alvaDesc 3.0

スタートアップガイド

株式会社アフィニティサイエンス

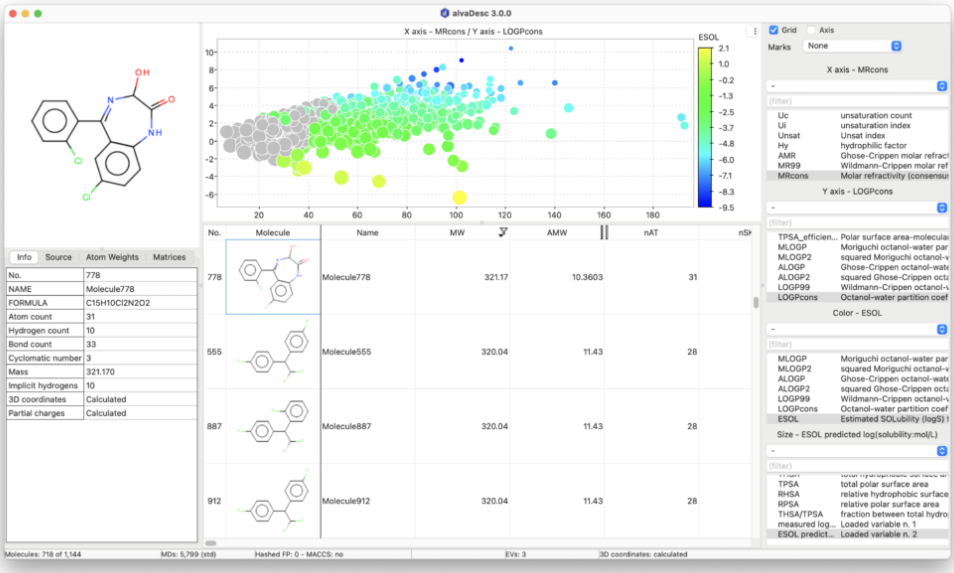
Revision: 2025 年 4 月 24 日 (1.0.0)

株式会社アフィニティサイエンス
〒141-0031 東京都品川区西五反田 1-11-1 アイオス五反田駅前

TEL:　03-6417-3695
FAX:　03-6417-3696
Email: help@affinity-science.com / sales@affinity-science.com
Web:　https://www.affinity-science.com

目次

1. はじめに



alvaDesc は、6000 種近くの分子記述子と 4 種のフィンガープリントを計算・解析することができるソフトウェアです。分子記述子のうち、3 次元座標なしで計算できるものは 4000 種以上、3 次元記述子も約 1600 種実装されています。

塩やイオン液体、金属錯体といった非完全結合型の分子に対しても計算が可能で、相関分析や主成分分析(PCA)、t-SNE 分析といった簡易的な分析機能なども備えています。また、バージョン 2.0 から 3.0 へのメジャーアップデートでは 3 次元座標の計算が可能となったことで 2 次元の構造ファイルからでも全ての記述子を計算できるようになりました。また、ユーザー定義による構造パターンの計算機能も追加されました。

alvaDesc のシンプルで直感的なインターフェースを使用して、これらの計算や解析を簡単に行うことができます。

このスタートアップガイドでは、初心者の方のスムーズな導入を目的として、GUI での操作を中心に、alvaDesc 3.0 の基本操作や機能についてチュートリアル形式で紹介する内容となっています。応用的な内容は省略していますので、より深く知りたいという方や CLI での操作方法を知りたい方は、alvaDesc のユーザーマニュアルをご参照ください。

このガイドでは、主に、分子記述子、フィンガープリント、構造パターンの計算とその解析・グラフ表示の操作方法を紹介していきます。末尾には参考情報も記載していますので、適宜ご参照ください。なお、本文中では分子記述子を簡単に『記述子』と表記させていただきます。

このスタートアップガイドを、皆様の業務や研究にお役立ていただければ幸いです。

※本資料は alvaDesc バージョン 3.0.6 を使用して作成しています。

2. 記述子・フィンガープリント計算の基本操作

この章では、分子構造ファイルの準備から、分子記述子やフィンガープリントの計算方法について説明します。

2.1 分子構造ファイルの入手・準備

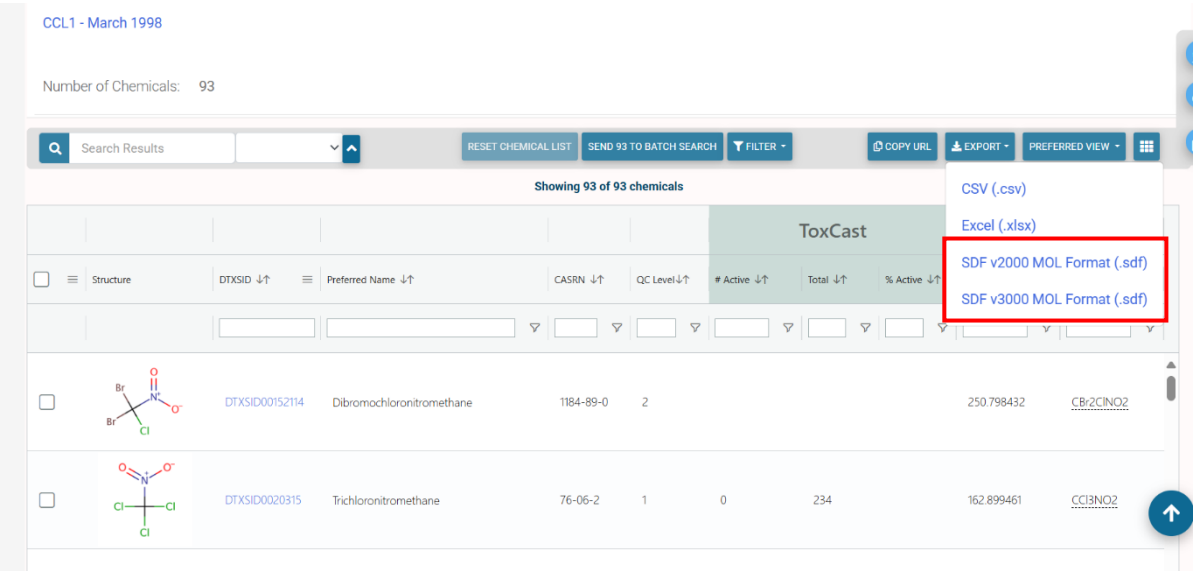
はじめに、計算に使用する分子構造ファイルを準備します。alvaDesc で分子構造として入力可能なファイルフォーマットは以下になります。

ファイル形式	拡張子
Sybyl ファイル	.ml2, .mol, .sm2, .mol2, .sy2
MDL ファイル	.mol, .mdl, .sdf, .sd
HyperChem ファイル	.hin
SMILES ファイル	.smi, .smiles
CML ファイル	.cml
Macromodel ファイル	.mmd, .mmod
SMILES が含まれたテキストファイル	.txt, .csv, .tsv

本資料では例として、アメリカ環境保護庁(EPA)のホームページ内で公開されている飲料水汚染物質候補リスト(CCL: Contaminant Candidate List)をダウンロードしてこの後の計算や解析に使用します。

下記の ULR にアクセスし、リスト右上の「EXPORT」ボタンからファイルフォーマットを選択します。CSV 形式も利用できますが、今回は SDF 形式を選択してファイルをダウンロードしてください。（v2000、v3000 形式のどちらでも構いませんが、このスタートアップガイドでは v3000 形式をダウンロードして使用します。）

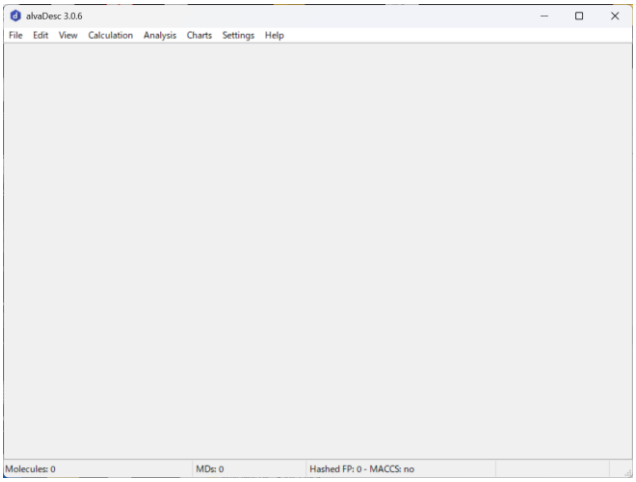
https://comptox.epa.gov/dashboard/chemical-lists/CCL



今回ダウンロードした SDF 形式ファイルには構造以外に含まれているプロパティ情報（以降、外部変数と称します）が含まれています。alvaDesc ではこのような外部変数を読み込み、解析に利用することもできます（【2.7 外部変数の読み込み】を参照）。メモ帳などのテキストエディタでファイルを開くと、外部変数の内容を確認できます。

2.2 alvaDesc（GUI）の起動

計算対象の分子構造ファイルの準備ができましたら、alvaDesc の GUI を起動し、計算を進めます。Windows の場合、「スタート」をクリックし、画面右上にある「すべて」をクリックします。一覧の中から、「Alvascience」、「alvaDesc」と順次選択し、ソフトウェアを起動します。



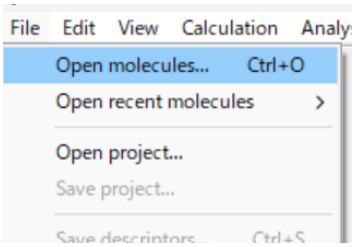
alvaDesc 起動時の画面

【Note】Linux/macOS 環境での alvaDesc GUI 起動方法
Linux 環境の場合、ターミナルから「alvaDescGUI」コマンドを入力・実行することで、GUI を起動することができます。通常、「/usr/bin/alvaDescGUI」としてインストールされますが、変更時は、ディレクトリを適宜読み替えてください。
macOS 環境では、アプリケーションフォルダ内の「alvaDesc」をダブルクリックすることにより、起動できます。

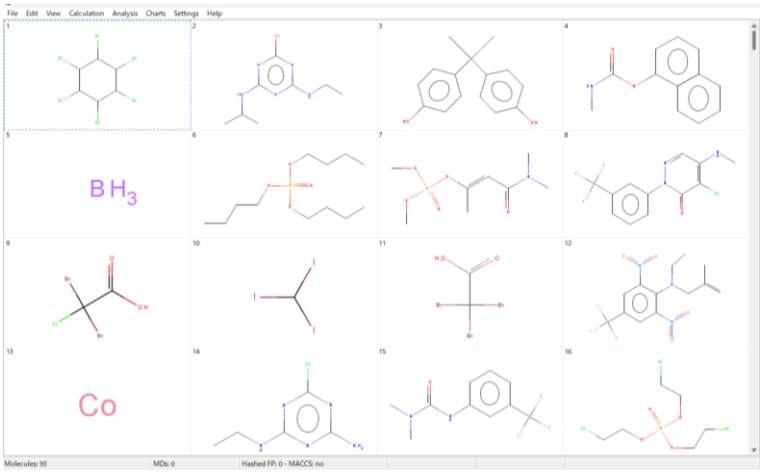
2.3 分子の読み込み・表示

2.3.1 分子構造ファイルのロード

(1) [File > Open molecules…] を選択し、計算対象の分子構造ファイルを開きます。ここでは例として、先ほど用意した SDF 形式の構造ファイルを選択して読み込みます。



(2) 読み込みが完了すると、画面に分子が表示されます（Molecule grid 表示）。



【Note】File メニューについて

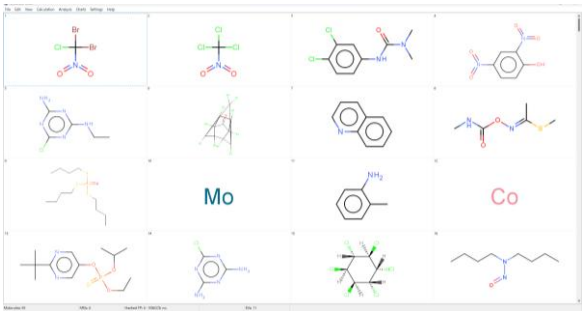
- [Open recent molecules] を選択すると、最近開いたファイルの履歴から選択できます。
- [Open project…] からはプロジェクト全体を保存したファイル（.adprj）を読み込むことができます。
- 分子構造を既に読み込んでいる状態で、さらに構造ファイルを読み込もうとすると、追加で読み込むか、新規に読み込むかを選択するダイアログが表示されます。
- 読み込むファイル形式によってはファイル選択後に外部変数の読み込みダイアログが表示されます。
- GUI 画面にファイルをドラッグアンドドロップして開くこともできます。

2.3.2 分子の表示

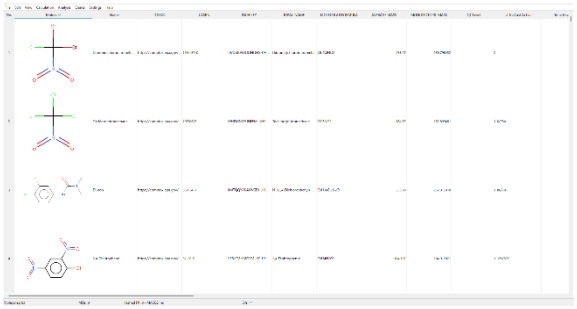
View メニューから分子の表示形式を変更できます。ここでは一部のメニューを紹介しますが、水素表示の設定やハイライト表示などを各種設定できます。View メニューから実行できる表示機能の一覧は、【3.3.2 分子の表示機能一覧（View メニュー）】にまとめていますのでそちらをご参照ください。

❑ View > Grid

分子をグリッド表示・ワークシート表示に切り替えます。Ctrl+D（macOS では command+D）キーで素早く切り替えることができます。



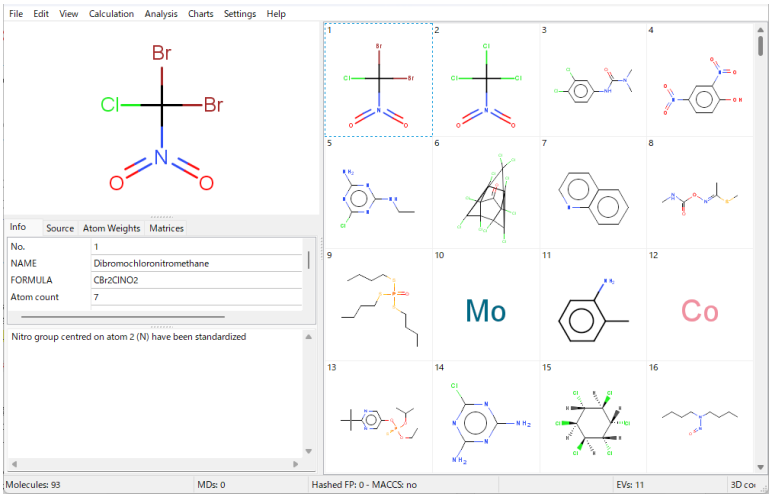
[Molecule grid] グリッド表示



[Molecule worksheet] ワークシート表示

❑ View > Molecule detail

画面左側に分子の詳細情報のパネルが表示されます。画面右側のグリッド／ワークシートで選択した分子の情報（Info/Source/Atom Weights/Matrices）を確認することができます。



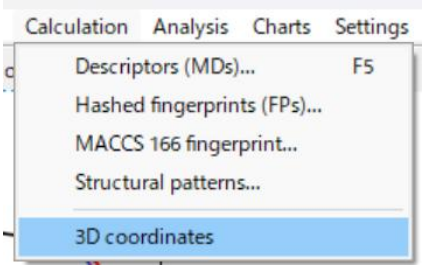
2.4 3D 座標の計算

alvaDesc バージョン 3.0 以降では、必要に応じて 3D 座標を計算することができます。これにより、読み込んだ分子が 3 次元座標を持たない場合でも、計算によって生成した座標を利用することで 3 次元分子記述子の計算が可能になります。なお、alvaDesc では分子ごとに 1 種類の立体構造を生成することができます。入力ファイルで絶対立体配置が指定されている場合は、その情報が 3D 座標計算に反映されます（未指定の場合でも、単一構造のみ生成されます）。

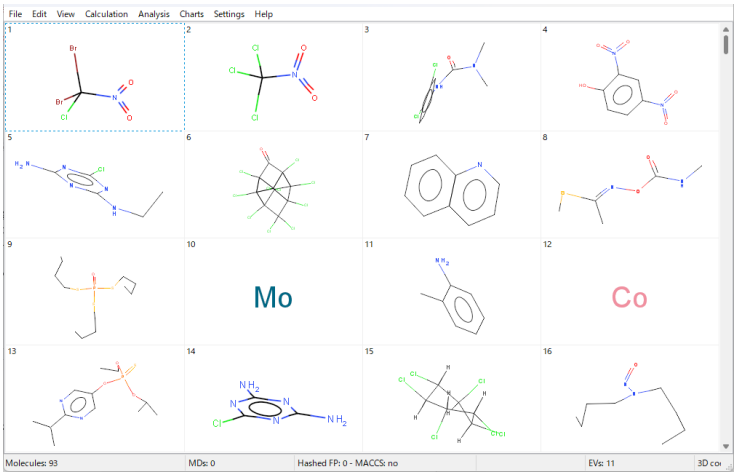
3D 座標計算のアルゴリズムの詳細については、alvaDesc ユーザーマニュアルの【4.5 3D coordinates】の項目をご参照ください。

前のステップで読み込んだ分子構造ファイルは、2 次元座標形式の SDF ファイルですので、次の操作により、3 次元座標を計算します（3D 記述子を計算しない場合、この操作はスキップしてかまいません）。

❑ [Calculation > 3D coordinates] をクリック



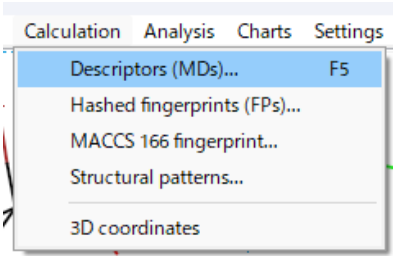
3D 座標の計算が行われます。計算後は、画面の分子構造の表示が 3D 表示になります。
（メニューの View > Coordinates から分子構造の表示座標形式（Use coordinates from file／2D／3D）を切り替えることができます）



2.5 分子記述子の計算

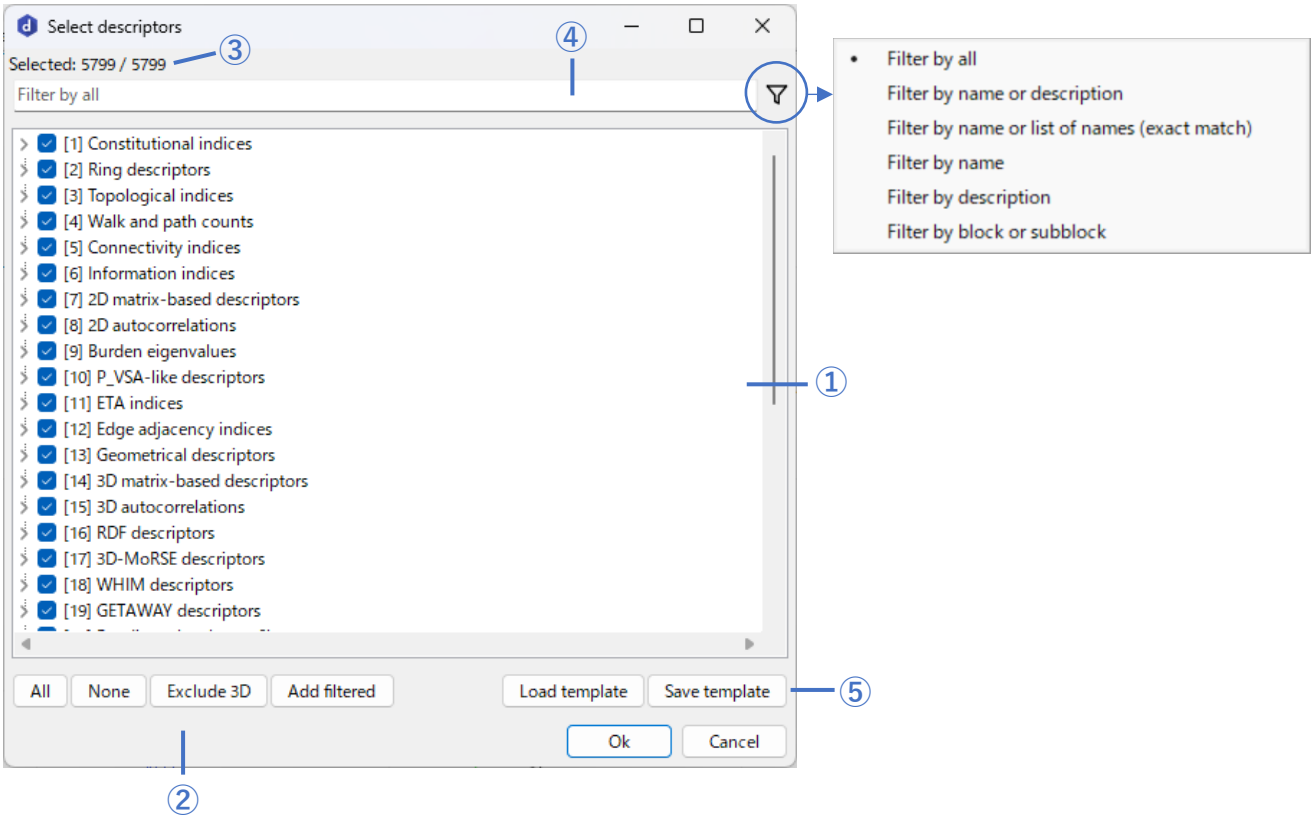
記述子の計算を行います。alvaDesc で計算可能な記述子の一覧は、[View > Descriptor list] から確認することができます。記述子の種類や詳細については参考情報の【5.2 分子記述子について】、または alvaDesc ユーザーマニュアルの【5. Descriptors】 をご参照ください。

(1) [Calculation > Descriptors (MDs)…] を選択



(2) 記述子選択ダイアログ（下記参照）で計算したい記述子を選択（今回は左下の All を選択）し、[OK]

＜記述子選択ダイアログ＞



① **記述子の選択**

初期表示では、記述子の各ブロックがリスト表示されています。ブロック名の左端の矢印のようなマークをクリックすると、ブロックに含まれている記述子が展開表示されます。チェックボックスをクリックすることで、選択/非選択を変更できます。ブロック（サブブロック）の左にあるチェックボックスをクリックすると、そこに含まれる全ての記述子を一括で選択/選択解除できます。ブロックの中の一部の記述子が選択されている状態では、チェックボックスが「-」のように表示されます。

② **選択オプション**

[All]　　：全て選択
[None]　　：全て非選択
[Exclude 3D]　：3D 記述子を除く
[Add filtered]　：絞り込み後の記述子を全て選択

③ **現在選択している記述子の数**（分母は計算可能な全記述子数）

④ **記述子リストのフィルタリング**

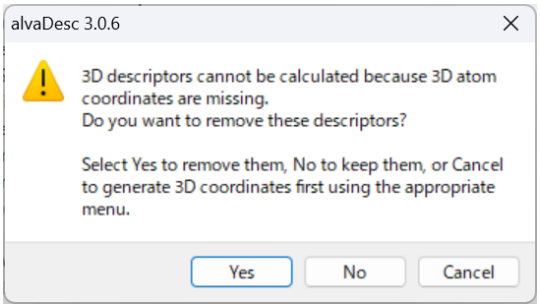
入力欄に記述子名などを入力すると、該当する記述子のみリストに表示されます。右のロートのアイコンをクリックすると、検索条件（記述子名から検索、記述子の説明を含めて検索、など）を設定できます。
絞り込み後の記述子を全て選択したい場合は、左下の [None] をクリックして選択をクリアした後、[Add filtered] をクリックします。既に選択済みの記述子があり、絞り込み後の記述子を追加したい場合は単に [Add filterd] をクリックします。入力欄をクリアすると、記述子の一覧画面に戻ることができ、選択した記述子を確認できます。

⑤ **記述子テンプレート**

[Save template]：選択した記述子のリストをテンプレートとして保存します（.adtpl 形式ファイル）。
[Load template]：保存した記述子リストのテンプレートを読み込みます。

【Note】3 次元座標を持たない SMILES で分子構造を入力した場合の Warning

3 次元座標を持たない SMILES で分子構造を読み込んでおり、3D 座標計算を行っていない場合、以下のような警告画面が表示されます。計算できない 3 次元記述子を除外するかどうかを尋ねられるので、除外する場合は、[Yes] を選択してください。なお、[No] を選択すると 3 次元記述子の値は欠測値（デフォルトでは"na"）として出力されます。



2.6 フィンガープリントの計算

alvaDesc では、以下のフィンガープリントの計算を行うことができます。

・Extended Connectivity Fingerprint（ECFP/ ECFPV3）

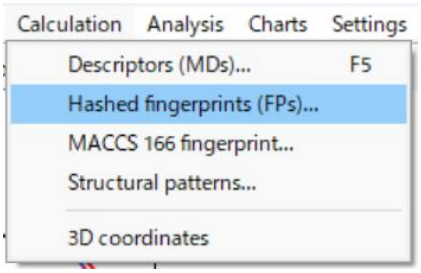
・Path Fingerprint（PFP）

・MACCS 166 Fingerprint

フィンガープリントの詳細については参考情報の【5.3 フィンガープリントについて】 、または alvaDesc ユーザーマニュアルの【6. Fingerprints】をご参照ください。

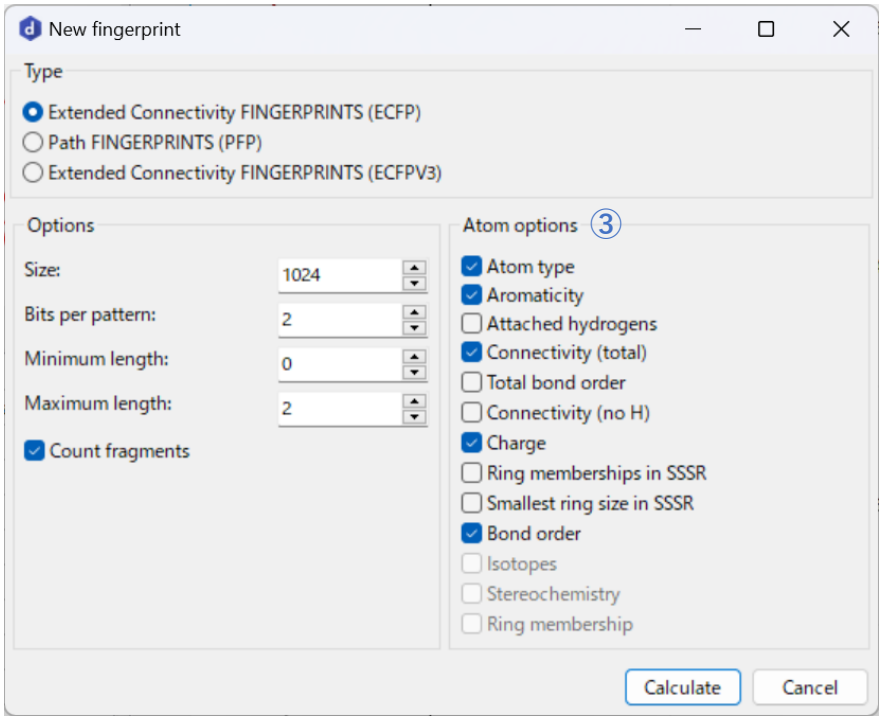
2.6.1 ハッシュ化フィンガープリント ECFP / ECFPV3 / PFP の計算

(1) [Calculation > Hashed fingerprints(FPs)…] を選択



(2) 計算オプションを設定し、[Calculate] をクリック（今回はデフォルト設定のまま、ECFP を実行）
計算時の各設定については alvaDesc ユーザーマニュアルの [6.2 Hashed fingerprints] に記載されているので、詳しく知りたい方はご参照ください。

＜計算オプションダイアログ＞



① **Type**
計算するフィンガープリントを指定

② **Options**

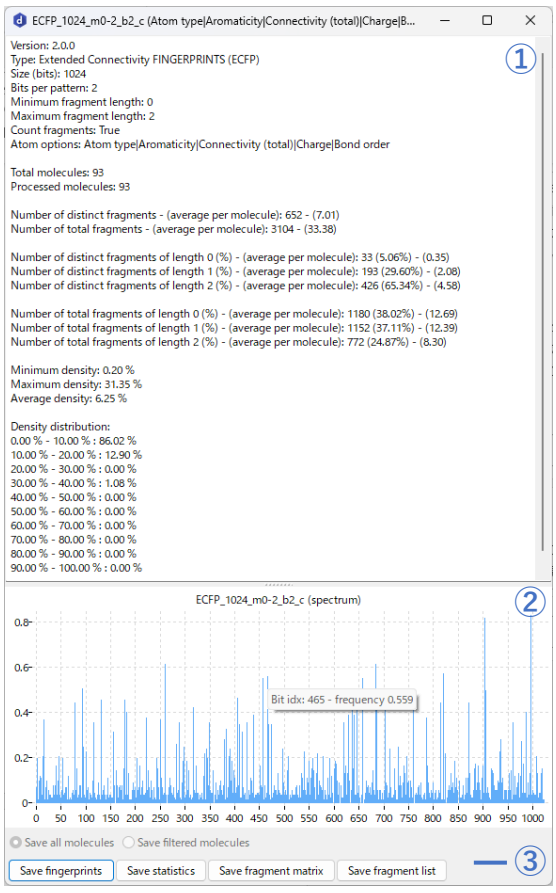
Size　：ブールベクトル（true/false のデータ列）の長さ
Bits per pattern　：サブ構造をエンコードする際に使用されるビットの数
Minimum length　：検出されるフラグメントの最小サイズ
Maximum length　：検出されるフラグメントの最大サイズ
Count fragments　：有効→分子サブ構造の出現回数を考慮する / 無効→分子サブ構造の存在の有無のみ出力

③ **Atom options**

Atom type　：原子の番号で区別するか
Aromaticity　：原子の芳香族性を区別するか
Attached hydrogens　：結合している水素の数で区別するか
Connectivity (total)　：結合している原子の総数で区別するか
Total bond order　：原子の総結合次数の違いを考慮するか
Connectivity (no H)　：結合している水素以外の原子数で区別するか
Charge　：電荷の違いを考慮するか
Ring memberships in SSSR　：原子がいくつの環に所属しているかを区別するか　※ECFPV3 では設定不可
Smallest ring size in SSSR　：原子が所属する環のサイズを考慮するか
Bond order　：結合次数を区別するか
Isotopes　：同位体を区別するか　※ECFPV3 のみ設定可
Stereochemistry　：立体化学を考慮するか　※ECFPV3 のみ設定可
Ring membership　：環に属する原子を区別するか　※ECFPV3 のみ設定可

(3) 計算結果が表示されるので、結果を確認する
結果画面のウィンドウは、×で閉じてしまっても[File > Save fingerprints > (フィンガープリント名)] から再度表示することができます。

＜結果画面＞



① **計算結果の統計情報**

計算実行時に設定したパラメーター情報や、認識されたフラグメントの数、ビット密度などの統計情報が出力されます。

② **フィンガープリントスペクトラム**

各ビット列の平均値がグラフに表示されます。チャートを右クリックすると画像のコピーや保存ができます。
グラフの棒にマウスオーバーするとビットの番号と値が表示されます。

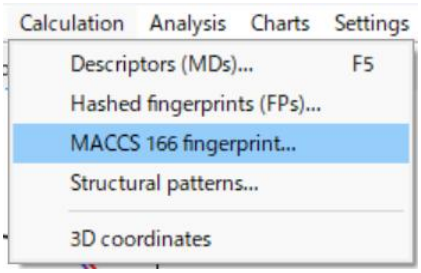
③ **各種保存ボタン**

[Save fingerprints]　：分子毎のフィンガープリントを保存（タブ区切り.txt）
[Save statistics]　：①の計算結果の統計情報を保存（.txt）
[Save fragment matrix]：分子ごとの SMARTS 表記されたフラグメントタイプ別ヒット数を保存（タブ区切り.txt）
[Save fragment list]　：SMARTS 表記されたフラグメントタイプごとの統計情報を保存（タブ区切り.txt）
※ECFPV3 は高速計算のため SMARTS 表記のフラグメントを使用していませんので、fragment matrix と fragment list の出力はありません。
※SMARTS 表記については、参考情報の【5.6 SMARTS 記法について】をご参照ください。

【Note】ECFP 計算の出力結果について
計算結果画面の上半分には、実行時のパラメーター情報や統計情報が表示されます。この中には、ビット密度（Density：化合物毎のフィンガープリントで「1」となっているビットの割合）の情報も出力されます。ビット密度が高い場合、フィンガープリント内の「1」のビットが多くなり、類似性評価において分子同士の違いが見えにくくなる可能性があります。特に、ハッシュ化フィンガープリントでは、異なるフラグメントが同じビットに割り当てられるビット衝突が発生することがあります。これは、フィンガープリントのサイズが小さい場合や、パターンごとのビット数（Bits per pattern）が少ない場合に起こりやすくなります。ビット密度は、フィンガープリントのサイズ、パターンごとのビット数（Bits per pattern）、最大フラグメント長（Maximum length）などのパラメーターを適切に調整することで、ビット衝突を抑えつつ、分子の特徴を効果的に表現できるフィンガープリントを作成できます。また、出力結果にはフラグメント数の情報なども記載されています。この結果からは、データセットの化学空間の多様性や分子構造の複雑さなどの情報を得ることができます。
結果画面の下半分には、フィンガープリントの各ビットの平均値（spectrum）がグラフ表示されます。このグラフでは、ビットの分布を視覚的に確認したり、データセット全体の特徴を分析したりすることができます。

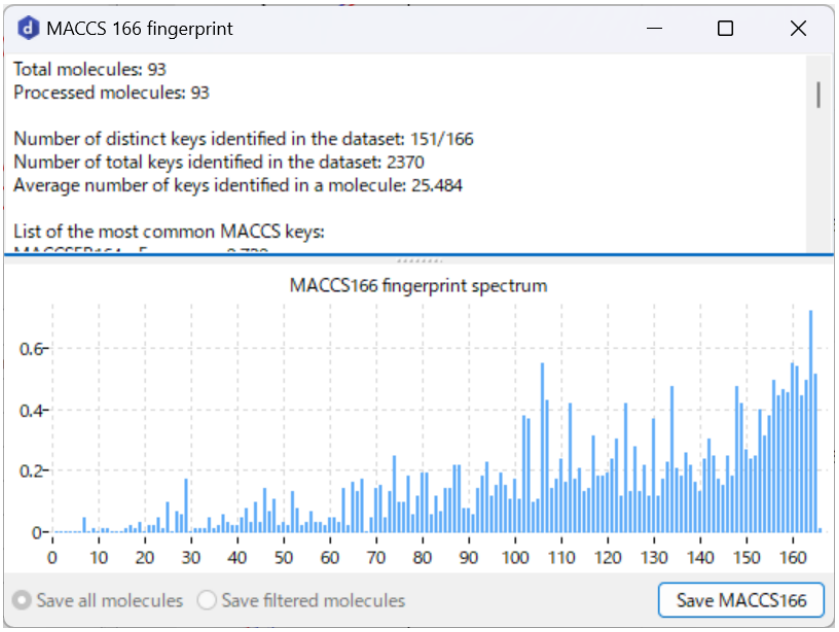
2.6.2 MACCS 166 フィンガープリントの計算

(1)　[Calculation > MACCS 166 finger print…] を選択



(2)　計算結果が表示されるので、結果を確認する

＜結果画面＞



結果画面の構成は前述（2.6.1 ハッシュ化フィンガープリント ECFP / ECFPV3 / PFP の計算）と同様です。
[Save MACCS166] をクリックすると分子毎の 166 ビットのフィンガープリントが保存できます（タブ区切り.txt）。

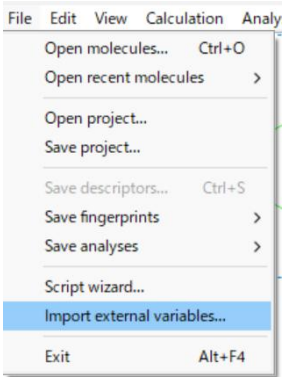
2.7 外部変数の読み込み

alvaDesc では外部変数（構造以外に含まれているプロパティ情報）を読み込むこともできます。この操作は任意で、対応しているファイル形式は、SMILES ファイル（.smi .smiles）、SDF ファイル（.mol .mdl . sdf .sd）、テキストファイル（.txt .csv .tsv）です。外部変数は、読み込んだ分子の順序と一致するように並んでいる必要があり、分子と同じ数のレコードが必要です。読み込んだ外部変数は、36 番目の記述子ブロックに格納され、計算された記述子と同じように、分析やグラフ表示、出力などを行うことができます。

先ほど読み込んだ SDF 形式の構造ファイルには外部変数が含まれているので、読み込んでみます。

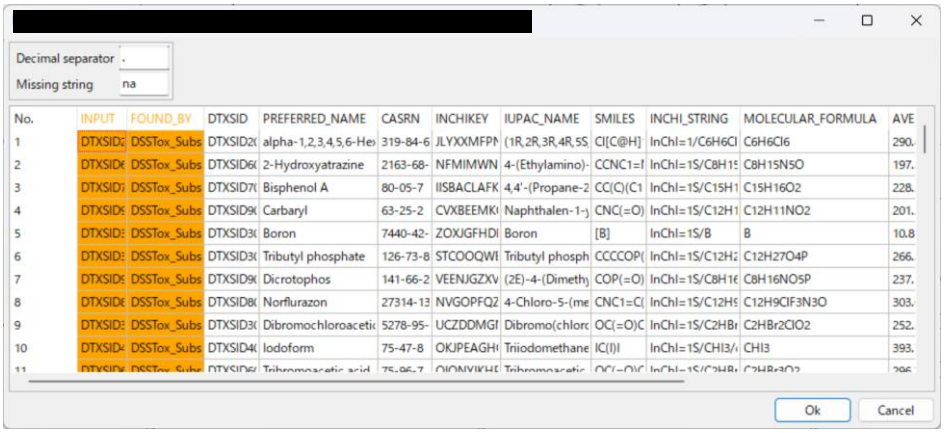
(1) [File > Import external variables…] を選択

外部変数を読み込みます。メニューを選択し、先ほど読み込んだ SDF ファイル（2.1 分子構造ファイルの入手・準備でダウンロードした SDF ファイル）を指定します。ファイルの選択画面では、拡張子がデフォルトでテキストファイル「All supported text file」になっているので、「MDL file format」や「All files」に変更してファイルを表示・選択してください。

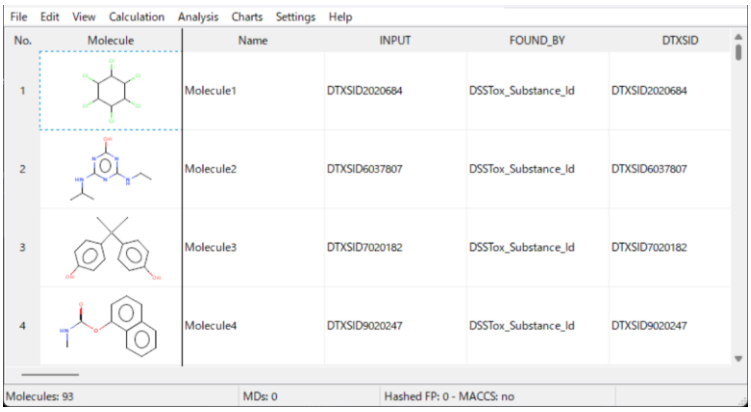


(2)　取り込み確認画面が表示されるので、内容を確認して[OK]
今回は特に設定せずにそのまま読み込みます。

除外したい外部変数がある場合は、下図のように読み込まないカラムのヘッダーをクリックしてオレンジ色の表示にしてください。



(3)　読み込みが完了すると画面に外部変数が追加表示されます（グリッド表示からワークシート表示に自動的に切替わります）。



【Note】分子構造ファイル形式と外部変数の読み込み

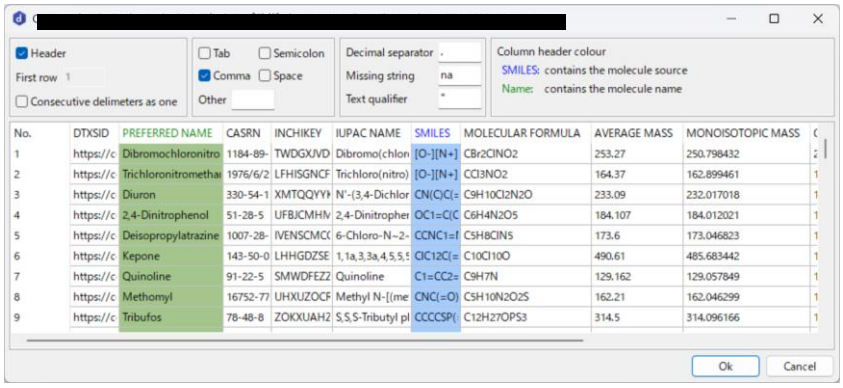
外部変数は [File > Import external variables…] メニューから読み込むことができますが、テキストファイル（.txt .csv .tsv）の場合は、 [File > Open molecules…] から分子構造ファイル読み込む際に外部変数を読み込むかどうかを尋ねるダイアログが表示されます。この時 [Yes] を選択すると、SMILES と NAME 以外の列が外部変数として自動的に読み込まれます。

例１）SMILES を含むテキストファイル（.txt .csv .tsv）

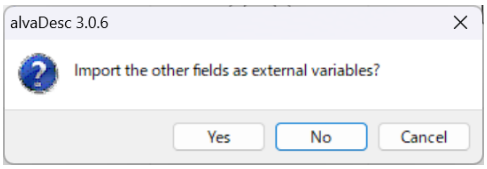
[File > Open molecules…] から SMILES を含むテキストファイルを選択すると、以下のような取り込み確認画面が表示されます。

NAME 列（緑）と SMILES 列（青）は自動認識されますが、カラムのヘッダーをクリックすると色が切り替わり、自分で指定することもできます。ヘッダーを含まないファイルを読み込みたい場合は、左上の「Header」のチェックを外してください。この画面からは、その他、読み込みの各種設定を行うことができます。

（[2.1 分子構造ファイルの入手・準備] で紹介した EPA のページからダウンロードした CSV 形式ファイルを読み込む場合、[INCHI STRING] 列に改行が含まれていることが原因で正しく読み込まれません。[INCHI STRING] 列を削除してしまうか、改行を除くなどの処理をしてからファイルの読み込みを行ってください。）

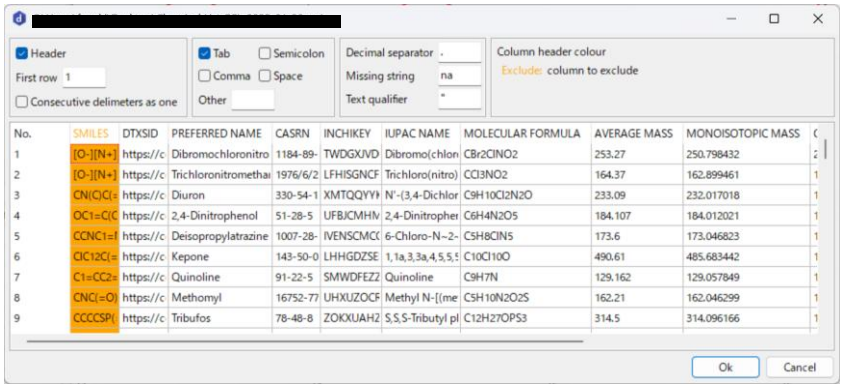


外部変数が含まれている場合は、[OK] をクリックして次へ進むと、外部変数を読み込むか尋ねるダイアログが表示されます。[Yes] を選択すると、全ての外部変数が自動で読み込まれます。



例２）外部変数を含む SMILES ファイル（.smi .smiles）

SMILES ファイルに含まれる外部変数を読み込むこともできます。拡張子がテキストファイル形式（.txt .csv .tsv）の場合は、例１と同様に分子の読み込み時に外部変数を読み込むことができますが、SMILES ファイル（.smi .smiles）形式の場合は、外部変数の読み込みダイアログは表示されないため、 [File > Import external variables…] メニューから別途読み込む必要があります。この場合、SDF ファイルの場合と同様に、次のような取り込み確認画面が表示されます。この画面では、ヘッダーをクリック（オレンジ色で表示）することで、読み込まない外部変数を指定することもできます。



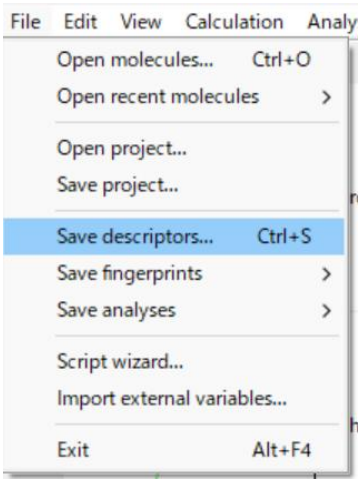
2.8 計算結果の保存

結果を保存する際の出力ファイルのフォーマットは以下になります。

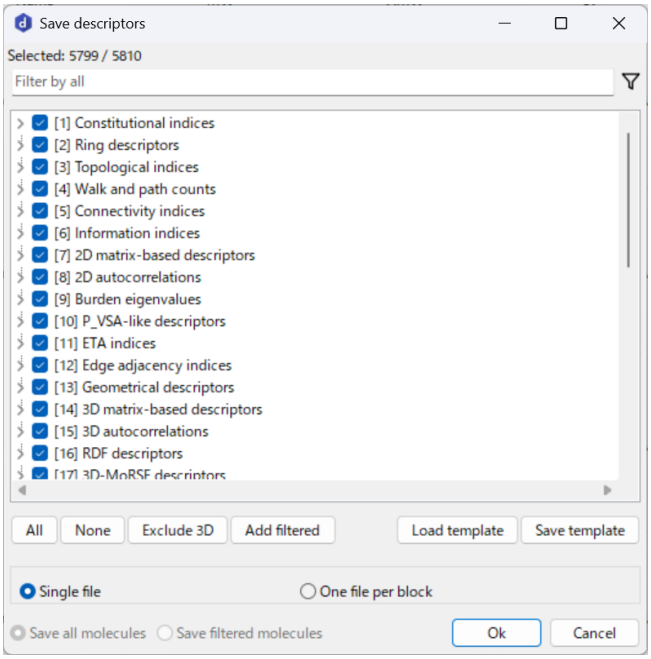
- ❑ 分子記述子計算結果
  - ▪ タブ区切りテキストファイル（.txt）（分子の名前と記述子）
  - ▪ ヘッダー付き SMILES ファイル（.txt）（各分子の SMILES を含む）
  - ▪ MDL ファイル(.sdf)（2D/3D）（V2000 形式）
  - ▪ エクセルワークブック(.xlsx)（分子の 2D 画像、SMILES を含む）
- ❑ フィンガープリント計算結果
  - • タブ区切りテキストファイル（.txt）
- ❑ プロジェクト全体
  - • 独自フォーマットファイル（.adprj）

2.8.1 記述子計算結果の保存

(1) [File > Save descriptors…] を選択



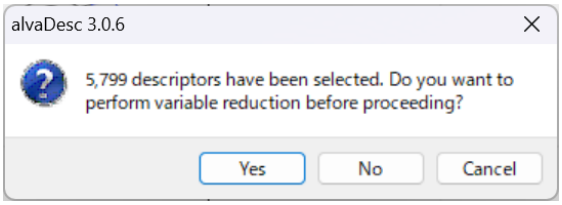
(2) 選択ダイアログで保存したい記述子を選択し、[OK]



【Note】記述子計算結果の選択ダイアログについて

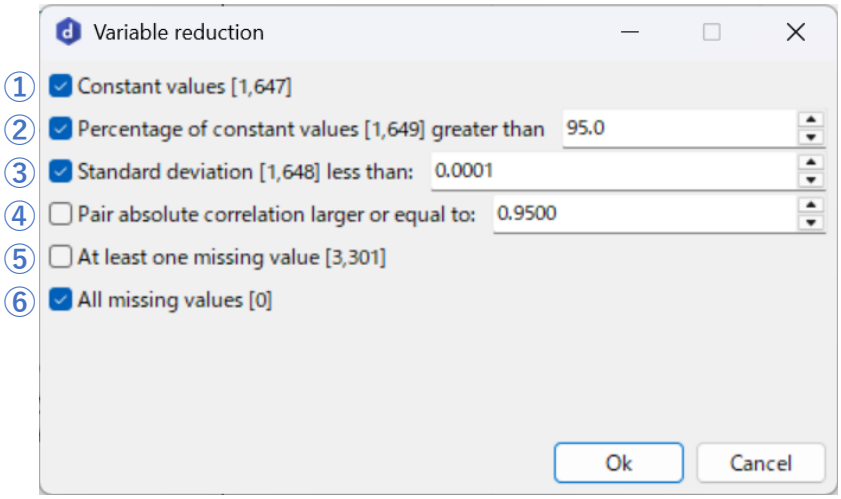
- 選択ダイアログの操作は【2.5 分子記述子の計算】を参照してください。
- 画面下部からは[Single file(単一のファイルとして保存)]または[One file per block(記述子ブロックごとにファイルを保存)]を選択できます。
- フィルタリング（【3.3.1 フィルタリングと並び替え】参照）を行った場合は、一番下の [Save all molecules]/[Save filtered molecules] が有効になり、後者をクリックして選択するとフィルタリング後の分子のみを保存することができます。
- 構造パターンを計算した場合は[35]、外部変数を読み込んだ場合は[36]の記述子ブロックに格納され、他の記述子と同様に選択することができます。

(3) 保存前に変数削減を実行するか尋ねられるので、実行する場合は [Yes] 、しない場合は [No] を選択
今回は [Yes] を選択して変数削減を実行してみます。変数削減の詳細については、【5.4 変数削減について】をご参照ください。



(4) (3)で [Yes] を選択した場合は、変数削減メニューが表示されるので、条件を設定して [OK] を選択
今回は例として、以下のスクリーンショットの設定で変数削減を実行してみます。

＜変数削減メニュー＞



① 全てが同じ値の記述子を削除
② 指定した割合以上が同じ値の記述子を削除
③ 標準偏差が閾値より小さい（値のばらつきがとても少ない）記述子を削除
④ ペア相関が閾値以上である記述子（の一方）を削除（※ 他の記述子との相関がより大きい方を削除）
⑤ 欠測値（計算結果が na）が 1 つでもある記述子を削除
⑥ 全てが欠測値（計算結果が na）である記述子を削除

【Tips】欠測値の出力は、Settings メニュー > Options…の Output タブ、Missing Value から別の数値・文字列へ変更できます（デフォルト：na）。

(5)　ファイル形式を選択して保存

以下のファイル形式を選択できます。

- Tabbed text file (*.txt)　　：分子名と記述子をタブ区切りテキストファイルとして保存
- SMILES file with header (*.txt)：各分子の SMILES を含むヘッダー付きのテキストファイルとして保存
- MDL file format (*.sdf)　　：2D または 3D の MDL ファイルとして保存
- Excel Workbook (*.xlsx)　　：分子の画像(含めるか選択可)と SMILES を含むエクセルワークブックとして保存

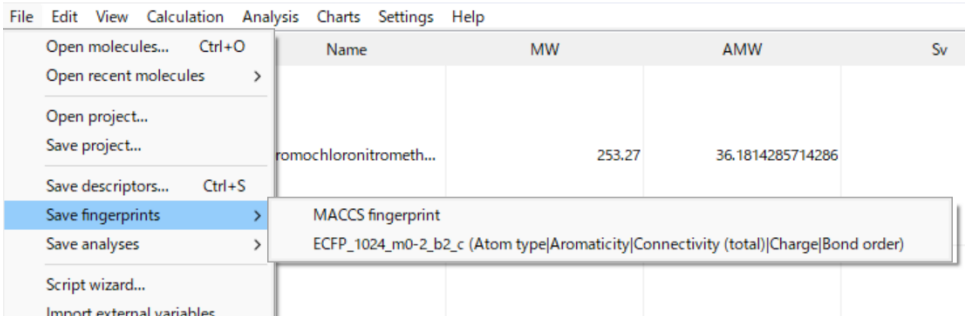
【Tips】変数削減済みの記述子テンプレートを保存する

変数削減を実行して記述子を保存した後、再度 [File > Save descriptors…] を選択して記述子選択画面を開くと、変数削減後の記述子が選択された状態になっているので、右下の [Save Template] ボタンからテンプレートとして保存しておくことで、変数削減済みの記述子セットを再度利用することができます。

2.8.2 フィンガープリント計算結果の保存

Save メニューからフィンガープリント計算結果画面を再表示して保存します。

- File > Save fingerprints > 保存したいフィンガープリントを選択



フィンガープリントの計算結果画面が表示されるので、下部にある保存ボタンから結果を保存できます。

詳細は、【2.6 フィンガープリントの計算】の項目をご参照ください。

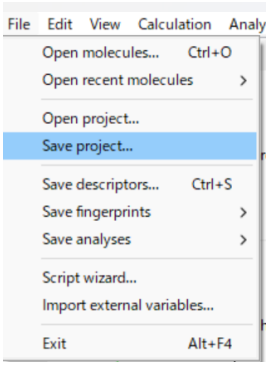
【Note】フィルタリング後の分子のみを保存する

フィルタリング（「3.3.1 フィルタリングと並び替え」参照）を行った場合は、計算結果画面一番下の [Save all molecules]/ [Save filtered molecules] が有効になり、後者をクリックして選択するとフィルタリング後の分子のみを保存することができます。

2.8.3 プロジェクト全体の保存

入力データ、記述子等の計算結果、また、分析結果等の全体をプロジェクトとして保存できます。ファイルは alvaDesc 独自の「alvaDesc project file（*.adprj ）」形式です。alvaModel をご利用の方は、このファイルをインポートし、QSAR/QSPR モデル構築に使用することができます。

(1)　[File > Save project...] を選択



(2)　ファイル名と保存先を指定して保存

3. 分析機能と可視化／構造パターン計算

この章では、記述子の計算結果の可視化や分析、構造パターン計算の手順を紹介します。

3.1 簡易分析

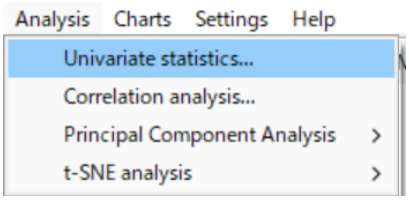
他のソフトウェアによる解析前の予備的な分析が可能です。[Analysis] メニューから、単変量解析、相関分析、主成分分析、t-SNE 分析を実行できます。簡易分析では、構造パターン（[3.4 構造パターンの計算]）や外部変数も解析対象に含めることができます。

> 【Note】分析前の変数削減
> 相関分析・主成分分析・t-SNE 分析では、分析実行前に自動的に変数削減が行われます。
> ここでは「全て定数／１つを除いて他が定数／標準偏差が 0.0001 未満」ではない記述子が分析対象となります。なお、標準偏差の閾値はオプション設定にて変更することもできます。（【5.4 変数削減について】を参照）

3.1.1 単変量解析

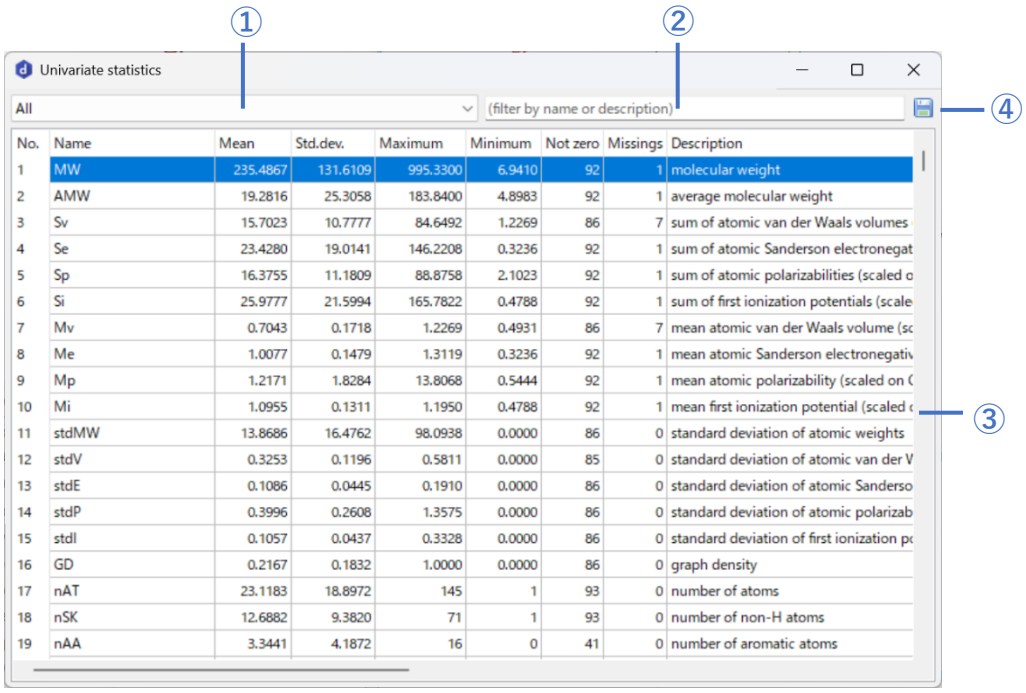
各記述子の平均値や標準偏差などの基本的な統計データを表示します。記述子と構造パターン（【3.4 構造パターンの計算】を参照） が解析対象です。

(1) [Analysis > Univariate statistics…] を選択



(2) 結果の確認

① 解析を実行すると、以下のような結果画面が表示されます。表示する記述子ブロックの選択



No.	Name	Mean	Std.dev.	Maximum	Minimum	Not zero	Missings	Description
1	MW	235.4867	131.6109	995.3300	6.9410	92	1	molecular weight
2	AMW	19.2816	25.3058	183.8400	4.8983	92	1	average molecular weight
3	Sv	15.7023	10.7777	84.6492	1.2269	86	7	sum of atomic van der Waals volumes
4	Se	23.4280	19.0141	146.2208	0.3236	92	1	sum of atomic Sanderson electronegat
5	Sp	16.3755	11.1809	88.8758	2.1023	92	1	sum of atomic polarizabilities (scaled o
6	Si	25.9777	21.5994	165.7822	0.4788	92	1	sum of first ionization potentials (scale
7	Mv	0.7043	0.1718	1.2269	0.4931	86	7	mean atomic van der Waals volume (sc
8	Me	1.0077	0.1479	1.3119	0.3236	92	1	mean atomic Sanderson electronegativ
9	Mp	1.2171	1.8284	13.8068	0.5444	92	1	mean atomic polarizability (scaled on C
10	Mi	1.0955	0.1311	1.1950	0.4788	92	1	mean first ionization potential (scaled c
11	stdMW	13.8686	16.4762	98.0938	0.0000	86	0	standard deviation of atomic weights
12	stdV	0.3253	0.1196	0.5811	0.0000	85	0	standard deviation of atomic van der W
13	stdE	0.1086	0.0445	0.1910	0.0000	86	0	standard deviation of atomic Sanderso
14	stdP	0.3996	0.2608	1.3575	0.0000	86	0	standard deviation of atomic polarizab
15	stdI	0.1057	0.0437	0.3328	0.0000	86	0	standard deviation of first ionization pc
16	GD	0.2167	0.1832	1.0000	0.0000	86	0	graph density
17	nAT	23.1183	18.8972	145	1	93	0	number of atoms
18	nSK	12.6882	9.3820	71	1	93	0	number of non-H atoms
19	nAA	3.3441	4.1872	16	0	41	0	number of aromatic atoms

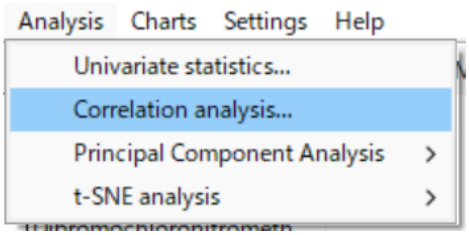
② 記述子名による絞り込み
③ 結果の一覧
Mean ：平均値
Std.dev. ：標準偏差
Maximum ：最大値
Minimum ：最小値
Not Zero ：0 以外の値をもつ分子の数
Missings ：欠損数（計算できなかった分子数）
Description：記述子の説明
Block ：記述子ブロック名
Sub-Block ：記述子サブブロック名
④ 保存アイコン

(3) 結果の保存：結果画面④の保存アイコンをクリック（タブ区切りテキストファイルとして保存可能）

3.1.2 相関分析

n 個の記述子間の相関マップを表示します。外部変数も分析対象に指定できます。

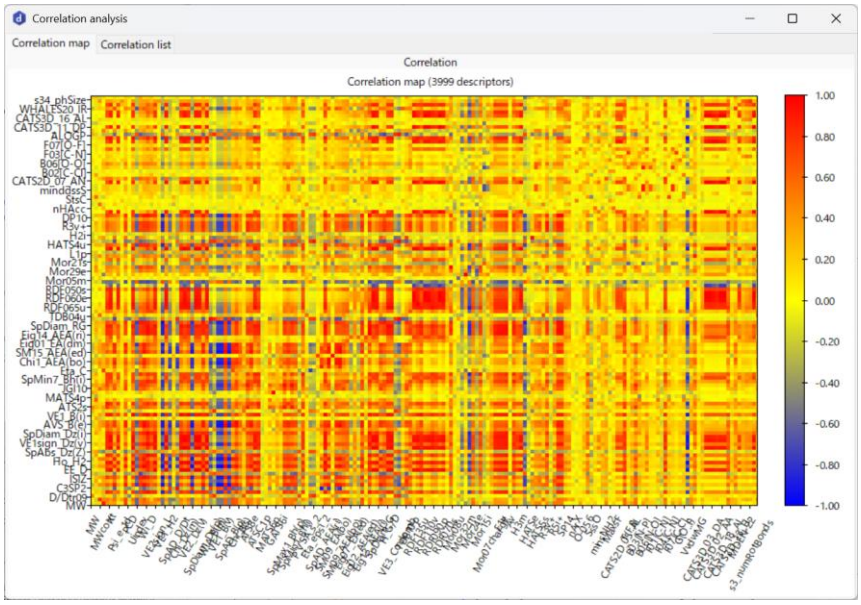
(1) [Analysis > Correlation analysis...] を選択



(2) 選択ダイアログで分析したい記述子を選択し、[OK]（ここでは例として All を選択）

(3) 結果の確認
解析を実行すると、結果画面が表示されます。左上のタブ [Correlation map] [Correlation list] から結果画面を切り替えられます。

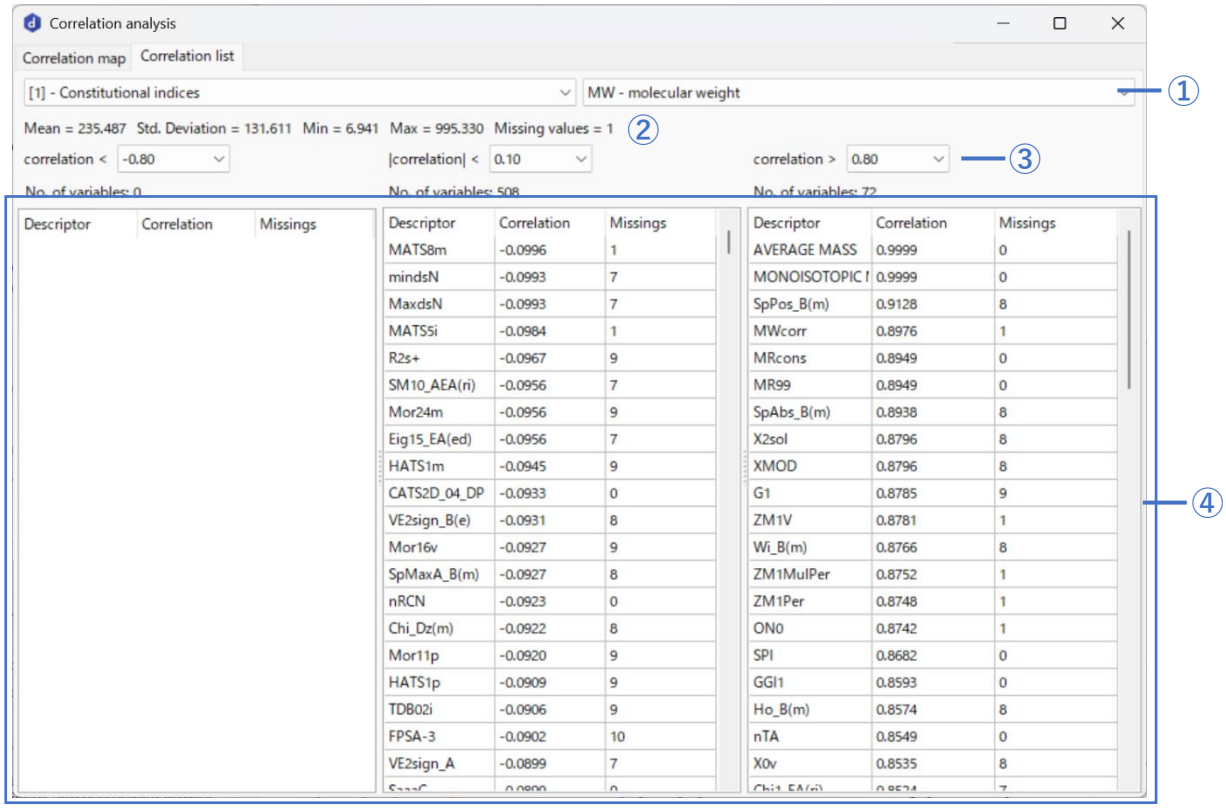
- 【Correlation map】タブ
  n × n の記述子毎の相関がマップに表示されます。マップにマウスオーバーすると、画面上部に記述子名と相関係数が表示されます。



赤色：正の相関（ある記述子の値が高いほど、もう一方の記述子の値も高くなる傾向）
黄色：相関なし
青色：負の相関（ある記述子の値が低くなる傾向）

- 【Correlation list】タブ

ある記述子との相関をリスト表示します。

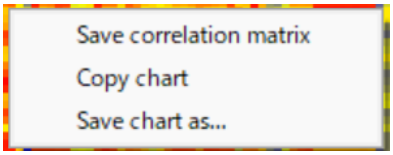


① 相関を見たい記述子のブロック、記述子名を選択
② 選択した記述子の単変量解析の値
③ 相関係数の閾値の設定（左から、負の相関、無相関、正の相関）
④ 相関別の記述子のリスト（左から、負の相関、無相関、正の相関）

(4) 結果の保存

【Correlation map】タブ

相関マップ上で右クリックすると、マップの保存メニューが表示されます。



- Save correlation matrix（相関行列をタブ区切りテキストファイルで保存）
- Copy chart（相関マップをコピー）
- Save chart as…（相関マップを.jpg/.bmp/.png 形式で保存）

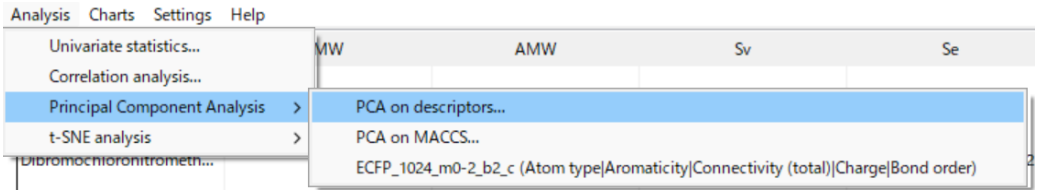
3.1.3 主成分分析

計算した記述子やフィンガープリントに基づく主成分分析（Principal Component Analysis, PCA）を実行します。PCA は、次元削減アルゴリズムの一つで、多次元のデータを低次元で表現する手法です。具体的には、データの分散が最大となる方向を見つけ、その方向に沿った新しい軸（第 1 主成分、PC1）を設定し、次にその軸と直交する軸の中で投影したデータの分散が最大となる軸（第 2 主成分、PC2）を設定し、…と繰り返します。PC1, PC2 といった寄与率の大きい主成分を用いることで、次元を落として 2D グラフ等に表現することができ、記述子間の関係を分析したり、外れ値や類似する分子のクラスターを識別したりするなど、データセット全体の傾向を把握するのに役立ちます。

(1) [Analysis > Principal Component Analysis >分析対象] を選択（ここでは例として PCA on descriptors…を選択）

分析対象として選択可能なデータ（いずれも計算が完了していることが条件）：

- 分子記述子（Descriptors）
- 各種フィンガープリント
  - Extended Connectivity Fingerprint（ECFP/ ECFPV3）
  - Path Fingerprint（PFP）
  - MACCS 166 Fingerprint

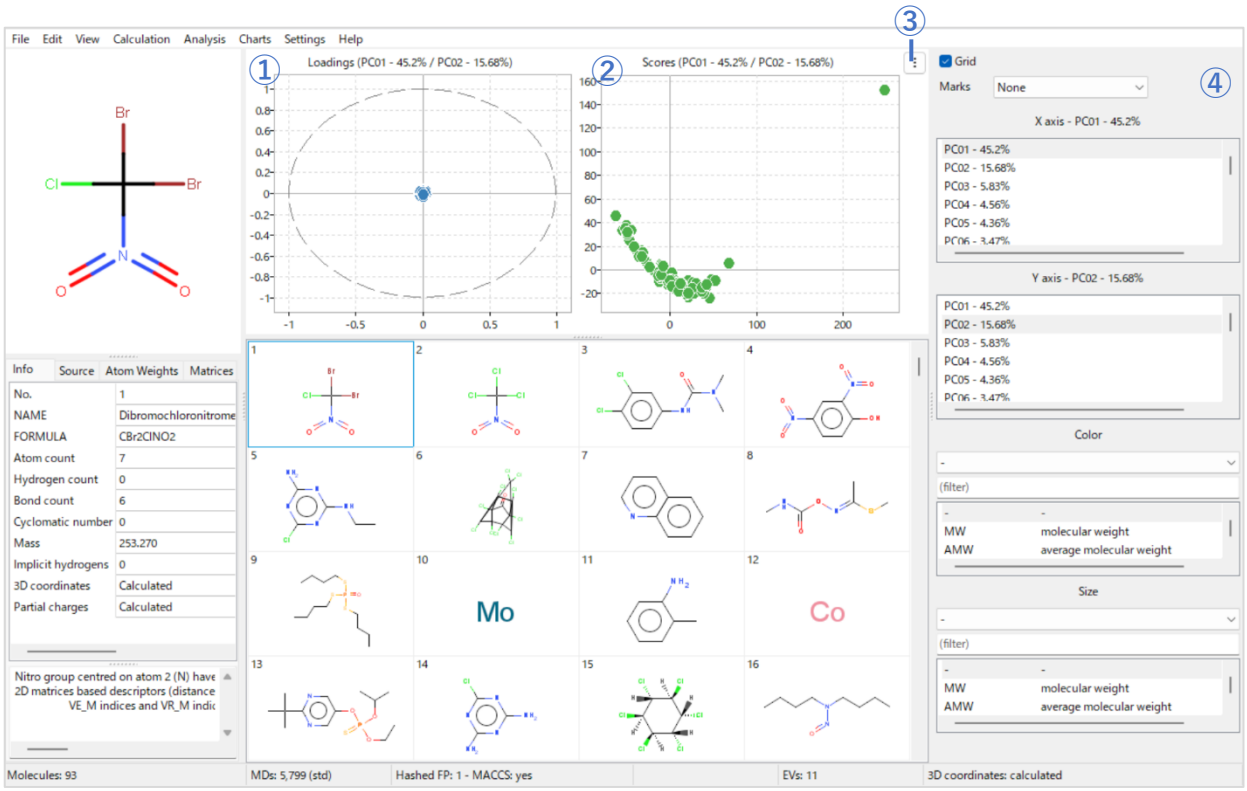


(2) 記述子選択ダイアログから、分析したい記述子を選択し、[OK]（ここでは例として All を選択）

(3) 結果の確認

解析を実行すると、以下のような結果画面が表示されます。

散布図上でマウスホイールを動かすと拡大/縮小ができ、各プロットにマウスオーバーすると記述子や分子の情報が表示されます。散布図上での各種操作の詳細については、[3.2 グラフ表示] の [グラフ画面上で可能な操作] をご覧ください。

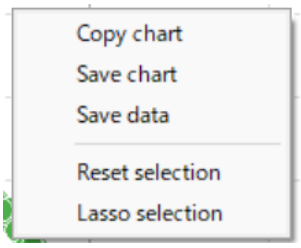


① 主成分負荷量のプロット（各記述子またはフィンガープリントのプロット）
② 主成分得点のプロット（各分子のプロット）
③ 画面切り替えアイコン（クリックすると④が非表示になり、プロット画面が拡大）
④ 表示設定

- Grid ：グリッド表示の ON/OFF
- Marks ：「Name」を選択するとプロットに記述子またはフィンガープリント（①）/分子名（②）を表示
- X axis ：X 軸に設定する主成分を選択（各主成分の隣は寄与率%）
- Y axis ：Y 軸に設定する主成分を選択
- Color ：選択した記述子の値を色で表示（②）
- Size ：選択した記述子の値をプロットのサイズに反映（②）

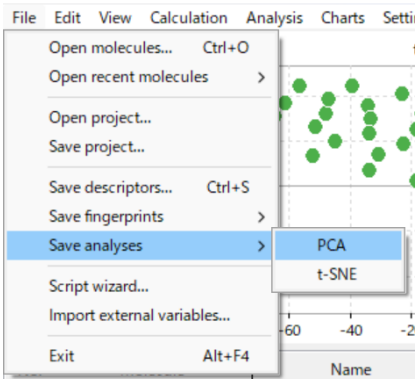
(4) 結果の保存

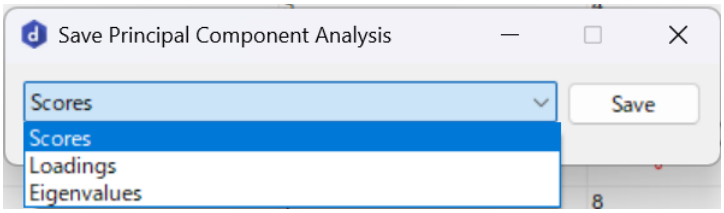
❑ 各散布図上で右クリックすると、保存メニューが表示されます。



- Copy chart（散布図をクリップボードにコピー）
- Save chart（散布図を.jpg/.bmp/.png 形式で保存）
- Save data（プロットの値をタブ区切りテキストファイルで保存）

❑ [File > Save analysis > PCA] メニューからも分析結果を出力できます。（タブ区切りテキストファイル形式）





- Scores（主成分得点）
- Loadings（主成分負荷量）
- Eigenvalues（固有値）

【Note】主成分分析の固有値と寄与率について
主成分分析における固有値は、各主成分に沿ったデータの分散の大きさを表し、値が大きい主成分ほどデータ全体の情報を多く保持しています。寄与率は、各主成分の固有値を全主成分の固有値の合計で割ることで求められ、その主成分が全体の情報に占める割合を表します。

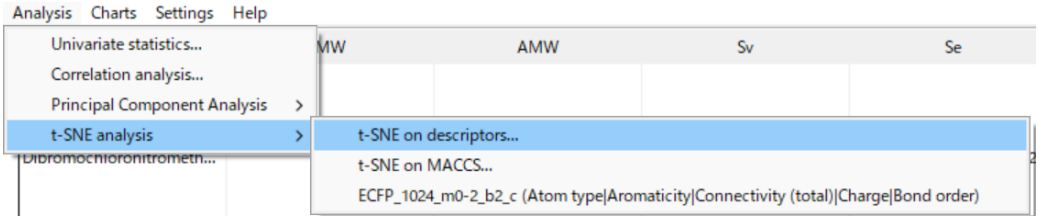
3.1.4 t-SNE 分析

計算した記述子やフィンガープリントに基づく t-SNE（t-distributed Stochastic Neighbor Embedding）分析を実行します。t-SNE 分析も次元削減アルゴリズムの一つです。高次元でのデータ点の類似度を確率的に表し、類似した化合物を近接した点、異なる化合物を離れた点で表現することで、指定された記述子に基づく高次元空間における各化合物の近接性、類似性を 2 次元平面上に可視化することができます。PCA 分析と比較して、非線形の特徴を抽出するのに優れています。また、PCA 分析は元の多次元空間全体を保持するのに対し、t-SNE 分析は元の多次元空間を 2 次元に削減して表すという違いもあります。

(1) [Analysis > t-SNE analysis > 分析対象] を選択（ここでは例として t-SNE on descriptors…を選択）

分析対象として選択可能なデータ（いずれも計算が完了していることが条件）：

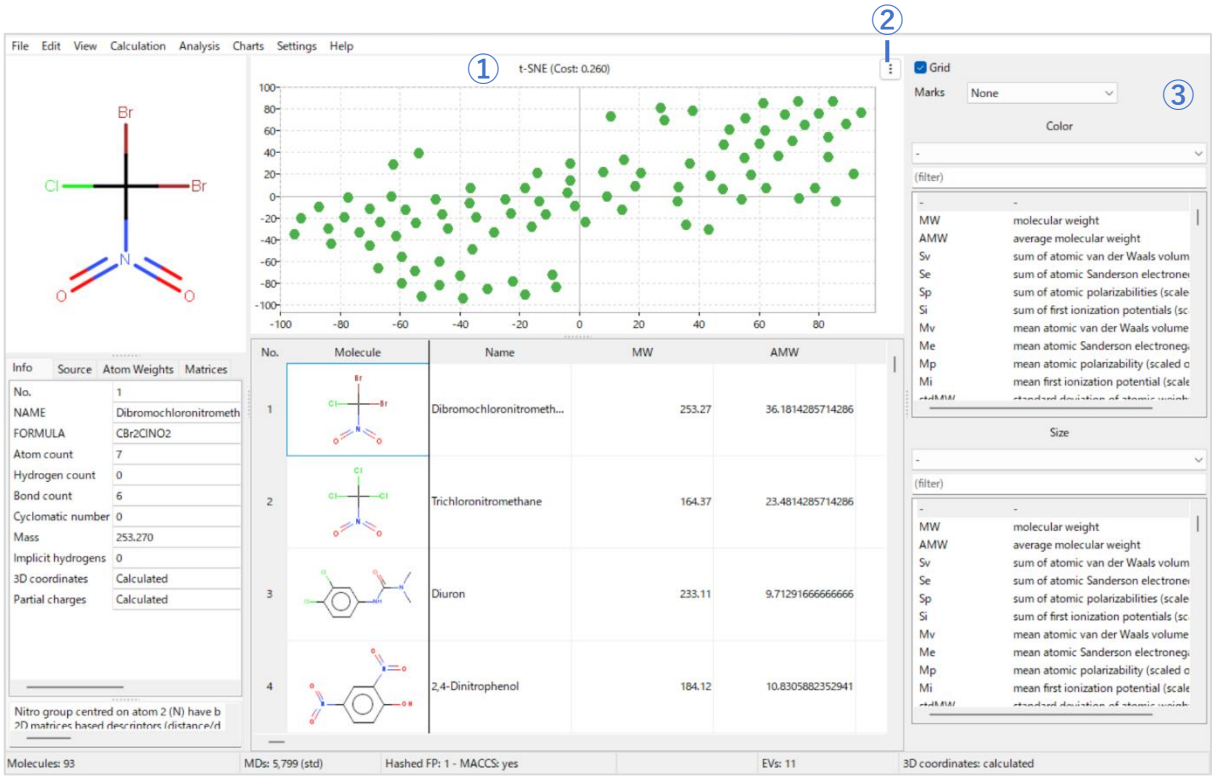
- 分子記述子（Descriptors）
- 各種フィンガープリント
  - Extended Connectivity Fingerprint（ECFP/ ECFPV3）
  - Path Fingerprint（PFP）
  - MACCS 166 Fingerprint



(2)　記述子選択ダイアログから、分析したい記述子を選択し、[OK]（ここでは例として All を選択）

(3)　結果の確認

解析を実行すると、以下のような結果画面が表示されます。



散布図上での各種操作は主成分分析と同様です。詳細は、[3.2 グラフ表示] の項目をご覧ください。

①　t-SNE プロット

②　画面切り替えアイコン（クリックすると③が非表示になり、プロット画面が拡大）

③　表示設定

Grid　　：グリッド表示の ON/OFF

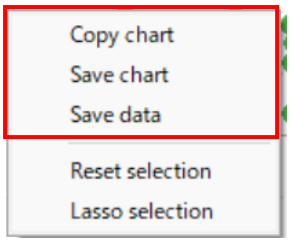
Marks　　：「Name（分子名）」「ID（化合物 No.）」「Value（計算値）」を選択するとプロットに表示

Color　　：選択した記述子の値を色で表示

Size　　：選択した記述子の値をプロットのサイズに反映

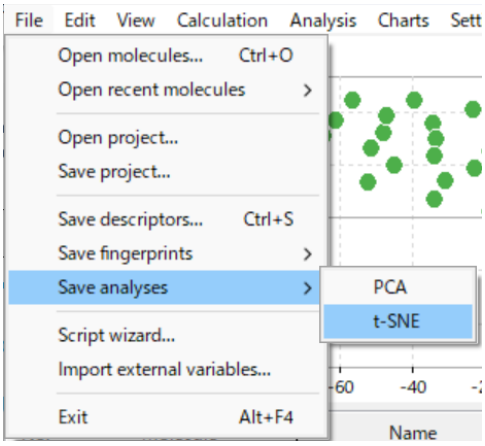
(5)　結果の保存

❑　散布図上で右クリックすると、散布図の保存メニューが表示されます。



- Copy chart（散布図をクリックボードにコピー）
- Save chart（散布図を.jpg/.bmp/.png 形式で保存）
- Save data（プロットの値をタブ区切りテキストファイルで保存）

❑ [File > Save analysis > t-SNE] メニューからも分析結果を出力できます。（タブ区切りテキストファイル形式）



3.2 グラフ表示

[Chart] メニューから、ヒストグラム（Histogram）、棒グラフ（Bar plot）、散布図（Scatter plot）、PCA プロット（PCA plot）、t-SNE 分析プロット（t-SNE plot）を表示できます。

3.2.1 グラフ画面の操作

ヒストグラムや棒グラフ、散布図などの表示画面上で可能な操作は基本的に共通なので、ここにまとめます。

【グラフ画面の操作】

❑ 拡大/縮小・プロット詳細情報の表示
マウスホイールで拡大/縮小、プロットにマウスオーバーすると記述子や分子の情報が表示されます。

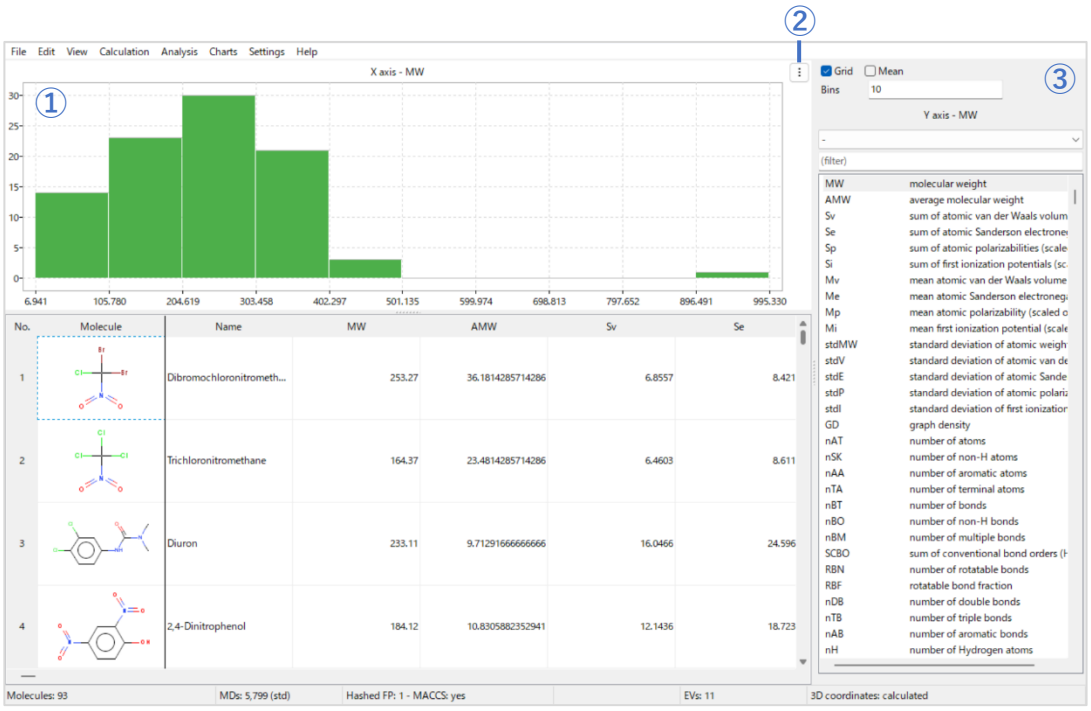
❑ 結果の保存（右クリックメニュー）
Copy chart ：グラフイメージをクリップボードにコピー
Save chart ：グラフイメージを.jpg/.bmp/.png 形式で保存
Save data ：グラフの値をタブ区切りテキストファイルで保存

❑ 分子の選択（フィルタリング）
個別選択 ：プロットや棒グラフを左クリック
矩形範囲選択 ：左クリックしたままドラッグ
投げ縄選択 ：右クリック > Lasso selection にチェックを入れ、左クリックしたままドラッグ
選択解除 ：右クリック > Reset selection を選択

3.2.2 ヒストグラム

❑ [Chart > Histogram] を選択して表示

単一の記述子に関する度数分布を表示します。



① チャート画面

② 画面切り替えアイコン（クリックすると③が非表示になり、プロット画面が拡大）

③ 表示設定

Grid ：グリッド表示の ON/OFF

Mean ：平均値の表示の ON/OFF

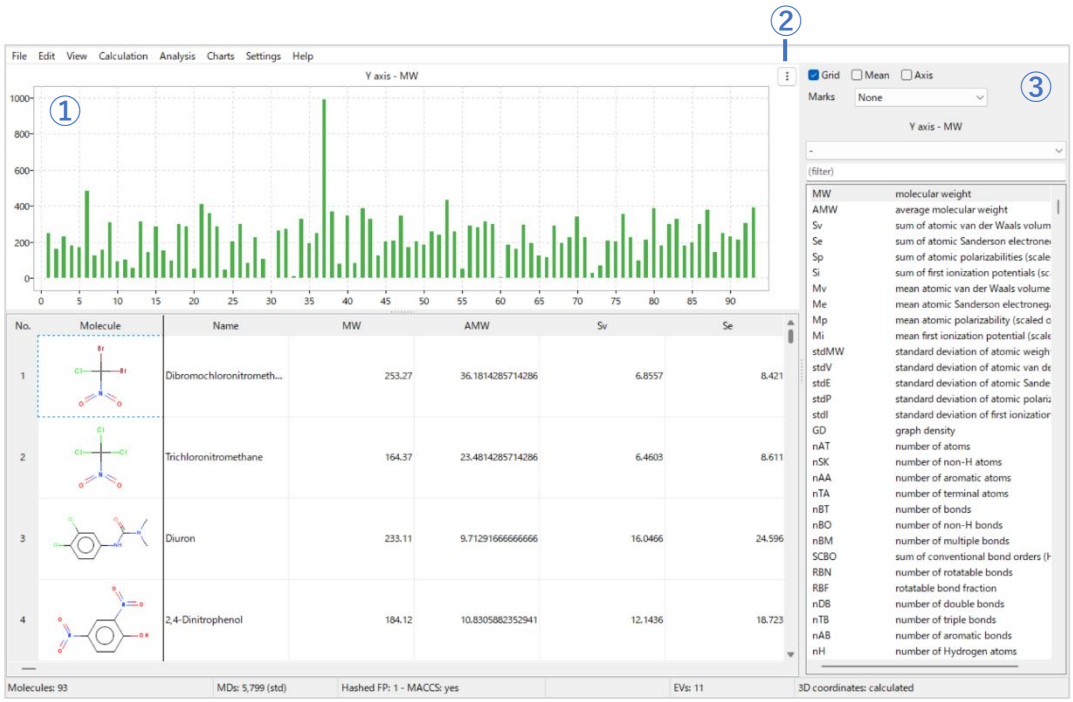
Bins ：区間（階級の数）の指定

Y axis ：ヒストグラムに表示する記述子を選択

3.2.3 棒グラフ

❑ [Chart > Bar plot] を選択して表示

記述子の個々の値を棒グラフに表示します。



① チャート画面

② 画面切り替えアイコン（クリックすると③が非表示になり、プロット画面が拡大）

③ 表示設定

Grid ：グリッド表示の ON/OFF

Mean ：平均値の表示の ON/OFF

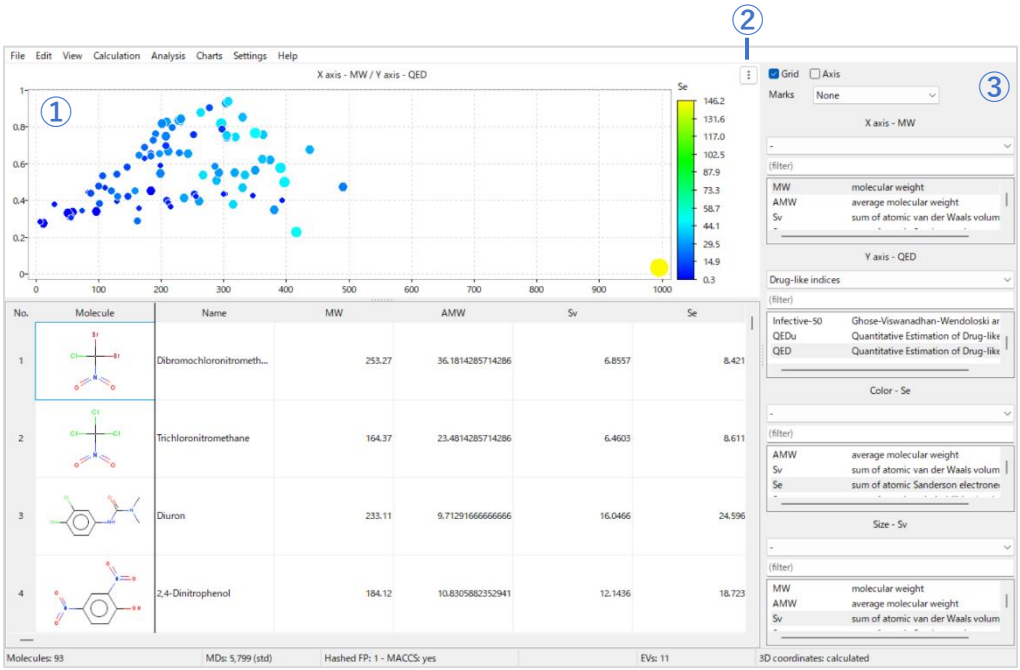
Axis ：X,Y 軸の表示の ON/OFF

Y axis ：ヒストグラムに表示する記述子を選択

3.2.4 散布図

❑ [Chart > Scatter plot] を選択して表示

2~4 つ（X 軸、Y 軸、色、プロットサイズ）の記述子の値を散布図に表示します。



① チャート画面

② 画面切り替えアイコン（クリックすると③が非表示になり、プロット画面が拡大）

③ 表示設定

Grid ：グリッド表示の ON/OFF

Axis ：X,Y 軸の表示の ON/OFF

X axis ：X 軸に設定する記述子を選択

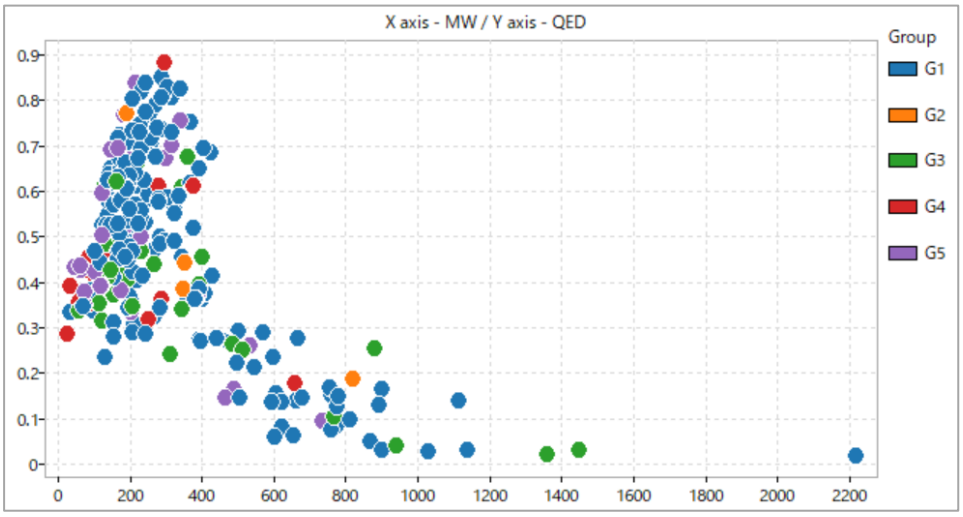
Y axis ：Y 軸に設定する記述子を選択

Color ：選択した記述子の値を色で表示

Size ：選択した記述子の値をプロットのサイズに反映

【Tips】質的変数によるカラーリング

[Color] の設定は、量的変数だけでなく質的変数も指定でき、グループごとに色分け表示することもできます（下図は他のデータセットの例です）。



3.2.5 主成分分析(PCA)プロット

❑ [Chart > PCA plot] を選択して表示
直近に行った分析の結果画面が表示されます。画面の詳細については、「3.1.3 主成分分析」をご覧ください。

3.2.6 t-SNE 分析プロット

❑ [Chart > t-SNE plot] を選択して表示
直近に行った分析の結果画面が表示されます。画面の詳細については、「3.1.4 t-SNE 分析」をご覧ください。

【Tips】プロットの表示／非表示の切り替え
Analysis や Charts メニューで表示したプロット画面から、元の分子の一覧画面に戻りたい場合は、Charts メニューから現在表示されているプロット名（✓がついているプロット名）を選択するとプロットの表示を OFF にできます。プロット名を再度選択することで表示を ON にすることができます。

3.3 フィルタリングと並び替え／分子表示

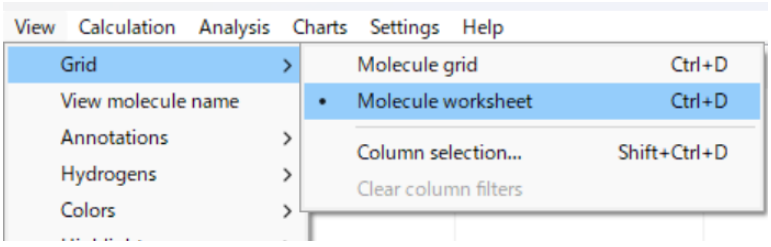
この章では、ワークシート表示画面での分子のフィルタリングや並び替え、表示メニュー（View メニュー）からできる表示設定などを紹介します。

3.3.1 フィルタリングと並び替え

ワークシート表示画面上では、分子のフィルタリングや並び替えを行うことができます。フィルタリングを活用することで、膨大な化合物データの中から特定の条件を満たす化合物を抽出することができます。例えば、医薬品のリード化合物のスクリーニングのためにリピンスキーのルールを満たす分子を選択したり、QSAR モデル作成の前処理として外れ値を除外したりすることができます。
分子のフィルタリング（選択）はグラフ表示画面から行うこともでき（【3.2.1 グラフ画面の操作】を参照）、この場合は、視覚的にクラスターを選択したり、記述子の特定の閾値に含まれる化合物を直感的に選択したりすることが可能です。
ここではワークシート表示画面上での操作について紹介します。

これ以降の操作はワークシート表示で行います。

❑ [View > Grid > Molecule worksheet] を選択



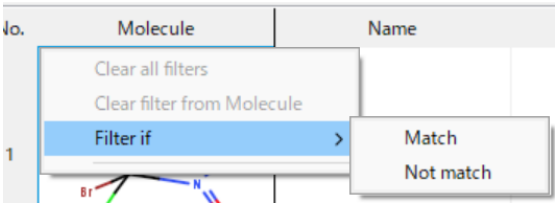
※前のセクションで表示したプロット画面が表示されている場合は、Charts メニューから現在表示されているプロット名（✓がついているプロット名）を選択して表示を OFF にしてください。

① 分子のフィルタリング

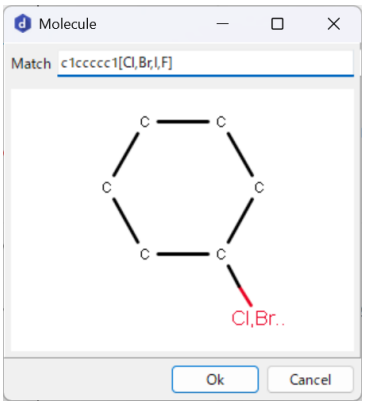
入力した分子の中から必要な分子のみを選択して絞り込みを行うことができます。分子の部分構造を指定したフィルタリング、分子名を利用したフィルタリング、記述子の値を使ったフィルタリングを行うことができます。

(1) **部分構造によるフィルタリング**：分子構造のヘッダー[Molecule] を右クリックし[Filter if > Match] 又は[Not match] を選択

「Match（指定した部分構造を含むもの)」、または、「Not match（含まないもの)」で絞り込みができます。今回は例として、[Match] を選択して、ベンゼン環にハロゲンが 1 つ置換した構造を指定してみます。



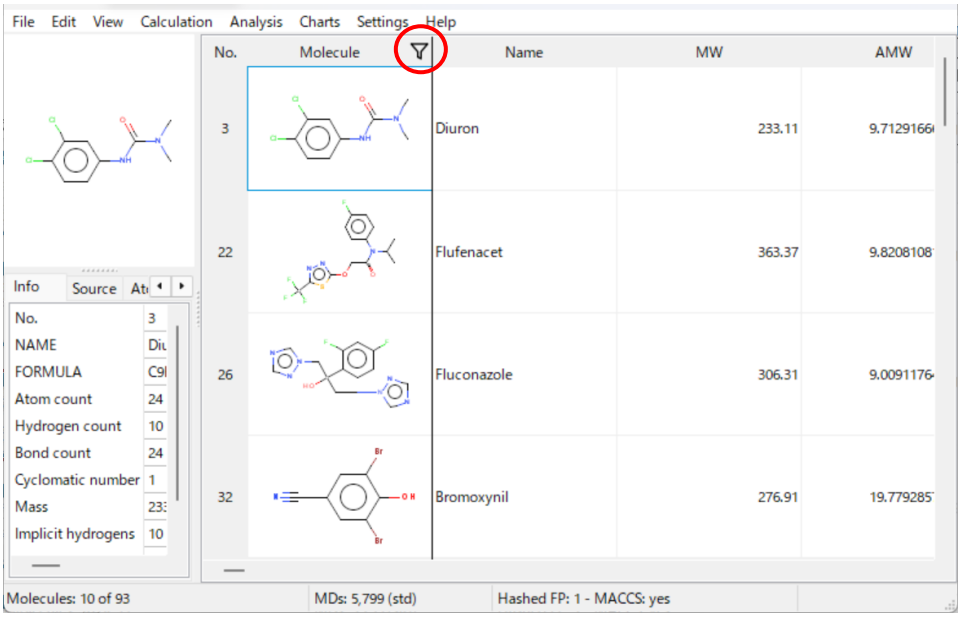
指定する部分構造は次の画面から SMARTS 形式で入力します。



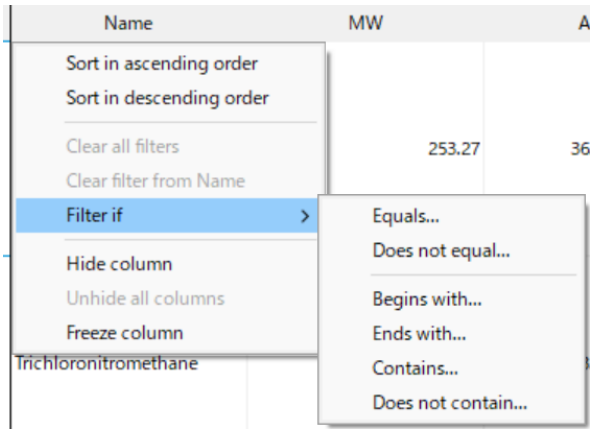
SMARTS 入力例：c1ccccc1[Cl,Br,I,F]

※SMARTS 表記については、参考情報の【5.6 SMARTS 記法について】をご参照ください。

(2)　[OK] をクリックして絞り込みを行うと、ヘッダーにロート（じょうご）のアイコンが表示されます



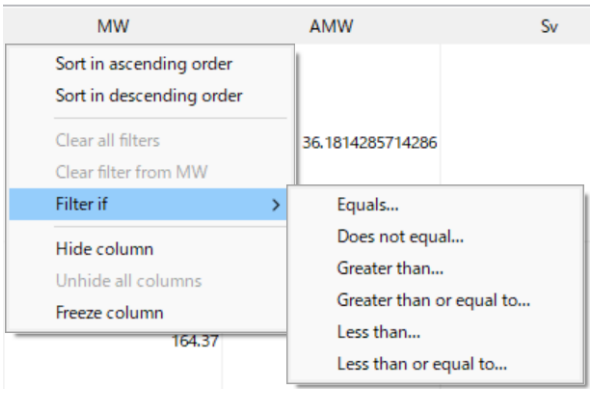
(3) **分子名/文字列によるフィルタリング**：分子名のヘッダー [Name] を右クリックし、[Filter if] から各種条件を指定
分子名などの文字列のカラムは、文字列を利用してフィルタリングを行うことができます。



Equals…	: ~と一致
Does not equal…	: ~と一致しない
Begins with…	: ~から始まる
Ends with…	: ~で終わる
Contains…	: ~を含む
Does not contain…	: ~を含まない

- 外部変数（テキスト）のカラムでも同様の操作が可能です。

(4) **記述子/数値によるフィルタリング**：記述子のヘッダーを右クリックし、[Filter if] から各種条件を指定
数値が入っているカラムは、特定の数値以上・以下などの条件を指定してフィルタリングを行うことができます。



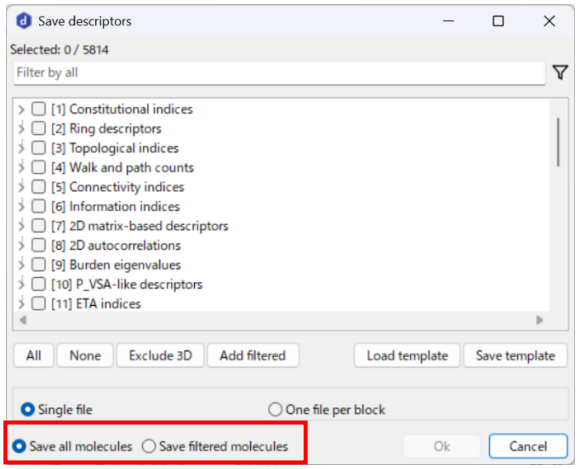
Equals…	: 等価（＝x）
Does not equal…	: 非等価（≠x）
Greater than…	: より大きい（＞x）
Greater the or equal to…	: 以上（≧x）
Less than…	: より小さい（＜x）
Less the or equal to…	: 以下（≦x）

- 外部変数（数値）のカラムでも同様の操作が可能です。

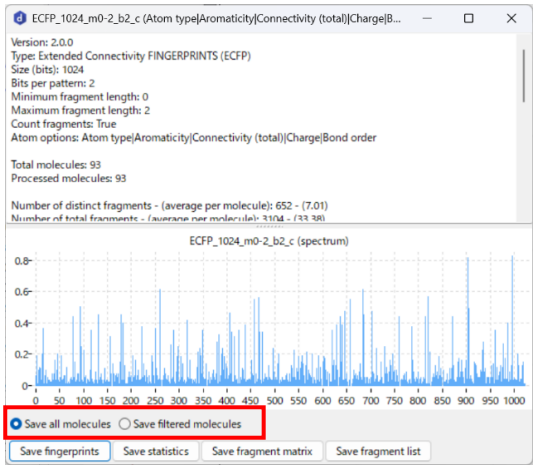
(5) フィルタリングを解除するには、ヘッダーを右クリック > [Clear all filters]（全てのフィルターを解除） 又は[Clear filter from] (ヘッダー名)（選択した列のフィルターを解除）を選択

【Note】フィルタリング後の記述子・フィンガープリント計算結果の保存
分子のフィルタリングを行った後に計算結果を保存する場合、保存画面や計算結果画面の一番下にある[Save all molecules]/ [Save filtered molecules] オプションが選択できるようになります。[Save filtered molecules] にチェックを入れると、フィルタリングをした分子の結果のみを保存できます。

記述子の保存画面



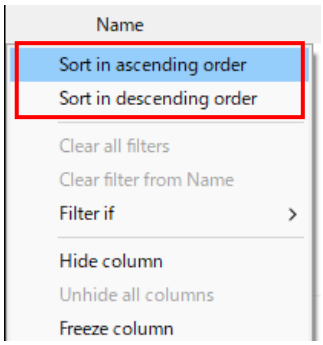
フィンガープリントの計算結果画面

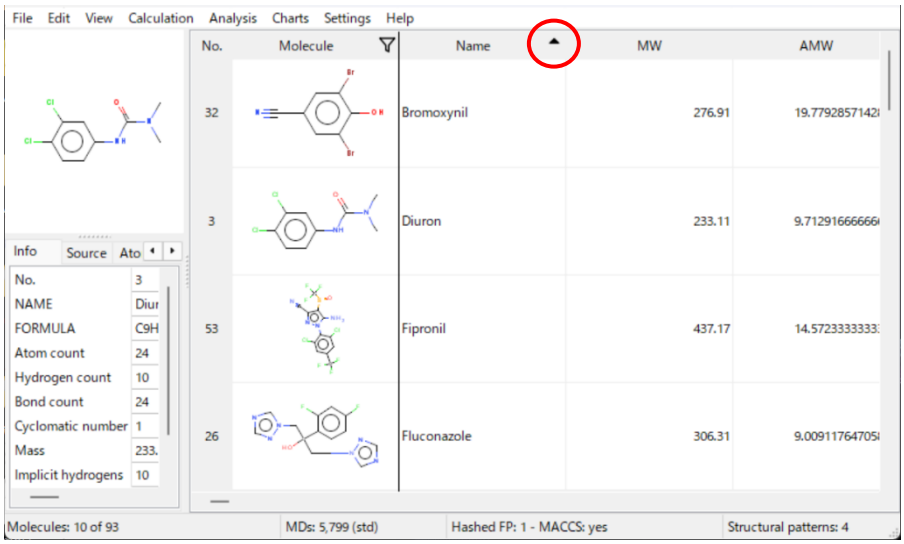


② 分子の並び替え

昇順または降順に並び替えができます。

❑ ヘッダーを右クリック > [Sort in ascending order]（昇順） 又は [Sort in descending order]（降順） を選択





昇順・降順の並べ替えを行うとヘッダーに▼▲マークが表示されます。

[Sort in ascending order]（昇順） 又は [Sort in descending order]を再度クリックすると、並び替えを解除することができます。

③ その他の右クリックメニュー（カラムの表示切替）

Hide column	：カラムの非表示（非表示にした列の場所には || のようなマークが表示されます）
Unhide column [ヘッダー名]	：[ヘッダー名]のカラムのみ再表示
Unhide all column	：全てのカラムを再表示
Freeze column	：カラムを固定する（Molecule 列の右に表示され、スクロールしても常時表示されます）

【Tips】カラムの選択表示

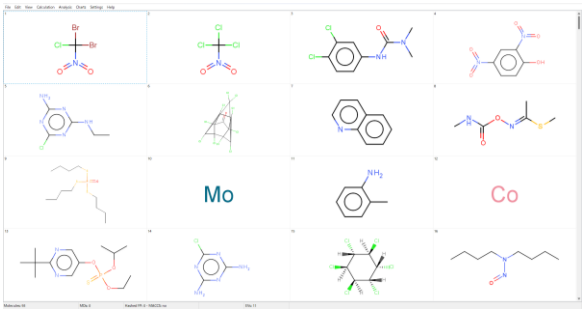
メニューの [View > Grid > Column selection…] を選択し、記述子の選択ダイアログから表示したい記述子を選択・表示することができます。

3.3.2 分子の表示機能一覧（View メニュー）

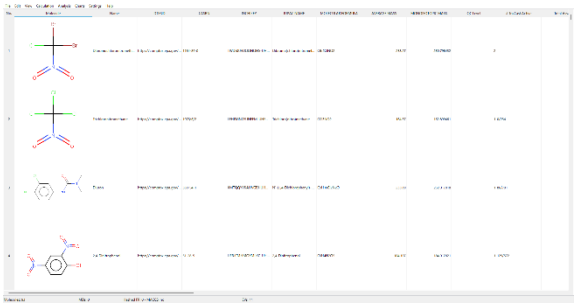
[View]メニューからは、分子の表示形式やハイライト表示などを設定することができます。

❑ [View > Grid]

グリッド表示/ワークシート表示の切り替えができます。



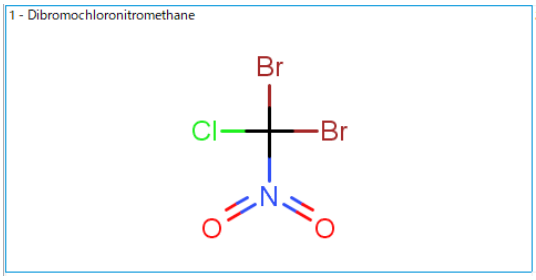
[Molecule grid] グリッド表示



[Molecule worksheet] ワークシート表示

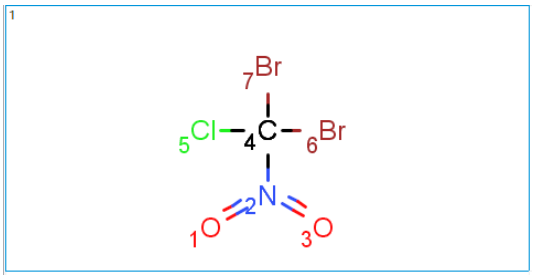
❑ [View > View molecule name]

分子構造のセルの左上に分子名を表示します。

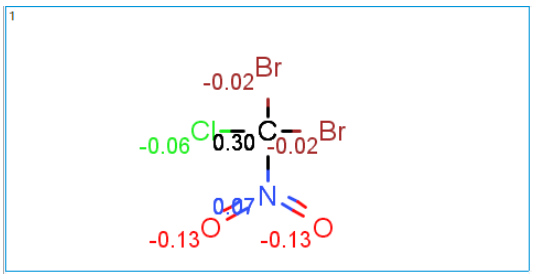


❑ [View > Annotations]

構造式にラベル（原子の番号/部分電荷）を追加します。



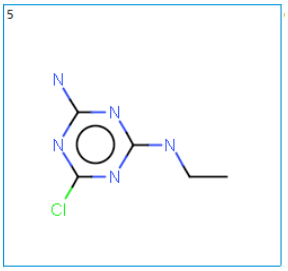
[Atom index] 原子の番号を表示



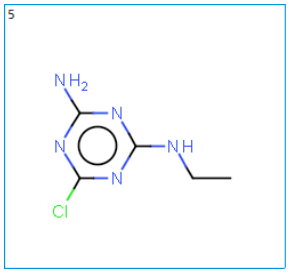
[Partial charge] 部分電荷を表示

❑ [View > Hydrogens]

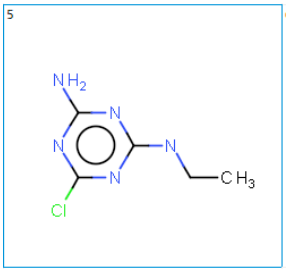
構造式の水素の表示形式を設定します。



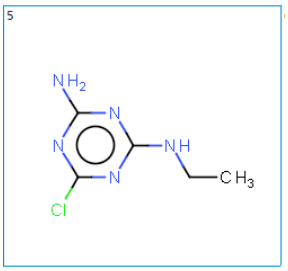
[Off] 表示なし



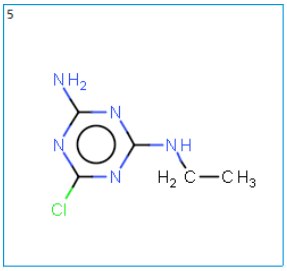
[On hetero atoms] ヘテロ原子上



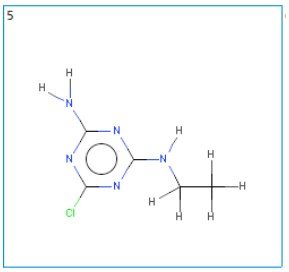
[On terminal atoms] 末端の原子上



[On hetero and terminal atoms]
ヘテロ原子と末端の原子上



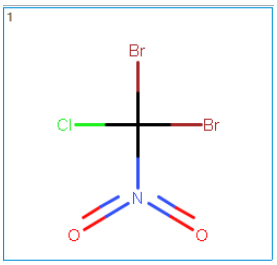
[On all] 全ての原子上



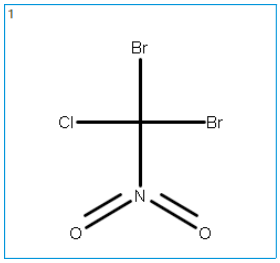
[Show H bonds] 水素の結合を表示

【Note】暗示的水素の自動付加について
デフォルト設定では分子の読み込み時に暗示的水素が自動的に付加されます。MDL ファイルまたは MOL2 ファイルで暗示的水素を自動付加したくない場合は、[Settings > Options… > Cluculation タブ]の[Add implicit hydrogens (MDL, MDL2)]のチェックを外し、構造ファイルを読み込んでください。（SMILES で読み込む場合は常に暗示的水素が付加されます。）

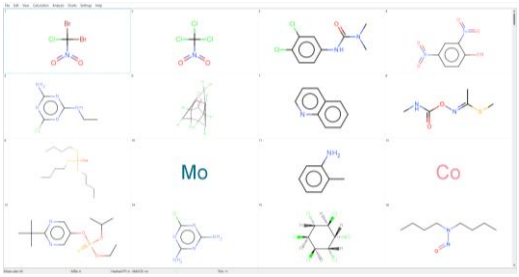
❑ [View > Colors]
構造式のカラーリングや背景色を変更します。



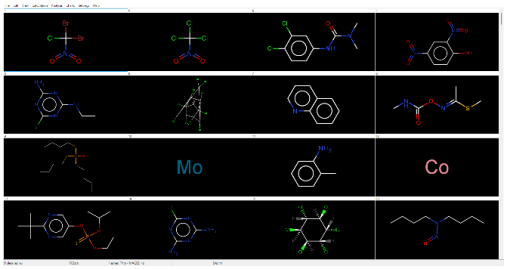
[CPK] CPK 表示



[Monochrome] 白黒表示

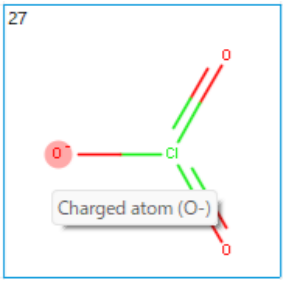


[White background] 白背景

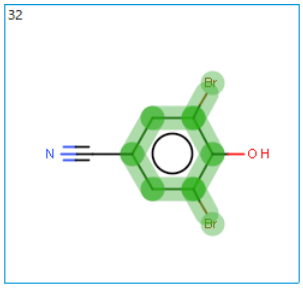


[Black background] 黒背景

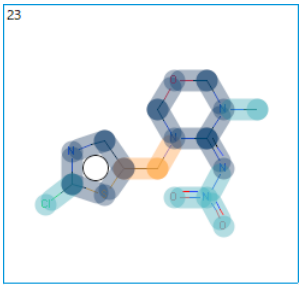
❑ [View > Highlights]
構造式のハイライト表示を設定します。Warning 表示の場合はハイライト部分にマウスオーバーすると警告内容が表示されます。



[Warnings] 警告の表示



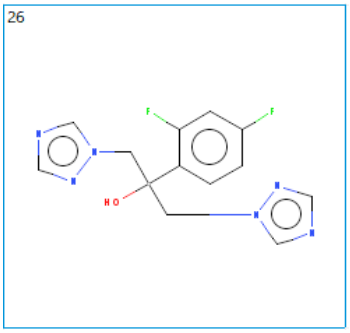
[Substructure]
SMARTS 入力した構造を強調表示



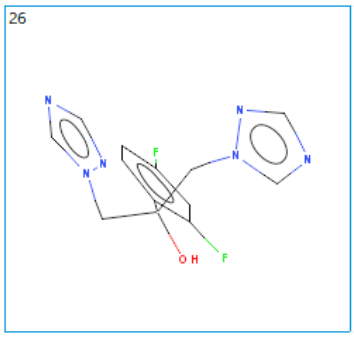
[Bemis-Murcko features]
Bemis-Murcko features 表示

❑ [View > Coordinates]

分子構造の 2D/3D 表示を切りえます。[Use coordinates from file] を選択すると、入力ファイルの情報がそのまま表示（2D で入力した場合は 2D 表示、3D 情報を含む場合は 3D 表示）されます。[3D] 表示は 3D 座標情報がある場合のみ立体表示されます。



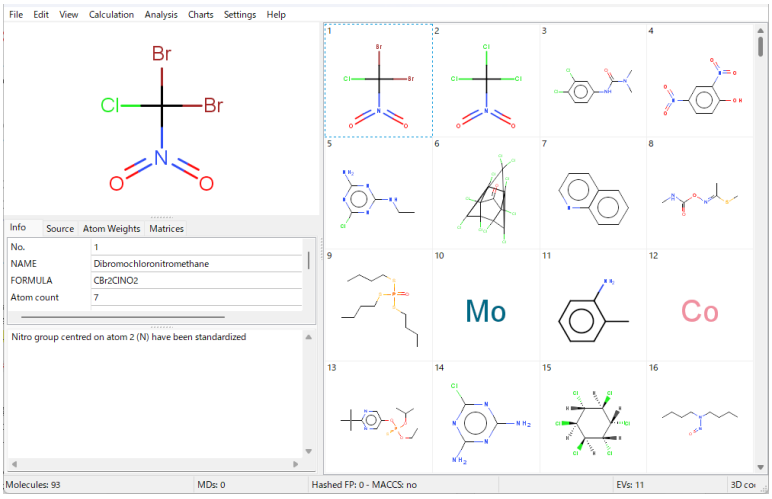
[2D] 2D 表示



[3D] 3D 表示（3D 座標がある場合）

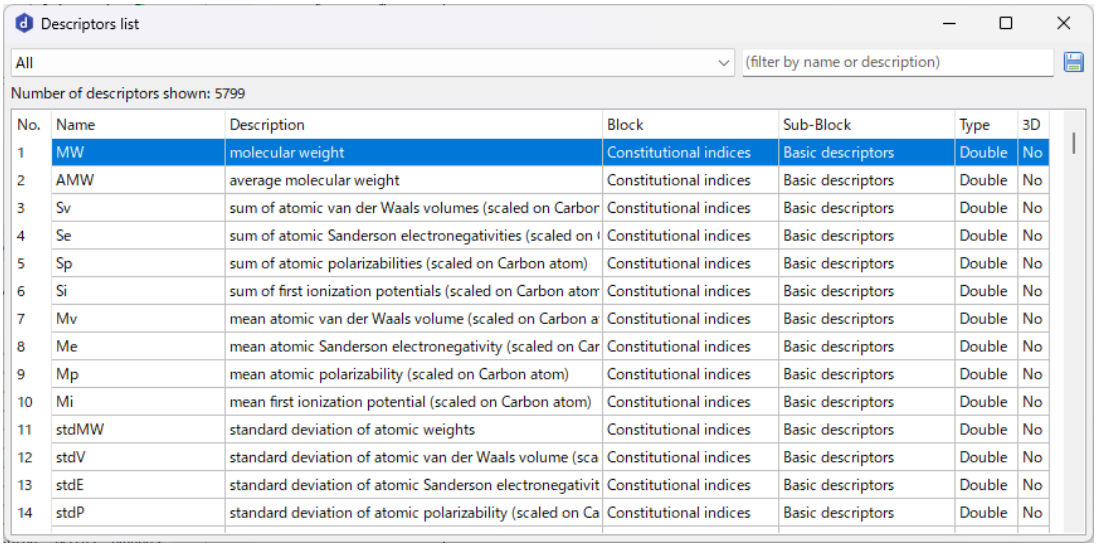
❑ [View > Molecule detail]

画面左側に分子の詳細情報のパネルが表示されます。



❑ [View > Descriptor list]

簡単な説明付きの分子記述子の一覧がリスト表示されます。

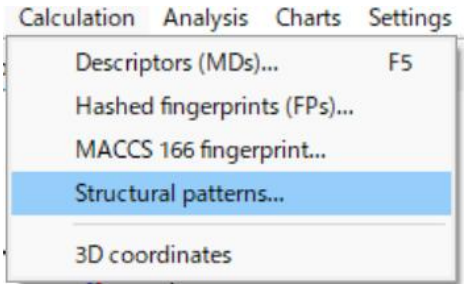


No.	Name	Description	Block	Sub-Block	Type	3D
1	MW	molecular weight	Constitutional indices	Basic descriptors	Double	No
2	AMW	average molecular weight	Constitutional indices	Basic descriptors	Double	No
3	Sv	sum of atomic van der Waals volumes (scaled on Carbor	Constitutional indices	Basic descriptors	Double	No
4	Se	sum of atomic Sanderson electronegativities (scaled on (	Constitutional indices	Basic descriptors	Double	No
5	Sp	sum of atomic polarizabilities (scaled on Carbon atom)	Constitutional indices	Basic descriptors	Double	No
6	Si	sum of first ionization potentials (scaled on Carbon atom	Constitutional indices	Basic descriptors	Double	No
7	Mv	mean atomic van der Waals volume (scaled on Carbon a	Constitutional indices	Basic descriptors	Double	No
8	Me	mean atomic Sanderson electronegativity (scaled on Car	Constitutional indices	Basic descriptors	Double	No
9	Mp	mean atomic polarizability (scaled on Carbon atom)	Constitutional indices	Basic descriptors	Double	No
10	Mi	mean first ionization potential (scaled on Carbon atom)	Constitutional indices	Basic descriptors	Double	No
11	stdMW	standard deviation of atomic weights	Constitutional indices	Basic descriptors	Double	No
12	stdV	standard deviation of atomic van der Waals volume (sca	Constitutional indices	Basic descriptors	Double	No
13	stdE	standard deviation of atomic Sanderson electronegativit	Constitutional indices	Basic descriptors	Double	No
14	stdP	standard deviation of atomic polarizability (scaled on Ca	Constitutional indices	Basic descriptors	Double	No

3.4 構造パターンの計算

指定した構造パターンが分子に含まれるかどうかを検出・同定し、特徴量として数値化することができます。構造パターンは SMARTS 形式で指定します。計算すると、構造パターンの有無や存在個数などの結果を含む記述子ブロックが 35 番目のブロックとして追加されます。

(1) [Calculation > Structural patterns…] を選択



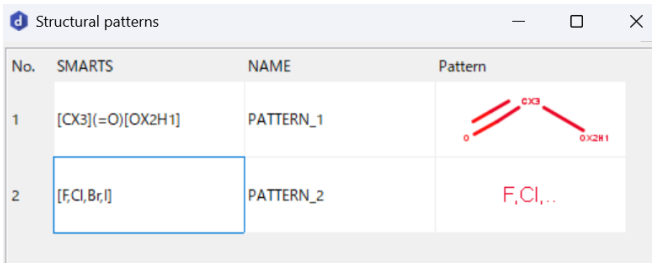
(2) 表示されるダイアログで構造パターンの入力と各種設定を行い、[OK]

ダイアログの操作例

- ❑ 左下にある[Add] ボタンをクリックし、新しい項目を追加します。
- ❑ 行が追加されるので、空白の SMARTS 列のセルをダブルクリックして、計算したい構造パターンを SMARTS 形式で入力します。ここでは例として、以下の 2 つの構造パターンを入力してみます。

  - カルボン酸　：[CX3](=O)[OX2H1]
  - ハロゲン　：[F,Cl,Br,I]

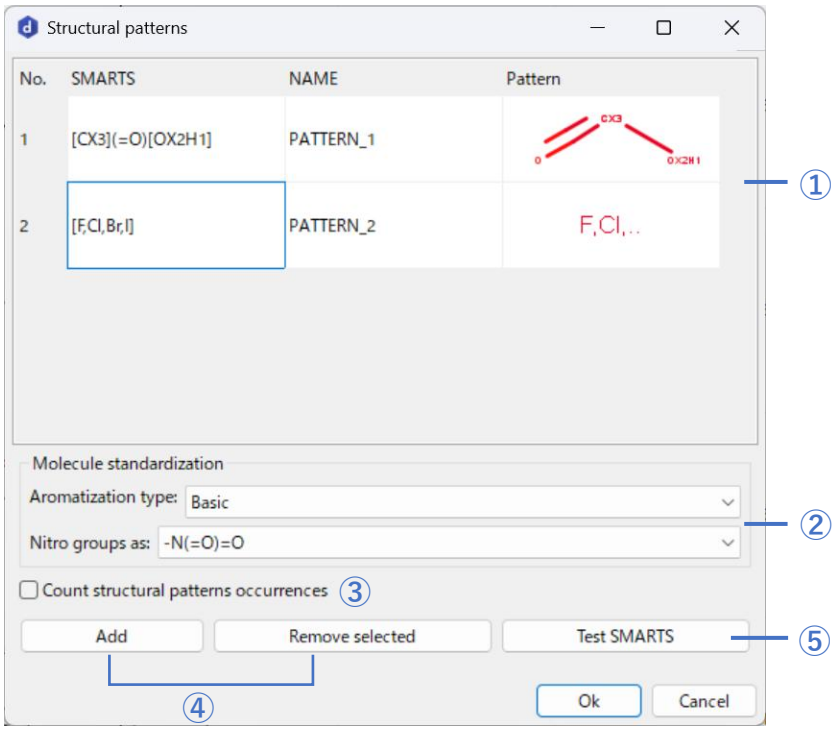
構造パターンの入力例



【Note】構造パターンの入力と削除について
- SMARTS を入力して Enter また入力セルの外部をクリックすると、Pattern 列に構造が表示されます。
- 複数の構造パターンを入力する際は、[Add] ボタンをクリックして順次追加してください。
- NAME 列に構造パターン名が自動入力されますが（PATTERN_1, PATTERN_2…）、セルをダブルクリックすることで文字列を編集可能です。
- 構造パターンを削除したい場合は、No.列の該当する番号をクリックして行全体を選択し、[Remove selected] ボタンをクリックします。No.ヘッダーをクリックすると全パターンが選択でき、一括削除が可能です。

- ❑ 必要に応じてダイアログ下部で各種設定を行い、[OK] をクリックして計算を実行します。
  今回はデフォルト設定のまま実行します。オプション等の詳細は、以下の【構造パターン設定ダイアログ】をご参照ください。

＜構造パターン設定ダイアログ＞



① **構造パターンのリスト**

[Add] ボタンで項目を新規追加し、SMARTS や構造パターン名を編集できます。SMARTS を入力すると、自動で Pattern 列に構造が表示されます。項目を削除したい場合は、No.列の番号をクリックして行全体を選択し、[Remove selected] ボタンをクリックします。

② **分子の標準化**

構造パターン同定前に、芳香族性とニトロ基を指定条件に標準化します。

- Aromatization type（芳香族性）：Basic／General

（Basic では芳香環に環外二重結合がある場合は芳香環と認識しません(ピリジン *N*-オキシドは除く)が、General では芳香環として認識されます。）

- Nitro group as（ニトロ基）　：-N(=O)=O／$-N^{+}(=O)=O^{-}$

※標準化のタイプの詳細については、alvaDesc ユーザーマニュアルの【4.1 Molecule representation】をご参照ください。

③ **構造パターンの出現回数をカウントするオプション**

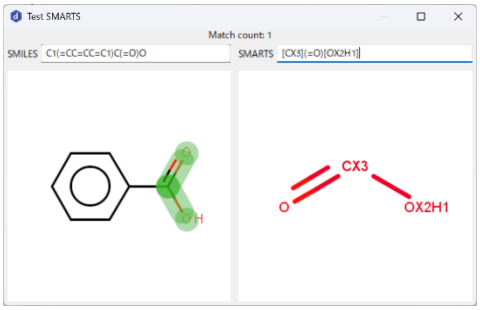
- 無効：構造パターンが存在するかが構造パターン変数に出力される（存在しない：0、存在する：1）
- 有効： 構造パターンが分子内に出現する回数を計算し、構造パターン変数に出力される

④ **Add/Remove selected ボタン**

[Add] ボタンで構造パターンを新規追加、No.列の番号をクリックして行全体を選択後、[Remove selected] ボタンを押すと項目を削除できます。

⑤ **Test SMARTS ボタン**

このボタンをクリックすると、Test SMARTS ダイアログが表示されます。このダイアログでは、SMILES 形式で入力された分子構造について、SMARTS による識別を評価できます。左側のフィールドにテスト対象の分子の SMILES を入力し、右側のフィールドに検出したい構造パターンの SMARTS を入力してください。ダイアログの上部には、検出されたパターンの数（Mach count）が表示されます。

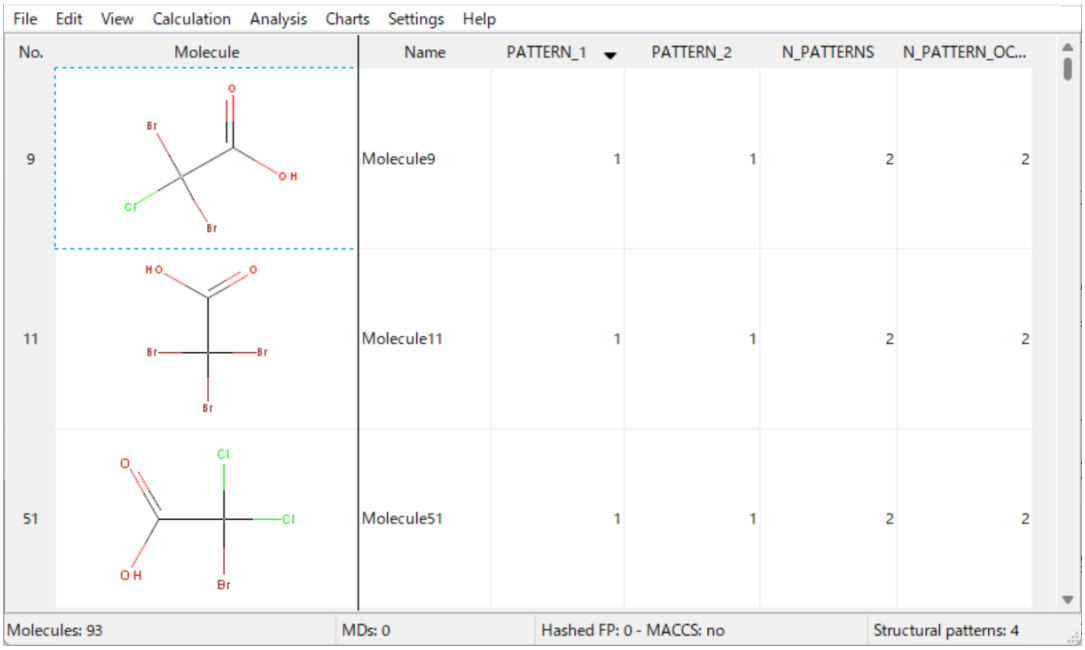


(3) 結果の確認

計算した構造パターンの結果は、35 番目の記述子ブロックとして追加されます。

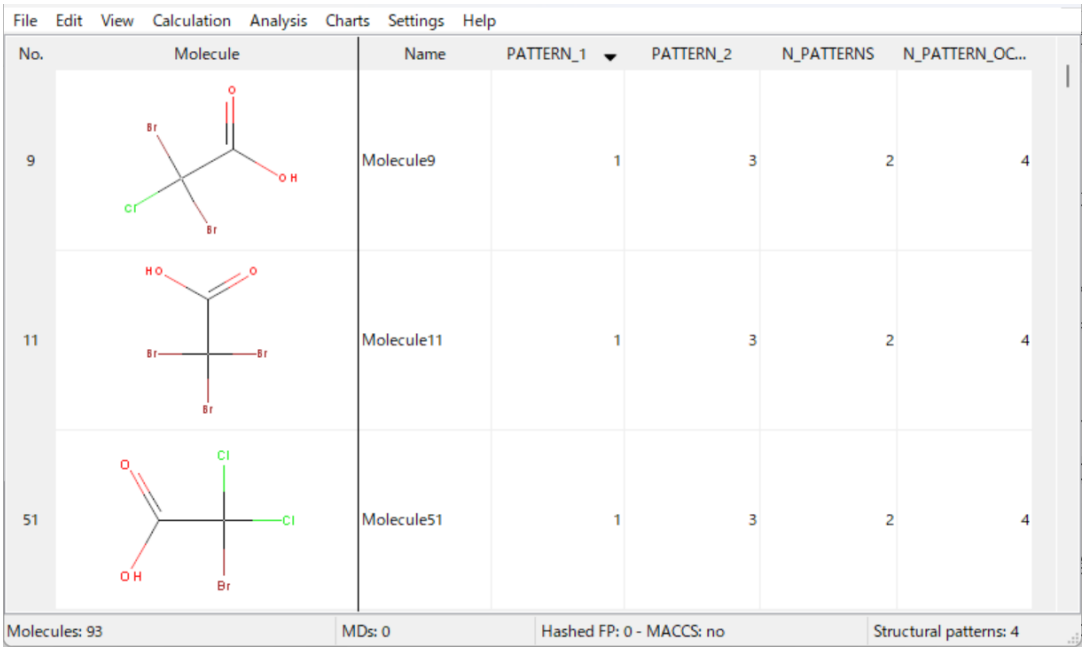
- 構造パターン名：構造パターンの存在の有無（出現回数をカウントするオプションを有効にした場合はその個数）
- N_PATTERNS：存在する異なる構造パターンの種類数
- N_PATTERN_OCCURRENCES：存在する構造パターンの数（出現回数をカウントするオプションを有効にした場合はカウントされた全ての構造パターン数、無効にした場合は N_PATTERNS と同じ値が出力）

・出現回数をカウントするオプションを無効にした場合の出力例



Molecule9 の化合物は、PATTERN_1（カルボン酸）が 1 つ、PATTERN_2（ハロゲン原子）が 3 つ含まれていますが、PATTERN_1 と PATTERN_2 の値は 1（各パターンの存在が示され、検出数は考慮されない）、N_PATTERN OCCURRENCES の値は 2（N_PATTERNS と同値）となっています。

・出現回数をカウントするオプションを有効にした場合の例

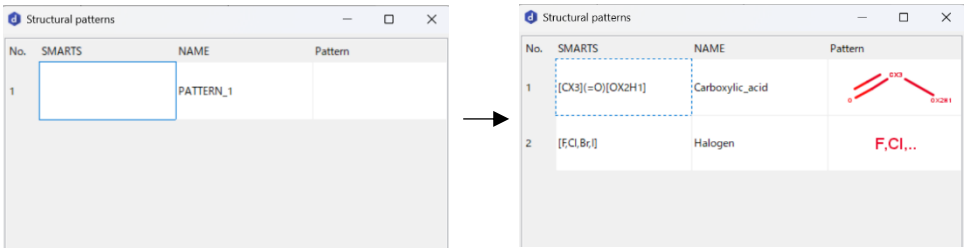


Molecule9 の化合物は、PATTERN_1 の値は 1（カルボン酸の検出数）、PATTERN_2 の値は 3（ハロゲン原子の検出数）、N_PATTERN OCCURRENCES の値は 4（各検出数を考慮した PATTERN_1 と PATTERN_2 の和）となっています。

【Tips】クリップボードを利用した構造パターンリストの入力

クリップボードを利用すると、SMARTS と構造パターン名を同時に入力 したり、複数構造パターンを含むリストを一度に作成 したりできます。

- 適当なセルをクリックしアクティブにした状態で、予めクリップボードにコピーしておいた SMARTS を Ctrl + V（macOS では command+V）でペーストして入力することができます。
- テキストエディタ等から SMARTS をコピーする際、「SMARTS（スペースまたはタブ区切り）構造パターン名」の形式にしておくと、ペースト時に 構造パターン名も自動的に [NAME] 列に入力されます。
  例）[CX3](=O)[OX2H1] Carboxylic_acid
- さらに、「SMARTS（スペースまたはタブ区切り）構造パターン名」の形式で一行ずつ記載し、複数行をまとめてコピーしておくと、複数の構造パターンを一括追加することもできます。
  例）[CX3](=O)[OX2H1] Carboxylic_acid
  [F,Cl,Br,I] Halogen



- コピー元となるリストはメモ帳などのテキストエディタで作成できますが、Excel を利用することもできます。Excel で 1 列目に SMARTS、2 列目に構造パターン名を入力したリストを作成し、該当セルをコピーすることで、そのまま貼り付けることができます。

4. おわりに

alvaDesc 3.0 スタートアップガイドをご覧いただき、誠にありがとうございました。

ここでは、alvaDesc 3.0 の基本的な使い方について、具体的なデータに基づいた GUI 画面例で説明してきました。表示されている画面には更に発展的な的な使い方の為に用意されている機能もありますが、できるだけ分かり易くご紹介いたしました。それらの機能を活用される場合、ぜひ alvaDesc ユーザーマニュアルをご覧になり、理解を深めて頂きたいと思います。

また、alvaDesc は分子の特徴量を計算するソフトですが、開発元である Alvascience 社のソフトウェアスイートには、alvaDesc で計算された特徴量に基づいて QSAR/QSPR のモデルを自動生成する alvaModel/alvaRunner や、alvaDesc の一部の記述子や alvaModel で作られたモデルを使ってゼロから新しい分子をデノボ生成する alvaBuilder、分子データセットのキュレーションを行う alvaMolecule というソフトウェア群があります。いずれも無償の評価ライセンスをご提供していますので、ご興味があれば営業担当までご連絡ください。技術的な質問やご不明点がございましたら、弊社ウェブサイトの FAQ をご覧いただくかテクニカルサポート係までご連絡ください。

営業担当
sales@affinity-science.com

テクニカルサポート係
help@affinity-science.com

alvaDesc よくある質問（FAQ）
https://www.affinity-science.com/alvadesc-faq/

なお、開発元のホームページには alvaDesc をはじめとする Alvascience 社ソフトが使用され、引用されている文献のリストも掲載されています。alvaDesc などがどのように使われているか、ご参考にしていただければと思います。

Citations - Alvascience
https://www.alvascience.com/citations/

ぜひ本ガイドの内容を実践の中で活用し、ご研究に役立てていただければ幸いです。

5. 参考情報

5.1 ヘルプ機能

[Help] メニューからは、alvaDesc ユーザーマニュアルを参照したり、バージョンやアップデート情報を確認したりすることなどができます。

❑ [Help > View help]

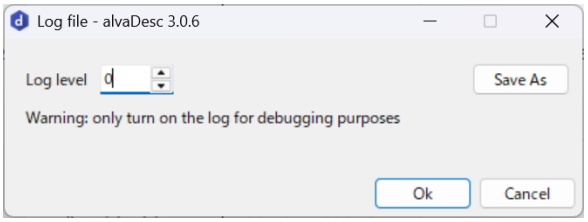
alvaDesc ユーザーマニュアルを表示します。ユーザーマニュアルは HTML 形式なので、ブラウザのお気に入りへ登録しておくと便利です。使い方の概略以外に、4.Molecules/5.Descriptors/6.Fingerprints/7.Structural patterns に記述子の定義などを含む理論的な解説が記載されています。

❑ [Help > Product license]

ライセンス情報が表示されます。

❑ [Help > Log file]

ログファイルを出力できます。デバッグの目的でのみログを有効にしてください、と注意書きが記載されている通り、通常は [Log level: 0 (無効)] に設定しておきます。ログレベルを変更する（ログを有効にする）場合は、[Log level]の値を変更し、[Ok] をクリックします。ログファイルを出力する場合は、[Save As] を選択します。



❑ [Help > About alvaDesc]

現在のバージョン情報を確認することができます。[Check for updates] にチェックを入れておくと、アップデート版がある場合にインストーラのダウンロードリンクが表示されます。（メジャーアップデートの場合には、ダウンロードリンクが表示されず、リセラー又は Alvascience 社にコンタクトするよう表示されます。）

[How to cite alvaDesc:] の記載は、論文などに alvaDesc を引用する際に使用してください。

5.2 分子記述子について

❑ 分子記述子とは

分子記述子は、分子の性質や構造的特徴を数値化して表した特徴量で、定量的構造活性相関（QSAR）や定量的構造特性相関（QSPR）でよく使用されています。Todeschini と Consonni の文献では、次のように定義されています。『分子記述子とは、分子の記号的表現にエンコードされた化学情報を、有用な数値や標準化された実験結果に変換する論理的・数学的手順の最終結果である』（Todeschini, R., & Consonni, V., 2000, Handbook of Molecular Descriptors, Wiley）。

近年多くの分子記述子が提案されるようになったことから、分子記述子が従うべき基本的なルールが以下のように定義されました。alvaDesc の分子記述子はこれらのルールに準拠しています。

1. 原子のラベリングとナンバリングに対して不変であること
2. 分子の回転並進に対して不変であること
3. 明確なアルゴリズムによって定義されていること
4. 分子構造に関して明らかに定義された応用性があること

なお、同じ定義の分子記述子でも以下のような要因により記述子計算ツールにより異なる計算値が得られることがあります。

- 記述子の実装に使用されるフレームワークやツールキットの化学モデルの違い
- 芳香族性の認識に対する異なるアルゴリズムやアルゴリズムの解釈の違い
- 記述子の計算に関係する可能性のあるパラメーターや、記述子の計算中に使用される参照データ値 (原子半径、電気陰性度の値など) の違い
- バグに起因する違い

❑ 分子記述子の分類

分子記述子にはさまざまな分類方法がありますが、一般的な方法の一つとして、化学構造から得られる情報を次元別に分類する方法があります。

- 0 次元記述子

  原子の結合に関する情報を考慮しない分子表現から得られる記述子です。

  例）分子量、原子タイプのカウント

- 1 次元記述子
  化学構造の全体ではなく一部のトポロジーを考慮します。
  例）官能基のカウント、原子中心のフラグメント、構造キー
- 2 次元記述子
  化学構造の 2D 表現から派生し、分子内の原子の組成や結合に関する情報を含みます。
  例）自己相関記述子、トポロジー指数
- 3 次元記述子
  分子グラフの 3D 表現を使用して計算され、原子間の結合だけでなく、原子が 3 次元空間内で占める位置も考慮します。
  例）WHIMs（Weighted Holistic Invariant Molecular descriptors）、GETAWAYs（Geometry, Topology, and Atom-Weights AssemblY descriptors）

❑ alvaDesc で計算可能な分子記述子

alvaDesc で計算可能な分子記述子は、ユーザーマニュアル（Help > View help）の[5. Descriptors] にブロックごとに説明が記載されています。また、[View > Descriptor list] からも記述子の一覧を表示することができます。開発元のウェブサイトにも記述子の一覧が記載されています（https://www.alvascience.com/alvadesc-descriptors/）。
記述子は現在 34 のブロックに分かれています。構造パターンを計算した場合は 35 番目に、外部変数を読み込んだ場合は 36 番目のブロックとして追加されます。

5.3 フィンガープリントについて

❑ フィンガープリントとは

フィンガープリントとは、分子構造における様々な特徴をベクトルに変換したものです。フィンガープリントは共通の部分構造を特定できるため、元々はデータベース検索を高速化するために導入されましたが、QSAR モデルの構築、特に kNN（K 近傍法）アプローチのような局所的類似性に基づくモデルの構築にも活用されています。
フィンガープリントは、構造キーとハッシュ化フィンガープリントの 2 種類に分類されます。alvaDesc では、構造キーのいくつかが記述子ブロックに含まれているほか、次の 3 種のフィンガープリントを計算できます。

構造キー
- MACCS 166 フィンガープリント

ハッシュ化フィンガープリント
- Extended Connectivity Fingerprints (ECFP and ECFPV3)
- Path FingerPrints (PFP)

❑ 構造キー

構造キーは、分子を区別するために準備された構造的特徴の集合として定義されます。alvaDesc では、様々な構造キーが含まれており、多くは記述子ブロックに含まれています（2D atom pairs ブロックの記述子や pharmacophore descriptors ブロックの CATS 2D 記述子など）。
また、alvaDesc では MACCS 166 フィンガープリントの計算が可能です。MACCS 166 フィンガープリントは、定義された 166 の構造的特徴の反映する固定サイズのブールベクトルです。ベクトルの各ビットは、特定の構造的特徴が分子内に存在するかどうかを 0 または 1 で示します。

❑ ハッシュ化フィンガープリント

ハッシュ化フィンガープリントは、事前に定義された構造的特徴リストを持たず、分子構造を探索してすべてのサブ構造を抽出します。サブ構造の数は事前に決まっておらず、ハッシュ関数を使って可変サイズのブールベクトルを固定サイズに変換します。
ハッシュ関数は決定的（deterministic）であるため、計算実行時のパラメーターが同じであれば、特定のフラグメントは常にフィンガープリント内の特定のビット集合に関連付けられますが、ハッシュ化されたフィンガープリントの

ビット集合から元のサブ構造を再現することはできません。
ハッシュ関数を利用することで、無数の構造的特徴のセットを固定長のベクトルに変換することができるというメリットがありますが、一方で、ビット衝突（異なるフラグメントがビットを共有する可能性がある）が発生するという欠点もあります。
ハッシュ化フィンガープリントの特徴として、他に、いわゆる「ダークネス (darkness)」と呼ばれるものがあります。ダークネスは、フィンガープリントの「1」のビットの割合を表します。データセットの平均的なダークネスの値が高いと類似が偽陽性となる可能性が高くなり、一方でダークネスが低いとフィンガープリントサイズを縮小できる可能性があることを示します。

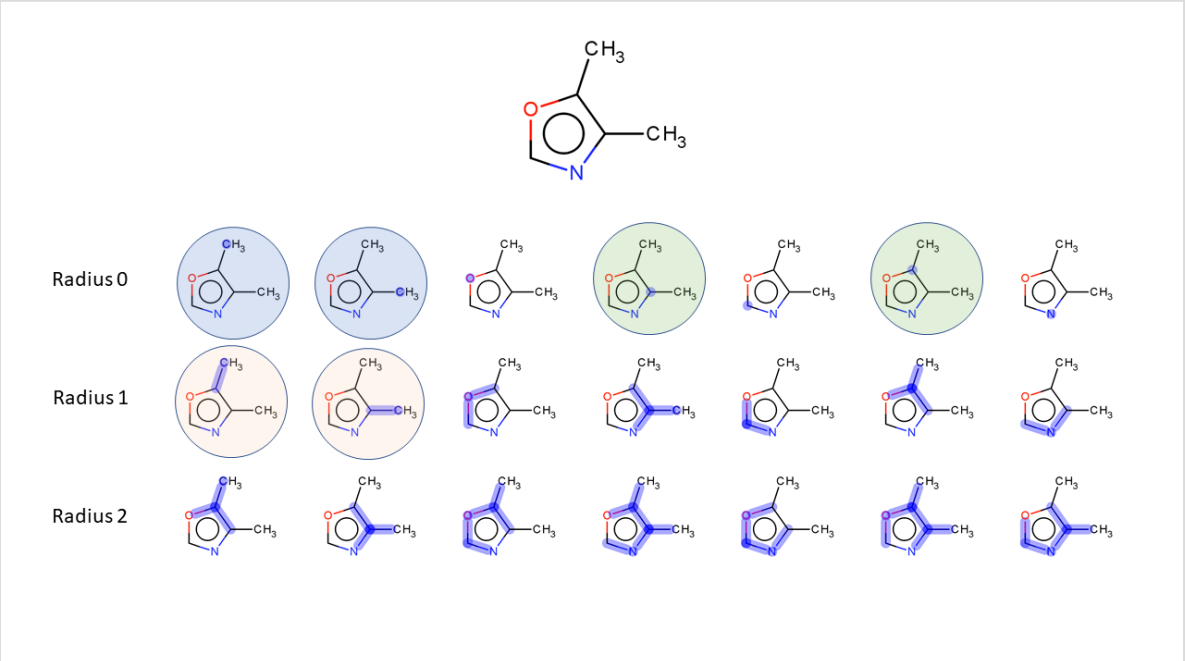
alvaDesc では、Extended Connectivity Fingerprints (ECFP、ECFPV3)と Path FingerPrints (PFP)の 2 つのハッシュ化フィンガープリントを計算できます。

- Extended Connectivity Fingerprints (ECFP、ECFPV3)

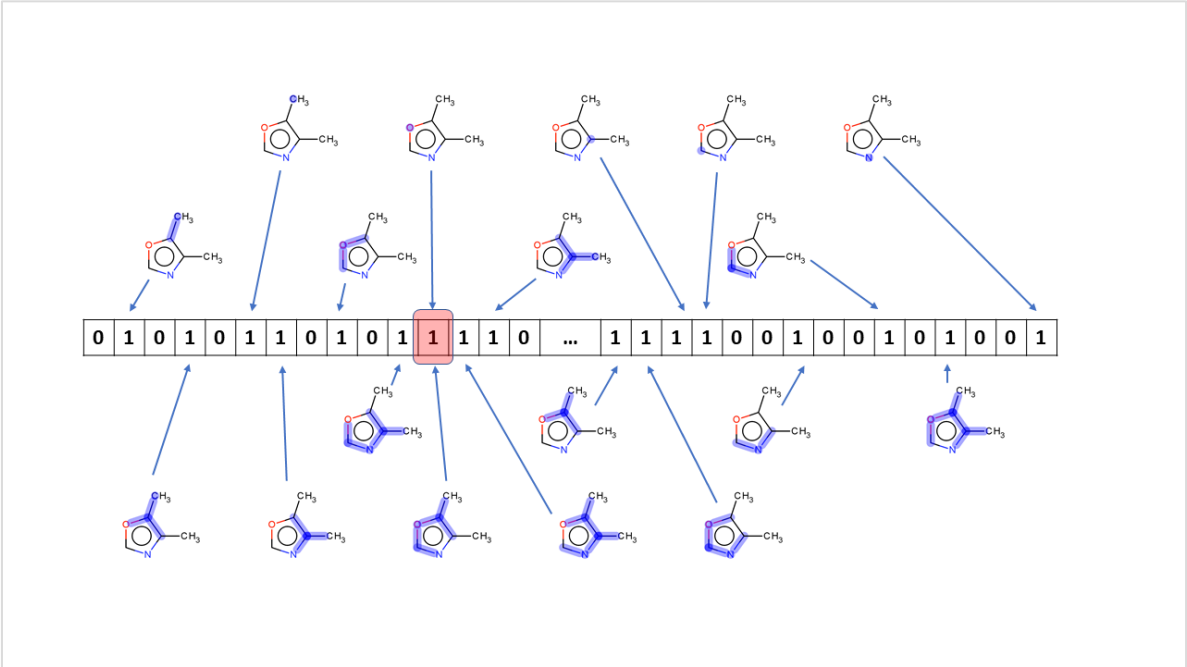
  ECFP と ECFPV3 の 2 つの異なる手法があります。ECFPV3 は alvaDesc バージョン 3.0 で新たに追加され、ECFP で利用できるオプションに加えて、同位体、立体化学、環構造の情報に関するオプションを利用でき、さらに、ECFP と比較して高速に処理を実行できます。

  Extended Connectivity Fingerprints (ECFP) は、水素以外の原子それぞれを中心として得られる円形フラグメントを、指定した半径（maximum length）まで列挙することで作成されます（分岐あり）。n を最大フラグメントの直径として、ECFPn と表すことができます。alvaDesc では[Maximum length] パラメーターが半径にあたるので、[Maximum length: 2] の場合は ECFP4 となります。

  次の図は、半径 0~2 の全フラグメント（青いハイライト部分）の識別例です。同じフラグメントは同じ色の円で表示しています。



  識別されたフラグメントは、以下のようにビットに表現されます。ハッシュ化フィンガープリントでは、赤い部分のように、ビット衝突が起こることがあります。

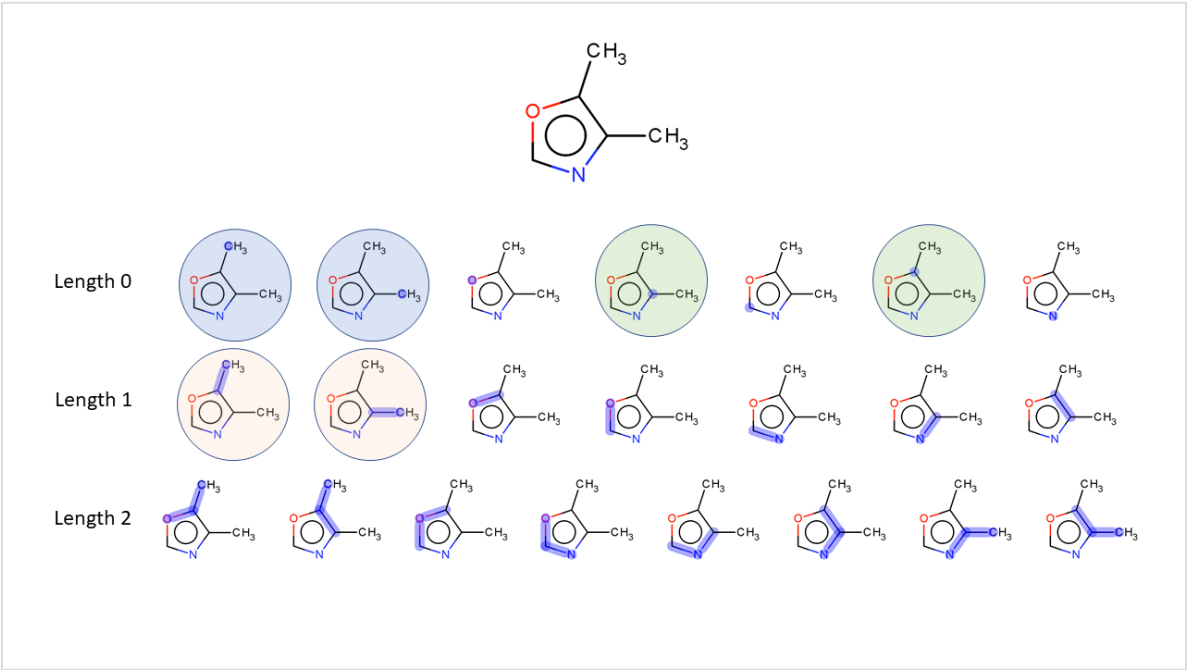


Extended Connectivity Fingerprints (ECFP) は、もともと構造活性モデリング専用に開発されましたが、現在ではハイスループットスクリーニング (HTS) やリガンドベースの仮想スクリーニング研究の分野など、さまざまな用途に広く使用されています。
共通の部分構造を識別する能力があるため、特に kNN アプローチや ADMET 特性の予測を含むリード最適化フェーズなどの局所類似性に基づく QSAR および QSPR モデルの構築にも使用されています。

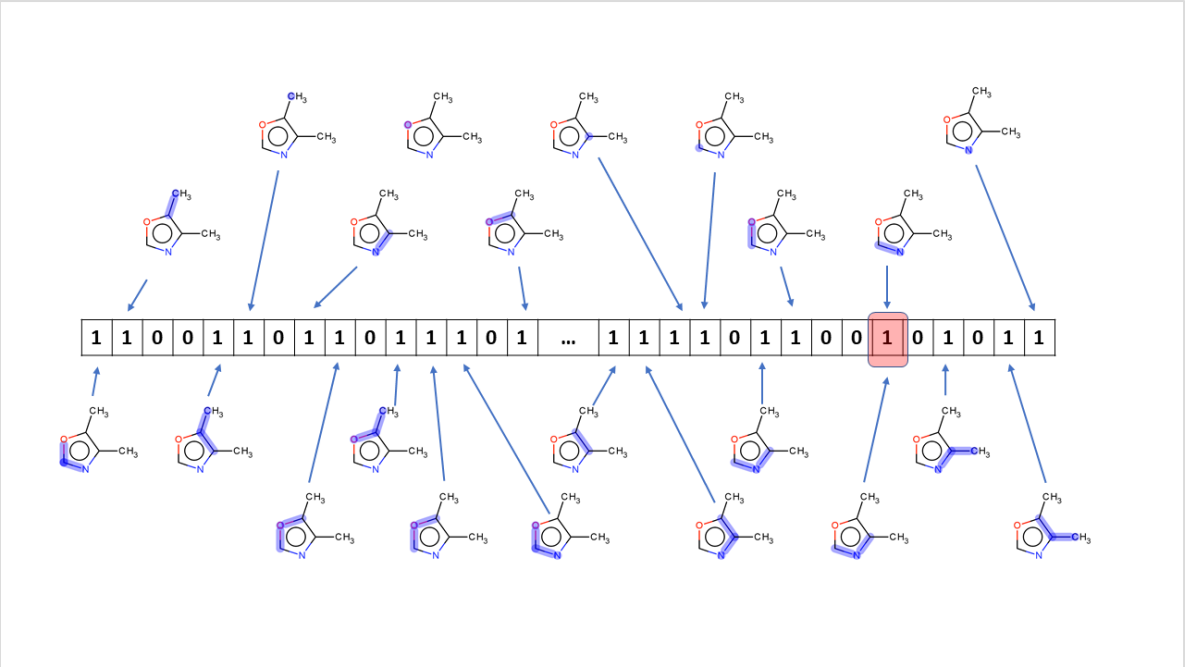
- Path FingerPrints (PFP)

  Path FingerPrints (PFP) は、分子内の水素以外の原子を起点として、指定された最大長（Maximum length）までのすべての直線的な経路（枝分かれなし）のフラグメントを系統的に列挙することで得られます。

  以下は、[Maximum length: 2] のフラグメント識別例です。



以下は、識別されたフラグメントのビット表現です。赤い部分はビット衝突です。



Path FingerPrints (PFP)は通常、完全な構造と部分構造の識別、類似性の検索、および化合物ライブラリの類似性分析を実行するために使用されています。

5.4 変数削減について

alvaDesc では多数の分子記述子を計算できるため、変数削減ツールを実装しています。alvaDesc の変数削減ツールには、空間充填設計から適応された変数削減手法である高速 V-WSP（Fast Variable Reduction Method Adapted from Space-Filling Designs）が含まれています。変数削減や V-WSP アルゴリズムについては、alvaDesc の FAQ ページに記載していますので、そちらもご参照ください。（https://www.affinity-science.com/alvadesc-faq-cat2/ [記述子を保存する際の変数削減（Variable reduction）オプションの詳細]）

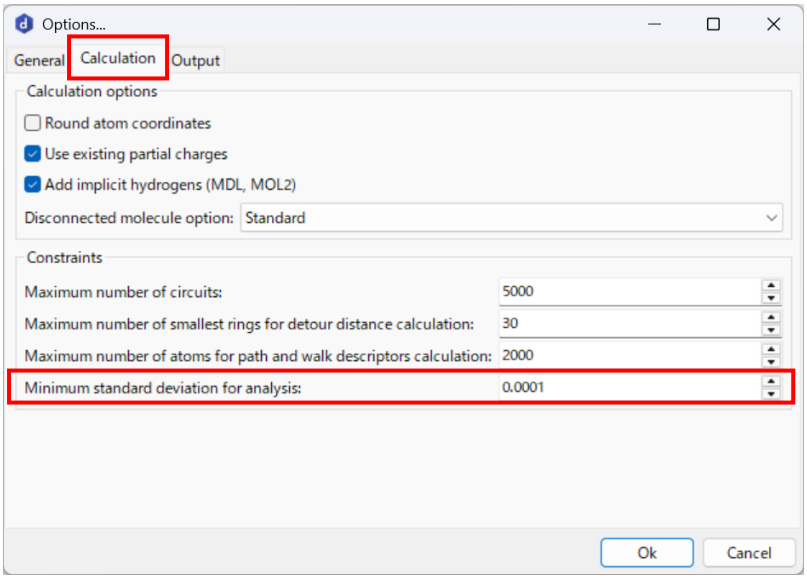
変数削減は、すべて同じ値の記述子や欠損値をもつ記述子を削除したり、記述子間の相関の閾値を設定したりすることで実行できます。計算した記述子の保存時に変数削減を実行できますが、この時に以下の設定が可能です。

- ❑ 全てが同じ値の記述子を削除
- ❑ 指定した割合以上が同じ値の記述子を削除
- ❑ 標準偏差が閾値より小さい（値のばらつきがとても少ない）記述子を削除
- ❑ ペア相関が閾値以上である記述子（の一方）を削除（※ 他の記述子との相関がより大きい方を削除）
- ❑ 欠測値（計算結果が n.a.）が 1 つでもある記述子を削除
- ❑ 全てが欠測値（計算結果が n.a.）である記述子を削除

ペア相関に基づいて変数削減を行う場合は、単に記述子を減らすだけでなく、データの冗長性や多重共線性を低減することができます。変数削減を行うことで、QSAR/QSPR モデルを構築する際により少数の意味のある記述子セットを出力することができます。

また、alvaDesc では、相関分析、主成分分析、t-SNE 分析時に、自動で変数削減が実行されるようになっています。ここでは、「全て定数／1 つを除いて他が定数／標準偏差が 0.0001 未満」 の記述子が削除されます。この標準偏差の閾値は、オプションから設定することができます（Settings > [Calculation]タブ > [Constraints]内、[Minimum standard deviation for

analysis])。



5.5 コマンドラインインターフェース（CLI）からの実行方法

CLI を利用することで、Windows のコマンドプロンプトなどのコマンドラインから alvaDesc のバッチ処理を実行できます。多量の分子を扱う場合や、他のアプリケーションと統合して使用する場合に役立ちます。
CLI を実行するには、alvaDescCLI 実行ファイルを使用します。デフォルトの保存先は以下になります。

OS	alvaDescCLI パス
Windows	C:¥Program Files¥Alvascience¥alvaDesc¥alvaDescCLI.exe
Linux	/usr/bin/alvaDescCLI
macOS	/Applications/alvaDesc.app/Contents/MacOS/alvaDescCLI

異なるディレクトリから実行する場合は、環境変数 Path を利用するとフルパスで指定せずに実行できるので便利です。

例 1) moleculars.smi に含まれる分子の MACCS166 フィンガープリントを計算する

```
> alvaDescCLI --input=molecules.smi --inputtype=SMILES --maccsfp
```

例 2) 標準入力に SMILES で入力された分子のすべての記述子を計算し、結果を output.txt に書き込む

```
> alvaDescCLI --input --inputtype=SMILES --descriptors=ALL --output=output.txt
```

例 3) alvaDescCLI のヘルプを表示する

```
> alvaDescCLI --help
```

また、スクリプトファイルを作成し、CLI から実行することで、様々なオプションを利用した処理も可能です。
CLI の利用方法や引数、スクリプトファイルなど、詳細については alvaDesc ユーザーマニュアルの[3. Command-line Interface] をご覧ください。

5.6 SMARTS 記法について

SMARTS（SMiles ARbitrary Target Specification）は、Daylight Chemical Information Systems, Inc. によって開発された、分子構造のパターンや特性を記述するための言語（表記方法）です。分子の 2D 構造を文字列として記述可能な SMILES を拡張したもので、より柔軟な構造検索やパターンマッチングが可能です。

【特徴】

- SMILES をそのまま拡張したルール
  - SMILES のすべての記号やルールをそのまま利用可能
  - SMILES に加え、論理演算子や追加の分子記述子を含む

- 柔軟な構造パターンの記述
  - SMARTS は、特定の分子構造を厳密に指定することも、一般的なパターンとして表現することも可能
    例：メタン（CH_4）を表現する場合
    ・特異性の高い SMARTS：[CH4]
    →4 つの水素を持つ脂肪族炭素原子のみマッチ
    ・特異性の低い SMARTS：C
    →任意の数の水素を持つ脂肪族炭素原子にマッチ（エタン、エテン、シクロペンタンなどもマッチする）

SMARTS 構文の詳細については、alvaDesc ユーザーマニュアル 7.1 SMARTS Syntax、または Daylight 社のウェブページをご参照ください。

・SMARTS Tutorial https://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html

5.7 参考文献

- 開発元による alvaDesc に関する文献
  Mauri, A. (2020). alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. In K. Roy (Ed.), Ecotoxicological QSARs (pp. 801–820). Humana Press Inc. https://doi.org/10.1007/978-1-0716-0150-1_32

- t-SNE 分析に関する文献
  Hinton, G., & Roweis, S. (2003). Stochastic neighbor embedding. Advances in Neural Information Processing Systems

- 変数削減手法の V-WSP アルゴリズムに関する文献
  Ballabio, D., Consonni, V., Mauri, A., Claeys-Bruno, M., Sergent, M., & Todeschini, R. (2014). A novel variable reduction method adapted from space-filling designs. Chemometrics and Intelligent Laboratory Systems, 136, 147–154. https://doi.org/10.1016/j.chemolab.2014.05.010

- MACCS 166 フィンガープリントに関する文献
  Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL Keys for Use in Drug Discovery. Journal of Chemical Information and Computer Sciences, 42(6), 1273–1280. https://doi.org/10.1021/ci010132r

- Extended Connectivity Fingerprints (ECFP) に関する文献
  Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5), 742–754. https://doi.org/10.1021/ci100050t