1）。上限にも近い行です（△）。$(6,1,1)$ は、各列の値（$n=6$、$s=1$、$b=1$）はどれも確かめた範囲の中で、s と b の組み合わせでは $s=b=1$ だけを確かめていません（炭素数まで含めて見れば、$n=6$ の資料値は 1-ヘキサノールだけなので、$n=6$ の二級・分岐も確かめていません。5.5 節）。境界からも 9 ℃ 以上離れています。

そこで、$(6,1,1)$ の分子を調べることにします。先に進む前に、$(6,1,1)$ になる分子、つまり炭素数 6 の二級アルコールで分岐点が 1 つのものを、できるだけ構造式で挙げてみてください。4 つあります。主鎖を 5 つにして分岐を 1 つ付ける場合と、主鎖を 4 つにして同じ炭素に枝を 2 つ付ける場合に分けて数えると、見つけやすくなります。三級（OH の炭素に炭素 3 つ）は対象の外です。

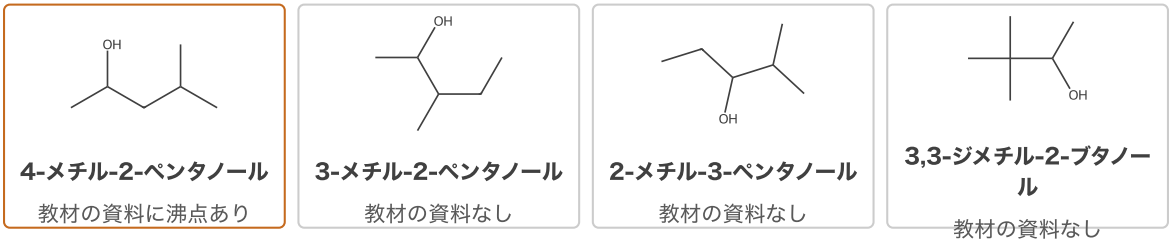
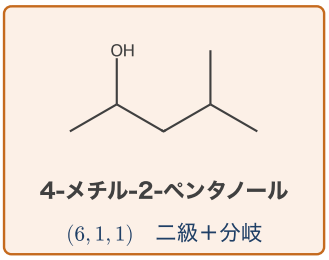


図 6-5　$(6,1,1)$ になる 4 つの分子です。M1 では、4 つとも同じ予測（131 ℃）になります。

4 つは、4-メチル-2-ペンタノール、3-メチル-2-ペンタノール、2-メチル-3-ペンタノール、3,3-ジメチル-2-ブタノールです（図 6-5）。同じ記述子の組なので、M1 は 4 つに同じ予測値を出し、区別できません（第 2 章の「同じ記述子になる異性体」）。どれを先に調べるかは、M1 の外の根拠で選びます。たとえば、6.1 節の仮説（枝が多い、OH が内側ほど水に溶けやすい）から、枝の多い 3,3-ジメチル-2-ブタノールや 2-メチル-3-ペンタノールを推すこともできます。資料が手に入りやすいかどうかも根拠になります。この本では、教材の資料に沸点がある 4-メチル-2-ペンタノールで進めます。なお、3,3-ジメチル-2-ブタノールは四級炭素をもつので、確かめた範囲の外です。

6.4　限られた測定機会で、次に何を確かめるか

追加の候補 4-メチル-2-ペンタノール（$CH_3CH(OH)CH_2CH(CH_3)_2$）について、いま分かっていることと、分かっていないことを整理します（図 6-6）。



物性	いまの値	未確認の点
沸点	予測 131 ℃	資料値ではない。二級＋分岐の組み合わせで M1 を確かめていない
水溶性 w_{25}	？(欠測)	25 ℃・水相・飽和溶液 100 g あたりの値が教材にない。別条件の値を w_{25} に使わない

図 6-6　追加の候補 4-メチル-2-ペンタノールの状態です。沸点は M1 の予測値（131 ℃）で、w_{25} は欠測です。

- **沸点**：M1 の予測は $20 \times 6 + 38 - 17 - 10 = 131$ ℃ です。資料値ではありません。しかも、二級と分岐の補正を同時に引く分子で、M1 を確かめたことはありません（未確認）。
- **w_{25}**：条件をそろえた値が、教材にありません（欠測）。資料には「2 g/100 mL」という値がありますが、温度が書かれておらず、分母も体積です。第 1 章 の記録の分け方で言えば「現状では判断できない」記録なので、$w_{25} = 2$ として使いません。

沸点の予測 131 ℃ は、下限から 16 ℃、上限まで 9 ℃ です。この余裕は、確かめた 5 分子での M1 の誤差（0〜5 ℃）より大きいので、沸点は範囲に入りそうだと言えます。それでも組み合わせは未確認なので、「沸点の要求には入りそうだが、未確認」と書きます。w_{25} は欠測なので、何も言えません。ですから、この候補を「採用」とは言えません。限られた測定の機会で、何を先に確かめるかを決めます。

確かめる前に決めておくこと

確かめる前に、次の 3 つを書いておきます。

1. **目的**：何のために確かめるか（有望な候補を確かめる、欠測を埋める、予測の前提を点検する など）。何を確かめるかは、残すか外すかの判断が変わる方を選びます。
2. **測るもの**：どの分子の、どの物性を、どの条件で測るか。
3. **結果ごとの行動**：結果がどうなったら、どうするか。

結果を見てから行動を決めると、都合のよい解釈になりやすいからです。

> **例題：境界に近い候補の、測り直しの計画**
>
> 3-ペンタノール（沸点 116 ℃、w_{25} 5.15）を候補に残すか決めたいとします。1 つ測れるとき、何を確かめますか。資料値にもばらつきがあることを踏まえて、確認の計画を立ててください。
>
> まず、判断が変わりそうな値を選び、目的を決めます。w_{25} は 4.0 より 1.15 大きいのに対して、沸点は下限まで 1 ℃ しかありません。そこで沸点を選び、測り直しても範囲内に入るかを確かめることを目的にします。
>
> 次に、測るものを決めます。3-ペンタノールの常圧の沸点です。要求値は常圧の沸点なので、測り直しも常圧でそろえます（第 1 章 の条件をそろえる考え方）。繰り返して測り、値の幅も書いておきます。
>
> 最後に、結果ごとの行動を決めます。115 ℃ 以上なら候補に残します。115 ℃ 未満なら、2-ペンタノールを優先します。
>
> 目的・測るもの・結果ごとの行動を、測る前に書いておくのが計画の型です。この型を、演習 2 で新しい候補に使います。

実際の工程で性能を確かめるには

ここまでの判断は、純物質の沸点と、25 ℃ の水溶性という 2 つの物性に基づいています。序章で断ったとおり、これだけでは、実際の工程での性能までは保証できません（序章の「今回の依頼：1-ペンタノールを改良する」）。候補を工程で使う前には、少なくとも次のことを考える必要があります。

- **混ざった液のふるまい**：工程では、溶媒は部品の油を溶かし込み、すすぎの水とも混ざります。純物質の値は、油や水が混ざった液の値とは違うかもしれません。
- **温度**：w_{25} は 25 ℃ の値です。すすぎの水の温度が違えば、溶けやすさも変わりえます。
- **乾かしやすさ**：乾かす段で影響するのは、沸点だけでなく、その温度での蒸発の速さや、風の当て方です。

- **ほかの要求**：油を落とす力（この授業では比べないと仮定しました）、引火のしやすさなどの安全性、においや毒性、価格や入手のしやすさも、溶媒の選択に関わります。

これらを確かめるには、候補を実際の部品と洗浄の条件で小さく試し、洗い上がりや、すすぎのあとに残る溶媒の量を、いまの 1-ペンタノールと同じ条件で比べます。比べるときに条件をそろえることは、第 1 章 で学んだことと同じです。

6.5　例題・演習：候補を選び、確認結果から判断を更新する

ここまでの考え方で、候補を選び、次の確認を計画し、最終提案をまとめます。

演習 1：既知の候補を比べる

授業のアプリ（点を押して余裕を見るアプリ）か、手元のノートで取り組みます。

1. 2-ペンタノール（119 ℃、4.46）と 3-ペンタノール（116 ℃、5.15）の余裕を、例題と同じ 3 つ（沸点の下限から・上限まで、w_{25} の 4.0 から）求めてください。
2. どちらを既知の候補に選びますか。どの余裕を重く見たかを、理由として 1 行で書いてください。
3.「余裕を足し合わせて点数にすれば、どちらがよいか決まる」と言えますか。

演習 2：次に確かめる 1 つを決める

授業のアプリ（選んで、行動を書いてから、結果を見るアプリ）か、手元のノートで取り組みます。4-メチル-2-ペンタノールの沸点は予測 131 ℃（余裕 9 ℃ 以上、未確認）、w_{25} は欠測です。1 つの物性だけ確かめられるとします（研究なら、測定にあたります）。

1. 残すか外すかの判断が変わる方を選んでください。w_{25} と沸点のどちらを確かめますか。目的を 1 行で書いてください。
2. 結果が要求を満たす場合と、満たさない場合の行動を、先に書いてください（例題の 3 つ目の型）。
3. 選んだ物性の結果を見たとして、「1 つの物性を確かめたので、この候補は確かめたことになる」と言えますか。

結果は 6.6 節で示します。問 1・2 を書いてから進んでください。

演習 3：最終提案をまとめる



> **工程担当者**
>
> 報告は 1 枚で。結果が違ったときの方針も書いておいて。

次の 4 つを、それぞれ 1〜2 行で書いてください。

1. 既知の候補と、根拠の資料値
2. 追加で調べる候補の、構造の特徴・予測（M1）・未確認の点（演習 2 で沸点を確かめた場合は、資料値も書く）
3. 追加の候補（または既知の候補）について、次に調べる物性と条件、優先する理由
4. 結果が要求を満たす場合と、満たさない場合の判断

演習 4：第 1 章の答えと比べる

第 1 章 の演習 3 の自分の答えと比べて、何を根拠に判断を変えたか（変えなかったか）を 1 行で書いてください。資料値、予測、未確認の語を使います。

演習の参考解答は付録 D にあります。

6.6　最終提案：候補、根拠、未確認事項、次の確認

演習 2〜4 に取り組んでから読んでください。

演習 2 の結果

4-メチル-2-ペンタノールの沸点を選んだ場合、資料値は 132 ℃ です（国際化学物質安全性カード）。M1 の予測 131 ℃ との差は 1 ℃ で、この 1 分子では予測が合いました。ただし、1 分子で合ったことは、二級と分岐の補正を同時に引くことが、ほかの $s = b = 1$ の分子でも正しいことの保証にはなりません。w_{25} は欠測のままです。

w_{25} を選んだ場合、条件をそろえた資料値は教材にありません。そこで、結果を受けて判断を更新する練習として、架空の値で考えます。もし w_{25} が 3.6 だったら（架空）、4.0 に 0.4 足りないので、いまの目標では優先度を下げます。もし 4.6 だったら（架空）、w_{25} は満たしますが、沸点が予測のままなので、採用は確定しません。3.6 も 4.6 も、4-メチル-2-ペンタノールの資料値や推定値ではありません。

どちらを選んだ場合も、「この候補は確かめたことになる」とは言えません。1 つの物性を確かめても、もう 1 つの物性は分からないままです。1 つ確かめたら、何

が分かり、何が残ったかを書きます。なお、132 ℃ は資料を見て知った値で、この本の中で新しく測った値ではありません。研究では、ここが実際の測定になります。

最終提案の例

最終提案は、成立するものが 1 つではありません。成立する例の 1 つを示します。

1. **既知の候補**：2-ペンタノール。沸点 119 ℃、w_{25} 4.46（どちらも資料値）。沸点の下限からの余裕（4 ℃）を重く見た。
2. **追加の候補**：4-メチル-2-ペンタノール $(6,1,1)$。枝分かれがあり、水溶性が上がる可能性がある（6.1 節の仮説）。沸点は M1 の予測 131 ℃（$s = b = 1$ の組み合わせは未確認。演習 2 で沸点を確かめた場合は資料値 132 ℃）。w_{25} は欠測。
3. **次に調べる**：4-メチル-2-ペンタノールの w_{25}（25 ℃・水相・飽和溶液 100 g あたり）。候補に残すかどうかを決めるのに欠けているから。
4. **結果ごとの判断**：4.0 以上なら候補に残す。4.0 未満なら優先度を下げ、既知の候補で進める。

演習 4 の例は、次のとおりです。第 1 章では、4 分子のデータしかなく、構造を変えた候補分子は仮説として書くしかなかった。いまは、条件をそろえた資料値と、確かめた予測とその範囲を根拠に、候補と次の確認を選べる。

既知の候補を 3-ペンタノールにしても（水溶性の余裕を重く見た）、次に 4-メチル-2-ペンタノールの沸点を確かめる計画にしても、理由が余裕や未確認の点と結び付いていれば成立します。成立しないのは、次のような提案です。

- 「131 ℃ が範囲内なので、4-メチル-2-ペンタノールを採用する」：予測で、組み合わせも水溶性も確かめていない。
- 「w_{25} が 2 g/100 mL で 4.0 に足りないので外す」：条件と分母の違う値を、w_{25} として使っている。
- 「3-ペンタノールは余裕の合計が大きいので選ぶ」：単位の違う余裕を足している。

この提案を、目標の平面に描くと 図 6-7 のようになります。

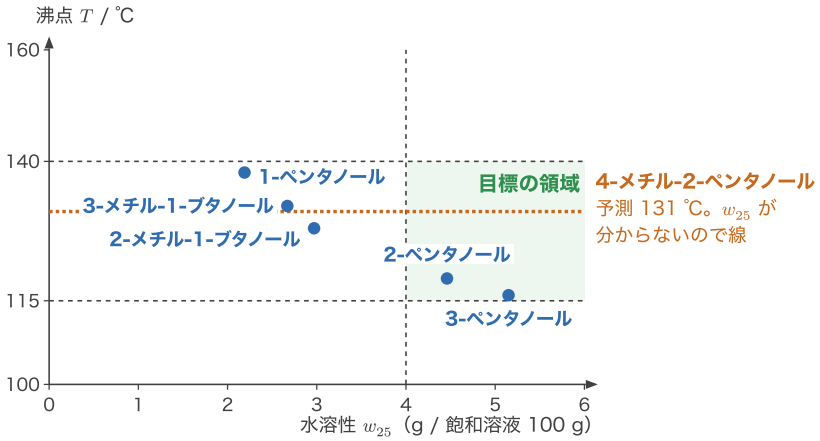


図 6-7　5 分子の資料値と、4-メチル-2-ペンタノールの沸点の予測（橙の点線）です。w_{25} が分からないので、点ではなく横の線で示しました。

6 章でしたこと

第 1 章 の演習 3 で見た 5 つの構造（出発候補の 1-ペンタノールと、候補の 4 分子。図 1-8）は、いまでは 5 つとも沸点と w_{25} の資料値があります（表 6-1）。第 1 章で「仮説として書く」しかなかった候補分子を、資料値で確かめ、予測と未確認の点を分けて提案できるようになりました。最初の候補分子が最終提案と違っていてもかまいません。大事なのは、何を根拠に判断を変えたかを言えることです。

各章の問いと答えを並べると、図 6-8 のようになります。

章	問い	答え
第 1 章　データ	出発候補の何が不足し、 どのデータが必要か	不足は水溶性。必要なのは、候補分子の w_{25} と沸点
第 2 章　記述子	分子のどの特徴を、表の列にするか	n・s・b を列にする。同じ行の分子は区別できない
第 3 章　似た分子	「似ている」をどう定義すると、 予測がどう変わるか	記述子と重みの決め方で、最近傍も予測値も変わる
第 4 章　予測式	構造と沸点の傾向を、 簡単な式にまとめられるか	炭素数の直線＋二級と分岐の補正（M1）に まとめられる
第 5 章　検証	予測がよくなったことと、 信用できることはどう確かめるか	使っていないデータで確かめ、言える範囲を限る
第 6 章　提案	複数の要求と未確認事項を踏まえ、 次に何を確かめるか	資料値で候補を選び、 判断を変える値を 1 つ確かめる

図 6-8　各章の問いと答えです。依頼を、データを整える → 構造を表す → 予測する → 確かめる → 候補と次の実験、の順に進めました。

2-ペンタノールで振り返ると、沸点の予測は、最近傍で 138・100・119 ℃（第 3 章）、M1 で 121 ℃（第 4 章）と変わり、資料値は 119 ℃（第 5 章）で、最後は既知の候補になりました（この章）。

序章の 3 つの進め方（序章の「分子・材料を設計する 3 つの進め方」）に戻ると、この依頼では 3 つを組み合わせました。データ駆動で候補を絞り（M1、逆設計）、理論と計算の代わりに化学の説明（分子間の力。係数の向き、沸点と水溶性の兼ね合い）で仮説を立て、経験（実験。ここでは資料値と次の測定）で確かめる、という組み合わせです。



返信メモ　工程担当者より

報告ありがとう。資料値のある候補で進めつつ、追加の候補は次の測定の結果で決めよう。

6.7　自分の研究で考える：構造・組成・調製条件・測定条件

この授業で持ち帰ってほしいのは、アルコールの式そのものではありません。データの作り方、比べ方、確かめ方、次の実験の選び方です。式は、アルコールのこの範囲のデータから作ったもので、無機材料や反応の収率にそのまま使えるものではありません。

この依頼で行ったことを、次の 5 つの項目に整理すると、自分の研究でも使えます（図 6-9）。

1. **良くしたい物性・収率など**
2. **変えられる構造・組成・条件**
3. **そろえる条件と、記録する条件**
4. **役立つ既知データと、不足している情報**
5. **次の確認実験と、結果に応じた判断**

項目	この授業では	自分の分野では
1. 良くしたい物性・収率など	沸点を保ちながら、水溶性を上げる	
2. 変えられる構造・組成・条件	OH の位置、分岐、炭素数	
3. そろえる条件と、記録する条件	25 ℃・水相・飽和溶液 100 g あたり、出典	
4. 役立つ既知データと、不足情報	似た分子の資料値。新しい候補の w_{25} が欠測	
5. 次の確認実験と、結果に応じた判断	4-メチル-2-ペンタノールの w_{25}。 4.0 未満なら優先度を下げる	

図 6-9　この依頼で行ったことを 5 つの項目に整理したものです。右の列に、自分の分野での例を書いてみてください。

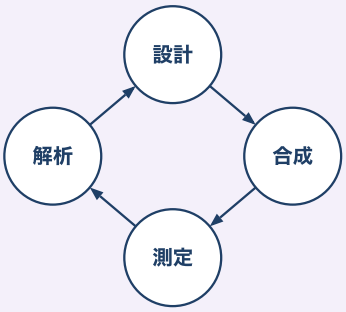
たとえば、塗料の乾燥時間を短くしたいとします。変えられるのは溶剤の組成、そろえるのは温度・湿度・膜厚、役立つのは似た組成の記録、次の確認は境界に近い組成で 1 つ測る、と整理できます。

材料では、分子の構造や組成だけでなく、調製の履歴（焼成の温度や時間）、微細な構造（粒の大きさ）、試料の状態（粉末か薄膜か）、測定の条件も、物性に関わりえます。表の列にする情報（第 2 章）や、そろえる条件（第 1 章）を決めるときに、これらを考えに入れます。

自分の研究（まだなら学生実験や身近な例）について、5 つの項目のうち 2〜3 項目でよいので、書いてみてください。第 1〜5 章の「実験系への接続」で考えたこと（そろえる条件、列にする情報、似た既知の例、式から言えること、確かめたデータ）が使えます。研究室の未公開データを書く必要はありません。

実際のケモインフォでは　予測と実験を繰り返す　試験の範囲外

実際の研究では、設計 → 合成 → 測定 → 解析を何度も繰り返します（創薬では Design–Make–Test–Analyze と呼ばれます）。測った結果をデータに加えて予測を直し、次に作るものを設計します。次の実験の選び方には、予測が不確かな所を測って少ない実験で予測をよくする方法（能動学習）や、予測のよさと不確かさの釣り合いを見て、よい候補を少ない実験で見つける方法（ベイズ最適化）があります。ほしい物性から構造を探す逆設計も、実際にはプログラムが多数の構造を作って予測します（図 6-10）。



・予測と実験を繰り返す（創薬の Design–Make–Test–Analyze など）
・能動学習：予測が不確かな所を測り、予測をよくする
・ベイズ最適化：予測のよさと不確かさの釣り合いで、
よい候補を早く見つける
・ほしい物性から、条件を満たす構造を探す（逆設計）

この授業：次に確かめる 1 つを、判断を変える値から選んだ

限られた実験の機会を、判断を変える値に使う

図 6-10　予測と実験を繰り返すサイクルと、次の実験を選ぶ方法の模式図です。

序章で紹介した有機 EL の発光分子の探索も、予測で候補を絞り、計算と実験で確かめた例でした。この章で、判断を変える値を 1 つ選んで確かめたことと共通しているのは、限られた実験の機会を、判断を変える値に使う、という点です。

分かったこと・まだ分からないこと

分かったこと

- 資料値で、2-ペンタノールと 3-ペンタノールが 2 つの要求を満たす。この 5 分子では、沸点が低いほど w_{25} が大きい（観察）。同じ構造の変化が両方を変えるという仮説が立つ。
- 余裕は物性ごとに、沸点は両側で見る。単位の違う余裕は足さない。
- 逆設計で $(6, 1, 1)$ の 4 分子が候補になる。4-メチル-2-ペンタノールの沸点は M1 の予測 131 ℃（演習 2 で沸点を確かめた場合は資料値 132 ℃）。w_{25} は欠測。
- 次の確認は、目的と結果ごとの行動を先に決めて選ぶ。1 つ確かめても、ほかは確かめたことにならない。

まだ分からないこと（次の実験で確かめること）

- 4-メチル-2-ペンタノールの w_{25}（25 ℃・水相・飽和溶液 100 g あたり）。
- $s = b = 1$ のほかの分子でも、M1 の予測が合うか。
- 候補が、実際の工程（油や水が混ざった液、工程の温度）でも 1-ペンタノールより残りにくいか。

まとめ

- 複数の要求は、資料値で 1 つずつ判定し、すべて満たすものを候補にします。この依頼では、2-ペンタノールと 3-ペンタノールが既知の候補になりました。
- 沸点と水溶性は、この 5 分子では逆向きに変化しました。同じ構造の変化が両方を変えるという仮説から、水溶性だけを上げ続けられない兼ね合いが分かります。
- 余裕は物性ごとに求め、範囲の条件は両側を見ます。単位の違う差は足さず、差から外れる確率も出せません。予測値の余裕は、予測の誤差の大きさと比べます。
- 逆設計では、ほしい物性から記述子の組と構造を探します。確かめた範囲の外の組（$b = 2$、四級炭素、$s = b = 1$）は、そのことを明記します。
- 次の確認は、目的・測るもの・結果ごとの行動を先に決めて選びます。最終提案には、候補、根拠（資料値・予測値の区別）、未確認事項、次の確認を書きます。

章末問題

手順を実行する

1. 架空の候補 F1（沸点 124 ℃、w_{25} 4.4）と F2（沸点 113 ℃、w_{25} 6.1）の余裕を、沸点の下限から・上限まで、w_{25} の 4.0 から、の 3 つ求め、2 つの要求を満たすか判定してください（値はすべて架空）。
2. 炭素数 7 の一級・二級アルコールについて、b が 0〜2 の記述子の組ごとに M1 で沸点を予測し、115〜140 ℃ に入る組があるかを答えてください。
3. 架空の候補 G は、沸点の資料値（架空）が 128 ℃ で、w_{25} は欠測です。次に確かめる物性と条件、目的、結果ごとの行動を書いてください。

理由と限界を説明する

1. 架空の候補 H の沸点を、ある予測の方法で 125 ℃ と予測しました。その方法の、使っていない分子での MAE は 2.4 ℃ で、H は確かめた範囲の中にあるとします。「下限から 10 ℃、上限まで 15 ℃ の余裕があり、MAE 2.4 ℃ より大きいので、H の沸点は必ず範囲に入る」は正しいですか。理由を説明してください。

2. 「2-ペンタノールは沸点の余裕 4 ℃ と w_{25} の余裕 0.46 で合計 4.46、3-ペンタノールは 1 ℃ と 1.15 で合計 2.15。だから 2-ペンタノールがよい」という判断の問題点を説明してください。
3. 「4-メチル-2-ペンタノールの沸点の資料値 132 ℃ は、M1 の予測 131 ℃ と 1 ℃ しか違わなかった。だから M1 は $s = b = 1$ の分子すべてで正しい」は正しいですか。理由を説明してください。

別の題材へ使う

1. 電池の電解液の粘度を下げたいとします（架空の課題）。5 つの項目のうち、「そろえる条件と、記録する条件」と「次の確認実験と、結果に応じた判断」の例を、それぞれ 1 つ書いてください。
2. 自分の研究（まだなら学生実験や身近な例）について、5 つの項目をすべて書いてください。

付録 A　計算とグラフの読み方

本文で使う計算を、ここにまとめます。どれも四則演算でできる計算です。例の数値は計算の練習のための数で、物性の値ではありません。

平均

いくつかの値を足して、値の個数で割ったものを（算術）平均と呼びます。値が $x_1, x_2, \ldots, x_N$ の N 個なら、

$$\text{平均} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

です。たとえば 2、4、9 の平均は、$\frac{2+4+9}{3} = \frac{15}{3} = 5$ です。

平均は、値のまん中あたりを 1 つの数で表したものです。平均が 5 でも、9 のように平均から大きく離れた値が含まれていることがあります。欠測（空欄）は値ではないので、平均に入れません（第 1 章）。

一次式：傾きと切片

$y = ax + c$ の形の式を一次式と呼び、グラフは直線になります。a を傾き、c を切片と呼びます。

- **傾き a**：x が 1 増えたときに、y がいくつ増えるか。
- **切片 c**：$x = 0$ のときの y の値。

2 つの点 (x_1, y_1) と (x_2, y_2) を通る直線は、次のように求めます。まず、傾きを

$$a = \frac{y_2 - y_1}{x_2 - x_1}$$

で求めます。次に、1 つ目の点を通るように、$y_1 = ax_1 + c$ から切片 $c = y_1 - ax_1$ を求めます。

たとえば点 $(1, 5)$ と $(3, 11)$ を通る直線は、傾き $a = \frac{11-5}{3-1} = 3$、切片 $c = 5 - 3 \times 1 = 2$ で、$y = 3x + 2$ です。

直線の式に x を入れれば、y の値をいくつでも計算できます。ただし、式を決めたデータの範囲の外で計算した値が当たるかどうかは、式からは分かりません（第 5 章）。

残差

データの値と、式などで出した予測値との差を残差と呼びます。この本では、引く向きを

$$\text{残差} = \text{資料値} - \text{予測値}$$

に固定します（第 4 章）。

- 残差が正：資料値が予測より高い。グラフでは、点が直線より上にある。
- 残差が負：資料値が予測より低い。グラフでは、点が直線より下にある。

残差の大きさは、グラフの上では、点と直線の縦のずれにあたります。

平方誤差と残差平方和

残差を 2 乗した値を平方誤差、各データの平方誤差を足した値を残差平方和（SSE）と呼びます。残差が $e_1, e_2, ..., e_N$ なら、

$$\text{SSE} = e_1^2 + e_2^2 + \cdots + e_N^2$$

です。たとえば残差が $+2$ と -1 なら、$\text{SSE} = 4 + 1 = 5$ です。残差の単位が ℃ なら、SSE の単位は ℃² です。

2 乗するのには、2 つの理由があります。

- 正と負の残差が打ち消し合わない。残差 $+2$ と -2 をそのまま足すと 0 になり、ずれがないように見えますが、2 乗して足すと 8 です。

- 大きな残差ほど重く数える。残差 3 の平方誤差は 9 で、残差 1 の 9 倍です。

式の係数をいくつか試し、SSE が最も小さくなるものを選ぶことで、データによく合う係数を選べます。SSE が最も小さくなる係数を計算で求める方法を、最小二乗法と呼びます（第 4 章 の発展の箱）。

絶対値と平均絶対誤差

ある数から符号を取ったものを絶対値と呼び、$|x|$ と書きます。$|+3| = 3$、$|-3| = 3$ です。残差の絶対値を絶対誤差と呼びます。外れの向きは問わず、外れの大きさだけを表します。

各データの絶対誤差の平均を、平均絶対誤差（mean absolute error, MAE）と呼びます。

$$\mathrm{MAE} = \frac{|e_1| + |e_2| + \cdots + |e_N|}{N}$$

たとえば残差が $+2$、-1、-3 なら、$\mathrm{MAE} = \frac{2+1+3}{3} = 2$ です。単位は残差と同じ（℃など）です。

MAE は平均なので、個々の絶対誤差がすべて MAE 以下になるわけではありません。この例でも、MAE は 2 ですが、絶対誤差 3 のデータがあります。

SSE と MAE の使い分け

SSE と MAE は、どちらも残差を 1 つの数にまとめますが、基準が違います。SSE は大きな残差を重く数え、MAE はすべての残差を大きさのとおりに数えます。この本では、係数を選ぶときに SSE を使い（第 4 章）、予測の成績をもとの単位でまとめるときに MAE を使います（第 5 章）。

付録 B　用語集

本文で使った主な用語を、五十音順（英字で始まるものは最後）に並べました。意味は短くまとめたので、詳しくは「主な章」の欄の章を読んでください。「（範囲外）」を付けた用語は、「実際のケモインフォでは」の箱で紹介したもので、試験の範囲外です。

あ行〜か行

用語	意味	主な章
当てはまり	式や参照を決めるのに使った（見ていた）データに、予測が合うこと。予測のよさの証拠にはならない	第 5 章
異性体	分子式が同じで、構造が違う分子。分子量が同じなので、分子量の列では区別できない	第 2 章
一級・二級（アルコール）	OH の結合した炭素に、炭素が 1 個結合していれば一級、2 個なら二級。3 個なら三級で、この授業の対象の外	第 2 章
同じ距離の最近傍	距離が最も小さい参照が複数あること。すべて使い、沸点の算術平均を予測値にする	第 3 章
重み	記述子の違い 1 つを、距離いくつと数えるかを決める数。自然に決まる値ではなく、何の違いを重く見るかの判断	第 3 章
過学習（範囲外の箱でも紹介）	決めるのに使ったデータに合わせすぎて、使っていないデータで外れるようになること。分子ごとに係数を付ければ当てはまりは完全になるが、予測はよくならない	第 5 章
架空値	練習のために作った値。必ず「架空」と書き添え、資料値と混ぜない	序章・第 1 章
仮説	データでまだ確かめていない予想。次に何を測るかを決める手がかりになる	第 1 章
観察	手元のデータで確かめられたこと。「この 4 例では」のように範囲を付けて書く	第 1 章

用語	意味	主な章
記述子	分子の構造の特徴を数にしたもの。データ表の構造の列になる	第 2 章
逆設計	ほしい物性から、それを満たしそうな記述子の組や構造を探すこと	第 6 章
距離	記述子の違いから決める、似ていなさの数。小さいほど似ている。この授業では $d = \lvert n - n' \rvert + 2\lvert s - s' \rvert + \lvert b - b' \rvert$	第 3 章
係数	説明変数に掛ける数。データから決める	第 4 章
欠測	条件のそろった値が手元にないこと。0 ではない。文献にまったく値がない、という意味でもない	第 1 章・第 6 章
ケモインフォマティクス	分子の構造や物性などの化学の情報をデータとして集め、計算機で表し、比べ、予測して、分子や実験の設計に使う分野	序章
交差検証（範囲外）	データを学習用と検証用に分け、検証用にする部分を入れ替えて何度も確かめ、成績を平均すること	第 5 章

さ行〜た行

用語	意味	主な章
最近傍	参照分子のうち、距離が最も小さいもの	第 3 章
最近傍による予測	最近傍の物性の値を予測値にする方法。同じ距離の最近傍が複数あれば、すべて使って算術平均する	第 3 章
最小二乗法	残差平方和が最も小さくなる係数を求める方法	第 4 章・付録 A
残差	資料値 − 予測値。引く向きはこの本を通して固定する。負なら式が高すぎる	第 4 章・付録 A
残差平方和（SSE）	各分子の残差を 2 乗して足した値。単位は ℃²。係数を選ぶときに使う	第 4 章・付録 A
尺度	記述子の値の幅。幅の違う記述子は、幅をそろえるか重みで調整してから距離を作る	第 3 章
出典	値がどの資料に載っているか。後から条件を確かめ直すために付ける	第 1 章
資料値	文献に載った測定の値。出典を付ける	序章・第 1 章
水溶性 w_{25}	25 ℃ でアルコールと水を混ぜて 2 層に分かれたときの、水相（水に富む相）の飽和溶液 100 g 中のアルコールの質量（g）	序章

用語	意味	主な章
絶対誤差	残差の絶対値。外れの大きさだけを表す	第 5 章・付録 A
説明変数	式に入れる列。この授業では構造の列（記述子）	第 4 章
データ駆動	既存のデータから構造と物性の関係を学び、予測して候補を絞る進め方。経験と試行錯誤、理論と計算と組み合わせて使う	序章
データ漏洩	確かめるためのデータを見て、式や重み、係数を決めたり直したりすること。確かめたつもりの成績が良く見えすぎる	第 5 章
適用範囲	式や予測の方法を使ってよいと考える範囲。この本では、資料値のある分子（式を決めた分子と、確かめた分子）が覆う範囲とする。列ごとの範囲と、列の組み合わせの両方で考える。実際の研究では、学習用のデータからの近さで決める	第 5 章
転載	ほかの資料の値を写して載せたもの。独立した確認には数えない	第 1 章

な行〜わ行

用語	意味	主な章
フィンガープリント（範囲外）	決めておいた多くの部分構造があるか（1）ないか（0）を並べて、分子を 0 と 1 の並びにしたもの	第 2 章
封をする	予測の記録を、資料値を見るまで書き換えないこと（この本の言い方）。資料値を見てから予測を直せないようにする	第 4 章・第 5 章
分岐点	炭素骨格で、3 個以上の炭素と結合する炭素。OH との結合は数えない。炭素 4 個と結合する炭素も 1 個と数える	第 2 章
分子グラフ	原子を点、結合を線として、分子をつながりとして見たもの。結合の長さや角度、立体的な形は入っていない	第 2 章
平均絶対誤差（MAE）	分子ごとの絶対誤差の平均。単位は ℃。平均なので、個々の誤差の上限ではない	第 5 章・付録 A
末端	炭素骨格で、ほかの炭素と 1 本だけ結合する炭素	第 2 章

用語	意味	主な章
末端までの距離	OH の炭素から最も近い末端までの C–C 結合の数。OH の炭素自身が末端なら 0	第 2 章
未確認	予測の前提（たとえば $s = b = 1$ の組み合わせで M1 が合うこと）を確かめていないこと	第 6 章
目的変数	予測したい物性の列。この依頼では沸点 T と水溶性 w_{25}	第 1 章
要求値（範囲の条件・下限の条件）	依頼で満たすべき値。沸点 115〜140 ℃ は両側に境界がある範囲の条件、w_{25} 4.0 以上は下限の条件	序章
予測値	M0・M1 や最近傍など、この授業で作る方法で出した値。測った値ではない	序章・第 3 章
予測の確認	使っていないデータで、予測を先に記録してから資料値と比べること	第 5 章
余裕	物性の値と要求値の境界との差。範囲の条件は両側を見る。単位の違う余裕は足さない	第 6 章

英字・記号

用語	意味	主な章
k 近傍法（範囲外）	似ている順に上位 k 個の分子を選び、その値の平均で予測する方法。この本の最近傍はその最小版	第 3 章
M0（炭素数だけの式）	$T = 20n + 38$。直鎖一級の両端の 2 点から決めた教育用の式	第 4 章
M1（補正式）	$T = 20n + 38 - 17s - 10b$。M0 に二級と分岐の補正の項を入れた式	第 4 章
(n, s, b)	記述子の組。n 炭素数、s OH の炭素が二級なら 1・一級なら 0、b 分岐点の個数	第 2 章
QSPR・QSAR	構造と物性（QSPR）や生理活性（QSAR）の関係を式で表したもの。定量的構造物性相関	第 4 章
SMILES（範囲外）	分子を 1 行の文字列で書く約束。1-ペンタノールは CCCCCO	第 2 章

付録 C　理解確認問題と総合問題

この付録では、各章の章末にある理解確認問題の使い方を案内し、定期試験の準備となる総合問題を 1 組示します。参考解答は、どちらも付録 D にあります。

章末の理解確認問題

各章の終わりに、理解確認問題を置きました。問題は、授業の到達目標の確かめ方(「手順を実行できる」「理由と限界を説明できる」「別の題材へ使える」) に合わせて、3 つの段に分けています。

表 C-1　章末の理解確認問題の 3 つの段です。

段	何を確かめるか
手順を実行する	換算、記述子、距離と最近傍、残差と SSE、MAE、余裕などの計算を、自分で最後までできるか
理由と限界を説明する	その計算や結論から何が言え、何が言えないかを、言い過ぎずに説明できるか
別の題材へ使う	同じ判断の流れを、合成・材料・自分の研究の例で使えるか

章を読み終えたら、まず「手順を実行する」を解き、解答で計算を確かめてから、残りの 2 段に進むとよいでしょう。小テスト（授業後のオンラインのドリル）は各章の核心を 1 問ずつ選ぶ形式で、章末問題は理由を自分の言葉で書く形式です。両方を使うと、選べることと説明できることの両方を確かめられます。

総合問題

定期試験と同じ 5 つの型の問題を 1 組示します。定期試験は 60 分・マークシート形式で、答えを選択肢から選びます。ここでは理由まで自分の言葉で書く形にしてあるので、試験より時間がかかります。試験の準備として、時間を計って解いてみてください。

- 目安は 90 分です。電卓を使ってかまいません。
- 1 つの共通の場面で、5 つの問いに答えます。必要な情報と式は問題に示します。物性の値を記憶から補う必要はありません。
- 判断を問う問いでは、答えは 1 つに決まらないことがあります。唯一の候補との一致より、根拠が一貫しているかを見ます。

この総合問題の場面・分子・数値は、定期試験とは別のものです。答えを覚えるのではなく、考え方の練習に使ってください。

共通事項

この問題の数値は、考え方を確かめるための**架空の値**で、実在の化合物の物性の値ではありません。

あなたは、ある工程で使う溶媒の候補を選ぶ担当者です。目標は、常圧の沸点が 125〜150 ℃、w_{25} が 3.0 以上の候補を提案することです。w_{25} は、25 ℃ でアルコールと水を混ぜて 2 層に分かれたときの、水相（水に富む相）の飽和溶液 100 g 中に含まれるアルコールの質量（g）です。

使う予測式は、次のとおりです。

- M0（炭素数だけの式）：$T = 20n + 38$
- M1（補正式）：$T = 20n + 38 - 17s - 10b$
- T は常圧の沸点 (℃)、n は炭素数、s は OH が結合する炭素が二級なら 1・一級なら 0、b は炭素骨格で 3 個以上の炭素と結合する炭素（分岐点）の個数（OH との結合は数えない）。

M1 は、炭素数 2〜6 の非環式・飽和で、OH を 1 つもつ一級または二級アルコールの資料値から作りました。M0 の直線は直鎖一級の 2 分子から、二級の補正（17）は二級の 2 分子から、分岐の補正（10）は分岐のある 1 分子から決めました。M1 の作成と確認に使った分子は、どれも s と b の少なくとも一方が 0 で、分岐点はどれも炭素 3 個と結合する炭素でした。

平均絶対誤差は、MAE = (|資料値 − 予測値| の合計)/ 分子数 です。

問 1　使えるデータを選ぶ（15 点）

ある候補の水溶性について、次の記録が見つかりました。A〜E の値はすべて架空です。

表 C-2　総合問題 問 1 の記録（架空）です。

記録	条件と値
A	25 ℃、水相、3.9 g アルコール／100 g 水
B	25 ℃、水相、3.6 g アルコール／100 g 飽和溶液
C	20 ℃、水相、4.0 g アルコール／100 g 飽和溶液
D	25 ℃、アルコール相、9.5 g 水／100 g 飽和溶液
E	条件の記載なし、値は空欄

1. w_{25} とそのまま比べられる記録を選び、理由を述べなさい。(3 点)
2. 記録 A を、w_{25} と同じ分母（飽和溶液 100 g）の値に直しなさい。直した値と B は一致しませんが、このことから「どちらかの記録が誤りである」と言えますか。理由も述べなさい。(5 点)
3. C と D を、B と同じものとして平均できない理由を、それぞれ述べなさい。(各 2 点、計 4 点)
4. E を 0 として平均に入れてよいですか。理由も述べなさい。(3 点)

問 2　分子構造を数値で表し、似た分子から予測する（20 点）

表 C-3　総合問題 問 2 の分子です。

分子	縮合構造式
K1	$CH_3CH(OH)CH_2CH_2CH(CH_3)_2$
K2	$CH_3CH_2CH_2CH(OH)CH(CH_3)_2$
K3	$HOCH_2CH_2C(CH_3)_2CH_2CH_3$
E2	$(CH_3CH_2)_2C(OH)CH_3$
F2	$CH_3CH_2CH_2OCH_2CH_2CH_3$

距離は $d = |n - n'| + \alpha|s - s'| + |b - b'|$ とします。(n, s, b) は予測したい分子、(n', s', b') は参照する分子、α は s の重みです。

1. K1、K2、K3 の n、s、b を求めなさい。(各値 1 点、計 9 点)
2. $\alpha = 2$ として K1 と K2 の距離を求めなさい。この値から、K1 と K2 は同じ沸点だと言えますか。この表現で捨てられている情報も挙げて説明しなさい。(3 点)
3. 次の 3 つの参照（沸点はすべて架空の値）から、K3 の沸点を最近傍による方法で予測しなさい。距離が最も小さい参照の沸点を予測値とし、同じ距離の最近傍が複数あるときは、それらの沸点の平均を予測値とします。$\alpha = 2$ のときと $\alpha =$

1 のときの予測値を求め、重みを変えて予測が変わった理由を説明しなさい。(4 点)

参照	n	s	b	沸点（架空）/ ℃
H1	7	0	0	172
H2	6	0	1	150
H3	7	1	1	155

4. E2 と F2 に M1 を使ってよいですか。それぞれ理由を述べなさい。(各 2 点、計 4 点)

問 3　使っていないデータで予測を確かめる（25 点）

次の 4 つの分子の資料値は、M0・M1 の作成や修正には使っていない、確認用の架空の値です。W1〜W4 は、この問だけの記号です。

表 C-4　総合問題 問 3 の確認用データ（架空）です。

分子	n	s	b	資料値（架空）/ ℃
W1	6	0	0	155
W2	6	1	0	138
W3	6	0	1	151
W4	4	1	0	103

1. 各分子について、M0 と M1 の予測値を求めなさい。(各値 1 点、計 8 点)
2. 各モデルの絶対誤差と MAE を求めなさい。(各 6 点、計 12 点)
3. この結果から、どちらのモデルを使いたいかを述べなさい。あわせて、この結果だけでは言えないことを 1 つ述べなさい。(5 点)

問 4　予測をどこまで信用するか（20 点）

次の主張を評価しなさい。問 3 の計算が終わっていなくても、答えられます。

1.「問 3 で M1 の MAE は 2.75 ℃ だった。だから、これから M1 で予測する分子の誤差は、どれも 2.75 ℃ 以下になる」。正しいですか。問 3 の数値か MAE の定義を使って説明しなさい。(8 点)
2.「問 3 の資料値を見て、W2 の誤差が 0 になるように c_s を 17 から 20 に直した。直した式で同じ 4 分子の MAE を求め直せば、改善したことを示せる」。問題点と、追加で必要な確かめ方を述べなさい。(6 点)

3.「記述子の組 $(7, 1, 1)$ の分子にも M1 は代入できるので、その予測は問 3 の分子と同じくらい信用できる」。問題点を 2 つ挙げなさい。(6 点)

問 5　候補を選び、次の確認を提案する（20 点）

出発の分子は、炭素数 7 の直鎖一級アルコールです。候補として、次の L1〜L3 を考えました。w_{25} はこの問題だけの架空の値です。

表 C-5　総合問題 問 5 の候補（w_{25} は架空）です。

候補	構造の特徴	n	s	b	w_{25}（架空）
L1	炭素数 6、直鎖の二級アルコール	6	1	0	3.4
L2	炭素数 7、分岐 1 か所の二級アルコール	7	1	1	不明
L3	炭素数 6、分岐 1 か所の一級アルコール	6	0	1	3.9

1. 各候補の M1 の予測値を求め、現時点の情報で目標を満たしそうか、何が未確認かを述べなさい。(6 点)
2. L1 と L3 について、沸点の予測の余裕（下限から・上限まで）と w_{25} の余裕を求めなさい。「余裕を足した合計が大きい方を選ぶ」という選び方の問題点も述べなさい。(4 点)
3. 次に、1 つの候補について 1 つの物性だけを確かめられます。候補、確かめる物性と条件、優先する理由、結果が期待と違ったときの対応を述べなさい。答えは 1 つではありません。(10 点)

予測値が目標の範囲に入ることと、実際に条件を満たすことを区別して答えなさい。

付録 D　演習の参考解答

各章の演習と章末問題、付録 C の総合問題の参考解答です。判断を問う問題では、答えは一つに決まらないことがあります。そのときは、成立する答えの例と、その理由を示します。成立しない答えの例も添えたので、自分の答えと照らし合わせてください。

第 1 章

演習 1：何が足りないかを確かめる

1. 炭素数の順は、エタノール（2）→ 1-ブタノール（4）→ 1-ペンタノール（5）→ 1-ヘキサノール（6）です。点は右上がりに並び、沸点は 78 → 117 → 138 → 157 ℃ と上がります。仮説（置く前）の例は「炭素数が増えると沸点が上がりそう」、観察（置いたあと）の例は「この 4 例では、炭素数が多いほど沸点が高い」です。観察に「この 4 例では」が付いているかを確かめてください。
2. 沸点は範囲の条件なので両側と比べます。$115 \le 138 \le 140$ で、満たします（上限まで 2 ℃）。w_{25} は下限の条件で、$2.19 < 4.0$ なので満たしません。差は $4.0 - 2.19 = 1.81$ です。

演習 2：記録を 3 つに分類する

1. R5 は「現状では判断できない」です。空欄は測っていない、または値が分からないことを表し、0 ではありません。判定を保留します。
2. R2 は「換算や追加確認が必要」です。温度が 30 ℃ で違うので、25 ℃ の値か、温度による変わり方の情報が必要です。R6 も「換算や追加確認が必要」です。分母が溶液 100 mL なので、同じ温度の飽和溶液の密度が必要です（1.3 節の発展の箱）。どちらも、一般的な温度の影響や水の密度を仮定して、手元で直してしまわないことが大切です。

演習 3：最初の構造変更案と、必要なデータ

1. 成立する候補分子の例は、OH を鎖の内側へ移す（2-ペンタノールなど）、枝分かれを付ける（3-メチル-1-ブタノールなど）です。炭素数 5〜6 の一級・二級アルコールで、表の 4 分子（図 1-2）の資料値から要求を外れると分かっていない候補分子なら、どれでもかまいません。成立しない例は「炭素数を 3 にする」（依頼の条件の外）や、1-ヘキサノール（表の 4 分子の 1 つで、沸点 157 ℃ が上限を超えると分かっている）です。
2. 必要なのは、候補分子の w_{25} と沸点です。構造を変えると沸点も変わりうるので、w_{25} だけでは足りません。w_{25} は 25 ℃・水相・飽和溶液 100 g あたりの値と、条件まで書きます。
3. 言えません。4 分子はどれも直鎖の一級アルコールで、OH の位置や枝分かれの違う分子のデータがないからです。「OH の位置を変えれば w_{25} も変わるかもしれない」のように、仮説として書くのはかまいません。「選んだ候補分子は w_{25} が 4.0 以上になる」のように、確かめていない値を結論として書くのは成立しません。

章末問題

手順を実行する

1. 溶液の質量は $100 + 3.0 = 103.0$ g なので、$\frac{3.0}{103.0} \times 100 = 2.91$ で、約 2.91 g アルコール／100 g 飽和溶液です。
2. 沸点は 143 ℃ で、上限の 140 ℃ を 3 ℃ 超えるので満たしません。w_{25} は 4.6 で、4.0 以上なので満たします。
3. U は「現状では判断できない」（相も溶質も違う）、V は「現状では判断できない」（欠測）、W は「換算や追加確認が必要」（温度が違う。25 ℃ の値か温度による変わり方が必要）です。

理由と限界を説明する

1. 言えません。4 例は直鎖の一級アルコールだけで、炭素数 5 の分子は 1-ペンタノール 1 つだけです。枝分かれや OH の位置が違う炭素数 5 の分子が、同じ沸点になるかは、このデータからは分かりません。

2. 言えません。PubChem の値が ICSC から転載されたものなら、同じ測定が 2 か所に書かれているだけで、確かめた測定は 1 つです。独立した確認と言うには、別の測定までたどる必要があります。
3. 空欄を 0 とすると、平均は $\frac{4.0+4.4+0}{3} = 2.8$ になり、値のある 2 つの平均 4.2 より大きく下がります。測っていない値を、測って 0 だった値として扱ってしまうからです。空欄は平均に入れず、保留します。

別の題材へ使う

1. 温度・時間（反応条件そのもの）と、収率の求め方（NMR で定量か、単離して秤量か）が違います。反応条件の良し悪しを比べるなら、求め方をそろえた記録が必要です。原料の量や純度、反応のスケールなども、記録としてそろえておきます。
2. 答えは人によって違います。たとえば、そろえる条件は測定の温度、記録しておく条件は試料の調製日や装置の型番、などです。

第 2 章

演習 1：C5 の分子の記述子を数える

1. 1-ペンタノールは $(5, 0, 0)$、2-ペンタノールは $(5, 1, 0)$ で、違うのは s だけです。1-ペンタノールの OH の炭素は炭素 1 個と結合（一級）、2-ペンタノールは 2 個と結合（二級）しています。2-ペンタノールの s を 2 と書くのは、OH の位置番号を入れてしまった誤りです。
2. $(5, 0, 1)$ です。OH の炭素（CH_2）に結合する炭素は 1 個なので一級で、$s = 0$ です。3 番の炭素が炭素 3 個と結合しているので、分岐点が 1 個です。枝分かれしているから $s = 1$、とするのは、枝分かれと二級を混ぜた誤りです。
3. 3-ペンタノールは、OH の炭素が炭素 2 個と結合（二級）し、分岐点はないので $(5, 1, 0)$ です。2-メチル-1-ブタノールは、OH の炭素（CH_2）が一級で、2 番の炭素が分岐点なので $(5, 0, 1)$ です。

演習 2：同じ行になる分子をまとめる

1. $(5, 1, 0)$ の 2-ペンタノールと 3-ペンタノール、$(5, 0, 1)$ の 2-メチル-1-ブタノールと 3-メチル-1-ブタノールの 2 組です。5 つの分子は 3 種類の行になります（図 2-6）。
2. 予測値は同じになります。3 列の値が、どちらも $(5, 1, 0)$ で同じだからです。実際の沸点が同じとは言えません。OH の位置の違いは 3 列に入っておらず、違うかどうかは資料値で確かめるまで分かりません（資料値は 第 5 章 で示します）。

演習 3：区別できる列を 1 つ足す

1. 1-ペンタノール 0、2-ペンタノール 1、3-ペンタノール 2、2-メチル-1-ブタノール 0、3-メチル-1-ブタノール 0 です（図 D-1）。2-ペンタノールと 3-ペンタノールは、OH が鎖の端から 2 番目か真ん中かの違いが数に表れて、区別できます。2-メチル-1-ブタノールと 3-メチル-1-ブタノールは、どちらも OH の炭素自身が末端なので 0 で、区別できません。
2. 2 つの違いは分岐点の位置なので、OH の炭素から分岐点までの C–C 結合の数（1 と 2）などを数えます（2.6 節の発展の箱）。成立しない答えの例は「末端までの距離を足せば、5 つとも区別できる」です。分岐の位置が違う組が残ります。
3. 言えません。区別できるようになっても、その違いが沸点の違いに合っているかは、資料値で確かめるまで分かりません。通し番号は沸点と関係がないので、予測がよくなる理由がありません。成立しない答えの例は「列を足すほど、予測はよくなる」です。確かめていないことを結論にしています。



図 D-1　末端までの距離です。緑の線が、OH の炭素（紺の丸）から最も近い末端までの経路です。

章末問題

手順を実行する

1. (a) 3-メチル-2-ペンタノールで、$(6, 1, 1)$ です。OH の炭素（2 番）は炭素 2 個と結合するので二級、3 番の炭素が炭素 3 個と結合するので分岐点です。(b) 2-エ

チル-1-ブタノールで、(6, 0, 1) です。OH の炭素は CH_2 で一級、その隣の炭素が炭素 3 個（OH の炭素と、2 つのエチル基の CH_2）と結合するので分岐点です。(c) 2,2-ジメチル-1-プロパノールで、(5, 0, 1) です。2 番の炭素は炭素 4 個と結合していますが、分岐点は 1 個と数えます。

2. どちらも、OH の炭素（2 番）の隣の 1 番の炭素が末端なので、距離は 1 です。この列では区別できません。2 つの違いは分岐点の位置で、分岐点までの距離なら 1 と 2 で区別できます。
3. (a) は 2-メチル-2-ブタノールで、三級アルコールなので使いません。s に三級を表す値がなく、OH の炭素自身が b の分岐点に数えられてしまいます。(b) は 1-メトキシブタンで、エーテルなので使いません。OH がないので s が決まらず、水素結合のしかたもアルコールと違うので、アルコールのデータで作った予測を当てはめる根拠がありません。

理由と限界を説明する

1. 言えません。記述子が同じなのは、n・s・b の 3 列では区別できない、という意味です。2 つは分岐点の位置が違い、その違いは 3 列に入っていません。沸点が同じかどうかは、資料値で確かめるまで分かりません。
2. 正しくありません。分子量は分子式だけで決まるので、分子式が同じ異性体どうしでは、分子量はすべて同じです。区別したい分子の間で値が変わらない列は、足しても区別の役に立ちません。
3. 2 つの点で言い過ぎです。1 つ目に、列を増やしても、すべての列で同じ値になる分子の組が残ることがあります（同じ行の分子がなくなるとは限りません）。2 つ目に、区別できることと、予測がよくなることは別で、予測がよくなったかは、資料値と比べて確かめるまで分かりません。

別の題材へ使う

1. 表に入れていない情報（省いた情報）が、2 つの試料で違っていた可能性があります。たとえば、粒の大きさ、試料の形（粉末か薄膜か）、調製の手順の細部、測定の温度や方法などです。足す列の候補は、そのうち記録できるもの（粒の大きさ、測定温度など）です。ただし、足した列が原因だと決まったわけではなく、確かめるには、その列だけを変えた試料で比べる必要があります。

2. 答えは人によって違います。たとえば、合成なら溶媒の乾燥の程度（含水量を測って数にする）、材料なら試料を保管した日数（日数をそのまま数にする）などです。数にしにくい情報は、あるかないか（1 か 0）で列にする方法もあります。

第 3 章

演習 1：s の重み 2 で予測する

1. 記述子の違いは、n が 0、s が 1、b が 0 です。$0 + 2 \times 1 + 0 = 2$ で、距離は 2 です。s の違いに重みを掛け忘れて 1 とするのは誤りです。
2. 2-ブタノールまでは $|5-4| + 2|1-1| + |0-0| = 1$、2-メチル-1-プロパノールまでは $1 + 2 + 1 = 4$ です。距離は 1-ペンタノール 2、2-ブタノール 1、2-メチル-1-プロパノール 4 なので、最近傍は 2-ブタノールで、予測は 100 ℃ です。
3. OH の炭素の級（s）の違いを距離に入れ、しかも重み 2 で、炭素 1 個の違いより重く数えたからです。炭素数だけでは 1-ペンタノールが距離 0 でしたが、重み 2 の距離では s の違いが距離 2 になり、2-ブタノール（距離 1）の方が近くなりました。重み 1 なら、2 つは同じ距離 1 になります（演習 2）。

演習 2：s の重み 1 で予測する

1. 距離は、1-ペンタノール $0 + 1 + 0 = 1$、2-ブタノール $1 + 0 + 0 = 1$、2-メチル-1-プロパノール $1 + 1 + 1 = 3$ です。同じ距離の最近傍が 2 つあるので、$\frac{138+100}{2} = 119$ で、予測は 119 ℃ です。
2. 変わります（100 ℃ → 119 ℃）。重みを変えただけで、最近傍と予測が変わります。
3. 決められません。2-ペンタノールの資料値をまだ見ていないので、3 つとも決め方から出た予測です。「119 ℃ は真ん中だから正しそう」は、真ん中にあることが正しさの理由にならないので成立しません。

演習 3：どの予測を採用するか

1. 距離は 1-ペンタノールが重みそのもの、2-ブタノールが 1 です。重み 0.9 なら 1-ペンタノールが最近傍で 138 ℃、1.1 なら 2-ブタノールが最近傍で 100 ℃ です。予測は急に変わります。119 ℃ は重みがちょうど 1 のときだけ出ます（図 3-6）。

2. 炭素数だけ：炭素数の違いを見た（138 ℃、範囲に入る）。重み 2：OH の炭素の級の違いを重く見た（100 ℃、下限を下回る）。重み 1：炭素数の違いと級の違いを同じに見た（平均 119 ℃、範囲に入る）。予測の判定で、資料値の判定ではありません。
3. 成立する答えの例は、138 ℃（炭素数が沸点に最も影響すると考えた）、100 ℃（OH の炭素の級の違いを、炭素 1 個の違いより重く見た）、119 ℃（炭素数の違いと級の違いを、同じ重みで見た）、決めない（資料値で確かめてから選ぶ）です。どれも、理由と、まだ確かめていないこと（2-ペンタノールの資料値）が書けていれば成立します。成立しない例は「119 ℃ は真ん中だから正しい」「重み 1 が最良の重み」「構造の情報を増やしたので、重み 2 の予測がよい」です。どれも確かめていないことを結論にしています。

章末問題

手順を実行する

1. 距離は、1-ヘキサノール $0+2\times1+0=2$、2-ブタノール $2+0+0=2$、1-ペンタノール $1+2\times1+0=3$ です。最近傍は 1-ヘキサノールと 2-ブタノールの 2 つ（距離 2）なので、$\frac{157+100}{2}=128.5$ で、予測は 128.5 ℃ です。
2. 距離は、1-ヘキサノール 1、2-ブタノール 2、1-ペンタノール 2 になります。最近傍は 1-ヘキサノールだけで、予測は 157 ℃ です。
3. 距離は、1-ヘキサノール α、2-ブタノール 2、1-ペンタノール $1+\alpha$ です。1-ペンタノールはいつも 1-ヘキサノールより遠いので、1-ヘキサノールだけが最近傍になるのは $\alpha<2$（$0\leq\alpha<2$）のときです。$\alpha=2$ では 2-ブタノールと同じ距離（予測 128.5 ℃）、$\alpha>2$ では 2-ブタノールが最近傍（予測 100 ℃）になります。

理由と限界を説明する

1. 言えません。1 つの分子で予測が資料値に近かったことは、その分子での一致にすぎません。重み 1 で 1-ヘキサノールが最近傍になるのは、$\alpha<2$ のどの重みでも同じなので、この 1 例は重みの範囲さえ決めていません。ほかの分子でも合うかは、別のデータで確かめる必要があります。
2. 正しくありません。距離が 0 なのは、記述子の組が同じということです。2-ヘキサノールと 3-ヘキサノールのように、記述子の組が同じでも、記述子に入っていない違い（OH の位置など）がある分子があります。その違いで物性が違えば、距離 0 の参照の値を使っても外れます。

3. 分子量は、炭素 1 個（CH_2）の違いで約 14 変わります。一方、s と b の違いは 0 か 1 なので、距離はほとんど分子量の違いで決まり、OH の炭素の級や分岐の違いは、ほとんど影響しなくなります。尺度（値の幅）が大きく違う列は、幅をそろえるか重みで調整してから、距離を作ります。

別の題材へ使う

1. 何を似ているとしたか（骨格、置換基の位置など）と、何が違うか（置換基の種類、反応の温度や時間、スケールなど）です。違う点が収率に影響しそうなら、参照の値をそのまま使わずに、その違いも書き添えておきます。
2. 答えは人によって違います。たとえば、参照にした文献の合成法について、似ている点は「同じ触媒と溶媒」、違う点は「基質の置換基と、反応温度」などです。

第 4 章

演習 1：M0 の残差を読む

1. M0 の予測は、エタノール $20 \times 2 + 38 = 78$、1-ブタノール 118、1-ペンタノール 138、1-ヘキサノール 158 ℃ です。残差は順に $78 - 78 = 0$、$117 - 118 = -1$、$138 - 138 = 0$、$157 - 158 = -1$ ℃ です。
2. 直鎖の一級の 4 分子では残差が 0 か −1 ℃ と小さいのに、二級・分岐の 3 分子では −10〜−18 ℃ と大きく、そろって負です。3 分子は、s か b が 1 です（二級か分岐）。M0 は直鎖の一級の傾向にはよく合いますが、二級・分岐の分子は高く予測します。
3. 予測値が同じになるのは、M0 が炭素数しか見ず、2 つとも炭素数 4 だからです（どちらも 118 ℃）。残差が違うのは、資料値が違う（100 ℃ と 108 ℃）からで、その違いは二級か分岐かという、M0 に入っていない構造の違いによると考えられます。

演習 2：残差平方和で c_s を選ぶ

1. 予測は $98 - 17 = 81$ ℃ と $118 - 17 = 101$ ℃、残差は $83 - 81 = +2$ と $100 - 101 = -1$ です。$\mathrm{SSE} = 4 + 1 = 5$ ℃^{2} です。

2. $c_s = 15$ では、予測 83・103 ℃、残差 0・−3 で、SSE は 9 ℃² です。15・17・20 の SSE は 9・5・29 ℃² なので、最小は $c_s = 17$ です。
3. $c_s = 16$ では、予測 82・102 ℃、残差 +1・−2 で、SSE は 5 ℃² です。17 と同じなので、17 だけが最もよい係数とは言えません。17 は、試した 3 つの中で最小というだけです（連続の値で最も小さいのは 16.5。4.6 節の発展の箱）。

演習 3：M1 を作る

1. M0 の予測 118 ℃ から c_b を引いて 108 ℃ になればよいので、$118 - c_b = 108$ から $c_b = 10$ です。1 分子なので、残差が 0 のとき SSE が最小です。
2. M1 は $T = 20n + 38 - 17s - 10b$ です。

演習 4：予測を記録する

1. 2-ペンタノールは $20 \times 5 + 38 - 17 = 121$ ℃、3-メチル-1-ブタノールは $20 \times 5 + 38 - 10 = 128$ ℃ です。記録の例：2-ペンタノールは、最近傍（炭素数だけ）138 ℃、最近傍（$\alpha = 2$）100 ℃、最近傍（$\alpha = 1$）119 ℃、M1 121 ℃。3-メチル-1-ブタノールは M1 128 ℃。
2. 言えません。3 分子の残差（+2、−1、0）が小さいのは、その 3 分子に合うように係数を選んだからです。合っても、使っていない分子で当たることの証拠にはなりません。成立しない例は「残差が小さいので、M1 は正しい式だ」です。

章末問題

手順を実行する

1. M0 の予測は $20 \times 6 + 38 = 158$ ℃、残差は $143 - 158 = -15$ ℃ です。
2. M0 の予測は Y1 が 158 ℃、Y2 が 178 ℃ なので、残差は Y1 が $c_s - 15$、Y2 が $c_s - 19$ です。SSE は、14 で $1 + 25 = 26$、16 で $1 + 9 = 10$、18 で $9 + 1 = 10$ ℃² です。16 と 18 が同じ SSE で最小になり、この候補からは 1 つに決まりません（連続の値で最も小さいのは、15 と 19 の平均の 17）。どちらを選んだかと、その理由を書いておきます。
3. 3-メチル-2-ブタノールは $20 \times 5 + 38 - 17 - 10 = 111$ ℃ で、下限の 115 ℃ を下回ります。2-エチル-1-ブタノールは $20 \times 6 + 38 - 10 = 148$ ℃ で、上限の 140 ℃ を上回ります。どちらも予測値で、資料値で確かめたものではありません。

理由と限界を説明する

1. 正しくありません。$c_b = 10$ は、分岐のある分子 1 つ（2-メチル-1-プロパノール）の残差を打ち消すように決めた値です。ほかの分子でも分岐で同じだけ沸点が下がるかは、確かめていません。係数は、決めたデータの範囲での平均的な補正で、法則ではありません。
2. 直鎖の一級 4 分子のうち、両端の 2 つは M0 を決めるのに使った分子なので、合うのは当然です。また、4 分子はどれも直鎖の一級で、二級の分子は入っていません。実際に、二級の 2-プロパノールでは残差が −15 ℃ でした。合ったデータと違う構造の分子にまで、当てはまりを広げることはできません。
3. 正と負の残差が打ち消し合うので、大きく外れていても和が 0 に近くなることがあります。たとえば残差が +10 と −10 なら和は 0 ですが、どちらの分子にも 10 ℃ 外れています。2 乗して足す（SSE）か、絶対値を足すことで、打ち消し合いを防ぎます。

別の題材へ使う

1. 言えることの例：40〜80 ℃ の範囲では、温度が高いほど収率が高い傾向があり、その範囲の中の温度での収率を見積もれる。言えないことの例：なぜ上がるのか（機構）、80 ℃ より高い温度でも同じ傾きで上がるか、温度を上げれば必ず収率が上がるか。
2. 答えは人によって違います。たとえば「濃度と吸光度の検量線」なら、測った濃度の範囲の外でも直線が成り立つかは、その式からは言えません。

第 5 章

演習 1：最近傍と M1 を、5 分子で比べる

1. 2-ペンタノールの資料値は 119 ℃ なので、炭素数だけは $|119 - 138| = 19$、$\alpha = 2$ は $|119 - 100| = 19$、$\alpha = 1$ は $|119 - 119| = 0$ ℃ です（M1 は 2 ℃）。
2. 5 分子の予測と絶対誤差は、次のとおりです（単位は ℃）。

分子	資料値	炭素数だけ	$\alpha = 2$	$\alpha = 1$
1-プロパノール	97	104（7）	123（26）	115.3（18.3）
2-ペンタノール	119	138（19）	100（19）	119（0）
2-メチル-1-ブタノール	128	138（10）	123（5）	123（5）

分子	資料値	炭素数だけ	$\alpha = 2$	$\alpha = 1$
3-メチル-1-ブタノール	132	138（6）	123（9）	123（9）
3-ペンタノール	116	138（22）	100（16）	119（3）
MAE		12.8	15.0	約 7.1

括弧の中が絶対誤差です。M0 は 11.6 ℃、M1 は 2.4 ℃ なので、この 5 分子では M1 の MAE が最も小さくなります。最近傍は参照分子が 3 つしかなく、炭素数 3 の 1-プロパノールは参照分子の範囲の外にあるため、大きく外れます。

演習 2：結果から言えること

1. 言えます（「この 5 分子では」の範囲で）。使っていない 5 分子で、予測を記録してから確かめ、MAE が 11.6 ℃ から 2.4 ℃ に小さくなりました。ただし、「M1 はどんな分子でもよい式だ」までは言えません。5 分子は、記述子の組み合わせで 3 種類しかありません。
2. 言えません。MAE は平均で、上限ではありません。この 5 分子でも、3-メチル-1-ブタノールは 4 ℃、3-ペンタノールは 5 ℃ 外れました。言える文の例は「この 5 分子では、M1 の絶対誤差は平均 2.4 ℃、最大 5 ℃ だった」です。
3. 言えません。資料値を見て係数を直した時点で、5 分子は係数を直すのに使ったデータになります（データ漏洩）。直した式は、別の使っていないデータで確かめます。

演習 3：組み合わせと係数の数

1. 言えません。列ごとの範囲と、組み合わせの実績は別です。$s = b = 1$ の分子は、確かめたデータに 1 つもありません。言える文の例は「$(6, 1, 1)$ は式で値は出るが、二級＋分岐の組み合わせは確かめていない」です。
2. 言えません。前半の、SSE を 0 にできることは正しいです。言えないのは「だから予測もよくなる」の所です。係数を分子の数だけもてば、どの資料値もそのまま再現できますが、それは当てはまりが完全になっただけです（過学習）。新しい分子にどの係数を使えばよいかが決まらず、予測はできません。

章末問題

手順を実行する

1. 絶対誤差は $|125-120|=5$、$|129-131|=2$、$|104-104|=0$ ℃ です。MAE は $\frac{5+2+0}{3}\approx 2.3$ ℃ です。
2. 絶対誤差は 7、3、6 ℃ で、MAE は $\frac{16}{3}\approx 5.3$ ℃ です。この 3 物質では、問 1 の方法の MAE の方が小さくなりました。言えるのは「この 3 物質では」までで、ほかの物質でも問 1 の方法がよいかは分かりません。3 物質はとても少ないことにも注意します。
3. $(4,0,0)$ だけです。$(8,0,0)$ は炭素数が範囲の外、$(6,1,1)$ は $s=b=1$ の組み合わせが範囲の外、$(5,0,2)$ は $b=2$ が範囲の外です。

理由と限界を説明する

1. 正しくありません。MAE は絶対誤差の平均なので、それより大きく外れる分子があってかまいません。たとえば絶対誤差が 0、0、9 ℃ でも、MAE は 3 ℃ です。
2. 言えません。5 分子の資料値を見て係数を直したので、5 分子はもう係数を決めるのに使ったデータです。同じ 5 分子で MAE が小さくなったのは当てはまりがよくなっただけで、予測がよくなったかは、別の使っていないデータで確かめないと分かりません。
3. $\alpha=1$ の最近傍が 2-ペンタノールで一致したのは、1 つの分子での結果です。同じ方法を 5 分子すべてで確かめると、MAE は約 7.1 ℃ で、M1 の 2.4 ℃ より大きく、1-プロパノールでは 18 ℃ 以上外れました。1 例の一致から、方法や重みが最良だとは言えません。

別の題材へ使う

1. 別のロット（別の日、別の原料ロット）で、同じ条件の収率を確かめます。ロット 1 での 85 % は、条件を決めたデータでの結果（当てはまり）なので、それだけでは「この条件なら 85 % が出る」とは言えません。報告するときは、何で確かめたかも書きます。
2. 答えは人によって違います。たとえば「同じ試料で測定条件を最適化し、同じ試料で性能を評価した」場合は、別に作った試料で同じ性能が出るかを確かめます。

第 6 章

演習 1：既知の候補を比べる

1. 2-ペンタノールは、沸点の下限から $119 - 115 = 4$ ℃、上限まで $140 - 119 = 21$ ℃、w_{25} は $4.46 - 4.0 = +0.46$ です。3-ペンタノールは、下限から 1 ℃、上限まで 24 ℃、w_{25} は $+1.15$ です。どちらも 2 つの要求を満たします。2-ペンタノールは沸点の下限に、3-ペンタノールは w_{25} に余裕があります。
2. どちらも選べます。成立する例は「沸点の下限が心配なので、下限から 4 ℃ 離れた 2-ペンタノール」「すすぎの水溶性を重視して、w_{25} の余裕が大きい 3-ペンタノール」です。成立しない例は「2-ペンタノールは予測でも合ったから」です（予測の当たり外れは、候補の物性の判定とは別です）。
3. 言えません。℃ の差と w_{25} の差は単位が違うので、そのまま足しても意味がありません。点数にするなら、尺度と重みを目的から決める必要があります（第 3 章 の尺度と重みと同じ考え方）。

演習 2：次に確かめる 1 つを決める

1. どちらも成立します。w_{25} を選ぶ目的の例は「欠測で、4.0 以上か未満かで残すか外すかが変わるので、候補に残せるかを決める」、沸点を選ぶ目的の例は「未確認の予測が外れていれば外すので、二級と分岐の補正を同時に引く M1 の予測を点検する」です。
2. w_{25} なら、4.0 以上で候補に残す（沸点は予測のまま）、4.0 未満で優先度を下げる。沸点なら、115〜140 ℃ に入れば沸点は満たす（w_{25} は欠測のまま）、範囲の外なら外すか、補正を見直す。
3. 言えません。沸点を確かめても（資料値 132 ℃）、w_{25} は欠測のままです。w_{25} を選んでも、沸点は予測のままです。1 つ確かめたら、何が分かり、何が残ったかを書きます（6.6 節）。

演習 3：最終提案をまとめる

6.6 節の「最終提案の例」が、成立する例の 1 つです。既知の候補を 3-ペンタノールにしたり、次に沸点を確かめたりする提案も、理由が余裕や未確認の点と結び付いていれば成立します。成立しない例も 6.6 節に挙げました（予測から採用を確定する、条件の違う値を w_{25} として使う、余裕を足す）。演習 4 は人によって違います。

第 1 章の候補分子が最終提案と違っていても、何を根拠に判断を変えたかが書けていれば成立します。

章末問題

手順を実行する

1. F1 は、沸点の下限から $124-115=9$ ℃、上限まで $140-124=16$ ℃、w_{25} は $4.4-4.0=+0.4$ で、2 つの要求を満たします。F2 は、沸点が $113-115=-2$ で下限に 2 ℃ 足りず（上限までは 27 ℃）、w_{25} は $+2.1$ です。w_{25} は満たしますが、沸点の要求を満たさないので、目標の領域の外です。
2. M1 は $178-17s-10b$ で、$(7,0,0)$ 178、$(7,0,1)$ 168、$(7,0,2)$ 158、$(7,1,0)$ 161、$(7,1,1)$ 151、$(7,1,2)$ 141 ℃ です。115〜140 ℃ に入る組はありません。最も近い $(7,1,2)$ の 141 ℃ も範囲の外で、しかも $b=2$ は確かめた範囲の外です。いずれも予測値で、炭素数 7 も確かめた範囲の外です。
3. 次に確かめるのは w_{25}（25 ℃・水相・飽和溶液 100 g あたり）です。目的は、欠測を埋めて候補に残せるかを決めることです。4.0 以上なら候補に残し（沸点は資料値で範囲内）、4.0 未満なら優先度を下げます。

理由と限界を説明する

1. 正しくありません。余裕が MAE より大きいことは、範囲に入りそうだと言う手がかりにはなります。しかし MAE は平均で上限ではなく、MAE より大きく外れる分子もあります。「範囲に入りそうだが、確かめていない」と書くのが妥当で、「必ず」とは言えません。外れる確率も、この情報からは出せません。
2. 単位の違う余裕（℃ と w_{25}）を、そのまま足しています。たとえば w_{25} を別の単位で表せば合計の大小は変わりうるので、この合計には意味がありません。どちらの余裕を重く見るかは、目的から決めて、理由として書きます。
3. 正しくありません。1 つの分子で予測が合ったことは、その分子での結果にすぎません。$s=b=1$ のほかの分子（3-メチル-2-ペンタノールなど）でも M1 が合うかは、確かめていません。

別の題材へ使う

1. 例：そろえる条件は測定の温度、記録する条件は電解液の調製日や含水量。次の確認実験は、粘度の目標の境界に近い組成を 1 つ測り、目標以下なら候補に残し、超えたら組成の方向を見直す、など。

2. 答えは人によって違います。6.7 節の塗料の例のように、5 つの項目を短く書けていれば十分です。研究室の未公開データは書く必要がありません。

付録 C の総合問題

配点の目安も添えます。判断を問う問いは、成立する答えの例を示しました。理由がデータや未確認の点と結び付いていれば、例と違う答えでも成立します。

問 1（15 点）

1. B です。25 ℃、水相、アルコールの質量／飽和溶液の質量という条件と定義が、w_{25} と一致します。（選択 1 点、理由 2 点）
2. 溶液の質量は $100 + 3.9 = 103.9$ g なので、$\frac{3.9}{103.9} \times 100 \approx 3.75$ g／100 g 飽和溶液です（3 点）。B の 3.6 とは一致しませんが、それだけでどちらかが誤りとは言えません。条件をそろえることと、値が信頼できることは別の問題で、測定の違いや資料ごとの差がありえます。どちらが確かかは、出典や測定の条件までたどって確かめます（2 点）。分母を読み替えて 3.9 のまま使う答えは、換算の 3 点を与えません。
3. C は温度が 20 ℃ で、w_{25}（25 ℃）と違います。溶けやすさは温度で変わるので、25 ℃ の値か、温度による変わり方の情報が必要です（2 点）。D はアルコール相の中の水の量で、相も溶質も違います。水相の中のアルコールの量は出せません（2 点）。
4. 入れてはいけません。空欄は、測っていない・分からないという意味で、測って 0 だったこととは違います。0 を入れると平均が不当に下がります（判断 1 点、理由 2 点）。

問 2（20 点）

1. 各値 1 点です。

分子	名称（参考）	n	s	b
K1	5-メチル-2-ヘキサノール	7	1	1
K2	2-メチル-3-ヘキサノール	7	1	1
K3	3,3-ジメチル-1-ペンタノール	7	0	1

K3 の 3 番の炭素は炭素 4 個と結合しますが、分岐点は 1 個と数えます。K2 の s に位置番号の 3 を入れる、K3 の b を 2 とする誤りに注意します。

2. 距離は 0 です（1 点）。距離 0 は、この記述子の上で同じという意味で、同じ沸点とは限りません（1 点）。OH の位置（2 位か 3 位か）や分岐の位置、OH の炭素の両側の鎖の長さが、表現に含まれていません（1 点）。
3. $\alpha = 2$ では、距離は H1 $0+0+1=1$、H2 $1+0+0=1$、H3 $0+2+0=2$ です。H1 と H2 が同じ距離なので、$\frac{172+150}{2} = 161$ ℃ です（1 点）。$\alpha = 1$ では、H3 も $0+1+0=1$ になり、3 つが同じ距離です。$\frac{172+150+155}{3} = 159$ ℃ です（1 点）。予測が変わったのは、H3 は K3 と s だけが違い、s の重みを 2 から 1 に下げると H3 までの距離が縮んで、最近傍に加わったからです（2 点）。どちらの予測が正しいかは、K3 の資料値で確かめるまで分かりません。
4. E2 は三級アルコールです。s に三級を表す値がなく、OH の炭素自身が分岐点に数えられてしまいます。M1 を作ったデータ（一級・二級）の範囲の外です（2 点）。F2 はエーテルで、OH がなく s が決まらず、水素結合のしかたもアルコールと違います（2 点）。

問 3（25 点）

分子	M0	M1	資料値（架空）	M0 の絶対誤差	M1 の絶対誤差
W1	158	158	155	3	3
W2	158	141	138	20	3
W3	158	148	151	7	3
W4	118	101	103	15	2

1. 上の表の M0・M1 の列です（各 1 点）。
2. M0 の MAE は $\frac{3+20+7+15}{4} = \frac{45}{4} = 11.25$ ℃、M1 の MAE は $\frac{3+3+3+2}{4} = \frac{11}{4} = 2.75$ ℃ です（絶対誤差 各 4 点、MAE 各 2 点）。符号のまま平均した答えは、MAE の 2 点を与えません。
3. この 4 分子では M1 の MAE が小さいので、M1 を使いたい（2 点）。言えないことの例：ほかの分子でも M1 の誤差が小さいこと、誤差がどれも 2.75 ℃ 以下になること、M1 が正しい式であること。4 分子は少なく、記述子の組み合わせも限られています（3 点）。

問 4（20 点）

1. 正しくありません（2 点）。MAE は絶対誤差の平均で、上限ではありません（3 点）。問 3 でも、MAE 2.75 ℃ に対して W1〜W3 の絶対誤差は 3 ℃ で、MAE より大きくなっています。これから予測する分子では、もっと大きく外れることもありえます（3 点）。
2. 確かめ用の 4 分子の資料値を見て係数を直したので、4 分子は係数を決めるのに使ったデータになりました（データ漏洩）。同じ 4 分子で MAE を求め直しても、予測がよくなったことは示せません（3 点）。直した式は、別の使っていないデータで確かめます（3 点）。なお、実際に $c_s = 20$ で計算し直すと、W2 の誤差は 0 になりますが、W4 の誤差は 2 ℃ から 5 ℃ に増え、MAE は 2.75 ℃ のままです。1 つの分子に合わせると、ほかの分子が外れることもあります。
3. 1 つ目に、炭素数 7 は、M1 を作ったデータの炭素数（2〜6）の範囲の外です（3 点）。2 つ目に、$s = b = 1$ の組み合わせは、M1 の作成にも確認にも使っていません（3 点）。式に入れて値が出ることと、その値を信用できることは別です。

問 5（20 点）

1. L1 は $20 \times 6 + 38 - 17 = 141$ ℃、L2 は $20 \times 7 + 38 - 17 - 10 = 151$ ℃、L3 は $20 \times 6 + 38 - 10 = 148$ ℃ です（各 1 点）。判断の例（各 1 点）：L1 は沸点の予測も w_{25} も目標を満たしますが、沸点は予測値です。L2 は沸点の予測が上限を 1 ℃ 超えますが、境界から 5 ℃ 以内なので外れると言い切れません。しかも炭素数 7 と $s = b = 1$ の組み合わせはどちらも確かめた範囲の外で、w_{25} も不明です。L3 は w_{25} を満たし、沸点の予測は範囲内ですが、上限まで 2 ℃ しかなく、入ると言い切れません。
2. L1 は、沸点の下限から 16 ℃、上限まで 9 ℃、w_{25} は +0.4 です。L3 は、下限から 23 ℃、上限まで 2 ℃、w_{25} は +0.9 です（2 点）。℃ の差と w_{25} の差は単位が違うので、足した合計には意味がありません。また、沸点の余裕は両側を見る必要があり、L3 は上限に近い点を見落としやすくなります（2 点）。
3. 次の観点で評価します（10 点）。候補と物性が明確か（2 点）、条件が適切か（沸点なら常圧、w_{25} なら 25 ℃・水相・飽和溶液 100 g あたり）（2 点）、優先する理由が数値・適用範囲・不足データと結び付いているか（3 点）、期待と違う結果のときの判断が具体的か（3 点）。

解答例 A：L3 の常圧の沸点を確かめる。予測 148 ℃ は上限まで 2 ℃ で、予測の誤差の内なので、範囲に入るかが判断を左右する。150 ℃ 以下なら L3 を候補に残し（w_{25} は 3.9 で満たす）、150 ℃ を超えたら L3 を外し、L1 を優先する。

解答例 B：L1 の常圧の沸点を確かめる。w_{25} と沸点の予測がどちらも目標を満たす候補で、沸点だけが予測値だから。125〜150 ℃ なら L1 を既知の候補として進める。範囲の外なら L1 を保留し、L3 の沸点を確かめる。

解答例 C：L2 の w_{25} を確かめるのは、優先度が低い。沸点の予測が上限を超え、確かめた範囲の外の組み合わせでもあるので、w_{25} が 3.0 以上でも採用に近づかない。このように、確かめても判断が変わりにくい候補を選ばない理由を書けていれば、評価する。

1 つを確かめても、もう 1 つの物性や、ほかの候補は確かめたことにならない点に触れていれば、なおよい答えです。

付録 E　記録の例

この授業で行った判断を、研究のノートに残すための記録の例を、6 つの場面に分けて示します。各表の右の列は空けてあるので、自分の研究（まだなら学生実験）について書いてみてください。研究室の未公開データは、ここに書く必要はありません。

どの記録にも共通する約束は、次の 3 つです。

- 値には、資料値・予測値・架空値の区別と、出典を書く。
- 空欄（欠測）は 0 と書かない。
- 決めたこと（列、重み、係数、計画）は、決めた理由と一緒に、結果を見る前に書く。

データを集めて表にする

比べる前に、条件と出典をそろえた表を作ります（第 1 章）。

表 E-1　データを集めて表にするときの記録の例です。

書く項目	この授業での例	自分の研究では
目的変数	常圧の沸点 T（℃）、w_{25}（25 ℃・水相・飽和溶液 100 g あたりのアルコールの g）	
行（1 行に何を書くか）	1 行に 1 分子	
そろえた条件	沸点は常圧、w_{25} は 25 ℃・水相・飽和溶液 100 g あたり。条件は列の見出しに書く	
値の出典	沸点は国際化学物質安全性カード、w_{25}は IUPAC の溶解度データ集に再録された 1937 年の測定値	
使わなかった記録と理由	温度や分母、相の違う記録（そのままでは比べられない）。換算できるものは換算のしかたを書く	
欠測	構造を変えた分子の値（取り寄せ中）。0 とは書かない	

構造や条件を表の列にする

何を列にし、何を省いたかを、表の横に書いておきます（第 2 章）。

表 E-2　構造や条件を列にするときの記録の例です。

書く項目	この授業での例	自分の研究では
列にした情報	炭素数 n、OH の炭素の級 s、分岐点の数 b	
数え方の決まり	炭素どうしの結合だけを数える。s は位置番号ではない。OH との結合は分岐に数えない	
使ってよい範囲	鎖状・飽和で OH を 1 つもつ一級・二級のアルコール	
省いた情報	OH の位置、分岐の位置	
同じ行になる組	2-ペンタノールと 3-ペンタノール、2-メチル-1-ブタノールと 3-メチル-1-ブタノール	

予測を記録する（封をする）

予測は、資料値を見る前に、決め方と一緒に記録します（第 3 章・第 4 章）。記録したら、資料値を見るまで書き換えません。

表 E-3　予測を記録するときの記録の例です。

書く項目	この授業での例	自分の研究では
予測する分子	2-ペンタノール $(5, 1, 0)$	
予測の方法と決め方	最近傍（参照 3 分子、s の重み 2・1、炭素数だけ）、M1（c_s は候補 15・17・20 から SSE で選んだ 17、c_b は 1 分子から 10）	
予測値	最近傍 138・100・119 ℃、M1 121 ℃（どれも予測値）	
決めるのに使ったデータ	参照の 3 分子、M0 の 2 分子、補正の 3 分子	
まだ見ていないデータ	2-ペンタノールの沸点の資料値	
採用する予測と理由	例：M1 の 121 ℃。二級の補正を、二級の 2 分子のデータで決めたから（確かめるまで、よいとは言わない）	

予測を確かめる

記録した予測を、使っていないデータと比べます（第 5 章）。

表 E-4　予測を確かめるときの記録の例です。

書く項目	この授業での例	自分の研究では
確かめに使ったデータ	式にも参照にも使っていない 5 分子の資料値	
当てはまりか、予測の確認か	予測の確認（予測を記録してから資料値を見た）。教員は値を知っていたので、教育用の確認	
成績	MAE は M0 11.6 ℃、M1 2.4 ℃。M1 の絶対誤差の最大は 5 ℃	
言えること	この 5 分子では、M1 の予測が M0 より小さい誤差だった	
言えないこと	どの分子でも誤差が 2.4 ℃ 以内とは言えない。$s = b = 1$ の組み合わせは確かめていない	
データを見て直したこと	なし（直したら、別のデータで確かめ直す）	

候補の状態と、次の確認を計画する

候補ごとに、物性の値の種類を区別して書き、次に確かめることを、結果を見る前に決めます（第 6 章）。

表 E-5　候補の状態と次の確認を計画するときの記録の例です。

書く項目	この授業での例	自分の研究では
候補と、物性ごとの値の種類	4-メチル-2-ペンタノール：沸点は M1 の予測 131 ℃（$s = b = 1$ は未確認）、w_{25} は欠測	
余裕	沸点の予測は下限から 16 ℃、上限まで 9 ℃（M1 の誤差 0〜5 ℃ より大きい）	
次に確かめるもの	4-メチル-2-ペンタノールの w_{25}（25 ℃・水相・飽和溶液 100 g あたり）	
目的	欠測を埋め、候補に残せるかを決める	
結果ごとの行動	4.0 以上なら候補に残す。4.0 未満なら優先度を下げ、既知の候補で進める	
確かめても残ること	沸点は資料値で確かめても、$s = b = 1$ のほかの分子で M1 が合うかは分からない	

自分の研究で考える

第 6 章の 5 つの項目（6.7 節）を、記録の形にしたものです。

表 E-6　自分の研究で考えるときの記録の例です。

書く項目	この授業での例	自分の研究では
良くしたい物性・収率など	沸点を保ちながら、水溶性を上げる	
変えられる構造・組成・条件	OH の位置、分岐、炭素数	
そろえる条件と、記録する条件	25 ℃・水相・飽和溶液 100 g あたり、常圧。出典を記録する	
役立つ既知データと、不足情報	似た分子の資料値。新しい候補の w_{25} が欠測	
次の確認実験と、結果に応じた判断	4-メチル-2-ペンタノールの w_{25}。4.0 未満なら優先度を下げる	

付録 F　資料と出典

この本で使った実在の分子の値（資料値）と、その出典をまとめます。本文の架空の値（練習用の記録や、架空の物質の値）は、ここには載せていません。

この本の資料値

沸点

沸点は、すべて国際化学物質安全性カード（International Chemical Safety Cards, ICSC）の値です。カード番号を出典の欄に書きました。カードは INCHEM のウェブサイト（https://www.inchem.org/）で公開されています。「初出」は、本文でその値を初めて示した章です。

表 F-1　この本で使った沸点の資料値です（常圧）。

分子	(n, s, b)	沸点 / ℃	出典（ICSC）	初出
エタノール	$(2, 0, 0)$	78	0044	第 1 章
1-ブタノール	$(4, 0, 0)$	117	0111	第 1 章
1-ペンタノール	$(5, 0, 0)$	138	0535	序章
1-ヘキサノール	$(6, 0, 0)$	157	1084	第 1 章
2-ブタノール	$(4, 1, 0)$	100	0112	第 3 章
2-メチル-1-プロパノール	$(4, 0, 1)$	108	0113	第 3 章
2-プロパノール	$(3, 1, 0)$	83	0554	第 4 章
1-プロパノール	$(3, 0, 0)$	97	0553	第 5 章
2-ペンタノール	$(5, 1, 0)$	119	1428	第 5 章
2-メチル-1-ブタノール	$(5, 0, 1)$	128	0506	第 5 章
3-メチル-1-ブタノール	$(5, 0, 1)$	132	0798	第 5 章
3-ペンタノール	$(5, 1, 0)$	116	0536	第 5 章
4-メチル-2-ペンタノール	$(6, 1, 1)$	132	0665	第 6 章

水溶性 w_{25}

w_{25} は、Ginnings と Baum が 1937 年に報告した測定を、IUPAC の溶解度データ集（Solubility Data Series, Vol. 15, *Alcohols with Water*, 1984）が再録した表の値です。同じデータ集には、評価者が丸めた推奨値も載っていますが、この本では、もとの測定値にそろえました。

表 F-2　この本で使った w_{25} の資料値です（25 ℃・水相・飽和溶液 100 g あたりのアルコールの g）。

分子	w_{25}	初出
1-ペンタノール	2.19	序章
2-ペンタノール	4.46	第 6 章
2-メチル-1-ブタノール	2.97	第 6 章
3-メチル-1-ブタノール	2.67	第 6 章
3-ペンタノール	5.15	第 6 章

使わなかった値

ICSC には水溶性の値も載っていますが、分子ごとに温度が違い（20 ℃、25 ℃、30 ℃ など）、分母も溶液の体積です。w_{25} とはそのまま比べられないので、この本では使いませんでした（第 1 章 の考え方）。2-ペンタノールについては、ICSC の水溶性と、上のデータ集に再録された測定値とは、単位を換算しても一致しません。原因は確かめていません。4-メチル-2-ペンタノールの ICSC の水溶性（2 g/100 mL）も、温度が書かれておらず分母も体積なので、$w_{25} = 2$ としては使っていません（第 6 章）。

PubChem のようなデータベースに ICSC の値が転載されていても、それは同じ値がもう 1 か所に書かれているだけで、独立した確認にはなりません（第 1 章）。

本文で引用した文献

- Ginnings, P. M.; Baum, R. “Aqueous Solubilities of the Isomeric Pentanols.” *J. Am. Chem. Soc.* **1937**, 59, 1111–1113. doi:10.1021/ja01285a043（w_{25} のもとの測定。この本では IUPAC Solubility Data Series, Vol. 15 の再録表で確かめた）
- IUPAC Solubility Data Series, Vol. 15, *Alcohols with Water* (1984).
- Alger, *J. Chem. Educ.* **1991**, 68, 939. doi:10.1021/ed068p939.1（2-ブタノールの水溶性の誤記が資料から資料へ広まったことを扱った教育論文。第 1 章）

- Stokes, J. M. et al. “A Deep Learning Approach to Antibiotic Discovery.” *Cell* **2020**, 180, 688–702. doi:10.1016/j.cell.2020.01.021（抗菌薬の探索。序章・第 5 章）
- Gómez-Bombarelli, R. et al. “Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach.” *Nat. Mater.* **2016**, 15, 1120–1127. doi:10.1038/nmat4717（有機 EL の発光分子の探索。序章・第 6 章）
- Delaney, J. S. “ESOL: Estimating Aqueous Solubility Directly from Molecular Structure.” *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1000–1005. doi:10.1021/ci034243x（水溶性の予測式 ESOL。第 4 章。この本では式の形と、係数を約 2,900 件の測定値から決めたことだけを紹介した）

資料値の示し方について

この本では、確かめるための資料値（構造を変えた分子の沸点）を 第 5 章 で、水溶性を 第 6 章 で初めて示しました。予測を記録してから資料値を見る順番で、予測がどれだけ当たったかを正直に確かめるためです。

ただし、この本は一冊にまとまっているので、後ろの章やこの付録を先に見ることもできます。また、著者は教材を作るときに、すべての資料値を見ています。分子の選び方も、教育のためのものです。ですから、第 5 章の確かめは、研究として完全に独立した検証ではなく、読者にとってまだ使っていないデータで確かめる、教育用の確認です（第 5 章）。

改訂履歴

本書は、授業の進行に合わせて順次改訂しています。お手元の版は、表紙の最下部と奥付の版番号で確かめられます。版番号の 3 つの数字は、左から順に、大きな改訂、内容の追加や差し替え、誤植などの修正を表します。

版	発行日	改訂内容
0.2.9	2026 年 11 月 3 日	第 6 章の演習を授業の流れにそろえて演習 3 を 2 つに分け、例題に確かめる値の選び方を追加。「確かめ済み」を「確かめたことになる」、「判断を変えうる点」を「判断を変える値」、「丸い」を「枝が多い、OH が内側」などに変更。
0.2.8	2026 年 11 月 3 日	第 5 章の演習を授業の流れにそろえて演習 2 を 2 つに分け、2-ペンタノールの予測と資料値の図を追加。「調整しすぎ」を「過学習」、単独の「参照」を「参照分子」などに変更。
0.2.7	2026 年 11 月 3 日	第 4 章の演習を授業の流れにそろえて演習 3 を 2 つに分け、「見落とす」を「表していない」、「物差し」を「基準」、「混み合う」を「立体障害」などに変更。
0.2.6	2026 年 11 月 2 日	第 3 章の演習を授業の流れにそろえ、「値を借りる」を「値を使う」、「支持する」を「採用する」、「跳ぶ」を「急に変わる」などに変更。
0.2.5	2026 年 11 月 2 日	第 2 章の演習を授業の流れにそろえ、「抜け落ちる」を「表せない違い」、「混み合い」を「立体障害」などに変更。
0.2.4	2026 年 11 月 2 日	用語を見直し、「案」を「候補分子」、「箱」を「分類」、「量る成分」を「溶質」、「範囲の型・下限の型」を「範囲の条件・下限の条件」などに変更。
0.2.3	2026 年 11 月 2 日	「はじめに」を追加。第 1 章の比較できるデータの節に、水相とアルコール相の値の違いと照らす 4 点の読み方を追加。
0.2.2	2026 年 11 月 1 日	書名を『分子設計』に確定。奥付に、公開ページの URL と利用の条件を追加。
0.2.1	2026 年 11 月 1 日	第 4 章で、係数の向きの化学的な説明が仮説であることを明記。第 5・6 章で、適用範囲の組み合わせの見方（s と b、炭素数まで含めた見方）を補足。

版	発行日	改訂内容
0.2.0	2026 年 11 月 1 日	序章と第 1〜6 章の図・説明・演習の解答を見直し、記号（重み α）と用語（適用範囲・水溶性の単位など）をそろえ、誤りと言い過ぎを修正。
0.1.0	2026 年 10 月 8 日	プロジェクトの雛形を作成。

分子設計

発行日　2026 年 11 月 3 日　第 0.2.9 版発行
著　者　山本 典史

© 2026　山本 典史

本書の著作権は著者が保有します。本書は、学習と教育のために、無償で閲覧・ダウンロード・印刷できます。著者の許可なく、本書の全部または一部を再配布・転載・改変・販売することはできません（著作権法が認める引用を除きます）。授業や勉強会で紹介するときは、ファイルを配らずに、下の公開ページを案内してください。

本書は、講義に合わせて改訂を続けています。引用するときは版番号を添えてください。

最新版と改訂履歴　https://yamlab.jp/books/mdes/

誤りの報告　https://yamlab.jp/access