1L3-4

準専門家クラウドソーシングによる化合物の合成可能性判定

Crowdsourcing Evaluation of Synthetic Accessibility

馬場 雪乃 *1　磯村 哲 *2　鹿島 久嗣 *1
Yukino Baba　Tetsu Isomura　Hisashi Kashima

*1京都大学大学院情報学研究科知能情報学専攻
Department of Intelligence Science and Technology, Kyoto University

*2株式会社地球快適化インスティテュート
The KAITEKI Institute, Inc.

A rapid method for the assessment of synthetic accessibility for a vast number of chemical compounds is expected to bring about a breakthrough in the drug discovery. Although several computational methods have been proposed, the compound evaluation has still been processed by medicinal chemists; however, the low throughput of the human evaluation due to the lack of chemists is a critical issue for handling a large number of compounds. We propose to use crowdsourcing for addressing this problem and we conducted an initial experiment to investigate the feasibility of incorporating semi-experts in the synthetic accessibility evaluation. Our experiment results show that we can obtain accurate assessments by statistical aggregation of the judgments from semi-experts.

1. はじめに

分子設計システムを用いた創薬では，合成可能性による化合物のスクリーニングが求められる．合成可能性の自動判定手法として，Molecular complexity に基づく手法 [Ertl 09]，Structure complexity に基づく手法 [Boda 07]，合成経路の設計システム [Pfoertner 03] の利用等が提案されているが，これらの手法は精度・計算時間の面で実用レベルには達していない．そのため合成可能性の判定は，専門家の直感や経験に基づく判断に未だ頼らざるを得ない．分子設計システムの進歩に従い多数の候補化合物が短時間で出力されるようになっても，人手によるスクリーニングがボトルネックとなり，創薬プロセスの高速化を目指す上での課題となっている．

クラウドソーシングのアプローチを利用することで，合成可能性判定の効率化が期待できる．クラウドソーシングとは，インターネット上で不特定多数の人に仕事を依頼する仕組みである．Amazon Mechanical Turk に代表されるクラウドソーシングプラットフォームの台頭をきっかけに，コンピュータ科学の分野では，従来は専門家によって行われていた画像や文章に対するアノテーション作業をクラウドソーシングで実施することで，アノテーション付きデータの大規模化が進められている．合成可能性判定においても，有機化合物の専門家以外，例えば有機化合物についてある程度の知識を有する準専門家に判定作業を依頼することで，スループットの向上が期待される．

専門家以外の人に合成可能性判定を依頼する場合，専門家に依頼するよりも判定の信頼性は低下する恐れがある．クラウドソーシングにおいては，回答の信頼性を向上するために複数人に同じ作業を依頼しその回答を多数決等で統合するという方法が広く用いられている．さらにまた，信頼性向上のために統計的な統合手法も複数提案されている．準専門家による合成可能性判定結果を統計的に統合することで，信頼性の高い判定結果を効率的に獲得できると考えられる．

我々は，合成可能性判定用のウェブインターフェースを実装し，準専門家に合成可能性判定を依頼し回答データを収集した．準専門家の回答データに対して統計的回答統合手法を適用し，準専門家の回答を統合することで専門家に匹敵する精度の判定結果を得られることを確認した．これまで，合成可能性判定の専門家間での一貫性の調査や [Takaoka 03, Lajiness 04]，専門家の判定基準の調査 [Kutchukian 12]，複数の専門家による判定結果の統合 [Oprea 09] などが行われてきたが，準専門家による合成可能性判定結果を収集し，統合時の精度を検討したのは本研究が初めてである．

連絡先: 馬場 雪乃, baba@i.kyoto-u.ac.jp

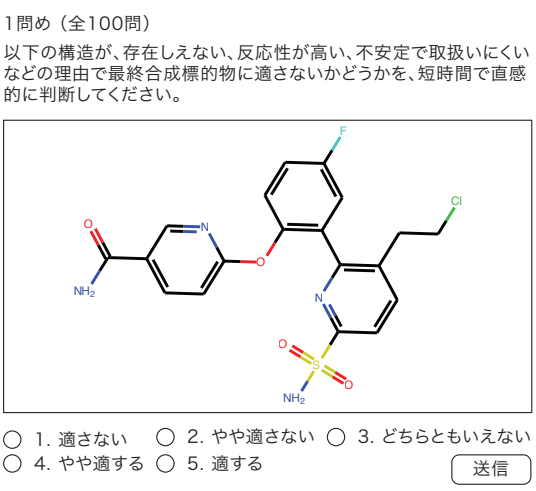


図 1: 合成可能性判定用のウェブインターフェース

2. 準専門家による合成可能性判定実験

合成可能性判定を準専門家に依頼し，その回答を統合することで合成可能性を正しく判定できるかを確認するために，実際に準専門家に判定作業を依頼し実験を行った．また，その判定精度を検討するために専門家と非専門家にも判定作業を依頼した．最終的に，以下の 3 群から成る合計 18 名の被験者が実験に参加した：(1) 創薬の「専門家」(5 名)，(2) 分子設計研究に携わる「準専門家」(9 名)，(3) 中枢薬理・行動薬理の研

究に携わる「非専門家」(4 名)．実験で用いる化合物として，以下の 2 群から成る 100 個の化合物を用意した：(1) 実在化合物（5 個），(2) 新規化合物（95 個）．実在化合物は，炎症関連蛋白質である COX-2 の阻害活性を有する化合物を公開データベースよりランダムに選択した．新規化合物は，COX-2 の阻害活性を有する化合物データセットをクエリとし，ランダムな部分構造置換により新規構造を発生させ作成した．

100 個の化合物それぞれについて，図 1 に示すウェブインターフェース上で合成可能性の判定を依頼した．被験者は，1（適さない）から 5（適する）までの 5 段階から一つを選択するよう依頼した．また，一つの化合物あたり数十秒〜数分で回答するように指示した．化合物が三つの群のどれに該当するのかは提示しなかった．インターフェースは一つの化合物の回答を送信すると次の化合物が出題されるように設計されており，回答の修正はできなかった．化合物の提示順は被験者ごとにランダムに並び替えた．

収集した回答に対して回答統合法を適用した．回答統合法は，各化合物 $i \in \mathcal{I}$ に対する各回答者 $j \in \mathcal{J}$ の回答の集合 $\{y_{ij}\}_{i,j}$ を入力として受け取る．$\mathcal{I}$ は化合物の集合，$\mathcal{J}$ は回答者の集合とし，$y_{ij} \in \{1, 2, \cdots, K\}$ とする．本実験では $K = 5$ である．出力として，化合物の合成可能性の予測結果の集合 $\{\hat{t}_i\}_i$ を返す．合成可能性の予測結果は $\hat{t}_i \in \{-1, 0, +1\}$ とし，$\hat{t}_i = -1$ は「合成標的物に適さない」，$\hat{t}_i = 0$ は「どちらともいえない」，$\hat{t}_i = +1$ は「合成標的物に適する」を表すとする．

回答統合法を 2 種類用意した．一つめは多数決である．多数決は，各化合物 i について回答の平均点 $\bar{y}_i = \frac{1}{J}\sum_{j=1}^{J} y_{ij}$ を計算し，以下のように各化合物 i に対する予測結果 $\hat{t}_i$ を出力する：

$$\hat{t}_i = \begin{cases} -1 & \text{if } \bar{y}_i < \frac{1+K}{2} \\ 0 & \text{if } \bar{y}_i = \frac{1+K}{2} \\ +1 & \text{if } \bar{y}_i > \frac{1+K}{2} \end{cases} \tag{1}$$

二つめはクラウドソーシングの回答統合に用いられている潜在段階クラス法である [Raykar 11]．この手法は，各回答者の信頼度も推定しながら回答統合を行う潜在クラス法 [Dawid 79] を段階ラベルに対して拡張したものである．この手法では，回答者 j の信頼性を二種類のパラメータ $\{\alpha_k^j\}_{k\in\{1,\cdots,K+1\}}$ と $\{\beta_k^j\}_{k\in\{1,\cdots,K+1\}}$ で表現する．ここで，化合物に対して，合成可能性の真の正解 $t_i \in \{-1, +1\}$ が割り当てられているものとする．ただし，真の正解は観測できない．α_k^j は，化合物 i の合成可能性の正解が $t_i = +1$ であるときに，回答者 j が k 以上の評点を与える確率であり，つまり：

$$\alpha_k^j = \Pr\left[y_{ij} \geq k \mid t_i = +1\right] \tag{2}$$

である．同様に，β_k^j は，化合物 i の合成可能性の正解が $t_i = -1$ であるときに，回答者 j が k 未満の評点を与える確率である：

$$\beta_k^j = \Pr\left[y_{ij} < k \mid t_i = -1\right]. \tag{3}$$

α_k^j は回答者 j の「感度」, つまり合成標的物に適する化合物に高い評点を与える能力を表現している．同様に β_k^j は，「特異度」，つまり合成標的物に適さない化合物に低い評点を与える能力を表現している．ここで，$\alpha_1^j = 1, \beta_1^j = 0$ であり，$\alpha_{K+1}^j = 0, \beta_{K+1}^j = 1$ である．このようなパラメータを導入すると，回答者 j の化合物 i に対する回答 y_{ij} が $k \in \{1, \cdots, K\}$ である確率は次式で表される：

$$\Pr\left[y_{ij} = k \mid t_i = +1\right] = \alpha_k^j - \alpha_{k+1}^j \tag{4}$$

$$\Pr\left[y_{ij} = k \mid t_i = -1\right] = \beta_{k+1}^j - \beta_k^j. \tag{5}$$

EM アルゴリズム [Dempster 77] を用いて，入力の回答集合 $\{y_{ij}\}_{i,j}$ から $\{\alpha_k^j\}_{j,k}$，$\{\beta_k^j\}_{j,k}$，$\{t_i\}_i$ を推定する．EM アルゴリズムは二つのステップから構成され，E ステップでは，$\{\alpha_k^j\}, \{\beta_k^j\}$ を現在の推定値に固定した上で $\{t_i\}$ の期待値を計算する．M ステップでは，期待値を用いて $\{\alpha_k^j\}, \{\beta_k^j\}$ を推定する．

推定結果を用いて，以下のように各化合物 i に対する予測結果 $\hat{t}_i$ を出力する：

$$\hat{t}_i = \begin{cases} -1 & \text{if } \Pr\left[t_i = +1 \mid \{y_{ij}\}_{i,j}\right] < 0.5 \\ 0 & \text{if } \Pr\left[t_i = +1 \mid \{y_{ij}\}_{i,j}\right] = 0.5 \\ +1 & \text{if } \Pr\left[t_i = +1 \mid \{y_{ij}\}_{i,j}\right] > 0.5. \end{cases} \tag{6}$$

3. 結果

被験者 18 名から 100 個の化合物に対する合成可能性の判定結果が 1,800 件集まった．準専門家・非専門家の回答の正しさを評価するため，まず専門家の回答から判定結果の正解を作成した．専門家全員の合意が得られている場合は合意解を正解とし，合意が得られない場合は「正解なし」とした．具体的には，ある化合物 i に対して「全員が 3 点以上と評価している」場合は，正解を $t_i = +1$,「全員が 3 点以下と評価している」場合は正解を $t_i = -1$ とし，それ以外の場合は正解なしとした．正解を多く集めるために，専門家に対しては一度それぞれが回答した後に，他者の回答の分布をフィードバックした上で二度目の回答を依頼した．1 回目では合意解が得られた化合物は 34 件だったが，2 回目では 62 件であった．以降の評価では，2 回目の回答から作成した正解を用い，合意解が得られた化合物 62 件だけを対象にして分析を行う．

正解を利用して，各被験者の正答率を計算する．まず，多数決の場合と同様に $\frac{1+K}{2}$ を閾値として，各被験者の予測結果 $y_{ij} \in \{1, \cdots, K\}$ を $\hat{t}_{ij} \in \{-1, 0, +1\}$ へと変換する．そのうえで，各被験者 j の正答率 a_j を次式により評価した：

$$a_j = \frac{1}{|\mathcal{I}_t|}\left(\sum_{i\in\mathcal{I}_t} \mathbb{1}\left(\hat{t}_{ij} = t_i\right) + \frac{1}{2}\sum_{i\in\mathcal{I}_t} \mathbb{1}\left(\hat{t}_{ij} = 0\right)\right). \tag{7}$$

ここで $\mathcal{I}_t$ は，合意解が得られた化合物の集合である．

表 1，表 2 に各被験者区分ごとの平均正答率を示す．被験者区分ごとの平均正答率は，高い方から専門家，準専門家，非専門家の順になっており，直感に沿った結果になっている．専門家の中でも正答率のばらつきはあり，最も正答率が高い人は正答率が 95% を超えているが，70% 程度の人も 5 名中 2 名いた．この 2 名は特に実在化合物に対する正答率が低いことから，今回扱った化合物の種類に慣れていなかったために正答率が低かったと考えられる．準専門家は，専門家以上に正答率のばらつきが大きかった．80% を超える人が 9 人中 2 人いる一方，50% を下回る人も 3 人いた．また非専門家は，正答率は全員 50% 前後だった．

表 3 に，準専門家の回答と非専門家の回答に対して回答統合手法を適用した際の正答率を示す．統合手法の正答率も，式 (7) を用いて算出した．多数決では，準専門家の回答を統合しても正答率は 78.2% で専門家の下位二人を多少上回る程度であるが，潜在段階クラス法を適用することで正答率は 90% を超え，専門家の上位者に匹敵する．これは潜在段階クラス法により回答者の信頼性を推定しながら回答統合を行うことで，能力の高い準専門家を見つけ出し，彼らに高い重みを与えることに成功したためと考えられる．実用上の対象である新規化合物については，潜在段階クラス法では正答率 89.8% であり，こ

れは専門家 5 名中 3 名を上回っている．一方，統計的回答統合手法がいつも上手くいくわけではないことが非専門家の回答に結果から確認される．非専門家の場合には，多数決では正答率 50.8% だったが，潜在段階クラス法では正答率は 48.4% に低下してしまった．潜在段階クラス法は，正答率が高い回答者がある程度存在しなければ正解と能力を正しく推定できない手法であり，非専門家の場合には全員の正答率が低かったために効果が得られなかったと考えられる．

図 3 に，潜在段階クラス法による準専門家の感度 (α_k^j) と特異度 (β_k^j) の予測結果と実際の値を示す．実際の値は，感度については $\frac{\sum_{i\in\mathcal{I}_t}\mathbb{1}\left(t_i=+1\wedge y_{ij}\geq k\right)}{\sum_{i\in\mathcal{I}_t}\mathbb{1}(t_i=+1)}$，と特異度については $\frac{\sum_{i\in\mathcal{I}_t}\mathbb{1}\left(t_i=-1\wedge y_{ij}<k\right)}{\sum_{i\in\mathcal{I}_t}\mathbb{1}(t_i=-1)}$ として計算した．値に多少の誤差はあるものの，感度と特異度のいずれについても各準専門家の傾向を予測できていることが確認できる．例えば，SE2 には「合成標的物に適する」化合物に対してもほとんど 5 点を付けない傾向があり，SE4 には，「合成標的物に適する」化合物に対して 2 点以上，3 点以上，4 点以上を付ける頻度がほぼ同じという傾向がある．これらの傾向は，潜在段階クラス法によって，合成可能性の正解が未観測であっても正しく推定されており，そのため回答統合により高い正答率を達成することができた．

図 2 に，準専門家の人数を変えた場合の回答統合法の正答率の変化を示す．化合物ごとに準専門家 x 人の回答をランダムサンプリングし，多数決と潜在段階クラス法をそれぞれ適用した．これを 100 回繰り返し，正答率の平均と標準偏差を計算した．準専門家の数が 2 人の場合には潜在段階クラス法の平均正答率は 69.5% だが，人数を増やすと正答率は向上し，3 人の場合は 73.1%，5 人の場合は 80.2% である．

以上，準専門家の回答に対して統計的回答統合手法を適用することで専門家に匹敵する精度の判定結果を得られることを確認した．これにより，準専門家の活用による化合物スクリーニングの効率化が期待される．

表 **1:** 各被験者の正答率

被験者	被験者区分	実在化合物 (3 問)	新規化合物 (59 問)	全化合物 (62 問)
E1	専門家	1.000	0.949	0.952
E2	専門家	1.000	0.907	0.911
E3	専門家	1.000	0.831	0.839
E4	専門家	0.000	0.780	0.742
E5	専門家	0.333	0.737	0.718
SE1	準専門家	1.000	0.847	0.855
SE2	準専門家	0.667	0.864	0.855
SE3	準専門家	0.667	0.780	0.774
SE4	準専門家	1.000	0.661	0.677
SE5	準専門家	1.000	0.619	0.637
SE6	準専門家	0.167	0.576	0.556
SE7	準専門家	0.667	0.475	0.484
SE8	準専門家	0.667	0.458	0.468
SE9	準専門家	0.667	0.381	0.395
NE1	非専門家	0.000	0.619	0.589
NE2	非専門家	0.833	0.508	0.524
NE3	非専門家	0.167	0.475	0.460
NE4	非専門家	1.000	0.381	0.411

表 **2:** 被験者区分ごとの平均正答率

被験者区分	実在化合物 (3 問)	新規化合物 (59 問)	全化合物 (62 問)
専門家	0.667	0.841	0.832
準専門家	0.722	0.629	0.634
非専門家	0.500	0.496	0.496

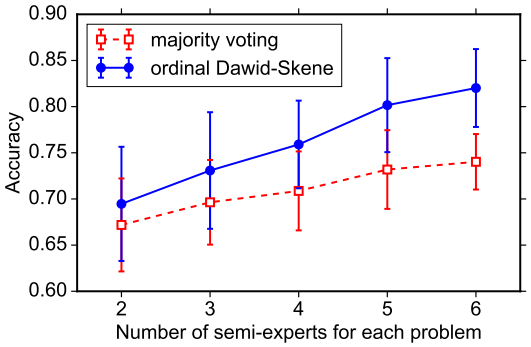


図 **2:** 準専門家の数を変化させたときの多数決 (majority voting) と潜在段階クラス法 (ordinal Dawid-Skene) の正答率の変化

4. むすび

化合物スクリーニングの効率化のため，クラウドソーシングによる合成可能性判定の方法を検討した．合成可能性判定を数十秒〜数分で実施するマイクロタスクにしても，専門家は高い精度で判定を行えることを確認した．また，複数の準専門家から得られた判定結果に対して統計的回答統合手法を適用することで，専門家に匹敵する精度で判定を行えることを確認した．以上の結果によりクラウドソーシングを活用することで，専門家による合成可能性判定の所要時間の短縮と，準専門家の活用による合成可能性判定のスループット向上が見込まれることを示した．化合物あたりの準専門家の人数を増やすほど回答統合法による精度は向上するが，人数を増やすと化合物あたりの処理速度は低下するため，精度とスループットの間にはトレードオフの関係がある．実用の際には，目的の精度とスループットに応じて，適切な人数を定める必要がある．

参考文献

[Boda 07] Boda, K., Seidel, T., and Gasteiger, J.: Structure and reaction based evaluation of synthetic accessibility, *Journal of Computer-Aided Molecular Design*, Vol. 21, No. 6, pp. 311–325 (2007)

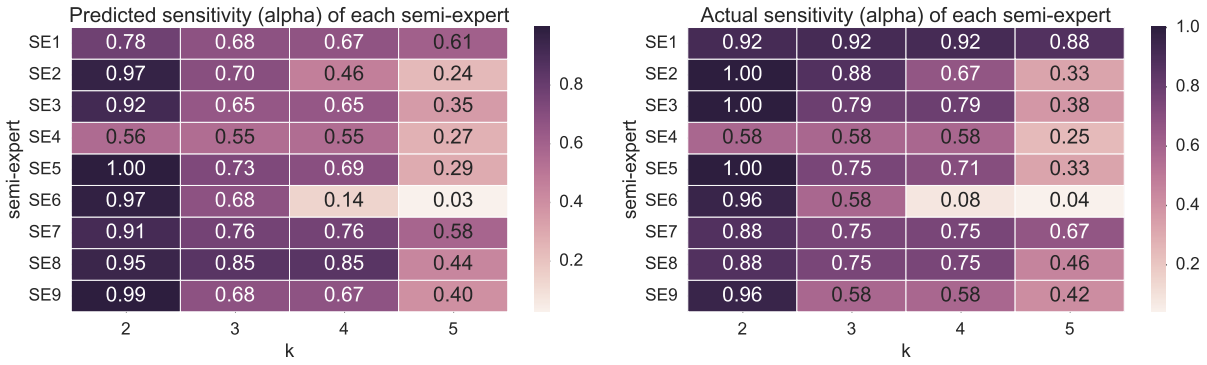
[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum likelihood estimation of observer error-rates using the EM algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statics)*, pp. 20–28 (1979)

[Dempster 77] Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38 (1977)

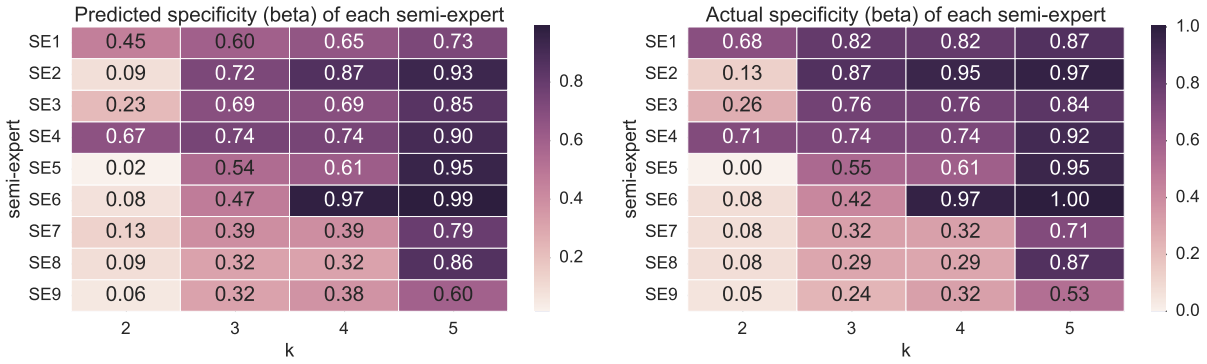
[Ertl 09] Ertl, P. and Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on

表 3: 回答統合手法の正答率

統合手法	対象被験者区分	実在化合物 (3 問)	新規化合物 (59 問)	全化合物 (62 問)
多数決	準専門家	1.000	0.771	0.782
多数決	非専門家	0.500	0.508	0.508
潜在段階クラス法	準専門家	1.000	0.898	0.903
潜在段階クラス法	非専門家	0.667	0.475	0.484



(a) 感度 (α_k^j) の予測結果（左）と実際の値（右）



(b) 特異度 (β_k^j) の予測結果（左）と実際の値（右）

図 3: 準専門家の感度と特異度の予測結果と実際の値

molecular complexity and fragment contributions, *Journal of Cheminformatics*, Vol. 1, No. 1, pp. 1–11 (2009)

[Kutchukian 12] Kutchukian, P. S., Vasilyeva, N. Y., Xu, J., Lindvall, M. K., Dillon, M. P., Glick, M., Coley, J. D., and Brooijmans, N.: Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery, *PLoS One*, Vol. 7, No. 11, p. e48476 (2012)

[Lajiness 04] Lajiness, M. S., Maggiora, G. M., and Shanmugasundaram, V.: Assessment of the consistency of medicinal chemists in reviewing sets of compounds, *Journal of medicinal chemistry*, Vol. 47, No. 20, pp. 4891–4896 (2004)

[Oprea 09] Oprea, T. I., Bologa, C. G., Boyer, S., Curpan, R. F., Glen, R. C., Hopkins, A. L., Lipinski, C. A., Marshall, G. R., Martin, Y. C., Ostopovici-Halip, L., et al.: A crowdsourcing evaluation of the NIH chemical probes, *Nature chemical biology*, Vol. 5, No. 7, pp. 441–447 (2009)

[Pfoertner 03] Pfoertner, M. and Sitzmann, M.: Computer-assisted synthesis design by WODCA, *Handbook of Chemoinformatics*, pp. 1457–1507 (2003)

[Raykar 11] Raykar, V. C. and Yu, S.: Ranking annotators for crowdsourced labeling tasks, in *Advances in Neural Information Processing Systems 24*, pp. 1809–1817 (2011)

[Takaoka 03] Takaoka, Y., Endo, Y., Yamanobe, S., Kakinuma, H., Okubo, T., Shimazaki, Y., Ota, T., Sumiya, S., and Yoshikawa, K.: Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition, *Journal of Chemical Information and Computer Sciences*, Vol. 43, No. 4, pp. 1269–1275 (2003)