医薬品原薬製造プロセス開発への機械学習の活用

明治大学大学院理工学研究科

２０２３年度

博士学位請求論文

医薬品原薬製造プロセス開発への機械学習の活用

Application of Machine Learning to the Development of Active Pharmaceutical Ingredients Manufacturing Processes

学位請求者　応用化学専攻

森下　敏治

目次

目次.....2
第 1 章　序論.....3
1.1　研究背景.....3
1.2　研究目的.....3
1.3　適用対象と検討課題.....4
1.4　本論文の構成.....5
1.5　参考文献.....6
第 2 章　自己促進分解温度の予測モデルの開発.....7
2.1　はじめに.....7
2.2　データセット.....8
2.3　記述子計算データセット.....32
2.4　予測モデル.....33
2.5　結果と考察.....35
2.6　まとめ.....46
2.7　参考文献.....47
第 3 章　クラスタリングを用いたベイズ最適化の初期条件の決定.....49
3.1　はじめに.....49
3.2　適応的実験計画法.....50
3.3　初期条件決定方法.....55
3.4　データセット.....58
3.5　結果と考察.....68
3.6　まとめ.....75
3.7　参考文献.....76
第 4 章　密度汎関数法(DFT)を用いたベイズ最適化探索性能向上.....78
4.1　はじめに.....78
4.2　データセット作成方法.....79
4.3　データセット.....80
4.4　ベンチマーク.....93
4.5　比較手法.....93
4.6　計算結果と考察.....94
4.7　まとめ.....114
4.8　参考文献.....114
第 5 章　結論.....115
謝辞.....116

第1章 序論

1.1 研究背景

一般的に新薬の研究・開発は十数年程度の期間と数百億円から数千億円の投資を必要とするが成功率は決して高くない。難病や希少疾患などのいわゆるアンメットメディカルニーズに対する治療薬を患者さんのもとへ届けるためには，低分子医薬だけでなく，抗体医薬をはじめ，核酸医薬や遺伝子治療薬，細胞医薬品のような様々なモダリティの医薬品開発が必要な状況であり，研究開発費は増加を続けている。このように日本だけでなくグローバルにおいても医療費の増加が大きな課題となっており，DX(Digital transformation)による業務効率化がすすめられている。製薬業界を取り巻く経営環境は目まぐるしく変化を続けており，研究・開発の難易度とリスク，開発コストは年々高まっている。

医薬品原薬の製造プロセス開発においても，開発スピードの加速化や検討の質の向上がこれまで以上に求められている。限られた開発期間の中で，環境，安全，品質，実験・製造コストなど，さまざまな観点で最適なプロセスの開発を行わなければならないが，研究開発の現場では研究者の勘と経験による試行錯誤で実験条件が決定されることも多い。医薬品原薬の製造プロセス開発に関しても，これまで以上に DX の推進および業務プロセスの大きな改革が求められている状況である。

1.2 研究目的

本研究では，医薬品原薬製造プロセス開発への機械学習の活用における課題の把握と対策の実行を行い，製造プロセス開発の効率向上による検討期間の短縮による上市までのスピードアップや開発コスト削減，安定生産へとつなげることを目的とする。

CMC(Chemistry, Manufacturing and Control)の原薬部門では，化学合成原薬およびバイオ原薬の製造プロセス開発，工業化研究，生産サイトへの技術移転や承認申請，製品化後のトラブル対応など，医薬品原薬のライフサイクルを通じて技術的な対応を実施している。製剤化研究，臨床試験をスケジュール通りに進めるためにはタイムリーな原薬供給が不可欠であり，高品質かつ効率的，堅牢な原薬製造プロセスを早期に構築することが求められている。そのようなプロセスを開発するために，研究者は効率的な実験を計画・実施し，その結果を分析機器で測定した後，測定結果を解析し，次の実験が必要かどうか，行う場合はどのような実験条件とすべきかの判断を繰り返す(図 1)。これらの作業は非常に属人的であり，手動操作も多く残されている状況である。また，反応予測，安全リスク，環境影響，品質・物性予測などに各種シミュレーション技術(量子化学計算，分子動力学(MD: Molecular

Dynamics)計算，流体解析(CFD: Computational Fluid Dynamics)，プロセスシミュレーションなど)が活用される場合もあるが，多様なモダリティに技術開発が追いついておらず，十分に活用できているとはいい難い状況である。予測モデルを用いて収率，物性，品質などを予測し，最良となる条件を少ない試行回数で探索することができれば，開発期間を大きく短縮させることができる。実験・分析・解析・判断のサイクルを，より迅速に，正確に，効率よく回すために，機械学習を用いた予測モデルの開発とベイズ最適化(BO: Bayesian Optimization)を用いた適応的実験計画法の活用を検討することとした。

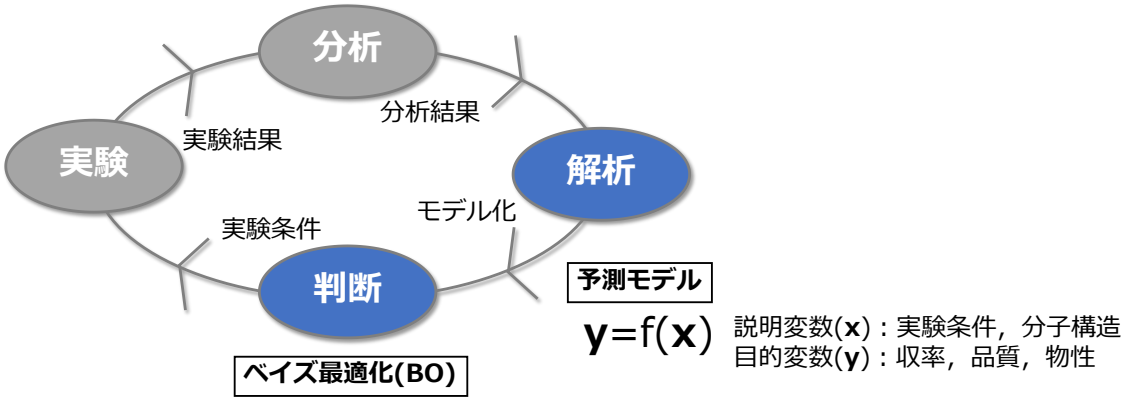


図 1　一般的な実験手順と機械学習の活用イメージ

1.3 適用対象と検討課題

機械学習予測モデルの開発対象として，有機過酸化物を対象とした自己促進分解温度(SADT: Self-accelerating Decomposition Temperature)予測モデルを構築することとした。熱的な危険性を把握することは化合物の開発や保管，輸送，製造等において非常に重要である。本モデル開発が医薬品開発の安定生産に貢献できると考え，従来法での課題であった計算負荷の削減と計算精度の向上を試みた。また，BO を用いた適応的実験計画法を有機合成反応条件探索に適用することとした。合成原薬の製造プロセスは，反応，分液，濃縮，抽出，晶析，ろ過，乾燥などの単位操作の組み合わせで構成される。反応にて目的物を得た後，その後の操作で不純物を除去して高い品質の化合物を得る。特に反応は最初に行われる最も重要な工程であり，この工程の完成度次第でその後の工程に要求されるレベルが大きく変化する。その中でも金属触媒を用いた反応は，特異的な挙動を示す場合があり，専門家でも反応挙動を正確に予測できないことが多い。最適条件を探索するためには網羅的にスクリーニングをする必要があり，大域的な最適解により少ない試行回数で到達することができる BO は，金属触媒反応における最適条件探索と相性が良いと考えた。一方で，BO を合成反応条件探索に適用する上で，以下のように様々な課題が残されていた。

- 初期条件の決定方法
- 化合物情報の記述子計算方法
- 多目的変数への対応
- 複数実験条件の提案
- スループット向上
- 研究者による実験空間の定義
- 計算リソース
- 利用者の心理的ハードル

多目的変数への対応, 複数実験条件の提案に関しては, 多くの解決方法が提案されており, 計算リソースの課題も含め, 計算アルゴリズムの改良や計算機のパフォーマンス向上により, 今後解決していくではないかと推察する。また, スループットに関してはロボットなどのオートメーションを活用することにより解決が可能である。研究者による実験空間の定義が必要な点に関しては, Generative Pre-trained Transformer-4(GPT4)などの生成AI(Artificial Intelligence)活用により, 実験空間の定義を含めた条件最適化の自動化ができるようになるかもしれない[1]。利用者の心理的ハードルは, 活用事例が増えて本技術への理解が深まれば下がっていくと思われる。BO を現場で活用する上で生じる数々の課題の中で, 初期条件の決定方法と化合物情報の記述子計算方法に絞り, 研究者のドメイン知識や記述子情報の適切な活用により, BO の最適解探索性能を向上することができないか可能性を模索した。性能評価のためには評価指標が必要となるが, BO の性能は初期サンプルの影響を大きく受け, 恣意的にサンプルを選択することはできない。そのため, 試行回数を十分に増やして, 探索結果の平均値やばらつきで評価を行うこととした。

1.4 本論文の構成

本論文は以下の 5 つのパートで構成される。

第 1 章「序論」では, 研究背景および研究目的, 本論文の構成を述べた。

第 2 章「自己促進分解温度の予測モデルの開発」では, 有機過酸化物の構造式から自己促進分解温度(SADT: Self-Accelerating Decomposition Temperature)を推算するモデルの構築を試みた。モデル構築時の前処理として分子力学 (MM: Molecular Mechanics)計算や密度汎関数(DFT: Density Functional Theory)計算, 変数選択に遺伝的アルゴリズム(GA: Genetic Algorithm)を用いたところ, 適用しない場合と比べて飛躍的に予測精度が向上した。

第 3 章「クラスタリングを用いたベイズ最適化の初期条件の決定法」では, BO の初期条件決定方法について考察を行った。BO で最適解を効率的に探索するには, ガウス過程回帰(GPR: Gaussian Process Regression)モデルを構築する際に適切な初期サンプルを用いる必

要がある。本研究では，目的変数に大きな影響を与える因子をカバーするクラスタリング情報に基づく初期サンプル選択手法を提案し，BO とのカップリング反応条件の最適化に適用した。クラスタを適切に形成し，各クラスタから初期サンプルを選択した場合，提案手法はランダムサンプリングや D 最適基準に基づくサンプリングよりも少ない実験回数で最適解に到達することを確認した。

第 4 章「DFT を用いたベイズ最適化探索性能向上」では，分子構造情報から密度汎関数理論(DFT: Density Functional Theory)計算で得られた記述子を利用する際に，ベイズ最適化の探索性能を向上させる方法について検討を行った。様々な組み合わせの基底関数・汎関数で DFT 計算された記述子を活用して，BO の探索性能を向上させることができる方法を開発した。本研究で提案した複数の記述子群を平均化する方法は，計算負荷が非常に小さく，量子科学計算に関する知識があまりない研究者でも簡単に利用可能である。

第 5 章「結論」では，本論文の内容に関するまとめを行った。

1.5 参考文献

1. Mahjour, B.; Hoffstadt, J.; Cernak, T. Designing Chemical Reaction Arrays Using Phactor and ChatGPT. Org. Process Res. Dev. 2023, 27, 8, 1510–1516.

第2章　自己促進分解温度の予測モデルの開発

2.1 はじめに

熱的な危険性を早期に把握することは化合物の開発や保管，輸送，製造等において非常に重要である。熱的危険性評価基準の一つに自己促進分解温度(SADT: Self-Accelerating Decomposition Temperature)と呼ばれる値がある。SADT とは有機過酸化物や自己反応性物質の自己加速分解が起こる可能性のある最低の温度と定義され，物質の貯蔵や輸送中の熱的危険を回避するために用いられる[2]。一般的には，各種熱分析から得られた情報をもとにリスク評価を実施するが，測定に必要な量は少なくないため，物量を確保できない開発初期は十分な評価を行うことができないことも多い。また，分析作業にはかなりの時間とコストを必要とする。化合物情報を用いた簡便で精度の高い SADT 予測法を開発することができれば，より安全，安定的に化合物を取り扱うことが可能となる。

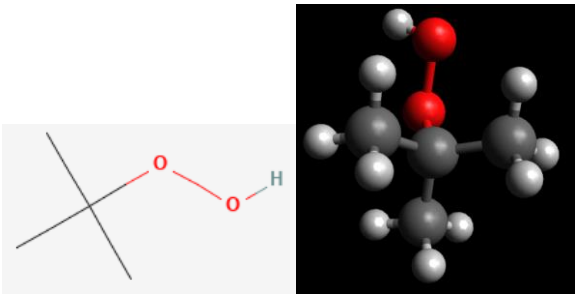
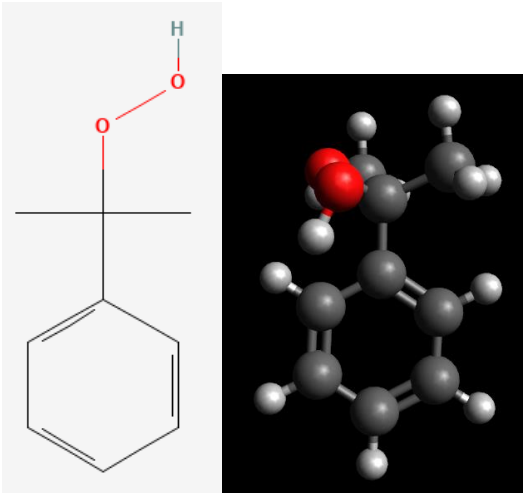
これまでの研究において，SADT を予測する構造物性相関(QSPR: Quantitative Structure-Property Relationships)モデルがいくつか提案されている。Wang ら[3]は Gaussian09 で DFT 計算(6-31G(d)/B3LYP)を行い，得られた記述子を用いて重回帰分析(MLR: Multiple Linear Regression)/サポートベクター回帰(SVR: Support Vector Regression)で SADT 予測モデルを構築した。量子化学的な記述子である最高被占軌道(HOMO: Highest Occupied Molecular Orbital)/最低空軌道(LUMO: Lowest Unoccupied Molecular Orbital)や結合乖離エネルギーを説明変数として加えている。また，HE ら[4]は前処理として半経験的分子軌道法(AM1)でジオメトリー最適化(Geometry Optimization)/振動計算(Frequency calculation)を行った分子構造を用いて DRAGON 6.0 で記述子を計算し，遺伝的アルゴリズム(GA: Genetic Algorithm)で変数選択を行って MLR/SVR で SADT 予測モデルを構築した。量子化学的な記述子は用いられていない。いずれの手法も比較的精度は良好であるものの，計算負荷，計算精度の面で実用上の課題が残されていた。

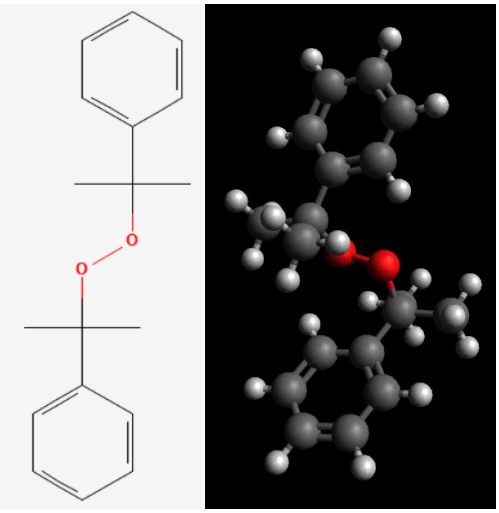
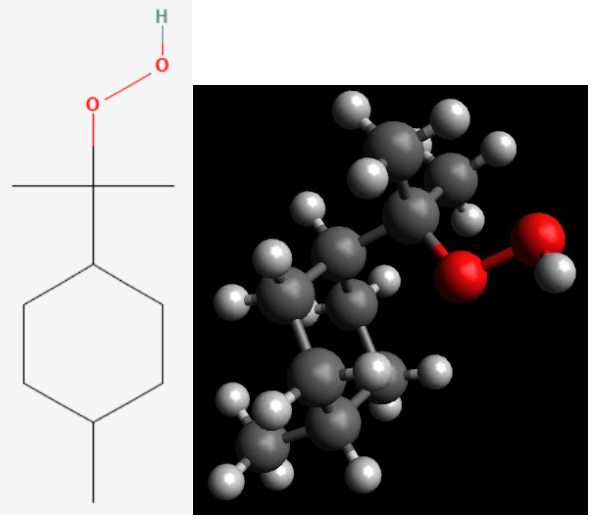
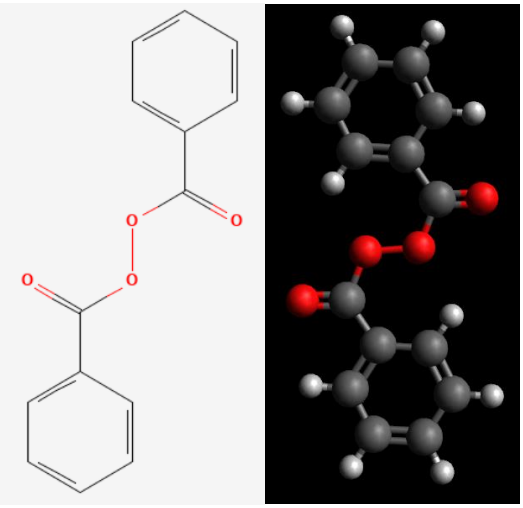
本研究では，有機過酸化物の構造式から SADT を推算するモデルの構築を行い精度向上と計算負荷削減を試みた。モデル構築時の前処理として分子力学(MM: Molecular Mechanics)計算や，基底関数を 6-31G，汎関数を B3LYP とした密度汎関数理論(DFT: Density Functional Theory)計算行い立体配座の最適化を実施した。分子記述子計算ソフトウェア alvaDesc 2.0.8[9]，CODESSA 3[10]を用いて記述子計算を実施した後，得られた記述子に対して GA-PLS(Genetic Algorithm-based Partial Least Squares)，GA-SVR(Genetic Algorithm-based Support Vector Regression)を適用して変数選択を行い，それぞれ部分的最小二乗回帰(PLS: Partial Least Squares)，SVR で予測モデルを構築して予測精度の検証を実施した。

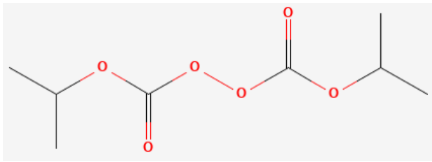
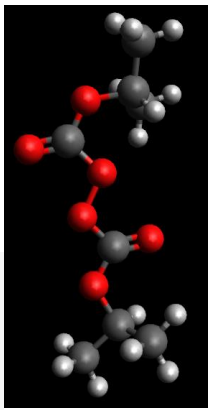
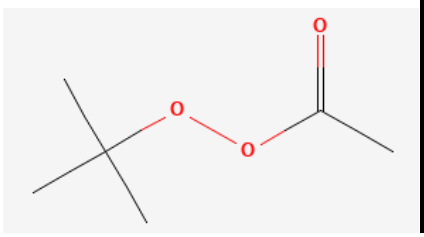
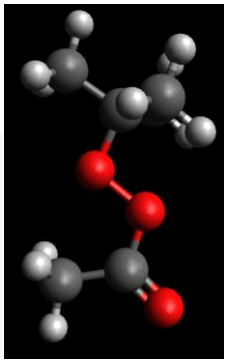
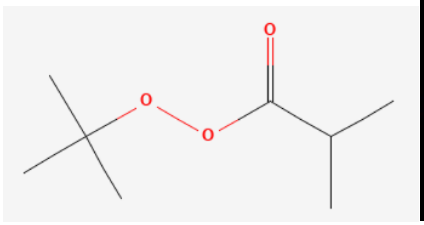
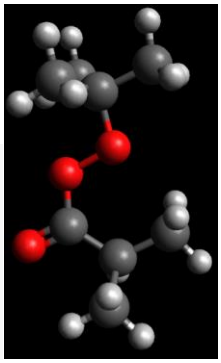
2.2 データセット

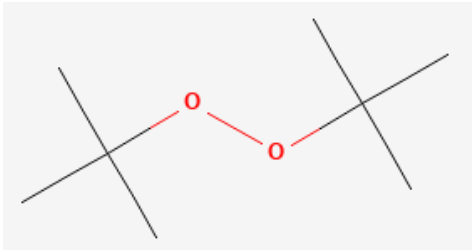
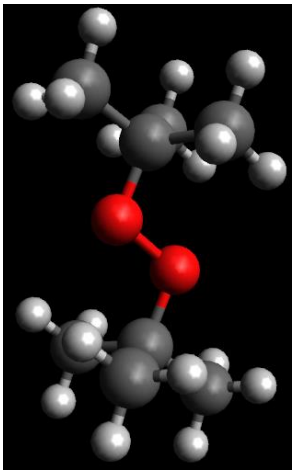
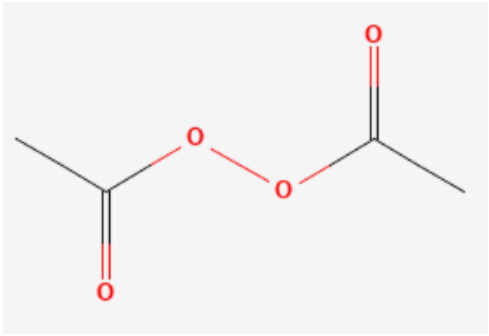
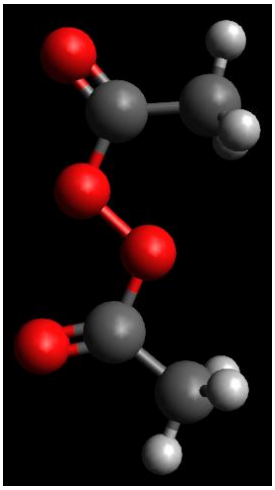
データセットには，文献[3,4,5]から得られた分子量 90.14 から 571.00，SDAT -5.0℃から 196.5℃の 65 種類の有機化合物が含まれ，異なる熱量測定法（TG-DSC と C80）を用いて決定された。以前に報告されたように，SADT は使用した測定方法に依存しない[5]。化合物には，ジアルキルペルオキシド，ジアシルペルオキシド，ヒドロペルオキシド，ペルオキシエステル，ケトンペルオキシド，ペルオキシカーボネート，ジペルオキシドなど，一般的に使用される有機過酸化物が含まれる。ここで使用した有機過酸化物とその実験的 SADT を表 1 に示す。データセットは論文[4]で報告された方法に従って，トレーニングセット（52 サンプル）とテストセット（13 サンプル）の 2 つのサブセットに分けて用いられた。

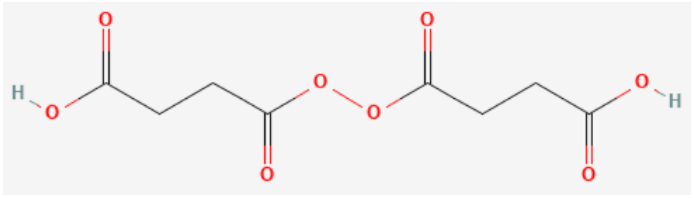
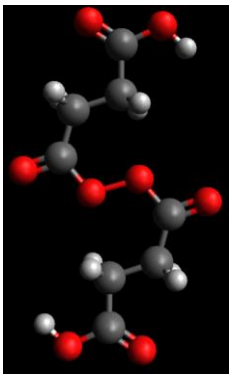
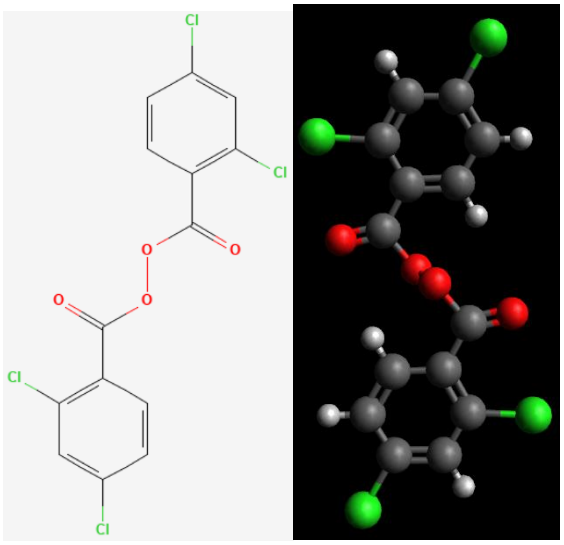
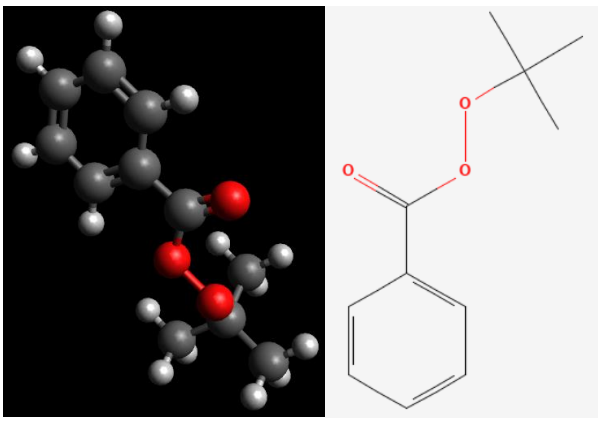
表 1　対象化合物および SADT 実測値

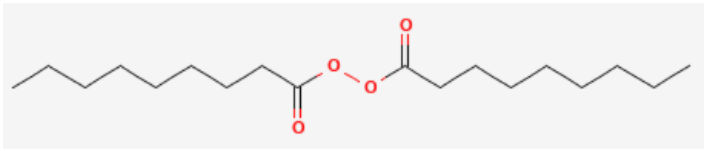
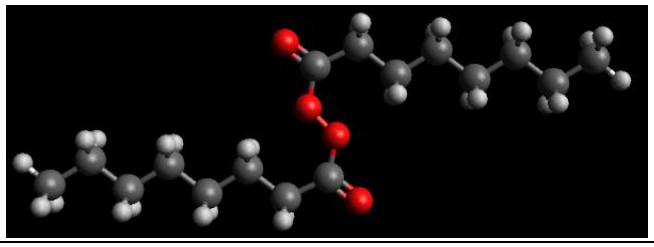
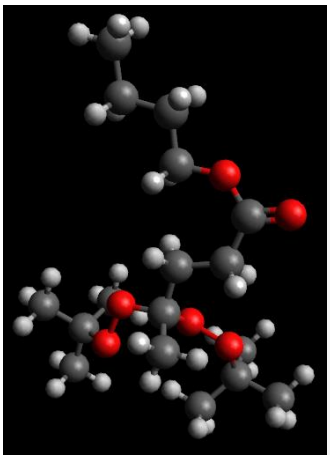
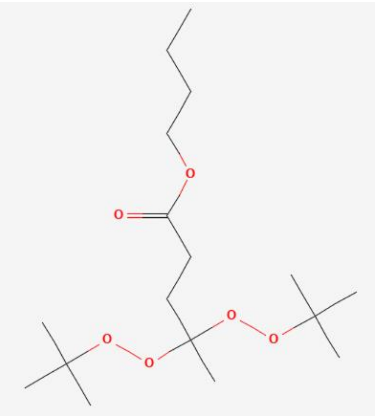
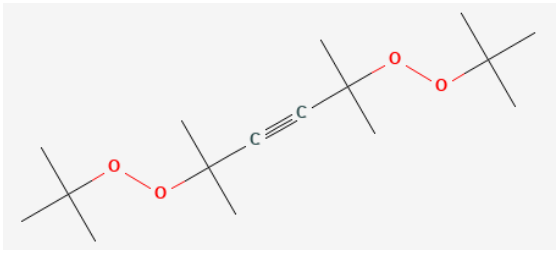
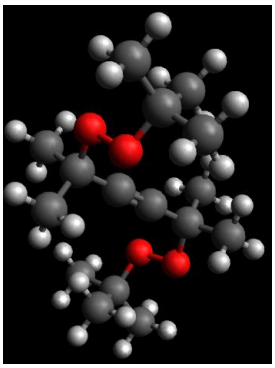
No.	Compound Name	CAS No.	MW	SADT [℃]
1	tert-Butyl hydroperoxide 	75-91-2	90.14	120.4
2	cumyl hydroperoxide 	80-15-9	152.21	79.0

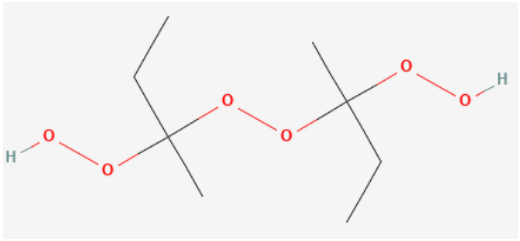
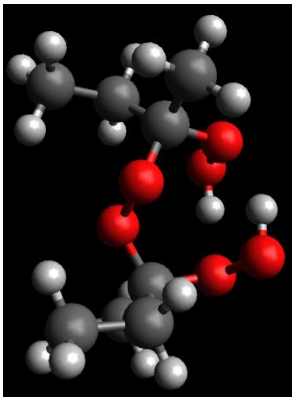
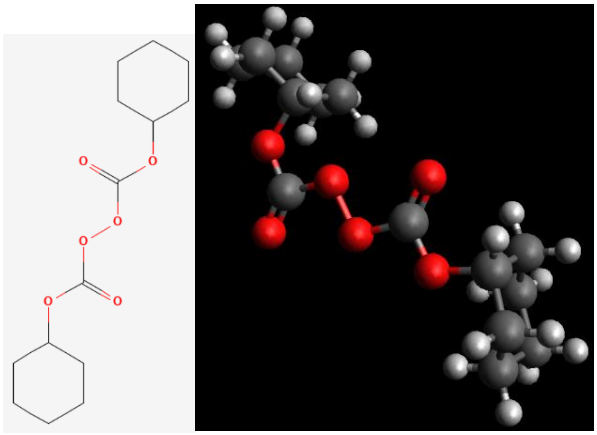
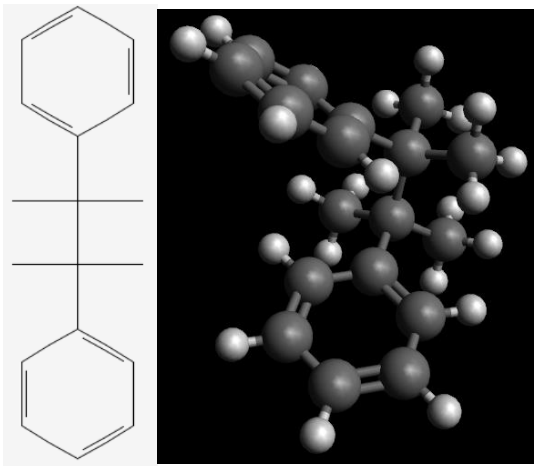
	dicumyl peroxide			
3		80-43-3	270.40	77.8
	p-Menthane hydroperoxide			
4		80-47-7	172.30	73.5
	Dibenzoyl peroxide			
5		94-36-0	242.24	80.0

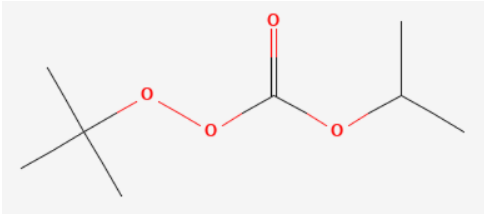
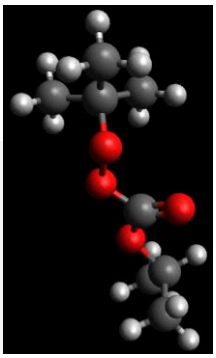
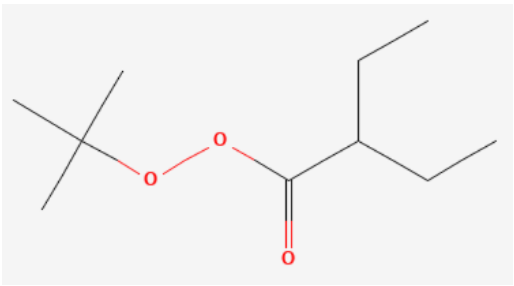
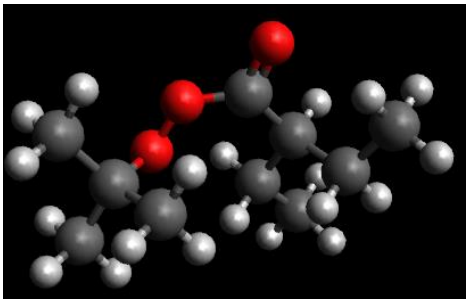
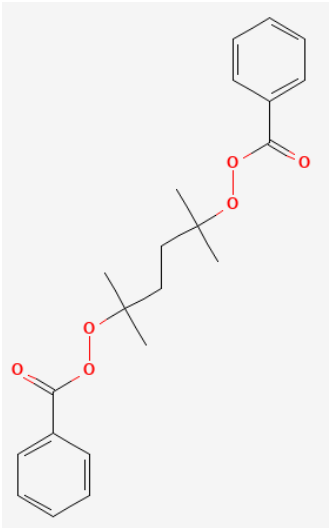
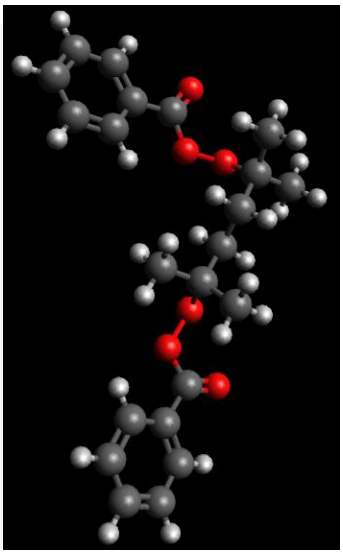
6	Diisopropyl peroxydicarbonate  		105-64-6	206.22	5.0
7	tert-butyl peroxyacetate  		107-71-1	132.18	65.0
8	tert-Butyl peroxyisobutyrate  		109-13-7	160.24	30.0

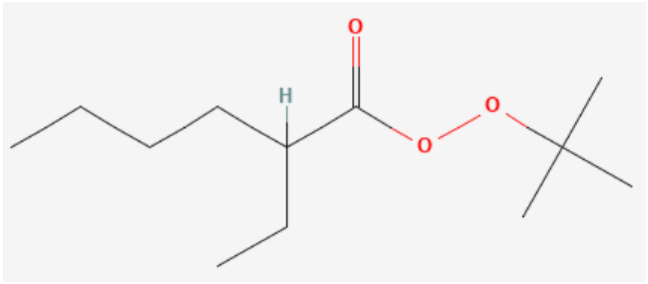
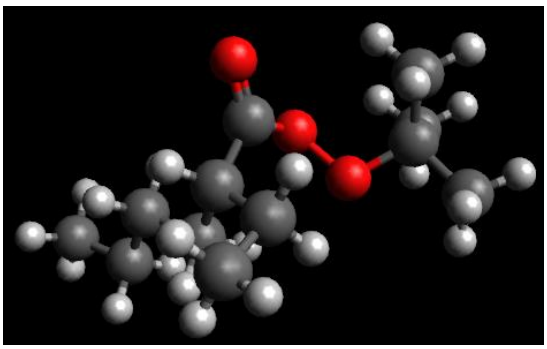
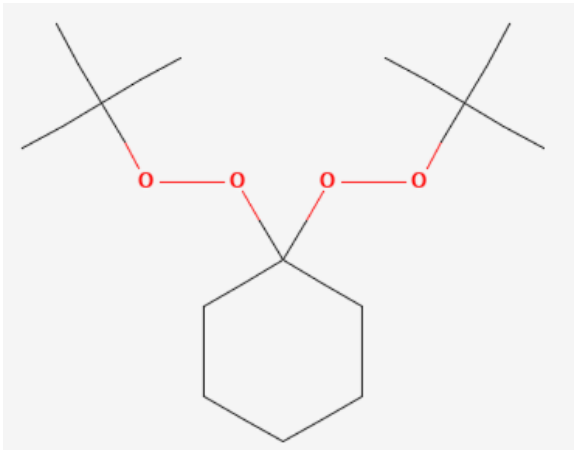
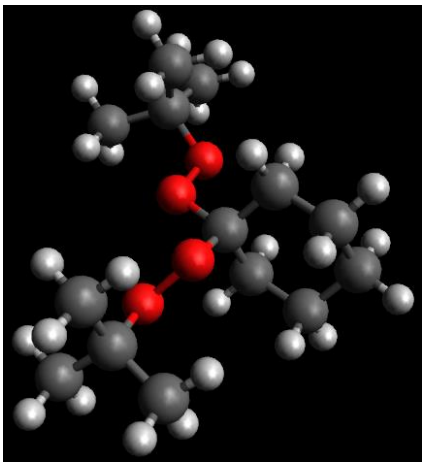
	di-tertbutyl peroxide  			
9		110-05-4	146.26	80.9
	Diacetyl peroxide  			
10		110-22-5	118.10	35.0

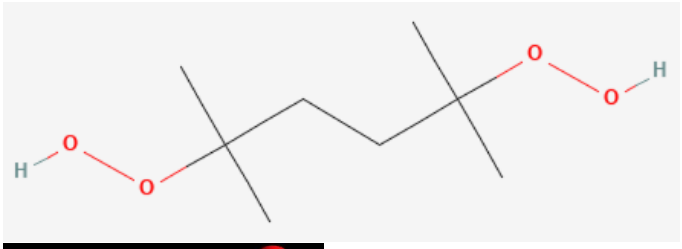
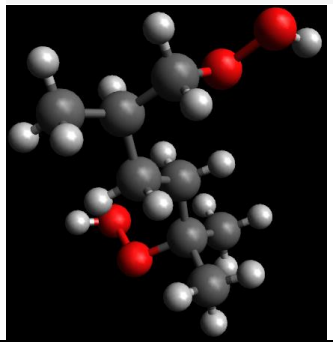
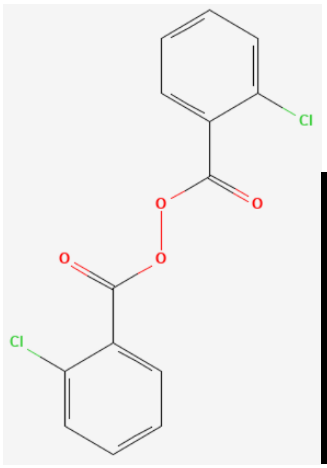
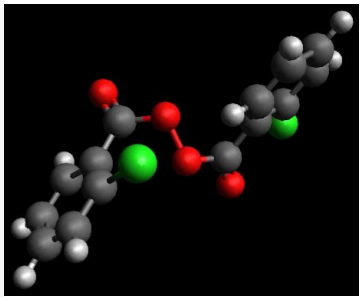
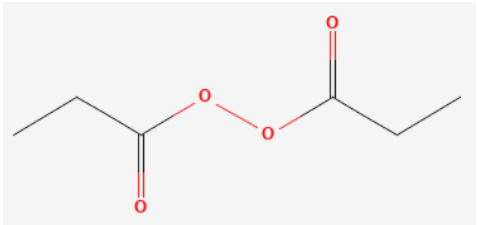
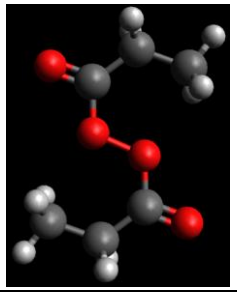
11	Disuccinic acid peroxide  	123-23-9	234.18	25.0
12	Bis(2,4-dichlorobenzoyl) peroxide 	133-14-2	380.00	60.0
13	tert-butyl peroxybenzoate 	614-45-9	194.25	65.8

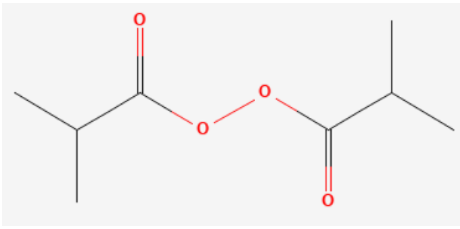
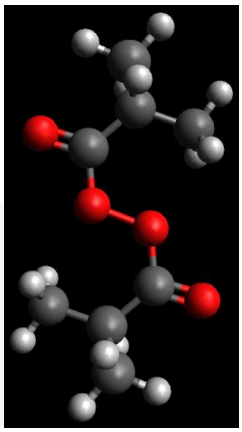
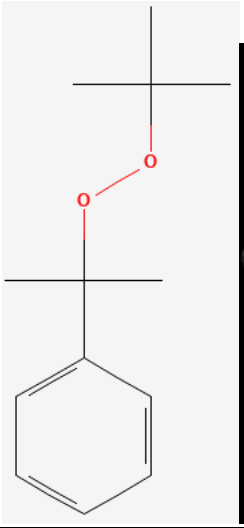
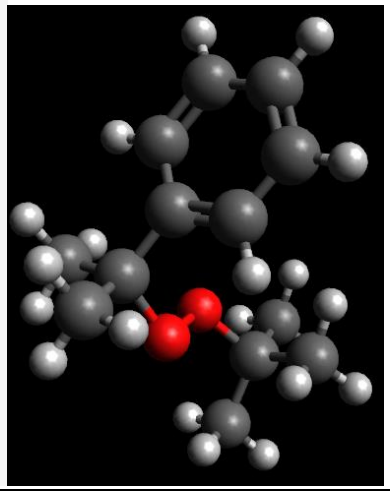
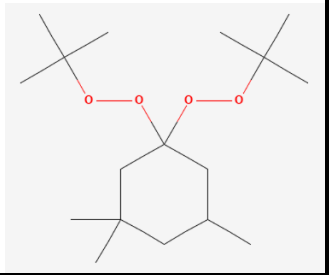
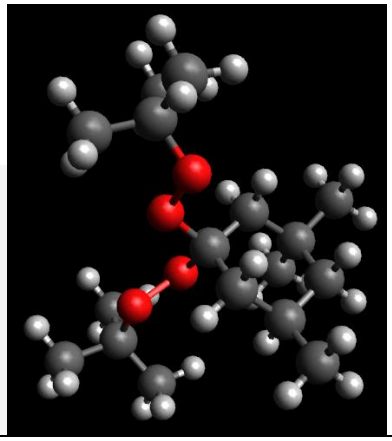
	Bis(1-oxononyl) peroxide			
14	 	762-13-0	314.52	20.0
	Butyl 4,4-bis(tert-butylperoxy) pentanoate			
15	 	995-33-5	334.51	55.0
	2,5-bis-(t-butylperoxy)-2,5-dimethyl-3hexyne			
16	 	1068-27-5	286.46	84.8

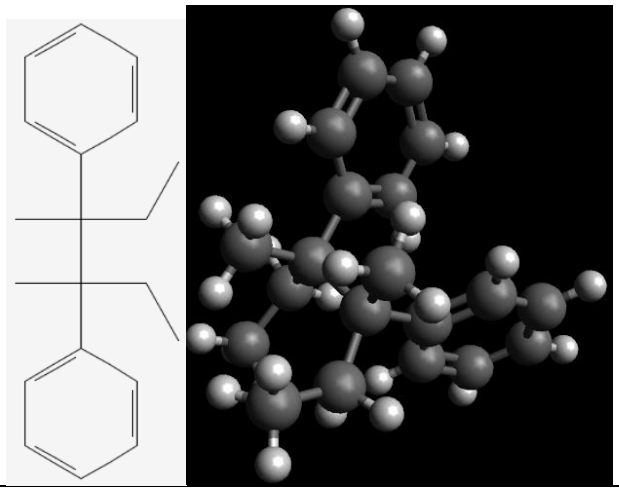
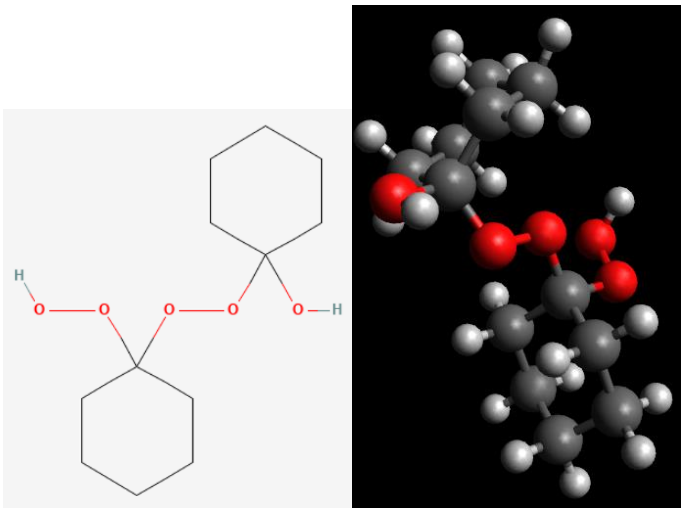
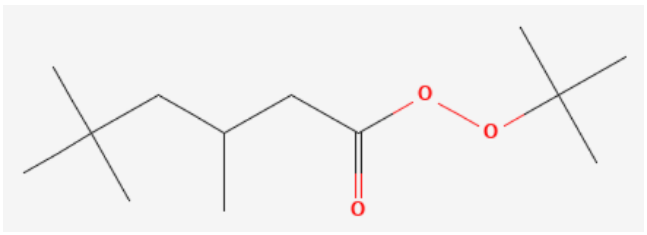
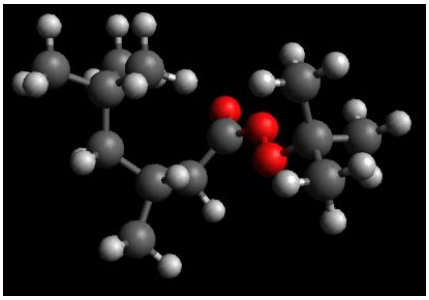
	Methyl ethyl ketone peroxide 			
17		1338-23-4	210.26	60.0
	Dicyclohexyl peroxydicarbonate 			
18		1561-49-5	286.36	25.0
	2,3-dimethyl-2,3-diphenylbutane 			
19		1889-67-4	238.40	196.5

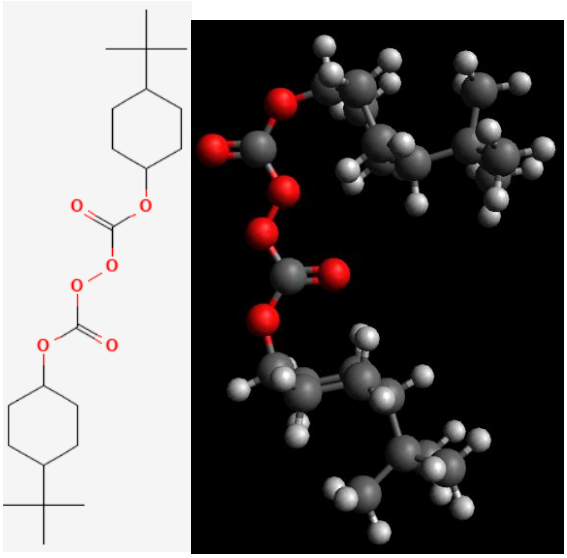
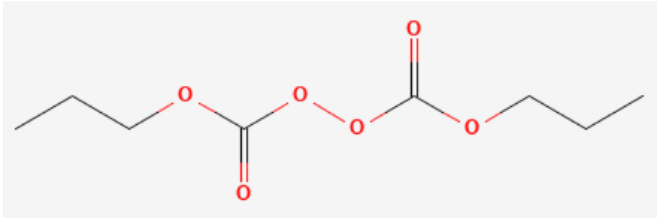
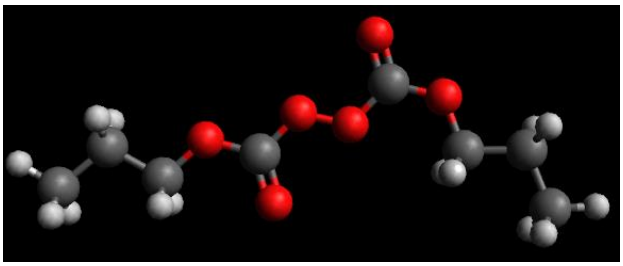
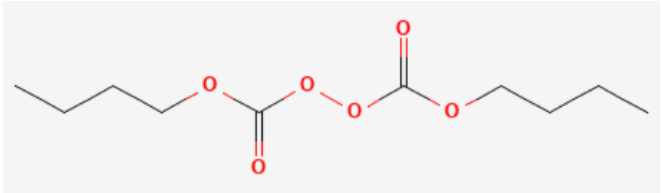
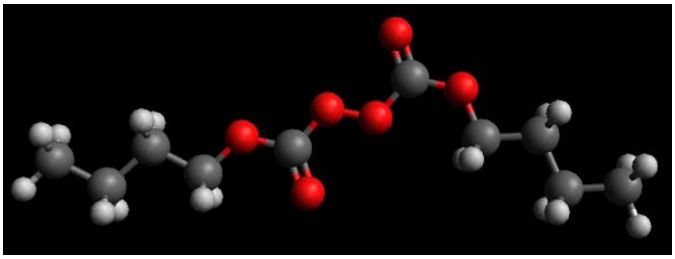
	tert-butyl peroxy isopropylcarbonate				
20			2372-21-6	176.24	62.2
	tert-butyl peroxy diethylacetate				
21	 		2550-33-6	188.30	35.0
	2,5-Dimethyl-2,5-di-(benzoylperoxy)hexane				
22			2618-77-1	386.48	69.0

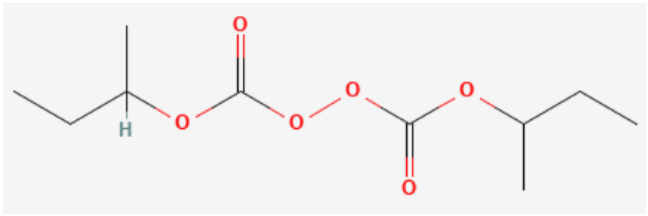
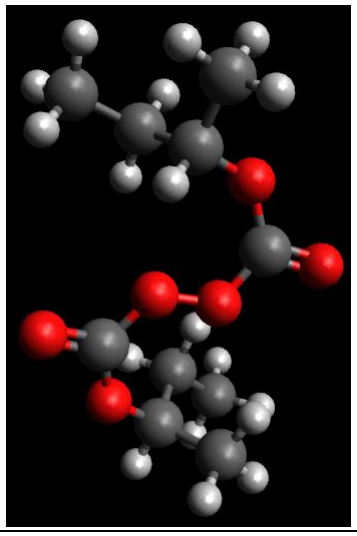
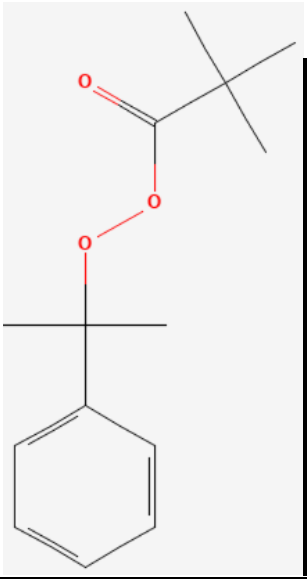
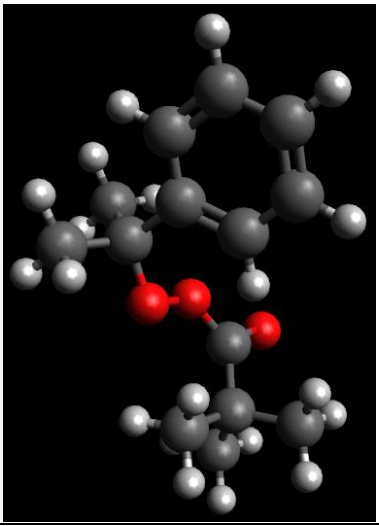
	tert-butyl peroxy-2-ethylhexanoate  				
23			3006-82-4	216.36	35.0
	1,1-bis-(tertbutylperoxy)cyclohexane  				
24			3006-86-8	260.42	60.0

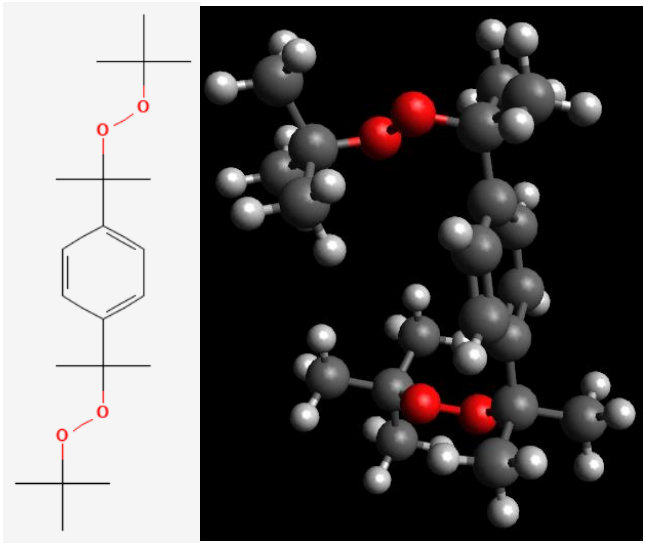
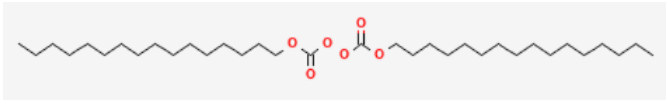
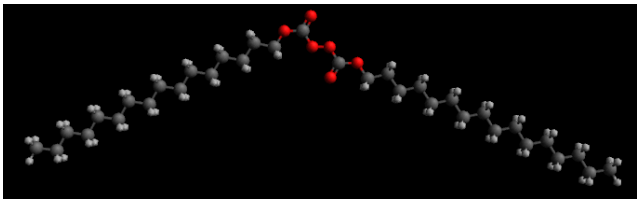
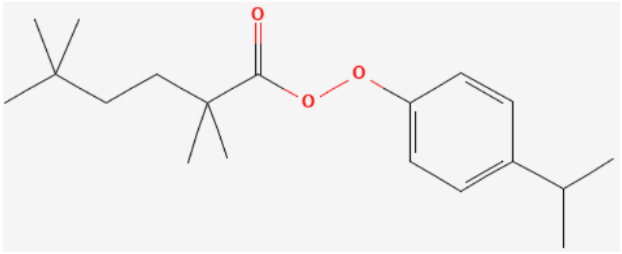
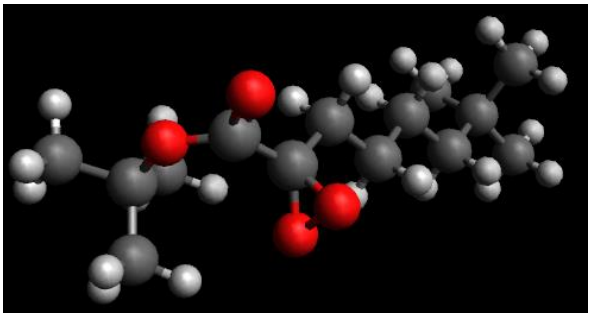
	2,5-Dimethyl-2,5-bis(hydroperoxy)hexane 			
25		3025-88-5	178.26	105.0
	Bis (2-chlorobenzoyl) peroxide  			
26		3033-73-6	311.12	51.3
	Dipropionyl peroxide 			
27		3248-28-0	146.16	30.0

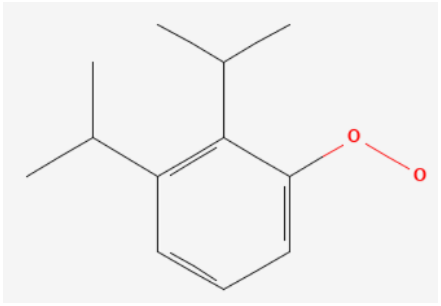
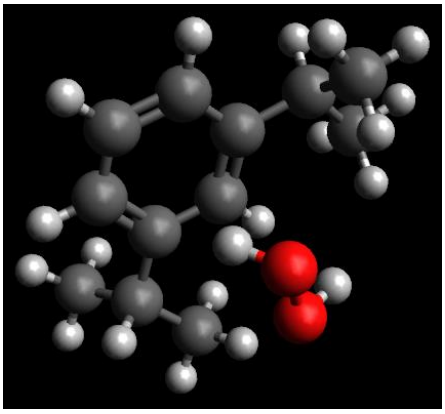
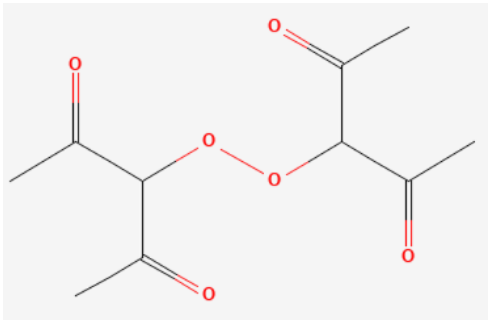
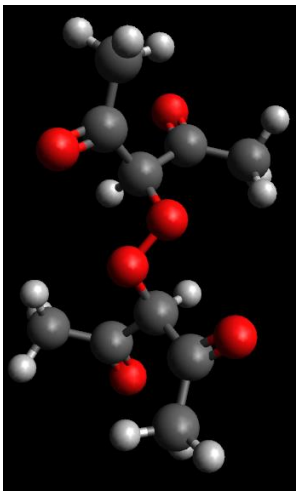
28	Diisobutyryl peroxide  		3437-84-1	174.22	0.0
29	tert-butyl cumyl peroxide  		3457-61-2	208.33	77.1
30	1,1-bis(tert-butylperoxy)-3,3,5- trimethylcyclohexane  		6731-36-8	302.51	60.0

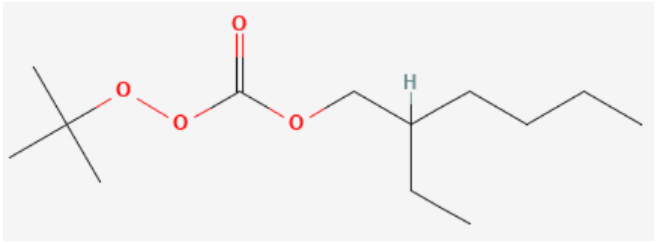
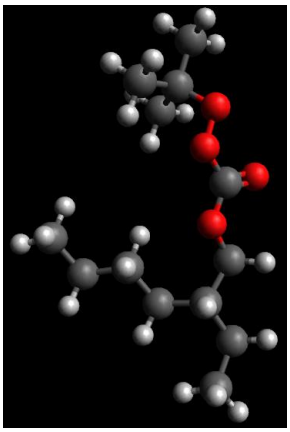
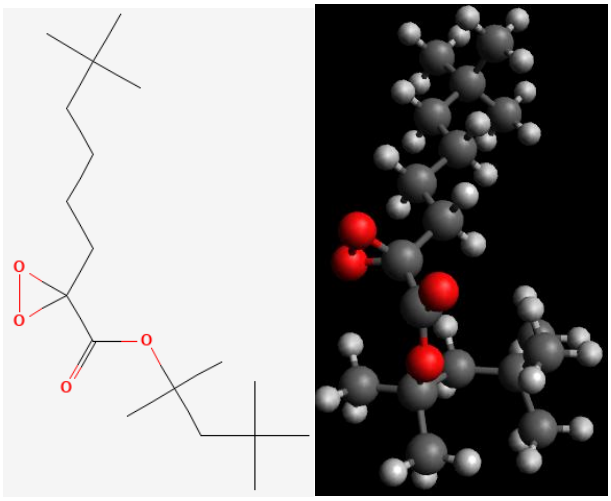
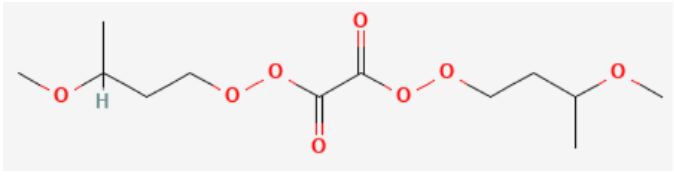
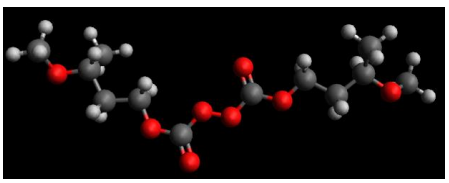
31	3,4-Dimethyl-3,4-diphenylhexane 	10192-93-5	266.46	158.6
32	Cyclohexanone peroxide 	12262-58-7	246.34	80.0
33	tert-Butyl 3,5,5-trimethylperoxyhexanoate  	13122-18-4	230.39	24.0

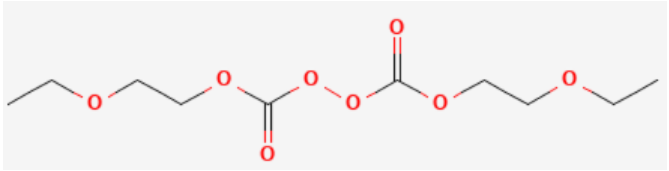
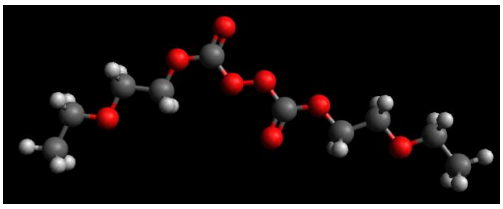
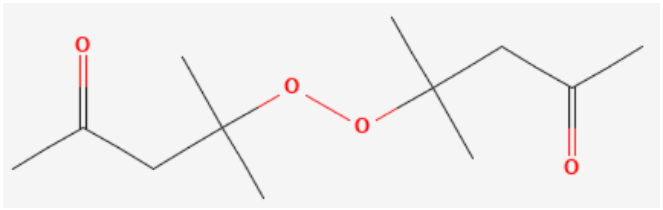
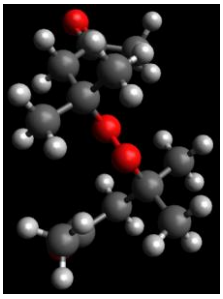
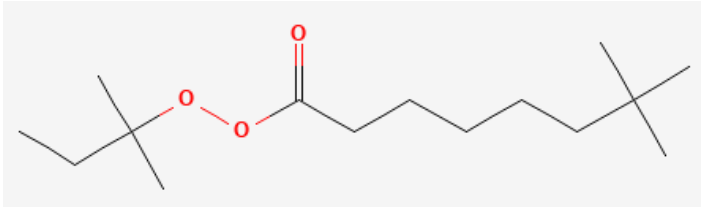
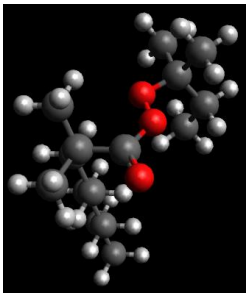
34	Bis(4-tert-butylcyclohexyl) peroxydicarbonate 	15520-11-3	398.60	40.0
35	Di(n-propyl) peroxydicarbonate  	16066-38-9	206.22	-5.0
36	Di-n-butyl peroxydicarbonate  	16215-49-9	234.28	5.0

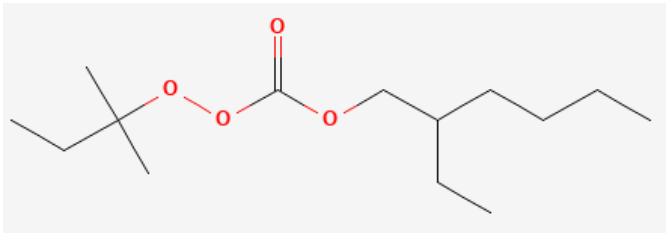
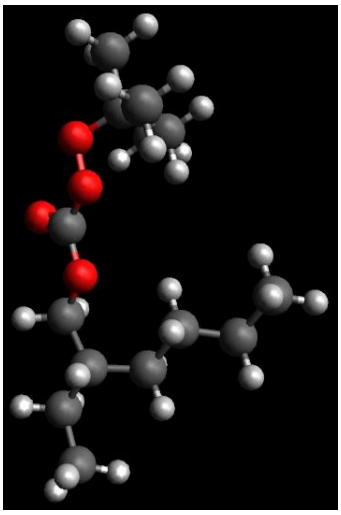
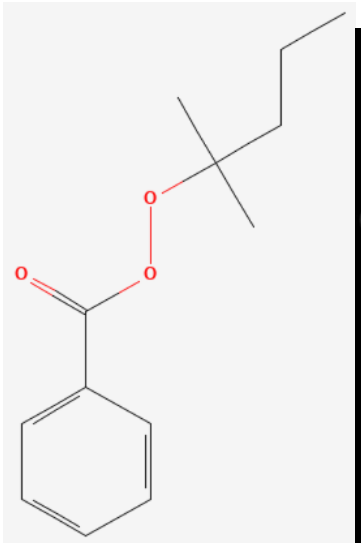
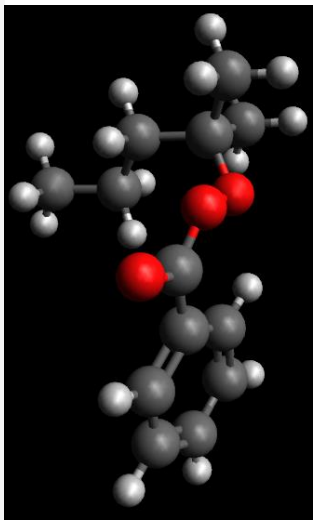
	Di-sec-butyl peroxydicarbonate 				
37			19910-65-7	234.28	0.0
	α , α -Dimethylbenzyl peroxypivalate  				
38			23383-59-7	236.34	15.0

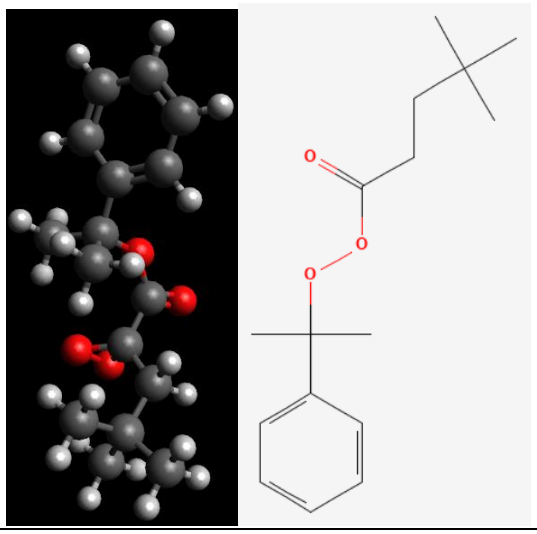
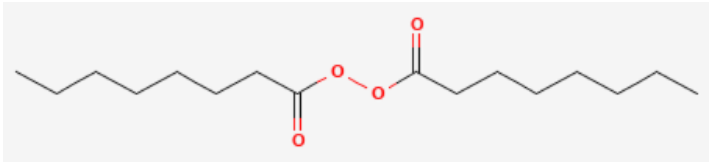
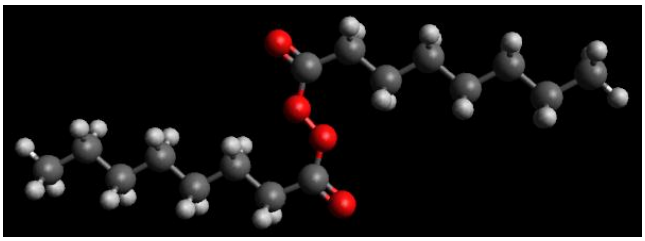
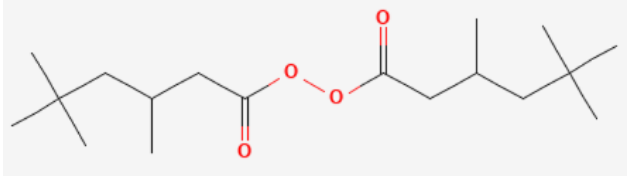
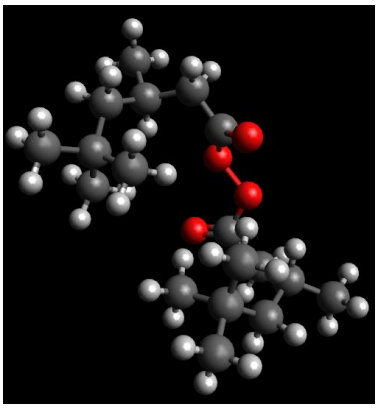
39	bis (tert-butyl peroxyisopropyl) benzene 	25155-25-3	338.54	80.8
40	Dihexadecyl peroxodicarbonate  	26322-14-5	571.00	37.5
41	Cumyl peroxyneodecanoate  	26748-47-0	306.49	7.8

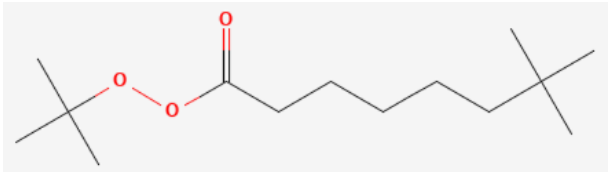
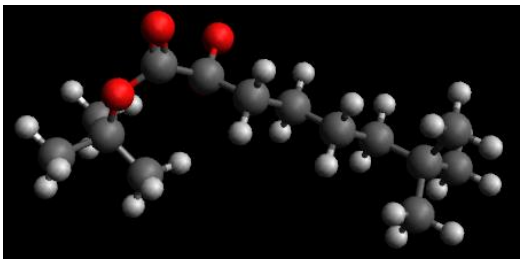
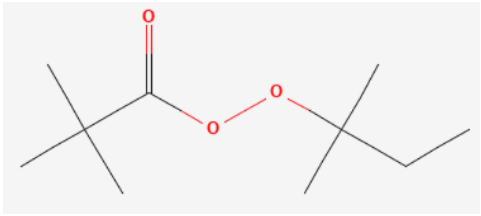
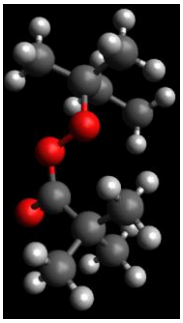
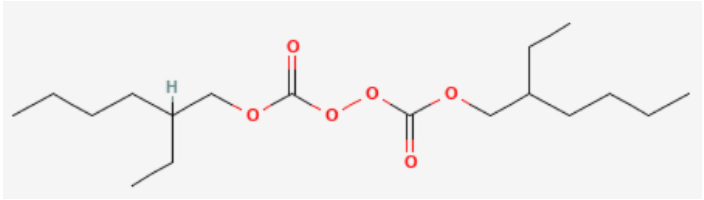
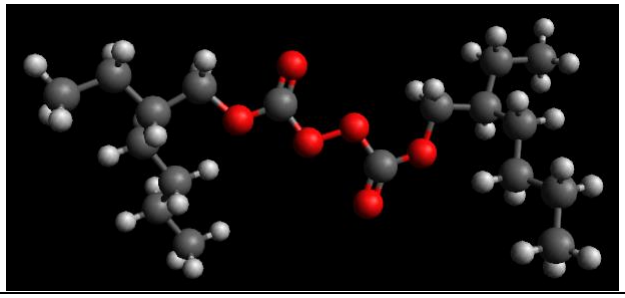
	Di-isopropylbenezene hydroperoxide 			
42		26762-93-6	196.32	65.0
	Acetylacetone peroxide 			
43		37187-22-7	230.24	64.7

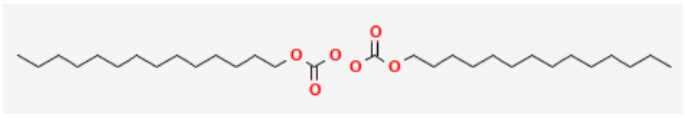
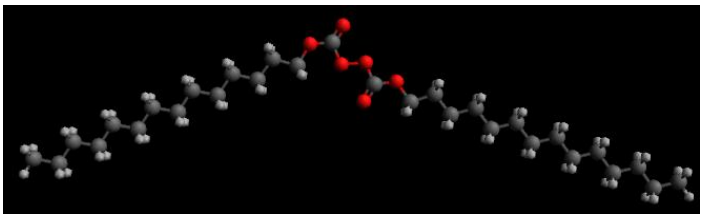
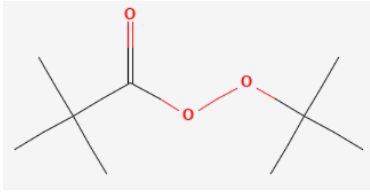
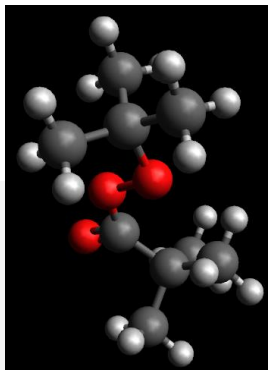
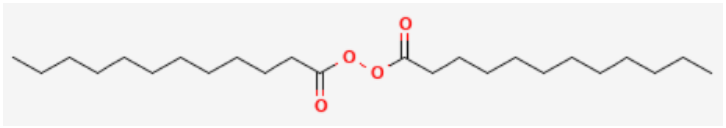
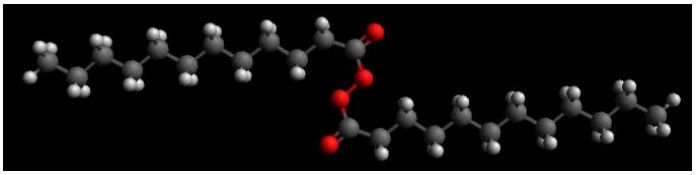
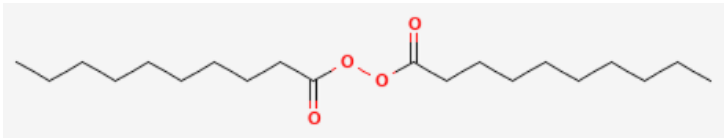
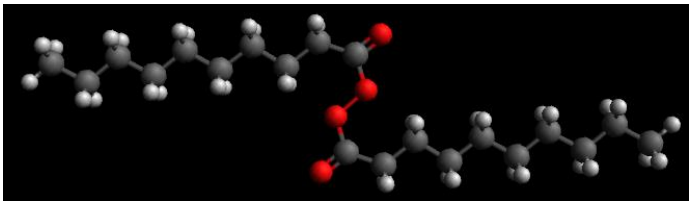
	tert-butyl peroxy-2-ethylhexyl carbonate 			
44		34443-12-4	246.39	51.0
	2,4,4-Trimethylpentyl-2-Peroxyneodecanoate 			
45		51240-95-0	314.52	18.1
	Bis(3-methoxybutyl) peroxydicarbonate 			
46		52238-68-3	294.34	15.0

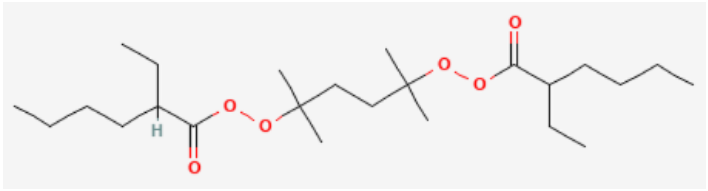
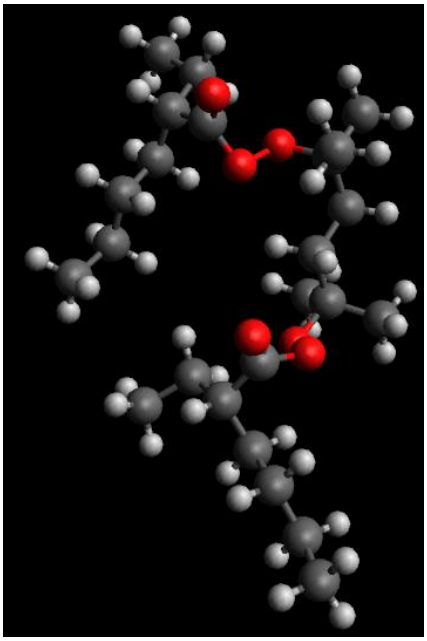
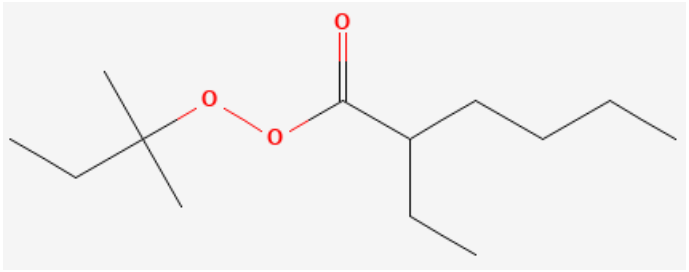
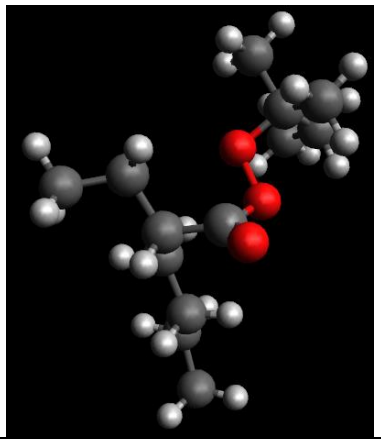
47	Di(2-ethoxyethyl) peroxydicarbonate  	52373-74-7	266.28	10.0
48	Diacetone alcohol peroxide  	54693-46-8	230.34	50.0
49	tert-Amyl peroxyneodecanoate  	68299-16-1	258.45	10.8

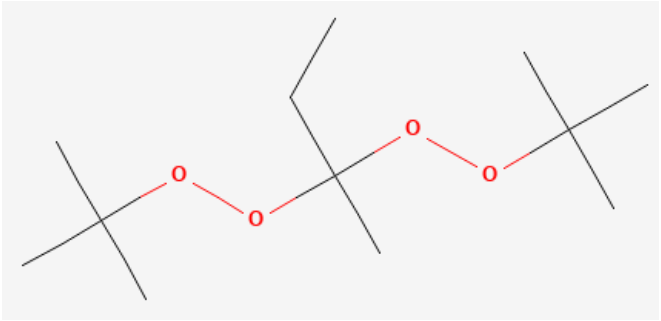
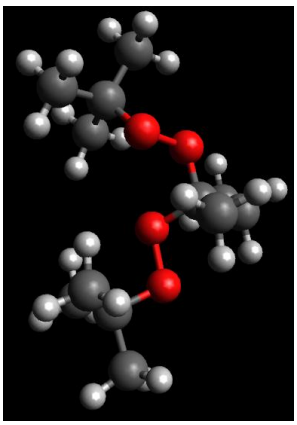
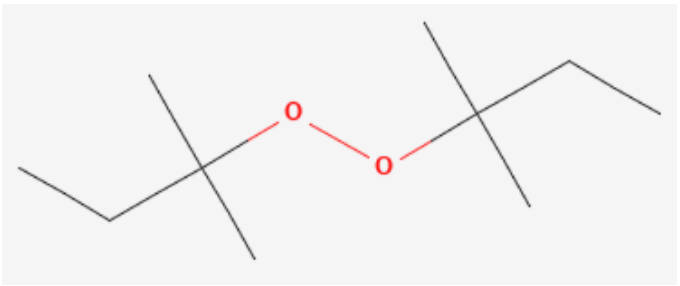
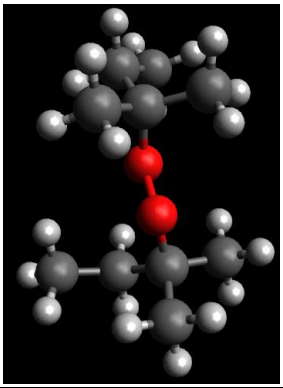
	Tert-Amyl peroxy 2-ethylhexyl carbonate 			
50		70833-40-8	260.42	55.0
	t-Hexyl peroxide benzoate  			
51		124350-67-0	222.31	62.2

	Cumyl peroxyneoheptanoate 			
52		130097-36-8	278.38	10.0
	Dioctanoyl peroxide  			
53		762-16-3	286.46	25.9
	bis(3,5,5-Trimethylhexanoyl) peroxide  			
54		3851-87-4	314.52	20.0

	tert-Butyl peroxyneodecanoate 			
55		26748-41-4	258.40	15.0
	tert-Amyl peroxypivalate 			
56		29240-17-3	188.30	21.6
	bis-(2-ethylhexyl) Peroxydicarbonate 			
57		16111-62-9	346.52	15.4

	dimyristyl peroxydicarbonate 			
58		53220-22-7	514.88	19.2
	tert-Butyl peroxypivalate  			
59		927-07-1	174.27	27.0
	di-Lauroyl peroxide 			
60		105-74-8	398.70	46.0
	di-Decanoyl peroxide 			
61		762-12-9	342.58	31.0

	2,5-bis-(2-ethylhexanoylperoxy)-2,5-dimethylhexane 			
62		13052-09-0	430.70	38.6
	tert-amyl peroxy-2-ethylhexanoate 			
63		686-31-7	230.39	35.0

	2,2-bis(tert-butylperoxy) butane 			
64		2167-23-9	234.38	70.0
	di-tert-amyl peroxide 			
65		10508-09-5	174.32	70.4

2.3 記述子計算データセット

65 化合物の分子構造は mol ファイル形式で作成され記述子の計算が行われた。分子構造に依存する記述子もあるため分子構造の最適化は非常に重要である。分子モデリングソフトウェアである Avogadro[6]を用いて，各 mol ファイルに対して自動構造最適化を行った。基本的な MM ポテンシャルエネルギー関数は，結合項(共有結合した原子間相互作用) と非結合項(長距離静電気力とファンデルワールス力)を含んでいる。ここでは， UFF(Universal Force Filed)を用いて不自然な結合長と結合角を改善し，最小エネルギーのコンフォメーションを得た。事前に MM 計算を行うことで，DFT 計算の収束性を向上させることができた。その後，MoCalc2012 [7] 経由で Firefly [8] を用いて DFT 計算を実施し，全化合物の立体構造の最適化を実施した。Job type はジオメトリー最適化，基底関数はスプリットバレンス基底系の 6-31G，汎関数は B3LYP 混成汎関数を選択した。一般的に有機化合物に対しては，水素以外の元素に d 軌道の関数を加えた分極基底系の 6-31G(d)を用いられることが多いが，計算負荷削減のために 6-31G を選択した。B3LYP 混成汎関数は，ハートリー・フォック法に不足する電子相関を補うために提案されたコーン・シャム方程式の交換・相関エネルギー汎関数項として，B88 交換汎関数，LYP 相関汎関数，LDA 交換・相関汎関数など複数の交換・相関汎関数を組み合わせたものであり，一般的に最も多く利用されている汎関数である。ジオメトリー最適化後，分子記述子計算ソフトウェアである alvaDesc 2.0.8 [9]，CODESSA 3 [10] を用いて記述子の計算を行った。それぞれ 5,666 種類，552 種類の記述子を得ることができたものの，SADT に寄与しない記述子も多く冗長であったため，標準偏差が小さい記述子，多重共線性が強い記述子を除去して記述子数を削減した。CODESSA 記述子には alvaDesc で計算されるような基本的な構造情報から計算される記述だけでなく，HOMO/LUMO などの量子化学的な記述子も含まれる。変数選択は遺伝的アルゴリズム(GA)を用いて行った。GA は実験条件を遺伝子として表現した個体を複数用意し，適応度が高い個体を優先的に選択して，交差・突然変異などの操作を繰り返しながら良好な適応度関数が得られる個体を継続的に探索するというものである。ここでは，適応度を PLS と SVR モデリングにおける 5-fold クロスバリデーション後の決定係数R^2とし，それぞれ GA-PLS と GA-SVR と呼ぶこととする。母集団の個体数は 100，交叉確率は 0.5，突然変異確率は 0.2，最大世代数は 200 とした。

2.4 予測モデル

PLS は線形回帰手法の一種である。説明変数の数が標本の数よりも有意に大きい場合に広く使用される。潜在変数と変数の間の共分散を最大化することにより，多重共線性の影響を小さくして回帰モデルを構築することができる。

$$\boldsymbol{X} = \sum_{a=1}^{A} \boldsymbol{t}_a \boldsymbol{p}_a^T + \boldsymbol{E} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{1}$$

$$\boldsymbol{Y} = \sum_{a=1}^{A} \boldsymbol{t}_a q_a + \boldsymbol{f} = \boldsymbol{T}\boldsymbol{q} + \boldsymbol{f} \tag{2}$$

$\boldsymbol{X}$は説明変数，$\boldsymbol{Y}$は目的変数，$\boldsymbol{T}$は潜在変数，Aは潜在変数の数，$\boldsymbol{t}_a$はa番目の潜在変数，$\boldsymbol{p}_a$はa番目の潜在変数のローディング，q_aはa番目の潜在変数の重み，$\boldsymbol{E}$, $\boldsymbol{f}$は$\boldsymbol{X}$，$\boldsymbol{Y}$のモデルで表現できない誤差項を表す。5-fold cross validation で得られる決定係数R^2が最も高い潜在変数数Aが予測モデルに使用された。

目的変数の予測への寄与度が大きな変数を識別するために，各変数の重要度（VIP: Variable Importance in Projection）[11]スコアを計算した。VIP スコアは，説明変数の重要度を新たす指標である。説明された$\boldsymbol{Y}$分散の比率によって重みづけされた潜在変数の重みの合計として定義される。

$$\boldsymbol{VIP_j} = \sqrt{\frac{\sum_{i=1}^{h} R^2(y,t_i)(w_{ij}/\|\boldsymbol{w}_i\|)^2}{(1/p)\sum_{i=1}^{h} R^2(y,t_i)}} \tag{3}$$

ここで，$\boldsymbol{VIP_j}$は変数$\boldsymbol{j}$の VIP スコア，hは潜在変数の数，$\boldsymbol{w}_i$は潜在変数における説明変数の重みベクトル，$R^2(y,t_i)$は，説明された$\boldsymbol{Y}$分散の比率である。特に，予測性能への貢献度の高い説明変数の VIP 値は 1 以上となる。

Support Vector Regression (SVR)はカーネル関数を用いて高次元への写像を行うことにより，非線形な変化にも対応可能な回帰手法である。i番目のサンプルの説明変数を$\boldsymbol{x}^{(i)}$とすると，目的変数の推定値$f\left(\boldsymbol{x}^{(i)}\right)$は以下のように表される。

$$f\left(\boldsymbol{x}^{(i)}\right) = \phi\left(\boldsymbol{x}^{(i)}\right)\boldsymbol{w} + b \tag{4}$$

$$\boldsymbol{w} = [w_1\ w_2\ \cdots\ w_k] \tag{5}$$

$f(\boldsymbol{x}^{(i)})$は$\boldsymbol{x}^{(i)}$の非線形写像，bは定数項，$\boldsymbol{w}$は重み，kは高次元空間での次元数を表す。また，誤差関数Eは下式で表現される。

$$E = C\sum_{i=1}^{n} max\left(0, \left|y^{(i)} - f(\boldsymbol{x}^{(i)})\right| - \varepsilon\right) + \frac{1}{2}\|\boldsymbol{w}\| \quad (6)$$

Cは正則化係数(予測誤差と回帰式の複雑さのバランスを決定する係数)，nは学習データのサンプル数，εは不感度係数(誤差の許容幅)を表す。学習データへの当てはまり(第 1 項)と汎化性能(第 2 項)のバランスが取れるようにパラメータを最適化する。なお，カーネル関数には Radial Basis Function(RBF)カーネルを用いた。

$$K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = exp\left(-\gamma\left\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right\|^2\right) \quad (7)$$

γは RBF カーネルパラメータ(モデルの複雑度)を表す。モデル開発と記述子の選択は，Python 3.7 を用いて行った。PLS と SVR の計算には，Python の機械学習ライブラリである Scikit-learn [12]を使用した。ハイパーパラメータの最適化には，Python のパラメータ推定器である GridSearchCV [13]を使用した。ハイパーパラメータであるC, ε, γに関しては，GridSearchCV で 5-fild cross validation を行い，高速最適化を実施した[14]。

ハイパーパラメータの最適化と記述子の選択は，トレーニングデータセットに対してのみ実行し，テストデータセットに対するモデルの性能を評価した。モデルの精度は，一般的な統計パラメータである二乗平均平方根誤差（RMSE: Root Mean Square Error），平均絶対誤差（MAE: Mean Absolute Error），決定係数R^2で評価した。これらのパラメータは以下のように計算された。

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{i,measured} - y_{i,predicted})^2}{n}} \quad (8)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_{i,measured} - y_{i,predicted}|}{n} \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,measured} - y_{i,predicted})^2}{\sum_{j=1}^{n}(y_{j,measured} - y_{mean})^2} \quad (10)$$

nはデータ数, $y_{i,measured}$はi番目の測定値, $y_{i,predicted}$はi番目の予測値, y_{mean}は$y_{i,measured}$の平均値を意味する。なお，回帰分析手法として PLS を選択する場合は GA-PLS，SVR を選択する場合は GA-SVR を用いて変数選択を実施した。

2.5 結果と考察

予測モデル構築前の構造最適化と変数選択を含む予測モデルと含まない予測モデルを構築し，既存のモデル（表 2）と予測精度を比較した。

表 2　テストデータに対する既存モデルの予測精度比較

	Wang [3]		HE [4]	
Geometry Optimization	DFT		MM+/MO	
	B3LYP/6-31G(d)		PM1	
Frequency calculation			-	
Variable selection	-		GA-MLR	
Number of descriptors	8		9	
Modeling method	MLR	SVR	MLR	SVR
Number of training data	40		57	
Number of test data	10		14	
RMSE	12.0	6.43	9.91	9.79

ケース 1 では，モデル開発の前に MM 計算のみが実行された。前処理後，5 つの潜在変数を持つ PLS と SVR(C=8.0, ε =0.00098, γ =0.00024)によりモデルを開発した。実際の値と予測値計算値のパリティ・プロットを図 1 に示す。訓練セットのモデルの統計パラメータは以下の通りである： R2=0.90, RMSE=12.35, MAE=9.45 (PLS), R2=0.99, RMSE=3.92, MAE=0.67（SVR）： PLS は R2=0.26, RMSE=22.40, MAE=17.46，SVR は R2=0.10, RMSE=24.69, MAE=19.95 であった。予測性能は既存のモデルより低かった。PLS では潜在変数の数を変更しても，SVR ではハイパーパラメータの値を変更しても，予測精度は向上しなかった。

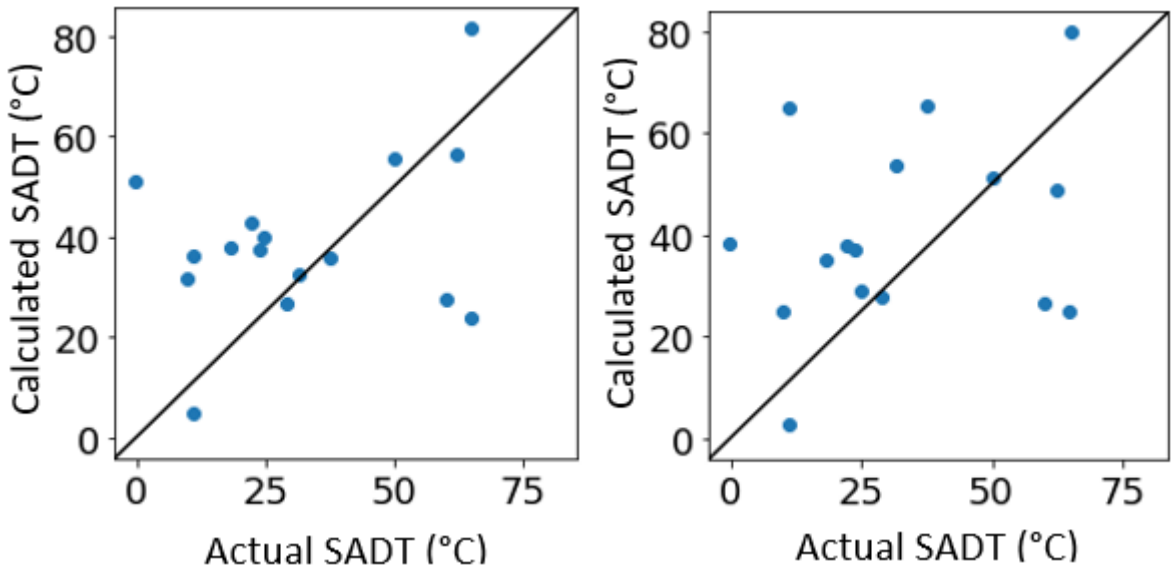


図 2　実測値と予測値の比較（Case 1，左: PLS，右: SVR）

ケース 2 では，MM と DFT の計算はモデル開発の前に行われた。前処理後，13 個の潜在変数を用いた PLS と SVR(C=4.0, ε =0.00098, γ =0.00049)によりモデルを開発した。実際の値と予測値のパリティ・プロットを図 2 に示す。訓練セットに対するモデルの統計パラメータは以下の通りであった： R2=0.99，RMSE=0.98，MAE=0.76（PLS），R2=0.99，RMSE=2.15，MAE=0.34（SVR）： PLS では R2=0.82, RMSE=9.80, MAE=7.70，SVR では R2=0.77, RMSE=11.07, MAE=8.68 であった。予測性能はケース 1 より有意に高く，既存のモデルと同等であった。両モデルの予測精度は，モデル構築前に DFT 計算を行い，量子化学記述子を追加することで改善される可能性がある。PLS では潜在変数の数を，SVR ではハイパーパラメータの値を変更しても，ケース 1 と同様に予測精度は向上しなかった。

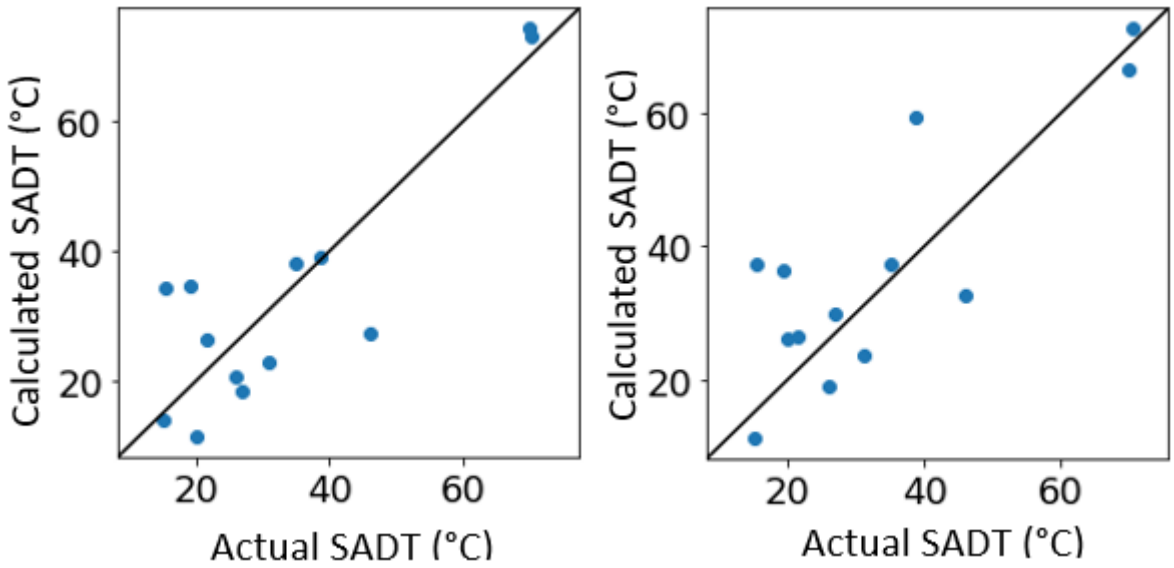


図 3　実測値と予測値の比較（Case 2, 左: PLS, 右: SVR）

ケース 3 では，前処理（ケース 2 と同様）に加えて，モデル構築の前に，GA-PLS と GA-SVR を用いて記述子を選択した。変数選択後，GA の適合関数は，PLS では R2=0.69 から R2=0.91 に，SVR では R2=0.38 から R2=0.90 に改善された。モデルは，11 個の潜在変数を持つ PLS と，SVR（C=466，ε =0.02343，γ =0.00000714）によって開発された。実際の値と予測値のパリティ・プロットを図 3 に示す。訓練セットに対するモデルの統計パラメータは以下の通りであった： R2=0.99，RMSE=1.57，MAE=1.14（PLS），R2=0.99，RMSE=3.69，MAE=1.78（SVR）： PLS では R2=0.95, RMSE=5.11, MAE=4.03，SVR では R2=0.91, RMSE=6.87, MAE=5.15 であった。ケース 2 と比較すると，予測性能は飛躍的に向上し，予測精度も既存モデルよりも向上した。このように，モデル構築前に変数を適切に選択することで，予測精度を向上させることができると言える。

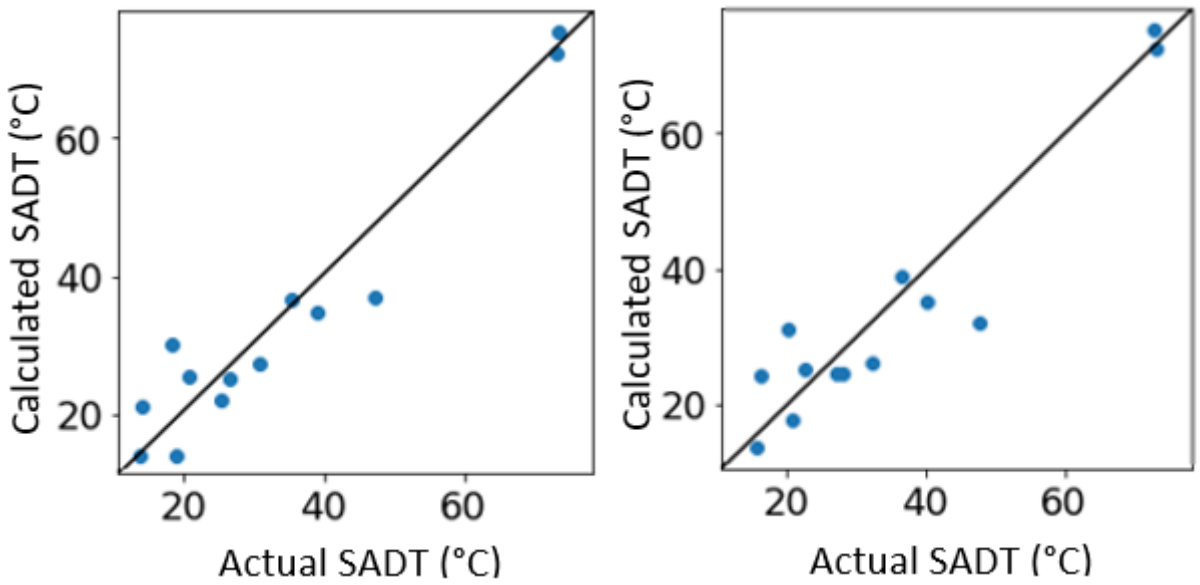


図 4　実測値と予測値の比較（Case 3，左: PLS，右: SVR）

PLS と SVR の予測精度の比較を表 3 に示す。" Descriptors"の行は前処理による記述子数の変化を示し，" Variable selection "の行は GA 適用後の記述子数を示す。予測精度は全体的に良好であり，本研究で使用したデータセットでは，SVR よりも PLS の方が高い精度を示した。MM/DFT 計算による形状最適化，量子化学記述子の追加，GA による変数選択が予測精度に影響を与えており，いずれの方法も実施をした方が精度は高くなる傾向があった。

表 3　モデル構築情報および検証結果

	Case1		Case2		Case3	
Descriptors	5889-> 2659		1216+553->1586		1216+553-> 1586	
Calculated by	alvaDesc 2		alvaDesc 2 +CODESSA 3		alvaDesc 2 +CODESSA 3	
Geometry optimization	MM (UFF)		MM (UFF) DFT (6-31G/B3LYP)		MM (UFF) DFT (6-31G/B3LYP)	
Variable selection	-		-		GA-PLS 1586->559	GA-SVR 1586->524
Modeling method	PLS	SVR	PLS	SVR	PLS	SVR
RMSE	22.4	24.7	9.8	11.1	5.1	6.9
MAE	17.5	20.0	7.7	8.7	4.0	5.2
R2	0.26	0.23	0.82	0.77	0.95	0.91

文献[3]のモデルは RMSE が小さく予測精度が高いが，DFT 計算のため計算負荷が非常に高い。一方，半経験的分子軌道法(AM1)を用いた文献[4]のモデルは，計算負荷は低いものの，相対的に RMSE が大きかった。提案モデルの予測精度は，量子化学記述子の追加と GA を用いた変数選択により大幅に向上した(表 4)。DFT 計算の基底セットを 6-31G に変更したが，予測精度を維持したまま計算負荷を軽減することができた。SADT 予測において，d 軌道の関数(分極関数)を加えても予測精度に大きな影響はないことがわかった。記述子計算に用いる分子構造の適切な最適化と量子化学記述子の追加は重要である。しかし，予測精度の向上はある時点で頭打ちとなるため，予測精度の向上と計算負荷のバランスを考慮し，効果的なモデルを構築する必要があると思われる。

表 4　提案手法と既存モデルにおける予測精度の比較

	Proposed method		Wang [3]		HE [4]	
Geometry optimization	MM/DFT 6-31G/B3LYP		DFT 6-31G(d)/B3LYP		MM+/MO PM1	
Frequency calculation	-				-	
Variable selection	GA-PLS	GA-SVR	-		GA	
Number of descriptors	559	524	8		9	
Modeling method	PLS	SVR	MLR	SVR	MLR	SVR
Number of training data	52		40		57	
Number of test data	13		10		14	
RMSE	5.11	6.87	12.0	6.43	9.91	9.79

VIP スコアの高い上位 15 の記述子を表 5 に示す。酸素結合に関連する記述子（結合次数，価数，電荷），量子化学記述子（LUMO，反発/引力エネルギー）が SADT の予測精度に影響を与えていることがわかった。

表 5　VIP スコア上位 15 記述子

No.	VIP	Descriptor	Calculated by	Explanation
1	2.248	AvgBondOrd_O	CODESSA	Average bond order for all atoms of O type
2	2.231	MaxOneCent-ElecElecRepEn	CODESSA	Maximum One-Center Electron-Electron Repulsion Energy
3	2.182	SM02_EA(dm)	alvaDesc	Spectral moment of order 2 from edge adjacency mat. weighted by dipole moment
4	2.176	SM08_EA(dm)	alvaDesc	Spectral moment of order 8 from edge adjacency mat. weighted by dipole moment
5	2.133	MinOneCent-	CODESSA	Minimum One-Center Core-Electron

		CoreElecAttrEn		Attraction Energy
6	2.082	SM07_EA(dm)	alvaDesc	Spectral moment of order 7 from edge adjacency mat. weighted by dipole moment
7	2.040	MaxTwoCent-TotEn_AB	CODESSA	Maximum Two-Center Total Energy, All Bonds
8	2.021	SpMax_B(s)	alvaDesc	Leading eigenvalue from Burden matrix weighted by I-State
9	2.004	AvgVal_O	CODESSA	Average valence for atoms of O type.
10	1.978	MaxTwoCent-CoreElecResEn_AP	CODESSA	Maximum Two-Center Core-Electron Resonance Energy, All Pairs
11	1.971	B02[O-O]	alvaDesc	Presence/absence of O – O at topological distance 2
12	1.958	AvgBondOrd_O_O	CODESSA	Average bond order among all bonds between atoms of type O and O.
13	1.949	qpmax	alvaDesc	Maximum positive charge
14	1.907	MinTwoCent-CoreElecAttrEn_AP	CODESSA	Minimum Two-Center Core-Electron Attraction Energy, All Pairs
15	1.861	LUMOEn	CODESSA	Energy of lowest energy unoccupied molecular orbital

図 5 のように，No.53，No.54，No.55，No.59，No.64，No.65 では，記述子計算の前に DFT 計算を行うことで予測精度が向上したが，No.57，No.58，No.60 では，DFT 計算による前処理を行っても予測精度はあまり向上しなかった。また，No.56，No.62，No.63 は最初から高い予測精度を示した。No.62 を除くすべてのケースで，GA による変数選択が予測精度を向上させたが，予測精度向上に対する全体的な影響は DFT 最適化の効果よりも小さかった。

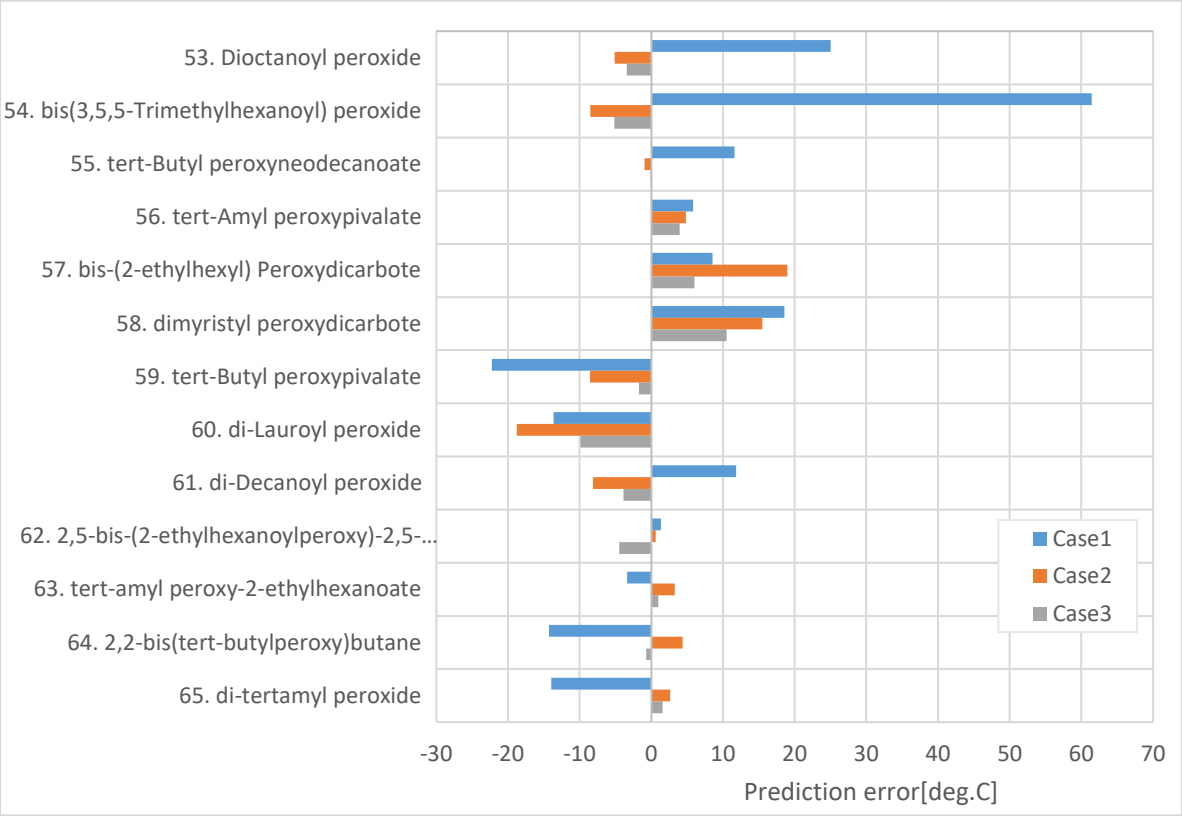


図 5　PLS モデルにおける各化合物の予測誤差の比較

表 6 に示すように，No.53, No.54, No.56, No.58 を比較したところ，各分子の O-O 結合付近のコンフォメーション（結合長と角度）は，DFT 計算後に大きく変化した。No.53 と 54 は No.56 と No. 58 に比べて相対的に大きな結合長の変化を示したが，これは No.56 と No.58 に比べて No.53 と No. 54 の初期状態と最適状態の差が大きいためである。No。56 はケース 1 において，わずか 18 回の反復にもかかわらず低い予測誤差を示したが，これは MM 計算のみが実行された最適な初期コンフォメーションによるものと考えられる。例外はあるが，反復回数が多いほど高い予測精度が得られた。No.58 の予測誤差はあまり減少しなかった。これは，基底関数として 6-31G(d)ではなく 6-31G を用いたため，DFT 計算の誤差に起因している可能性がある。このように，分子構造の適切な最適化と量子化学記述子の追加は，SADT 予測に多大な影響を与えることがわかった。

表 6　代表 4 化合物における結合長，角度の比較

Element	Unit	No.53	No.54	No.56	No.58
Improvement of prediction accuracy	℃	19.9	52.9	1.0	3.1
	%	79.6	86.2	17.3	16.6
Iterations	times	21	51	18	25

Length (O1-O2)	angstrom	+0.263	+0.237	+0.162	+0.174
Length (C1-O1)	angstrom	+0.059	+0.064	+0.020	+0.024
Length (C2-O2)	angstrom	+0.059	+0.063	+0.065	+0.046
Angle(∠C1O1O2)	degree	-9.9	-12.5	-7.6	-13.5
Angle(∠O1O2C2)	degree	-9.9	-9.9	-2.9	-12.5

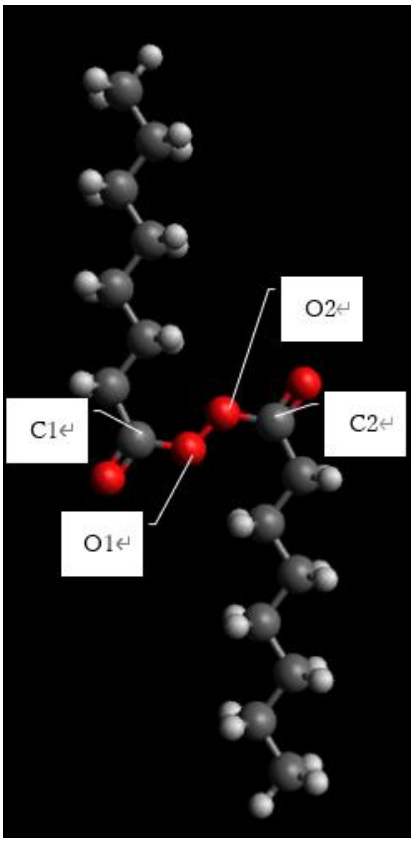


図 6　No.53 の分子構造(dioctanoyl peroxid)

表 7　No.53 の代表部分における構造最適化前後の結合長および角度

Element	Unit	Before optimization	After optimization	Difference
O1-O2	angstrom	1.268	1.531	+0.263
C1-O1	angstrom	1.359	1.418	+0.059
C2-O2	angstrom	1.359	1.418	+0.059
∠C1O1O2	degree	124.5	114.6	-9.9
∠O1O2C2	degree	124.5	114.6	-9.9

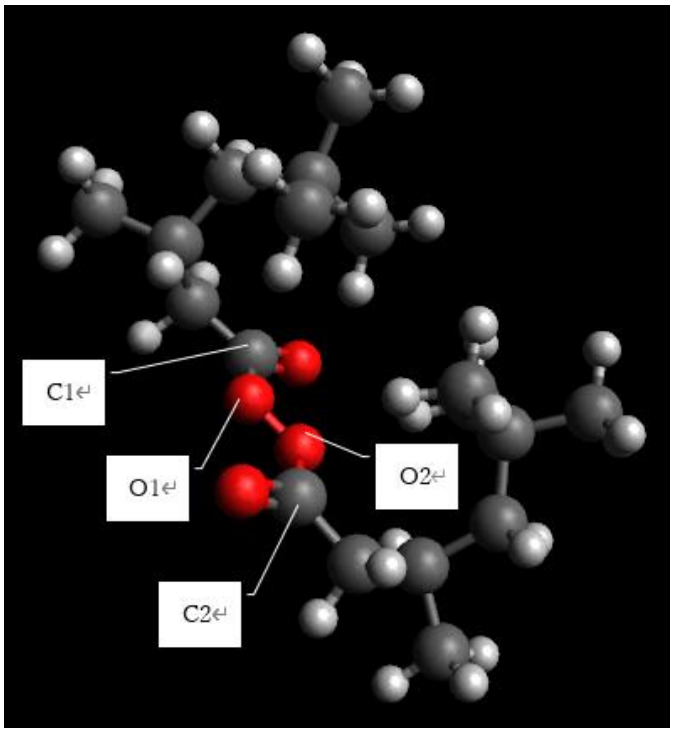


図 7　No.54 の分子構造(bis(3,5,5-trimethylhexanoyl) peroxide)

表 8　No.54 の代表部分における構造最適化前後の結合長および角度

Element	Unit	Before optimization	After optimization	Difference
O1-O2	angstrom	1.273	1.510	+0.237
C1-O1	angstrom	1.352	1.416	+0.064
C2-O2	angstrom	1.353	1.416	+0.063
∠C1O1O2	degree	122.3	109.8	-12.5
∠O1O2C2	degree	120.5	110.4	-9.9

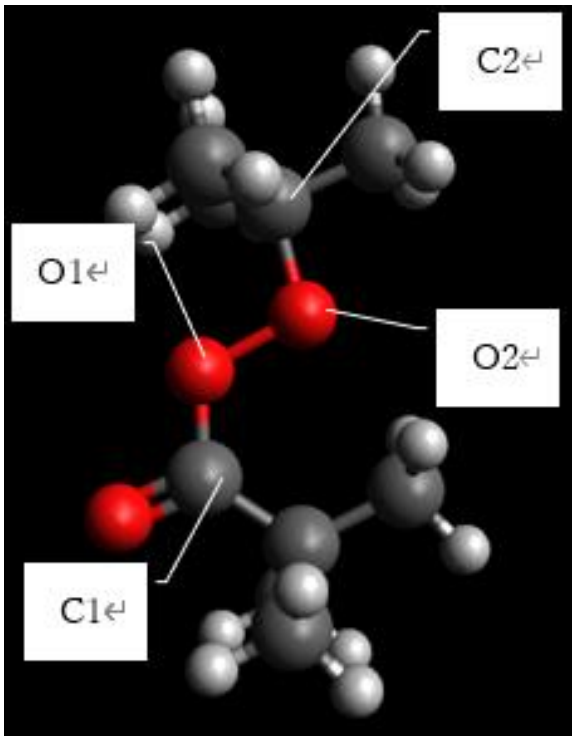


図 8　No.56 の分子構造(tert-amyl peroxipivalate)

表 9　No.56 の代表部分における構造最適化前後の結合長および角度

Element	Unit	Before optimization	After optimization	Difference
O1-O2	angstrom	1.297	1.459	+0.162
C1-O1	angstrom	1.360	1.380	+0.020
C2-O2	angstrom	1.417	1.482	+0.065
∠C1O1O2	degree	125.0	117.4	-7.6
∠O1O2C2	degree	108.2	105.3	-2.9

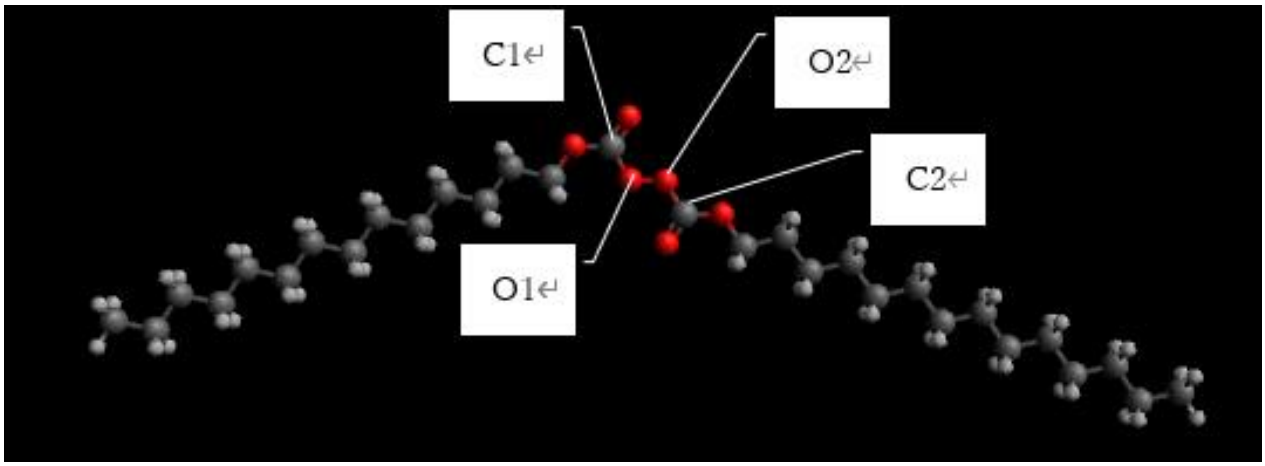


図 9　No.58 の分子構造(dimyristyl peroxydicarbonate)

表 10　No.58 の代表部分における構造最適化前後の結合長および角度

Element	Unit	Before optimization	After optimization	Difference
O1-O2	angstrom	1.272	1.446	+0.174
C1-O1	angstrom	1.354	1.378	+0.024
C2-O2	angstrom	1.352	1.398	+0.046
∠C1O1O2	degree	121.7	108.2	-13.5
∠O1O2C2	degree	121.1	108.6	-12.5

テストデータの温度範囲は 15～70℃であり，モデルがより広い温度範囲に適用できるかどうかは不明であった。高温域での外挿性（汎化性能）を検証するため，ケース 3 のデータ，変数，モデル構築条件を用いたダブルクロスバリデーションを実施した。内側クロスバリデーションでは，ハイパーパラメータを 5-fold クロスバリデーションで決定し，外側クロスバリデーションでは，Leave-one-out クロスバリデーションを行いモデルの汎化性能の評価を行った。PLS モデルは R2=0.76，RMSE=17.74，MAE=13.29 であり，SVR モデルは R2=0.74，RMSE=18.11，MAE=13.50 であった。低温時と高温時の予測誤差に大きな差は見られず，両モデルとも広い範囲の SADT を精度よく予測できた。

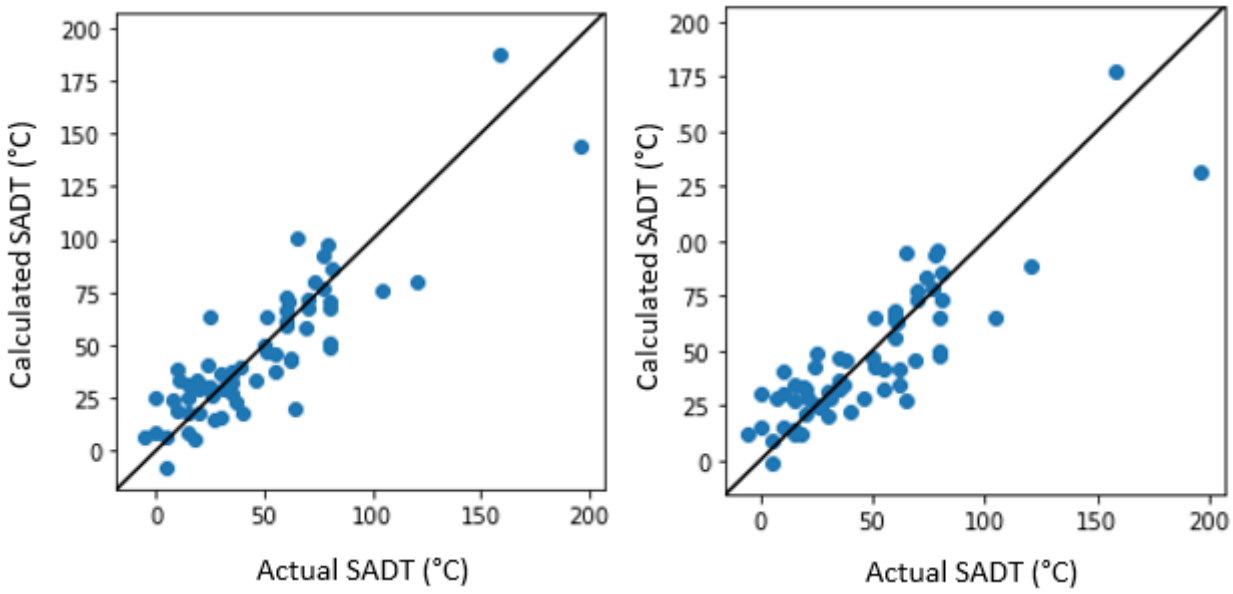


図 10　実測値と予測値の比較（左: PLS，右: SVR）

2.6 まとめ

本研究では，有機過酸化物の構造式から自己促進分解温度(SADT)を推算するモデルの構築を試みた。化合物の構造式から記述子を計算し，偏最小二乗 (PLS: Partial Least Squares) 回帰，サポートベクター回帰(SVR: Support Vector Regression)を適用した結果，比較的精度良く SADT を予測できるモデルを構築することができた。モデル精度が向上することにより，化合物の開発や保管，輸送，製造において，早期に正確に熱的危険性を把握することが可能となる。モデル構築時の前処理として分子力学 (MM: Molecular Mechanics)計算や密度汎関数(DFT: Density Functional Theory)計算，変数選択に遺伝的アルゴリズム(GA: Genetic Algorithm)を用いたところ，適用しない場合と比べて飛躍的に予測精度が向上した。既存のモデルと比較しても遜色ない精度で予測をできていることを確認した。精度が良いモデルを構築するためには十分な前処理と適切な変数選択が重要である。SADT の予測精度を向上させるためには，より厳密に化合物構造の最適化を実施してから記述子の計算を行う方が望ましいが，計算負荷が非常に重くなるため，予測精度と計算負荷バランスの最適化を考える必要がある。予測精度のさらなる向上のために，SVR 以外の機械学習モデルの適用や，記述子計算のための前処理方法，SADT 予測に対する寄与度が高い記述子探索などについて検討を継続する。また，対象を有機過酸化物に限定するのではなく，発熱のメカニズムが異なるような化合物であっても一般的に予測可能なモデルの開発を検討したい。

2.7 参考文献

2. Chen W. T.; Chen W. C.; You M. L.; Tsai Y.T.; Shu C. M. Evaluation of thermal decomposition phenomenon for 1,1-bis(tertbutylperoxy)-3,3,5-trimethylcyclohexane by DSC and VSP2. *J. Therm. Anal. Calorim.* **2015**, *122* (3), 1125–1133. DOI: https://doi.org/10.1007/s10973-015-4985-2

3. Wang B.; Yi H.; Xu K.; Wang Q. Prediction of the self-accelerating decomposition temperature of organic peroxides using QSPR models. *J. Therm. Anal. Calorim.* **2017**, *128*, 399–406. DOI: https://doi.org/10.1007/s10973-016-5922-8

4. He P.; Pan Y.; Jiang J. C. Prediction of the self-accelerating decomposition temperature of organic peroxide based on support vector machine, *Procedia Engineering* **2018**, *211*, 215–225. DOI: https://doi.org/10.1016/j.proeng.2017.12.007

5. Sun J.; Li Y.; Hasegawa K. A study of self-accelerating decomposition temperature (SADT) using reaction calorimetry. *J. Loss Prev. Process Ind.* **2001**, *14* (5), 331–336. DOI: https://doi.org/10.1016/S0950-4230(01)00024-9

6. *Avogadro home page*. https://avogadro.cc/ (accessed 2021-10-11).

7. *SourceForge MoCalc2012 download page.* https://sourceforge.net/projects/mocalc2012/ (accessed 2021-10-11).

8. *Firefly computational chemistry program home page.* http://classic.chem.msu.su/gran/gamess/ (accessed 2021-10-11).

9. *AlvaDesc home page.* https://www.alvascience.com/alvadesc/ (accessed 2021-10-11).

10. *Semichem page for Codessa III.* http://www.semichem.com/codessa/codessa-new.php (accessed 2021-10-11).

11. Akarachantachote N.; Chadcham S.; Saithanu K. CUTOFF THRESHOLD OF VARIABLE IMPORTANCE IN PROJECTION FOR VARIABLE SELECTION, *Int. J. Pure Appl. Math.* **2014**, *94* (3), 307–322. DOI: https://dx.doi.org/10.12732/ijpam.v94i3.2

12. *Scikit-learn home page.* https://scikit-learn.org/stable/ (accessed 2021-10-11).

13. *Scikit-learn GridSearchCV page.* https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed 2021-10-11).

14. Kaneko H.; Funatsu K. Fast optimization of hyperparameters for support vector regression models with highly predictive ability. *Chemom. Intell. Lab. Syst.* **2015**, 142, 64–69. DOI: https://doi.org/10.1016/j.chemolab.2015.01.001

第3章　クラスタリングを用いたベイズ最適化の初期条件の決定

3.1 はじめに

機械学習，ニューラルネットワーク，ロボット工学，航空宇宙工学，実験計画など，さまざまな科学技術分野で最適解の探索は重要課題の一つであり，その解決法の一つとしてBO[15,16] が広く検討，活用されてきている。化学反応器最適設計にてパレート最適解を求めるために用いられ，CFD 計算の実行回数を減らすことができ，開発された攪拌槽反応器において消費電力の最小化とガス保持量の最大化に貢献した[27]。有機化学分野への適用例も増えており，新規分子探索において，変分オートエンコーダの潜在空間に対して BO を適用した場合に，無効な分子構造を生成しやすいという課題に対して，制約付き BO を用いる事で影響の緩和を試みている[18]。また，BO は最安定な分子コンフォーマーの探索にも用いられている。分子コンフォーマーの探索は探索空間の次元が高く，最適構造を決定するために実施される量子化学計算に膨大な時間を要するという課題があった。BO と量子化学計算を組み合わせることにより量子化学計算コストを約 90%削減することができた[19]。このように, BO を用いた適応的実験計画法は様々な分野で検討・利用されている。一方で，反応条件最適化への適用事例はあまり多くはない。2021 年にベイズ反応最適化のためのフレームワークとオープンソースのソフトウェアツールが開発され, 有機化合物の合成実験の最適化に BO が適用されたという事例が紹介された[17]。その研究ではパラジウム触媒による直接アリール化反応の大規模ベンチマークデータをハイスループット実験(HTE: High-Throughput Experimentation)で収集して，BO と熟練者との最適解探索速度の比較実験を行ったところ，BO は有機合成の熟練者よりも高速に最適条件に到達することができたと報告された。

このように，ベイズ最適化を用いた適応的実験計画法もしくは能動学習(active learning)はさまざまな分野で検討，利用されてきている。一方で，効率的に最適解を探索するためには，ガウス過程回帰モデル構築時に初期サンプルを適切に与えることが重要である。一般に実験前は目的変数や説明変数と目的変数間の関係は未知であるため，説明変数の情報のみを用いて初期サンプルを選択する必要がある。実験実施前にはどの条件が最適解か未知であるため，空間上で可能な限り広く散らばった条件を選択する方が望ましいという考え方が一般的である。初期条件探索方法としてランダムサンプリングまたは D 最適基準に基づいたサンプリング，Latin Hypercube Sampling などが採用されることが多い[15, 25, 28, 33, 34, 35]。

説明変数に化合物を含まない実験計画法の場合，例えば D 最適基準が大きくなるようにサンプルを選択すれば，実験条件同士に相関がないような初期サンプルを選択でき，モデル構築のための実験済み初期サンプルが効率よく得られる。実験条件の一つに化合物を含む実

験計画法の場合，初期サンプルを選択することは他の実験条件に加えて，化合物もしくは化合物の組み合わせを選択することに対応する。化合物の説明変数は化学構造から計算される分子記述子である事が多いため，説明変数間には必ず相関関係がある。D 最適基準に基づいて，説明変数間の相関が小さくなるように初期サンプルを選択する方法が機能しないと考えられる。さらに，化合物を含むサンプルでは，構造的に類似した化合物郡ごとにクラスタを形成することが多い。実験条件が類似しないような初期サンプル獲得を目指す場合，初期サンプルは各クラスタから満遍なく選択される事が望ましいが，D 最適基準ではクラスタの情報が考慮されていないため適切なサンプルを選択することができないと考えられる。D 最適基準のように実験条件間の相関を考慮するよりも，全てのクラスタから初期条件を選択する方が，空間上から取りこぼす事なく，満遍なく初期サンプルを選択できるため，ベイズ最適化における探索効率は向上する可能性がある。

そこで本研究では，構造的に類似した化合物郡ごとにクラスタを形成する特徴を考慮して，クラスタリング後の情報に基づいて初期サンプルを選択する方法を提案する。具体的には，説明変数情報に基づいてクラスタリングを行い，サンプルが化合物群ごとのクラスタを形成したことを確認した後，各クラスタから少なくとも 1 つのサンプルを選択する。クラスタリングに基づく手法は以前からいくつか提案されている。例えば，ファジィクラスタリングに基づく 3 つの初期学習データ選択法が，能動学習の性能向上のために提案されている [36]。一方，化合物の組み合わせに対するベイズ最適化において，クラスタリングに基づく初期サンプリング法の有効性を確認した事例は少ない。本研究では説明変数に化合物構造を含むベイズ最適化の適用対象として，カップリング反応の条件最適化を選択した。実験条件データに前処理を行なった後に，ランダムサンプリング，D 最適基準に基づいたサンプリング，クラスタリングに基づいたサンプリングにて初期サンプルを決定し，それぞれ最適条件に到達するまでの実験数がどのように変化するかを確認することにより提案手法の検証を行った。

3.2 適応的実験計画法

実験計画法とはある因子の条件を複数の水準で変えて実験を行う際に，効率良く，かつ漏れのないようにデータを取得して，統計的に結果を判定する解析手法である。得られた実験データを用いて予測モデルを構築し，良好な目的変数$\boldsymbol{y}$が得られる実験条件$\boldsymbol{x}$を求めることで次の実験条件を提案することができる。一方で，モデルを用いた実験条件やパラメータの選択，実験やシミュレーションの実施，実験やシミュレーション結果を用いたモデル構築を，目標を達成する実験条件が発見できるまで何度も繰り返す手法を適応的実験計画法と呼ぶ。その際，予測モデルとしてガウス過程回帰(GPR: Gaussian Process Regression)モデルを用いて，目的変数の推定値だけでなく，ばらつきも考慮して次の条件を選択する手法が BO である。

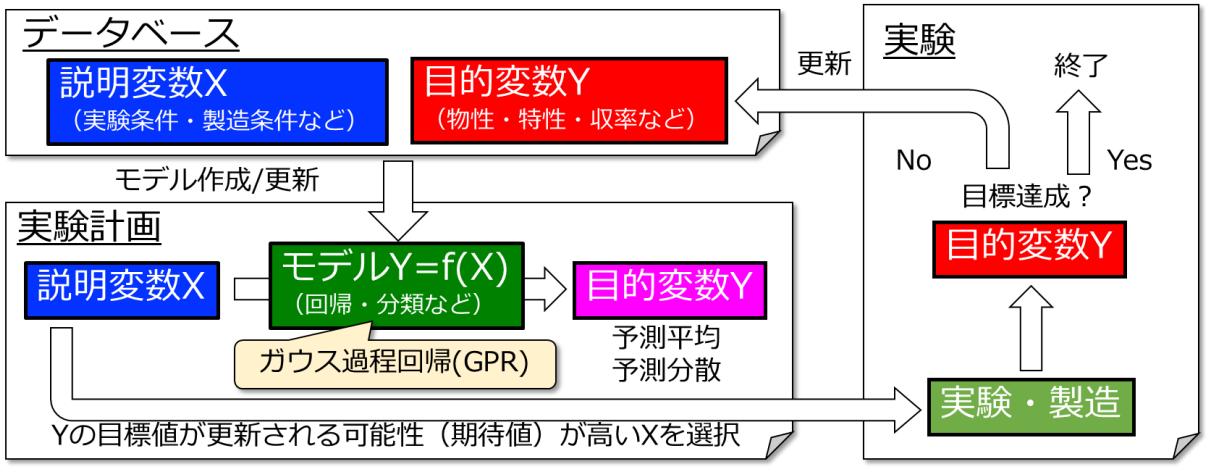


図 11　適応的実験計画法およびベイズ最適化

GPR は統計学や機械学習の分野で使用される回帰手法の一つである。回帰分析において確率的なアプローチを採用し，不確実性をモデル化するために利用される。ガウス過程は平均関数とカーネル関数で定義される確率過程と定義する。モデルの複雑性を表現することができるカーネル関数は論文[17]で利用されている Matern52 を用いた。BO および GPR のハイパーパラメータは当該論文で最適化された値を用いた。ガウス過程f，カーネル関数$k(r)$は以下の通りである。

$$f \sim GP(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')) \tag{11}$$

$$k(r) = \alpha\left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)e^{-\frac{\sqrt{5}r}{l}} \tag{12}$$

$$r = |\boldsymbol{x} - \boldsymbol{x}'| \tag{13}$$

$\mu(\boldsymbol{x})$は平均関数，$k(\boldsymbol{x}, \boldsymbol{x}')$はカーネル関数，$r$実験条件間の距離，$\alpha$，$l$はハイパーパラメータである。実験条件$\boldsymbol{x}$におけるガウス過程の事後分布における平均$\mu(\boldsymbol{x})$は次式で与えられる。

$$\mu(\boldsymbol{x}) = k(\boldsymbol{x})^T(K_\theta + \sigma_n^2 I)^{-1}\boldsymbol{y} \tag{14}$$

$k(\boldsymbol{x})$は実験条件$\boldsymbol{x}$と実験条件間の共分散ベクトル，K_θは全実験条件間の共分散行列，σ_n^2は

推定されたノイズの分散，Iは単位行列，$\boldsymbol{y}$は対応する目的変数ベクトルである。実験条件$\boldsymbol{x}$におけるガウス過程の事後分布における分散は以下で表される。

$$\sigma^2(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}') - k(\boldsymbol{x})^T (K_\theta + \sigma_n^2 I)^{-1} k(\boldsymbol{x}) \tag{15}$$

GPR モデルでは出力が正規分布で表されるため，最尤推定法でハイパーパラメータを計算する事ができる。以下の対数尤度関数が最大となるようにハイパーパラメータを決定する。

$$\log p(\boldsymbol{y}|\theta) = -\frac{1}{2}\boldsymbol{y}^T (K_\theta + \sigma_n^2 I)^{-1} \boldsymbol{y} - \frac{1}{2}\log|K_\theta + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \tag{16}$$

BO における獲得関数は更新幅の期待値である EI(Expected Improvement)を用いた。更新幅$I(\boldsymbol{x})$現在得られている目的変数$f(\boldsymbol{x})$の最良値f^+に対する$f(\boldsymbol{x})$の増加分を表す。

$$I(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) - f^+ - \delta & f(\boldsymbol{x}) > f^+ \\ 0 & f(\boldsymbol{x}) < f^+ \end{cases} \tag{17}$$

実験条件$\boldsymbol{x}$における更新幅$I(\boldsymbol{x})$の期待値$EI(\boldsymbol{x})$は次のような形となる。

$$EI(\boldsymbol{x}) = \begin{cases} I(\boldsymbol{x})\Phi\left(\frac{I(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right) + \sigma(\boldsymbol{x})\varphi\left(\frac{I(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right) & \sigma(\boldsymbol{x}) > \delta \\ 0 & \sigma(\boldsymbol{x}) < \delta \end{cases} \tag{18}$$

ここで$I(\boldsymbol{x})$ は事後分布における予測平均$\mu(\boldsymbol{x})$の更新幅，$\sigma(\boldsymbol{x})$は事後分布における標準偏差を意味する。Φとφはそれぞれ標準正規分布の累積分布関数と確率密度関数である。閾値を表すパラメータδは 0.01 に設定した。

次の実験条件の選択は$EI(\boldsymbol{x})$の値によって行われる。実験条件は各種化合物，温度などの組み合わせで表現されるため，実験空間は有限となり，期待値が最も高くなるような実験条件を選択する事ができる。

$$arg\ max_{\boldsymbol{x}\in X}\ EI(\boldsymbol{x}) \tag{19}$$

$\boldsymbol{x}$は実験条件，Xは実験空間，$EI(\boldsymbol{x})$は実験条件$\boldsymbol{x}$における更新幅$I(\boldsymbol{x})$の期待値である。

実験を複数並列に実施する場合は，Kriging believer アルゴリズム[32]を用いて，$EI(\boldsymbol{x})$が最大となる$\boldsymbol{x}$を繰り返し計算する。既知のデータにガウス過程の事後分布における平均$\mu(\boldsymbol{x})$を追加し，予測平均値を実測値とみなして GPR モデルを更新することにより行われる。BO を実施する際の具体的な流れは以下の通りである(図 12)。

1. 実験空間(溶媒，配位子，温度など)を定義し，初期サンプルを選択する。既にサンプル $\mathbf{x}$ および対応する $\mathbf{y}$ の情報がある場合は 3 に進む。
2. 選択されたサンプルに基づき実験を行う。
3. サンプル $\mathbf{x}$ と対応する $\mathbf{y}$ の情報を用いてガウス過程回帰(GPR)モデルを構築する。計算では規格化された $\mathbf{x}$ および $\mathbf{y}$ が用いられる。
4. 期待改善度 EI が最も大きなサンプルを次の実験条件として選択する。実験を複数並列に実施する場合は繰り返し計算する。
5. 選択されたサンプルに基づき実験を行う。
6. 目的変数 $\mathbf{y}$ が目標値に到達するまで 3~5 を繰り返す。

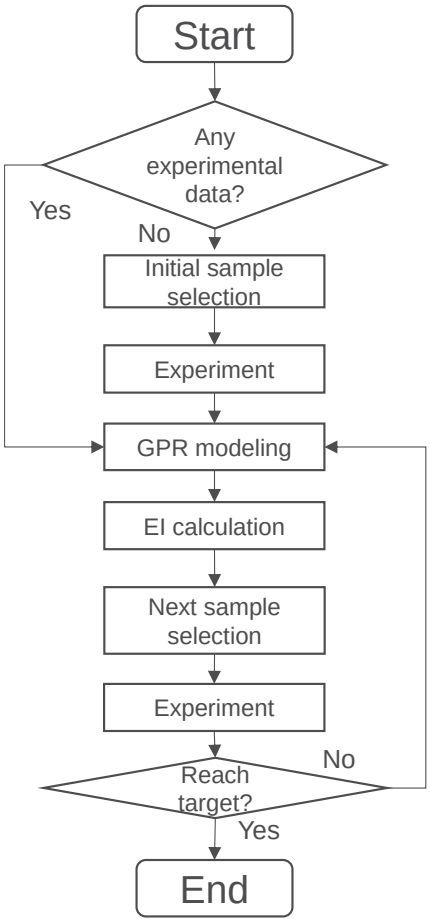


図 12　ベイズ最適化の実行手順

3.3 初期条件決定方法

I. D 最適基準に基づく初期条件決定法

D 最適基準は実験計画法などの実験条件の選択時に一般的に用いられる。一般的な線形回帰モデルは以下の式で表される。

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} \tag{20}$$

$\boldsymbol{y}$ は目的変数ベクトル，$\boldsymbol{X}$は説明変数行列，$\boldsymbol{w}$は回帰係数ベクトルである。回帰係数ベクトル$\boldsymbol{w}$は下式で計算できる。

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{21}$$

同じ実験回数であれば，推定量の分散ができる限り小さく，共分散が 0 に近いサンプル選択が望ましいため，分散共分散行列($\sigma^2\boldsymbol{X}^T\boldsymbol{X}$) の成分ができるだけ小さくなる(($\boldsymbol{X}^T\boldsymbol{X}$)$^{-1}$ を最小化する or $\boldsymbol{X}^T\boldsymbol{X}$ の行列式を最大化する)ようにサンプルを選択すればよい。この$\boldsymbol{X}^T\boldsymbol{X}$ の行列式のことを D 最適基準と呼ぶ。

実験計画を作成する際，実験因子および操作範囲，カテゴリーなどを定義した後にそれらの組み合わせによりサンプル群を生成する。実験条件が類似しないような初期サンプル獲得を目指す場合，初期サンプルは実験空間から満遍なく選択される事が望ましいが，初期サンプルをランダムにサンプリングした場合，選択されたサンプルの中に，偶然似たようなサンプルが含まれる可能性がある。選択されたサンプルに対して D 最適基準を繰り返し計算して，値が最大となる組み合わせを初期サンプルとして採用する。具体的な流れは以下の通りである(図 13-a)。

1. 定義された実験空間内で各因子を組み合わせて複数のサンプルを作成する。
2. 得られたサンプルからランダムに初期サンプルを選択する。
3. 選択されたサンプルの D 最適基準の値を計算する。
4. D 最適基準の最大値が更新されないようになるまで 2~3 を繰り返す。

実験空間の類似性を利用した類似手法としては，Kennard-Stone サンプリング[25]などがあるが，本研究ではそれらの代表的な手法として D 最適基準に基づくサンプリングを用いた。

II. クラスタリングに基づく初期条件決定法

初期サンプル選択時の前処理としてクラスタリングを用いる。クラスタリング手法として PAM(Partitioning Around Medoids)[30]と DBSCAN(Density-based spatial clustering of applications with noise)[31]を用いた。

PAM においてクラスタは medoid で代表される。medoid はクラスタ内のデータ点でその点以外のクラスタ内の点でまでの非類似度の総和が最小になる点である。クラスタを $X_i = \{x\}$，データ間の距離を$d(x,y)$とした時，medoid は次式で表現される。

$$\arg\ min_{x \in X_i} \sum_{y \in (X_i - \{x\})} d(x,y) \tag{22}$$

最初はランダムに k 個の medoid を選び，各 medoid とそれ以外の点とを交換し，評価値が改善するように交換を繰り返す。評価値が改善されなくなったら終了とする。クラスタ数kはあらかじめ与える必要がある。

DBSCAN は密度準拠クラスタリングである。DBSCAN では，次の条件に基づいて，各データ点に特別なラベルを割り当てる。

- 半径 ε 以内に少なくとも MinPts 個以上の隣接点がある点はコア点とみなされる。
- 半径 ε 以内の隣接点の個数が MinPts に満たないが，コア点の半径ε以内に位置するような点は，ボーダー点とみなされる。
- その他全ての点はノイズとみなされる。

その後，コア点ごとにクラスタを形成し，各ボーダー点を最近接コア点のクラスタに割り当てることによりクラスタリングを行う。k-means や k-medoids と異なり，クラスタ数を指定せずに実行する事が可能であるが，クラスタの判定に用いる半径距離 ε やコア点とみなされる隣接点の閾値 MinPts をあらかじめハイパーパラメータとして与える必要がある。

本研究ではクラスタリング後の情報に基づいて初期サンプルを選択する方法を提案する。実験計画を作成する際，実験空間を定義した後にそれらの組み合わせによりサンプル群を生成する。実験条件が類似しないような初期サンプル獲得を目指す場合，初期サンプルは実験空間から満遍なく選択される事が望ましい。説明変数情報に基づいてクラスタリングを行い，サンプルが化合物群ごとのクラスタを形成したことを確認した後，各クラスタから少な

くとも 1 つのサンプルをランダムに選択することで, 実験条件が類似しないような初期サンプル選択が可能となる。目的変数に大きな影響を与える因子については，ドメイン知識を活用してクラスタリングを行う。クラスタ数は初期標本数より少なくなければならない。また，形成されたクラスタから満遍なく初期サンプルを選択する事が望ましいため, 初期サンプル数はクラスタ数の n 倍(n=1,2,…)とすることが望ましい。初期サンプル数がクラスタ数の n 倍でない場合であっても, 各クラスタから取得するサンプル数は可能な限りばらつきがないようにすべきである。クラスタリング手法として適切なハイパーパラメータを持つ DBSCAN を採用する事によりクラスタ数を自動的に決定することができる。k-means や k-medoids などの方法はクラスタ数をあらかじめ与える必要があり, 形成されたクラスタに属するサンプル数に偏りが生じるなどの理由から本手法には不向きである。どのようなクラスタが形成されるかは適用対象次第ではあるが，経験上 1 または 2 つの要因で(溶媒，配位子などの化合物ごとに)クラスタを形成する事が多い。具体的な流れは以下に示す通りである(図 13-b)。

1. 定義された実験空間内で各因子を組み合わせて複数のサンプルを作成する。
2. 得られたサンプルに対してクラスタリングを行う。
3. 各クラスタからサンプルを選択し，それらを初期サンプルとする。

サンプリングの方法はランダムサンプリングだけでなく，D 最適基準に基づくサンプリングとすることも可能である。D 最適基準に基づいたサンプリングを行う場合は上記に加え以下の 4, 5 を追加で実施する(図 13-c)。

4. 選択されたサンプルの D 最適基準の値を計算する。
5. D 最適基準の最大値が更新されないようになるまでを 3~4 を繰り返す。

同様のクラスタリング手法には，k-means，階層的クラスタリング[26]，DBSCAN，PAM などがあるが，本研究ではそれらの代表として PAM を使用した。

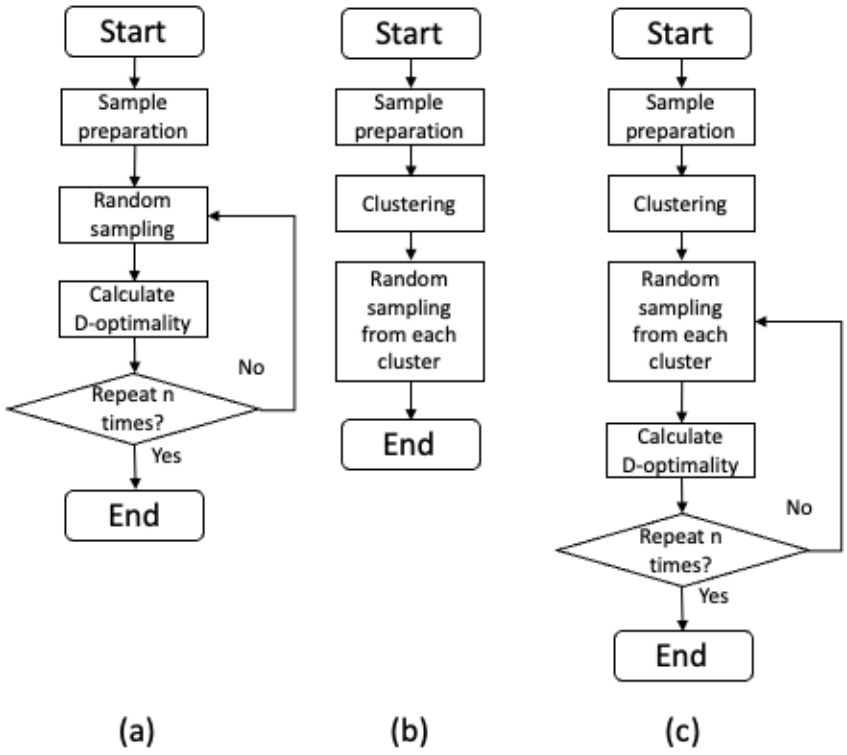
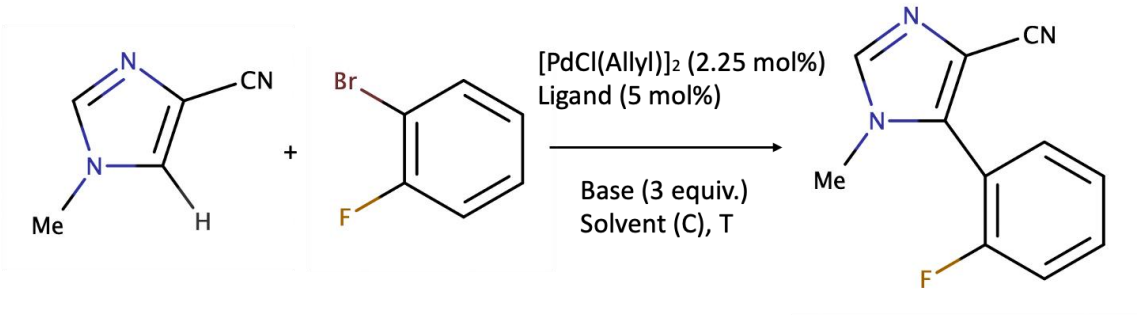


図 13　初期サンプル決定方法（a: D 最適基準によるサンプリング, b: クラスタリングを基にしたランダムサンプリング, c: クラスタリングを基にした D 最適基準によるサンプリング）

3.4 データセット

本検討では直接的アリール化反応(スキーム 1)に対してベイズ最適化を適用した論文[17]にて報告された実験データを用いた。当該論文では HTE を行い，短時間で大量の実験データを取得した。実験条件は反応基質，触媒と配位子の当量は固定し，反応温度 3 種類，基質濃度 3 種類，配位子種 12 種類，溶媒種 4 種類，塩基種 3 種類の組み合わせとなる全 1,728 通り。収率が 95%を超える条件は 10 通りで全体の 0.58%であった。配位子，溶媒，塩基は化合物であり，カテゴリーデータであるため，MORDRED[21, 22]を用いて mol ファイルの分子構造を 0 次元，1 次元, 2 次元記述子に変換して用いた[24]。化合物を取り扱う場合，一般的に説明変数の数が非常に多くなる事が多い。約 5,800 変数からなる説明変数は高次元データのままでは解釈が難しく，クラスタリングの妥当性の確認が困難となるため，クラスタリングの前処理として数値は全て平均 0，標準偏差 1 に標準化され，主成分分析および t-SNE(t-Distributed Stochastic Neighbor Embedding)[29]を用いて潜在変数が作成された。主成分分析では累積寄与率がほぼ 1 となる第 20 主成分まで変数数を削減した。t-SNE のハイパーパラメータである Perplexity を 10 から 1,000 まで変化させてみたものの，形成され

たクラスタに大きな変化はなかったため，本研究では Perplexity 85 の結果を用いて検証を実施した。

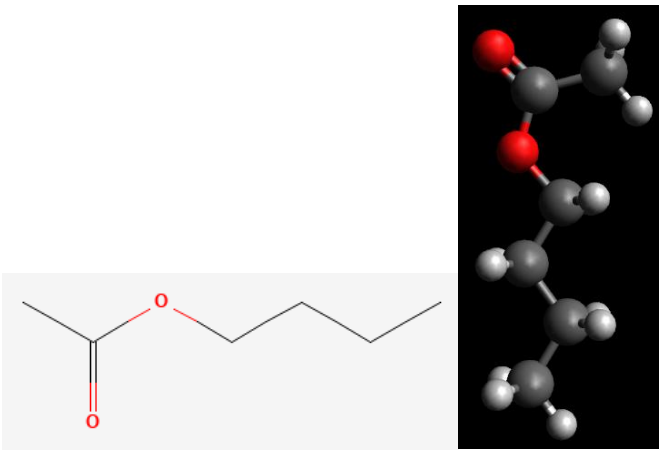
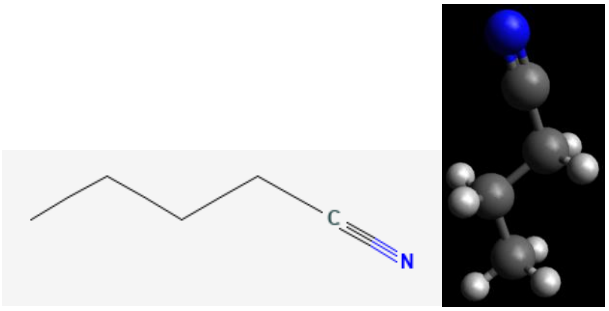
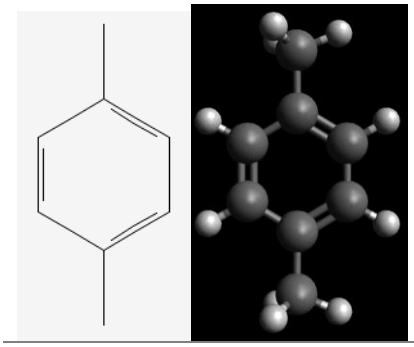
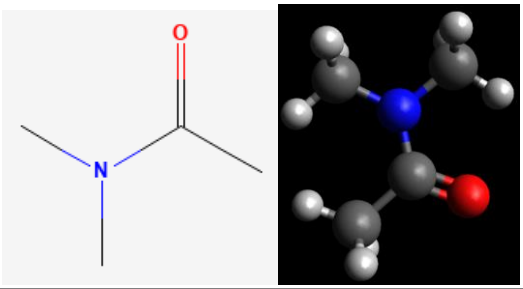


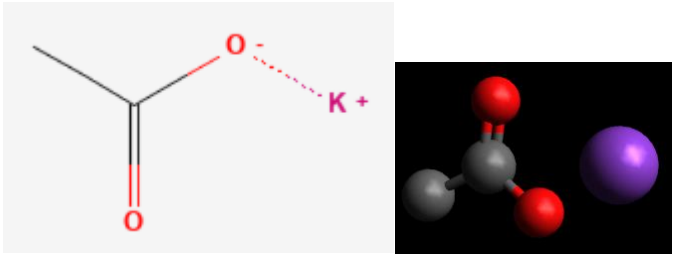
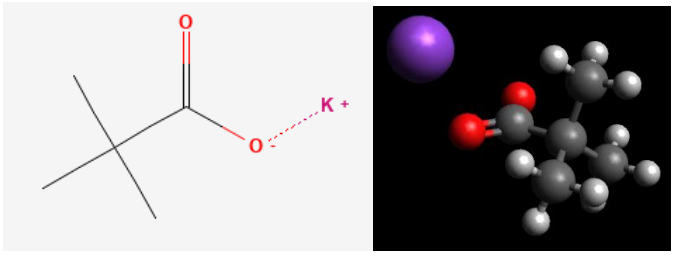
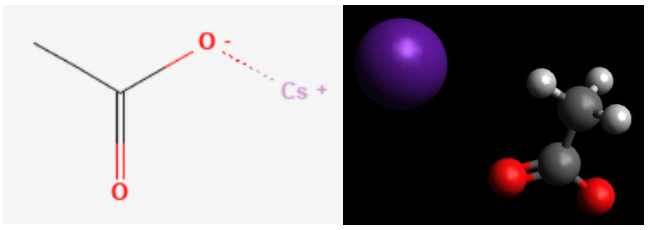
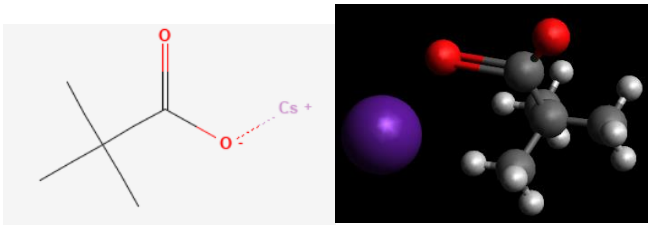
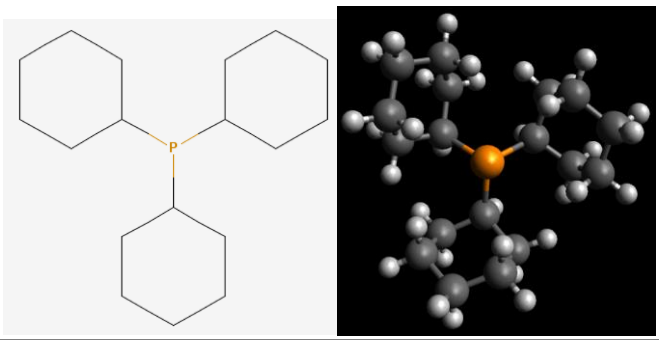
スキーム 1. パラジウム触媒による直接的アリール化反応

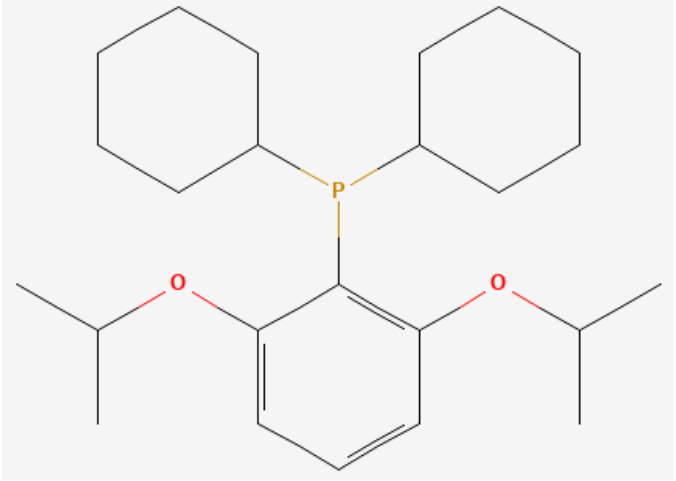
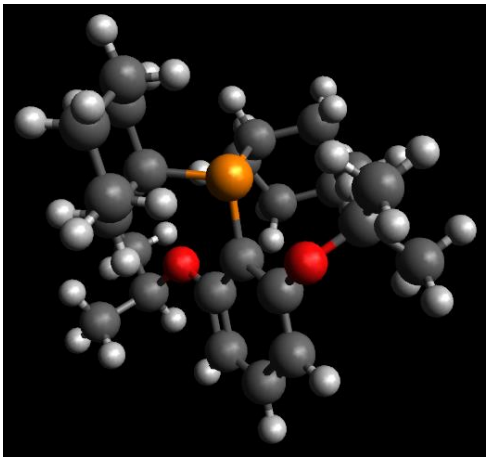
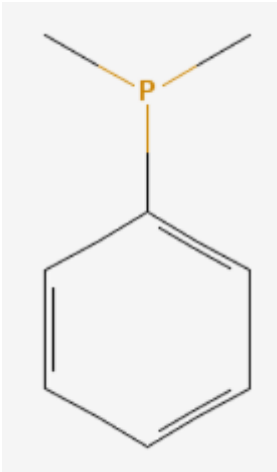
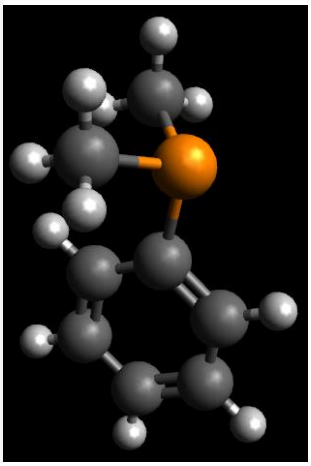
表 11　実験条件(パラジウム触媒による直接的アリール化)

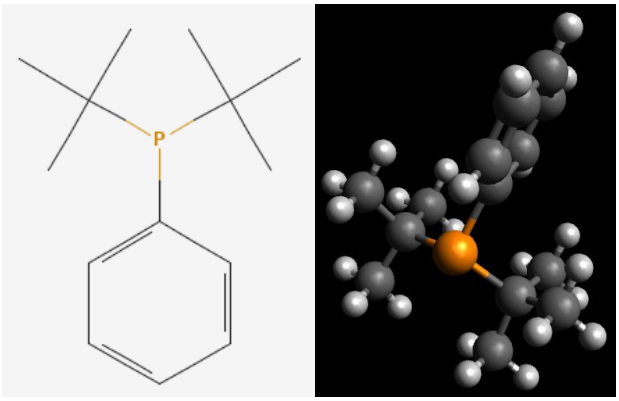
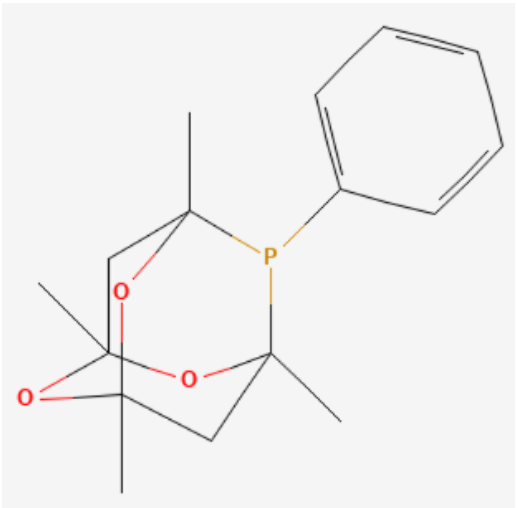
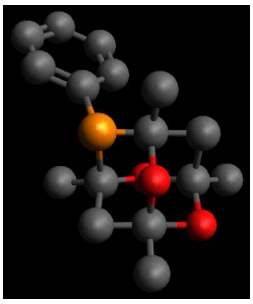
配位子	塩基	溶媒	温度	濃度
PCy3	KOAc	BuOAc	90	0.057
GorlosPhos	KOPiv	BuCN	105	0.100
PPhMe2	CsOAc	p-Xylene	120	0.153
PPht-Bu2	CsOPiv	DMAc		
CgMe-PPh				
XPhos				
BrettPhos				
t-BuPh-CPhos				
PPh2Me				
PPh3				
P(fur)3				
JackiePhos				

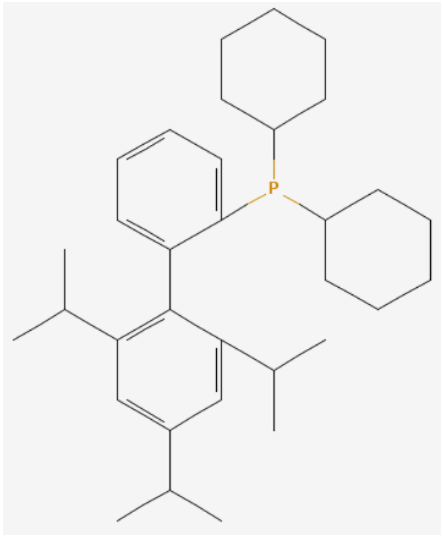
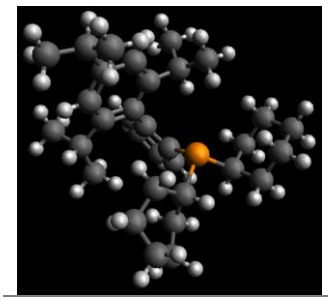
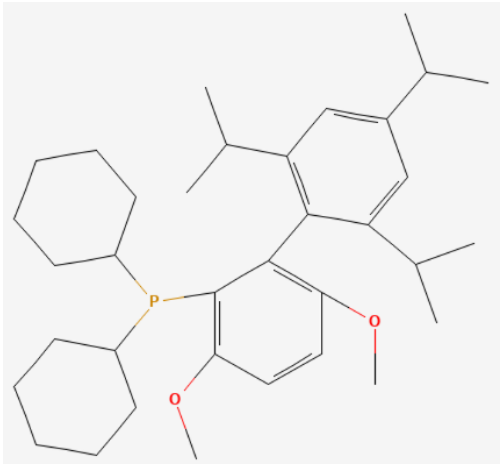
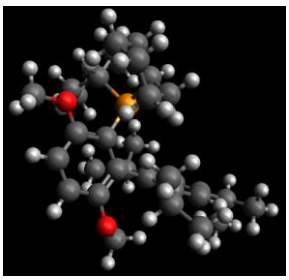
表 12　反応試剤情報(パラジウム触媒による直接的アリール化)

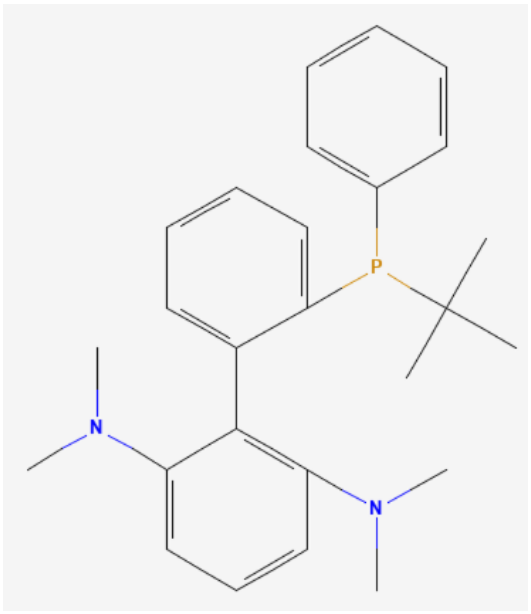
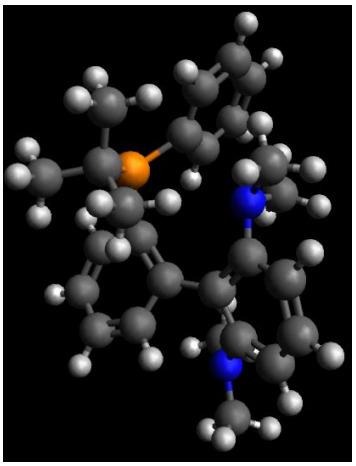
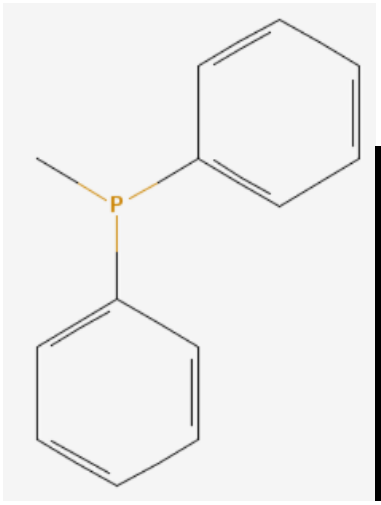
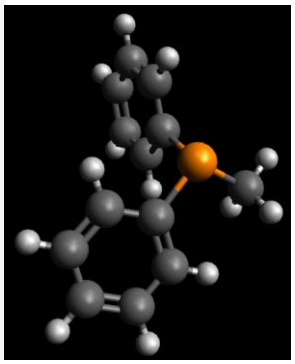
Name	SMILE
BuOAc 	CCCCOC(C)=O
BuCN 	CCCC#N
p-Xylene 	Cc1ccc(C)cc1
DMAc 	CC(=O)N(C)C

KOAc 	CC(=O)[O-].[K+]
KOPiv 	CC(C)(C)C(=O)[O-].[K+]
CsOAc 	CC(=O)[O-].[Cs+]
CsOPiv 	CC(C)(C)C(=O)[O-].[Cs+]
PCy3 	C1CCC(P(C2CCCCC2)C2CCCCC2)CC1

GorlosPhos  	CC(C)Oc1cccc(OC(C)C)c1P(C1CCCCC1)C1CCCCC1
PPhMe2  	CP(C)c1ccccc1

PPht-Bu2  	CC(C)(C)P(c1ccccc1)C(C)(C)C
CgMe-PPh  	CC12C[C@@]3(C)O[C@](C)(C[C@](C)(O3)P1c1ccccc1)O2

XPhos  	CC(C)c1cc(C(C)C)c(-c2ccccc2P(C2CCCCC2)C2CCCCC2)c(C(C)C)c1
BrettPhos  	COc1ccc(OC)c(P(C2CCCCC2)C2CCCCC2)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C

t-BuPh-CPhos  	CN(C)c1cccc(N(C)C)c1-c1ccccc1P(c1ccccc1)C(C)(C)C
PPh2Me  	CP(c1ccccc1)c1ccccc1

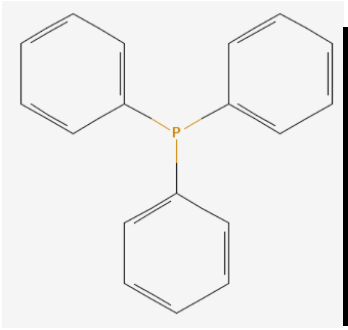
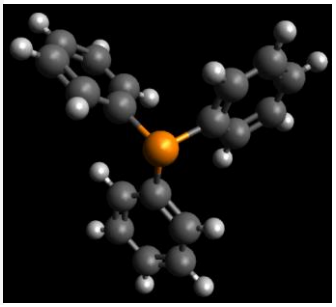
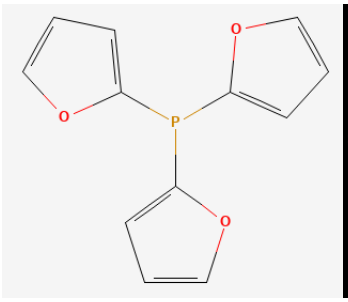
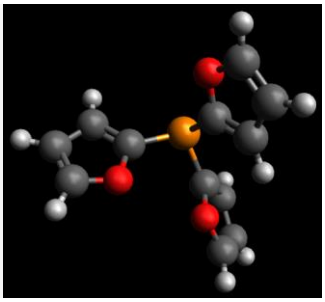
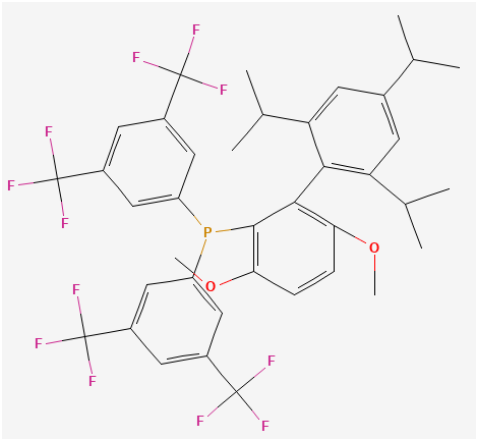
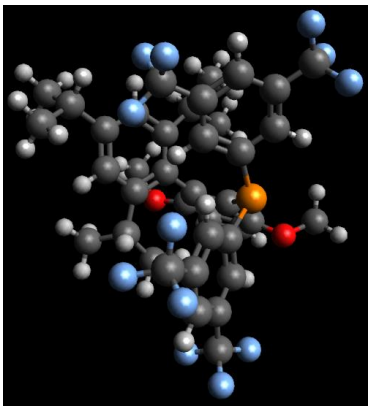
PPh3  	c1ccc(P(c2ccccc2)c2ccccc2)cc1
P(fur)3  	c1coc(P(c2ccco2)c2ccco2)c1
JackiePhos  	COc1ccc(OC)c(P(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C

表 13　収率上位 10 条件(パラジウム触媒による直接的アリール化)

No.	配位子	塩基	溶媒	濃度(M)	温度[℃]	収率[%]
1	CgMe-PPh	CsOAc	DMAc	0.153	105	100
2	CgMe-PPh	CsOPiv	DMAc	0.153	105	100
3	CgMe-PPh	CsOAc	BuCN	0.153	120	99.98
4	CgMe-PPh	KOPiv	DMAc	0.153	120	99.81
5	CgMe-PPh	CsOAc	DMAc	0.153	120	99.22
6	CgMe-PPh	KOPiv	DMAc	0.153	105	98.49
7	CgMe-PPh	KOAc	DMAc	0.153	120	98.38
8	CgMe-PPh	KOAc	DMAc	0.057	120	96.64
9	CgMe-PPh	CsOAc	BuCN	0.153	105	96.38
10	CgMe-PPh	CsOAc	DMAc	0.057	120	95.48

反応条件のベイズ最適化では，化合物種や温度，濃度など離散的な値の組み合わせを最適化することになるため，実験条件が空間上に離散的に分布し，使用する化合物ごとに，所属する条件数がほぼ均一なクラスタを形成する。今回のケースでは配位子ごとに 12 個のクラスタが構築され，特定のクラスタに収率が高い実験条件が集中した(図 14)。配位子はカップリング反応において最も重要な因子であり，配位子によりクラスタが形成されたことは有機合成化学者として理解できる。t-SNE では次元圧縮したい高次元データを低次元空間上の点に対応付ける。その際，高次元空間におけるサンプル同士の近さが低次元空間におけるサンプル同士の近さに反映されるよう学習が行われる。今回は配位子が高次元空間上において相対的に化合物間の距離(非類似度)が大きく表現されていたため，配位子ごとのクラスタが形成されたのだと思われる。

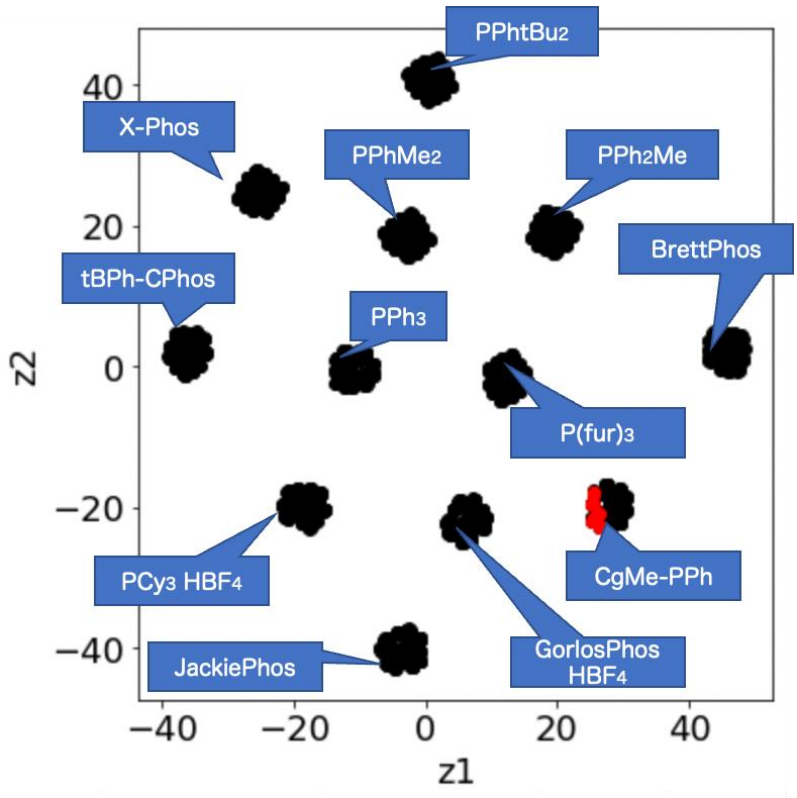


図 14　t-SNE による可視化結果での各クラスタの分類(赤点: 収率上位 10 条件)

3.5 結果と考察

初期サンプル選択方法およびクラスタリングが探索結果に与える影響を確認するために，以下の 6 種類のサンプリング法を用いて初期条件を決定した場合の，ベイズ最適化における探索性能の比較を行った。

- ランダムサンプリング(Random)
- D 最適基準に基づくサンプリング(D-optimal)
- DBSCAN によるクラスタリング後に各クラスタからランダムサンプリング(DBSCAN+Random)
- DBSCAN によるクラスタリング後に各クラスタから D 最適基準に基づくサンプリング(DBSCAN+D-Optimal)
- PAM(Partitioning Around Medoids)によるクラスタリング後に各クラスタからランダムサンプリング(PAM+Random)
- PAM(Partitioning Around Medoids)によるクラスタリング後に各クラスタから D 最適基準に基づくサンプリング(PAM+D-Optimal)

計算手順を図 15 に示す。全てのケースにおいて，1 ラウンドあたりの実験数やクラスタからのサンプリング数を変化させて収率 95%以上に到達するまでに必要なラウンド数の平均値と標準偏差の変化を確認した。いずれも小さい方が探索性能は高いと判断できる。信頼できる結果を得るために，各条件の実行結果が収束するまで 10,000 回以上繰り返し計算を行った。クラスタリングを用いてサンプリングを行う際のクラスタ数は t-SNE で形成されたクラスタ数と同数の 12 とした。

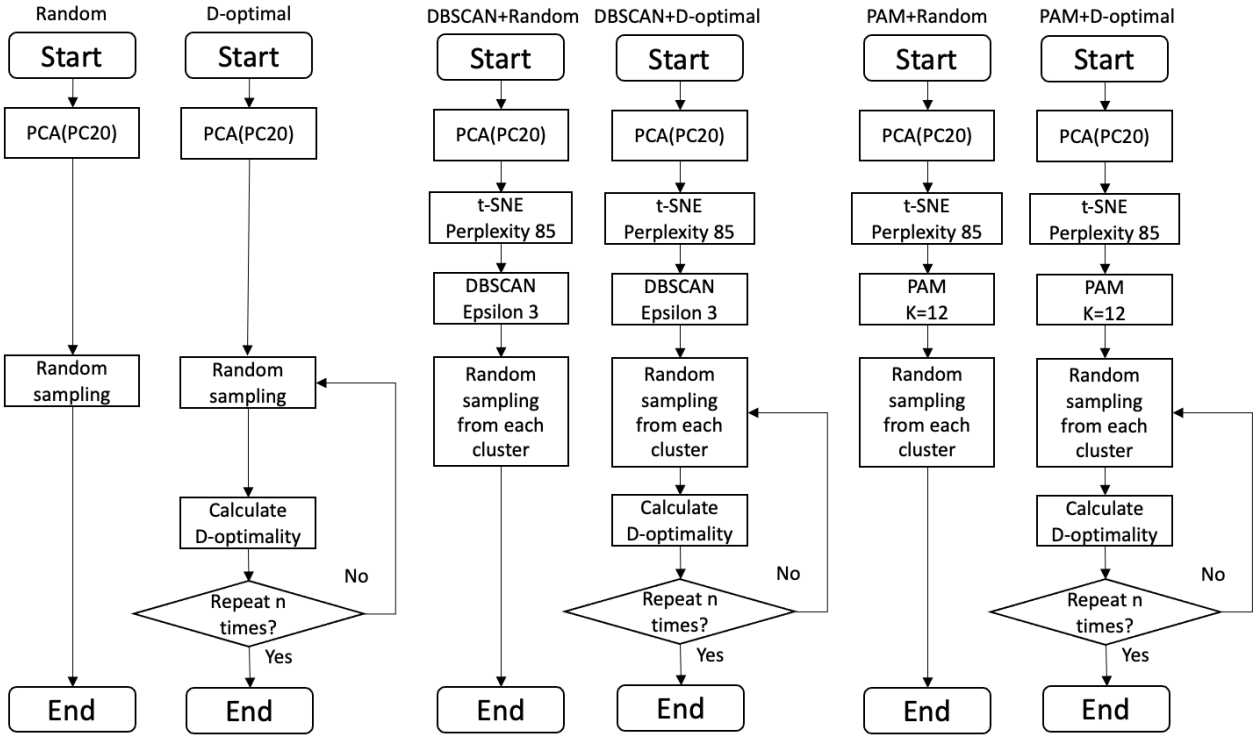
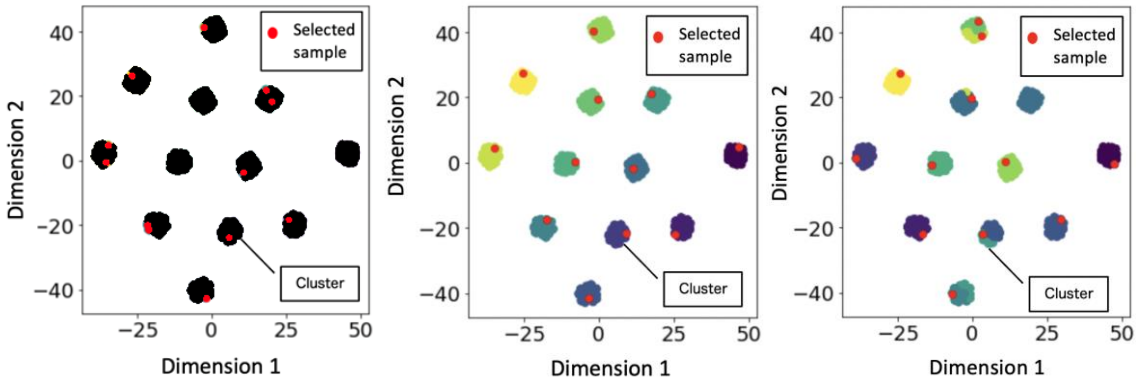


図 15　初期サンプル選択方法一覧

それぞれの手法にて初期条件を決定した際にどのような違いが見られるかを確認するために，クラスタリングした際の結果を色分けして t-SNE で可視化を行った。図 16(a)～(c)の色分けは異なるクラスタを表しており，赤点は選択された初期サンプルを示している。クラスタ情報を活用せずに D 最適基準に基づいたサンプリングを行なった場合(D-Optimal)，同一のクラスタから 2 つ以上初期サンプルを選択しており，サンプルを選択できていないクラスタが存在する場合がみられた（図 16(a)）。DBSCAN でクラスタリングを行った後に各クラスタからランダムサンプリングした場合(DBSCAN+Random)，全てのクラスタから満遍なく初期サンプルを選択できた(図 16(b))。PAM でクラスタリングを行い，各クラスタからランダムサンプリングした場合(PAM+Random)，クラスタリングの結果が t-SNE で実施した可視化の結果と一致しておらず，全くサンプリングされていないクラスタや 1 つのク

ラスタから 2 サンプル選択されているケースがあることがわかった。また，各クラスタに所属する実験条件数が均一でなく偏りが生じていた(図 16(c))。



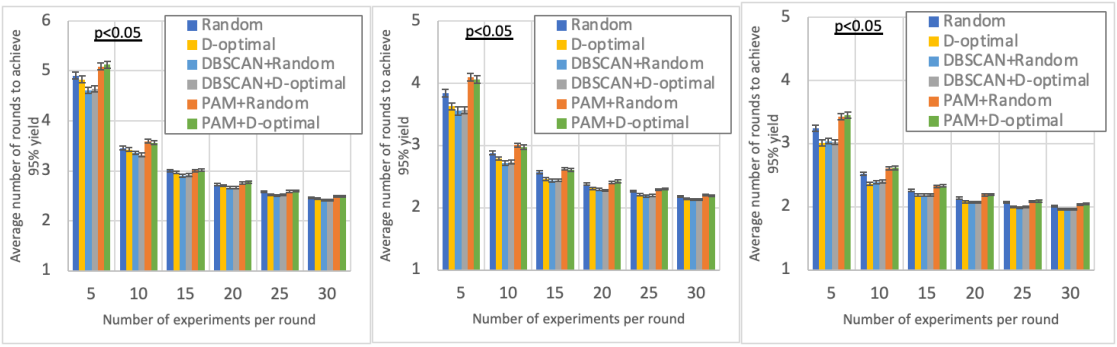
(a) D-Optimal　　(b) DBSCAN+Random　　(c) PAM+ Random

図 16　t-SNE による可視化結果での初期条件の配置(赤点:初期条件，色:クラスタリング結果)

1 ラウンドあたりの実験数を変化させて，収率 95%以上の実験条件に到達するまでに要したラウンド数の平均値及び標準偏差を確認した。その結果を図 17(a)～(c)に示す。グラフの縦軸は歩留まり 95%に達するまでに要したラウンド数の平均値，横軸は 1 ラウンドあたりの実験回数を示している。各図のエラーバーは平均値の標準誤差の 95%信頼区間を示す。すべての場合において，平均値の差は統計的に有意であった。DBSCAN+Random は Random や D-optimal よりも少ないラウンド数で, 95%以上の収率に達することができた。これはカップリング反応において重要な因子である配位子に関して, 全ての化合物を少なくとも 1 つ以上初期条件として選択する事ができたためである。

初期条件から全ての配位子を網羅できた事で，収率が高い条件を早い段階から探索できた。特にラウンドあたりの実験数，初期サンプリング数が少ないほど，その効果は大きくなった。また，ラウンド数や初期サンプリング数が大きくなるほど，サンプリング手法間の差は小さくなった。ラウンドあたりの実験数や初期サンプリングが多くなるほど，実験条件が選択されないクラスタが少なくなるためと推察する。提案手法はランダムサンプリングや D-optimality に基づくサンプリングに比べ，最大 5%少ない実験回数で最適解に到達することを確認した。DBSCAN クラスタリング後に D 最適基準に基づいたサンプリング(DBSCAN+D-Optimal)を実施したが，クラスタリング後にランダムサンプリングした場合(DBSCAN+Random)と大きな違いは見られなかった。DBSCAN クラスタリング後に D 最適基準に基づいたサンプリングを行った時点で満遍なく初期条件を選択できていたため, D 最適基準に基づいたサンプリングの効果が相対的に小さくなったように見えているのではないかと推察する。D 最適基準に基づいたサンプリング(D-Optimal)，クラスタリング後に

サンプリングする方法(DBSCAN+Rand, DBSCAN+D-Optimal)，いずれの方法であってもランダムサンプリング(Random)よりも良好な結果を示した。一方で，PAM でクラスタリングを行った場合(PAM+Random, PAM+D-Optimal)，ランダムサンプリング(Random)よりも探索に要する実験数の平均値や標準偏差が 3～5 %程度大きくなった。クラスタ間に実験条件や数に偏りが生じていることに加え，配位子種でクラスタを形成できていないため，各クラスタから 1 つずつサンプルを選択しても，全ての配位子種を網羅することができていない。このように，最適条件が所属するクラスタから実験条件をサンプリングできた場合であっても，クラスタリングの結果(例えばクラスタに所属するサンプル数，サンプル種など)次第で，探索性能がよくなる可能性もあれば悪くなる可能性もあることがわかった。実際に最適化を行う際には，事前にどのクラスタに最適解が存在するかを知ることはできないため，最適解付近の実験条件をサンプリングする確率を意図的に上げることは難しい。一方で，クラスタに属する条件数を同程度とすることで，探索に要する実験数のばらつきを小さくすることは可能である。



(a)12 initial conditions　(b) 24 initial conditions　(c) 36 initial conditions

図 17　収率 95%以上に到達するまでに要した平均ラウンド数と 1 ラウンドあたりの実験数

表 14　収率 95%到達に要する実験ラウンド数平均(初期条件数 12)

Exp. num/round	Random	D-optimal	DBSCAN Random	DBSCAN D-optimal	PAM Random	PAM D-optimal
5	4.90	4.83	4.61	4.64	5.09	5.12
10	3.46	3.43	3.36	3.32	3.59	3.56
15	3.01	2.97	2.90	2.92	3.00	3.01
20	2.72	2.70	2.66	2.66	2.75	2.78
25	2.59	2.53	2.50	2.52	2.59	2.60
30	2.46	2.45	2.41	2.41	2.49	2.49

表 15　収率 95%到達に要する実験ラウンド数標準偏差(初期条件数 12)

Exp. num/round	Random	D-optimal	DBSCAN Random	DBSCAN D-optimal	PAM Random	PAM D-optimal
5	3.47	3.43	3.33	3.32	3.54	3.56
10	1.81	1.76	1.73	1.70	1.87	1.84
15	1.29	1.28	1.24	1.23	1.29	1.26
20	1.03	1.02	0.98	0.99	1.05	1.04
25	0.88	0.87	0.84	0.84	0.87	0.88
30	0.77	0.76	0.75	0.75	0.77	0.77

表 16　収率 95%到達に要する実験ラウンド数平均(初期条件数 24)

Exp. Num/round	Random	D-optimal	DBSCAN Random	DBSCAN D-optimal	PAM Random	PAM D-optimal
5	3.84	3.63	3.55	3.57	4.10	4.11
10	2.88	2.79	2.71	2.73	3.01	2.98
15	2.57	2.46	2.44	2.44	2.62	2.61
20	2.38	2.31	2.30	2.28	2.40	2.42
25	2.26	2.21	2.19	2.20	2.29	2.30
30	3.84	3.63	3.55	3.57	2.20	2.20

表 17　収率 95%到達に要する実験ラウンド数標準偏差(初期条件数 24)

Exp. num/round	Random	D-optimal	DBSCAN Random	DBSCAN D-optimal	PAM Random	PAM D-optimal
5	2.85	2.70	2.58	2.58	3.03	3.05
10	1.56	1.45	1.40	1.42	1.64	1.61
15	1.17	1.08	1.03	1.03	1.18	1.18
20	0.94	0.89	0.85	0.84	0.94	0.95
25	0.81	0.76	0.75	0.74	0.81	0.82
30	0.71	0.68	0.67	0.67	0.72	0.71

表 18　収率 95%到達に要する実験ラウンド数平均(初期条件数 36)

Exp. Num/round	Random	D-optimal	DBSCAN Random	DBSCAN D-optimal	PAM Random	PAM D-optimal
5	3.25	3.01	3.04	3.02	3.42	3.49
10	2.52	2.36	2.39	2.40	2.61	2.62
15	2.26	2.18	2.18	2.19	2.32	2.32
20	2.14	2.08	2.07	2.07	2.19	2.19
25	2.07	2.00	1.99	2.00	2.08	2.07
30	2.01	1.96	1.96	1.96	2.04	2.05

表 19　収率 95%到達に要する実験ラウンド数標準偏差(初期条件数 36)

Exp. num/round	Random	D-optimal	DBSCAN Random	DBSCAN D-optimal	PAM Random	PAM D-optimal
5	2.46	2.20	2.21	2.16	2.63	2.76
10	1.38	1.22	1.21	1.24	1.42	1.48
15	1.00	0.92	0.92	0.92	1.06	1.05
20	0.83	0.78	0.76	0.77	0.86	0.87
25	0.72	0.66	0.68	0.67	0.74	0.75
30	0.67	0.62	0.62	0.62	0.69	0.67

また，今回は配位子ごとに 12 個のクラスタが構築され，特定のクラスタに収率が高い実験条件が集中した場合で検証を行ったが，化合物およびクラスタリング方法，ハイパーパラメータの設定によっては別の因子(温度，溶媒，塩基，触媒など)でクラスタが形成されることもありえる。その場合，収率が高いサンプルが特定のクラスタに集中せず，複数のクラスタに散らばる可能性もある。その場合，影響が大きな因子でクラスタリングされた場合と比べると探索性能が低下することは明らかである。例えば今回のケースにおいて，配位子ではなく溶媒でクラスタリングされたと仮定すると，各クラスタから 1 サンプル選択することにより全ての溶媒種から 1 種類ずつ選択することはできるが，今回の反応において最も影響が大きな因子である配位子を網羅できるとは限らないためである。配位子の情報を意図的に削除してクラスタリングした例を図 18 に示す。溶媒と塩基の組み合わせでクラスタが形成され，収率上位 10 の実験条件は単一のクラスタに集中せず，複数のクラスタに分散している。この場合，各クラスタから初期条件をサンプリングすると，全ての塩基と溶媒の組み合わせは網羅できるが，全ての配位子を網羅することはできない。よって，探索性能が向上することもない。

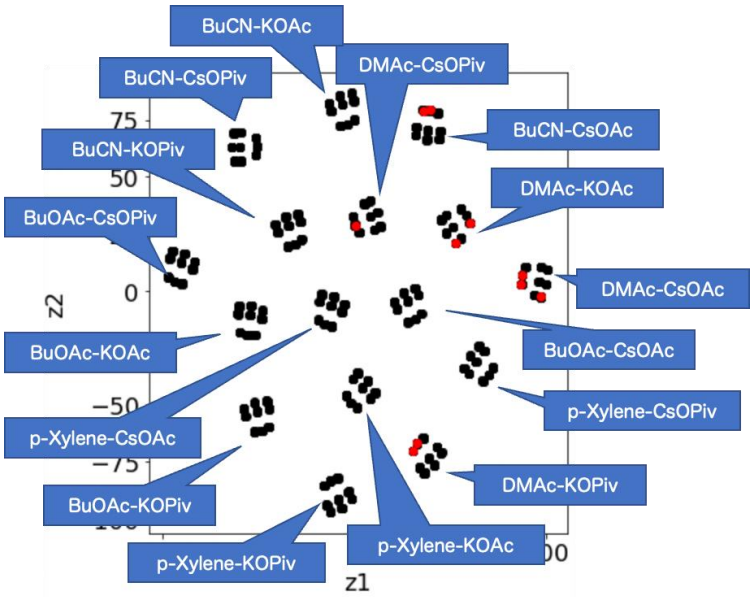


図 18　配位子情報を除いて DBSCAN でクラスタリングを行った結果を t-SNE で可視化した例(Perplexity 20)

本研究で提案した，クラスタリング情報を活用した初期条件決定方法は，適切にクラスタリングができれば，ランダムサンプリングや D 最適基準に基づいたサンプリングと比較して，最適条件に到達するまでに要する実験回数を削減できる。適切なクラスタリングとは，所属するサンプル数がほぼ均一であり，目的変数への寄与が大きな因子でクラスタを形成できる場合のことである。反応条件最適化のような化合物の組み合わせ最適化であれば，実験条件が空間上に離散的に分布し，使用する化合物ごとに，所属する条件数がほぼ均一なクラスタを形成する。一方で，目的変数への寄与が大きな因子でクラスタを形成するためには，有機合成化学者の知識が必要となる。実験前に実験結果を知ることはできないため容易ではないが，ドメイン知識を活用して適切に形成されたクラスタ情報を用いて初期条件を決定する事ができれば，ベイズ最適化の探索性能をさらに向上させる事が可能となる。また，今回対象とした反応条件最適化は説明変数の数が非常に多かったため，高次元データのままでは解釈が難しく，クラスタリングの妥当性の確認が困難となると考え，t-SNE による可視化ならびに DBSCAN でのクラスタリングを実施した。クラスタリングや可視化手法のハイパーパラメータは試行錯誤で決定する必要があるように思われたが，本研究の結果，これらはそれほど大きな問題ではないことが示された。適切な初期サンプルは，対象変数に大きな影響を与える要因から少なくとも 1 つの化合物で構成される必要がある。特定の要因から 1 つずつサンプルを選択するだけであれば，特定の機械学習手法 (k-means，PAM など) を用いることなく実現可能である。因子によるクラスタリングを用いて実験条件を決定する方法は従来からある方法であり，実験者にとって最も身近で利用しやすい方法である。

3.6 まとめ

BO で最適解を効率的に探索するには，GPR モデルを構築する際に適切な初期サンプルを提供する必要がある。化合物を用いた実験デザインの場合，化学構造から計算される分子記述子間には常に高い相関が存在し，類似した構造を持つ化合物は化学空間においてクラスタを形成する。そのため，実験条件の情報が最大となる初期試料を得るためには，各クラスタから一様に初期試料を選択することが望ましい。D 最適基準のような実験空間の類似性を利用したサンプリング手法は，相関性の高い分子記述子ではうまく機能せず，また，サンプル選択にクラスタの情報を考慮しない。本研究では，目的変数に大きな影響を与える因子でのクラスタリング情報に基づく初期サンプル選択手法を提案し，BO とのカップリング反応条件の最適化に適用した。その結果，クラスタを適切に形成し，各クラスタから初期サンプルを選択した場合，提案手法はランダムサンプリングや D 最適基準に基づくサンプリングよりも少ない実験回数で最適解に到達することを確認した。さらに，1 ラウンドあたりの実験回数が少ない場合，提案手法の効果はクラスタ情報を含まない他の手法よりも大きく，探索に必要なラウンド数を削減できることもわかった。適切なクラスタリングとは，クラスタに属するサンプル数がほぼ一様であり，対象変数に大きく寄与する要因によってクラスタを形成できる場合を指す。クラスタリングは教師なし学習であり，目的変数（実験結果）に関する情報を実験前に知ることはできない。これらの情報を結びつけるためには，付加的な情報（専門家の知識）が必要となる。研究現場では，ベイズ最適化と専門家の知識を組み合わせて使いたいという要望が多い。ドメイン知識を活用して適切なクラスタを形成し，初期条件を決定することができれば，反応条件最適化のような化合物の組み合わせの場合，BO の検索性能をさらに向上させることができることが確認できた。本研究は，BO の初期試料選択手法に貢献するものであり，領域知識を活用して適切に形成されたクラスタ情報を用いて初期試料を決定できれば，提案手法は様々な科学技術分野において BO の検索性能を向上させることができると思われる。今後は，因子数が多い場合（組み合わせ数が膨大な場合），複数の因子が目的変数に与える影響が大きい場合の取り扱い，多目的最適化の場合などを検討する予定である。

3.7 参考文献

15. Greenhill, S.; Rana, S.; Gupta, S.; Vellanki, P.; Venkatesh, S. Bayesian Optimization for Adaptive Experimental Design: A Review. *IEEE Access*. 2020 *8*, 13937–13948.
16. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*. 2015, *104*(1), 148–175.
17. Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature.* 2021, *590*(7844), 89–96.
18. Griffiths, R. R.; Hernández-Lobato, J. M. Constrained Bayesian Optimization for Automatic Chemical Design Using Variational Autoencoders. *Chem. Sci.* 2020, *11*(2), 577–586. DOI: 10.1039/C9SC04026A
19. Fang, L.; Makkonen, E.; Todorović, M.; Rinke, P.; Chen, X. Efficient Amino Acid Conformer Search with Bayesian Optimization. *J. Chem. Theory Comput*. 2021, *17*(3), 1955–1966.
20. https://github.com/b-shields/edbo (accessed 2022-9th-July).
21. https://github.com/mordred-descriptor/mordred (accessed 2022-9th-July).
22. Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform*. 2018, *10*(1), 1–14.
23. Nakayama, R.; Shimizu, R.; Haga, T.; Kimura, T.; Ando, Y.; Kobayashi, S.; Yasuo, N.; Sekijima, M.; Hitosugi, T. Tuning of Bayesian Optimization for Materials Synthesis: Simulation of the One-dimensional Case. *Sci. Technol. Adv. Mater.* 2022, *2*(1), 119–128.
24. https://mordred-descriptor.github.io/documentation/master/descriptors.html (accessed 2022-2nd-September)
25. Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. Technomet- rics 1969, 11, 137–148.
26. Rännar, S.; Andersson, P. L. A Novel Approach Using Hierarchical Clustering To Select Industrial Chemicals for Environmental Impact Assessment. J. Chem. Inf. Model. 2010, 50, 30–36.
27. Park, S.; Na, J.; Kim, M.; Lee, J. M. Multi-objective Bayesian Optimization of Chemical Reactor Design Using Computational Fluid Dynamics. *Comput. Chem. Eng.* 2018, *119,* 25–37.
28. Klein, A.; Falkner, S.; Bartels, S.; Hennig, P.; Hutter, F. Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA.* JMLR, 528–536.
29. Van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE, *J. Mach. Learn. Res*. 2008, *9*(11) 2579–2605.

30. https://en.wikipedia.org/wiki/K-medoids (accessed 2022-9th-July).
31. https://en.wikipedia.org/wiki/DBSCAN (accessed 2022-9th-July).
32. Yang, K.; Palar, P. S.; Emmerich, M.; Shimoyama, K.; Bäck, T. A Multi-point Mechanism of Expected Hypervolume Improvement for Parallel Multi-objective Bayesian Global Optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference.* 2019, 656–663.
33. Zhang, Y.; Bahadori, M. T.; Su, H.; Sun, J. FLASH: Fast Bayesian Optimization for Data Analytic Pipelines. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, 2065–2074.
34. Nandiwale, K. Y.; Hart, T; Zahrt, A. F.; Namibiar, A. M.; Mahesh, P. T.; Mo, Y.; Nieves-Remacha, M. J.; Johnson, M. D.; Garcia-Losada, P.; Mateos, C.; Rincon, J.A.; Jensen, K. F. Continuous stirred-tank reactor cascade platform for self-optimization of reactions involving solids. In *react. Chem. Eng.* 2022, 1315-1327.
35. Iwama, R; Kaneko, H. Design of ethylene oxide production process based on adaptive design of experiments and Bayesian optimization. In Journal of Advanced Manufacturing and Processing, Volume 3, Issue 3
36. Yuan, W; Han, Y; Guan, D; Lee, S; Lee, Y, Initial Training Data Selection for Active Learning. ICUIMC '11: Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, Article No. 5, 1-7.

第4章　密度汎関数法(DFT)を用いたベイズ最適化探索性能向上

4.1 はじめに

機械学習，ニューラルネットワーク，ロボット工学，航空宇宙工学，実験計画など，さまざまな科学技術分野で最適解の探索は重要課題の一つであり，その解決法の一つとしてBO[15,16] が広く検討，活用されてきている。例えば，材料開発の分野では開発のスピードを加速させる目的で利用されており，探索性能の向上や多目的最適化に関する検討が行われている[37,38]。また，各種分野におけるシミュレーションの計算回数削減や機械学習のハイパーパラメータの最適化目的にも用いられており良い結果が報告されている。数値流体力学(CFD)を用いた化学反応器最適設計にてパレート最適解を求めるために用いられ，CFD 計算の実行回数を減らすことができた。開発された攪拌槽反応器において消費電力の最小化とガス保持量の最大化に貢献した[27]。有機化学分野への適用例も増えており，新規分子探索において，変分オートエンコーダの潜在空間に対して BO を適用した場合に，無効な分子構造を生成しやすいという課題に対して，制約付き BO を用いる事で影響の緩和を試みている[18]。また，BO は最安定な分子コンフォーマーの探索にも用いられている。分子コンフォーマーの探索は探索空間の次元が高く，最適構造を決定するために実施される量子化学計算に膨大な時間を要するという課題があった。BO と量子化学計算を組み合わせることにより量子化学計算コストを約 90%削減することができた[19]。このように，BO を用いた適応的実験計画法は様々な分野で検討・利用されている。一方で，反応条件最適化への適用事例はあまり多くはない。2021 年に有機化合物の合成実験の最適化に BO が適用された事例が紹介された。本研究ではパラジウム触媒による直接アリール化反応の大規模ベンチマークデータを HTE で収集して熟練者との比較実験を行った。BO は有機合成の熟練者よりも高速に最適条件に到達することができ，分子構造から DFT 計算にて得られた記述子を用いた場合，化合物を One-hot-encoding (OHE)で扱った場合と比較して探索性能が高くなったとの報告がなされた[17]。

BO で化合物を取り扱う場合，分子構造情報から計算された記述子を説明変数として利用することができる。一般的な化合物であれば各種データベースから構造及び記述子情報を得ることができるが，例えば医薬品中間体や原薬のような秘密性が高い化合物は，データベースから検索・取得することはできない。特に量子科学計算にて化合物の安定構造・電子状態を求める必要がある場合，記述子情報を得るには多大な計算時間を要する。計算負荷が比較的軽い DFT を活用して記述子情報を得ることは可能だが，実験前に最良な基底関数・汎関数を選択することは難しく，一般的な組み合わせが慣習的に選択されることが多い。BO の探索性能に対するこれらの影響について議論されている事例は少なく，良好な探索性能を得られるかどうかは適用対象次第である。そこでいくつかの組み合わせの基底関数・汎関

数で計算された記述子を活用して， BO の探索性能を向上させる方法を開発した。Shields ら[**エラー! 参照元が見つかりません。**17]で研究された 2 つの反応，直接アリール化と鈴木-宮浦カップリング反応について記述子を計算し，それぞれの記述子セットから作成したデータセットを用いて BO の検索性能を比較した。その後，BO に最適な記述子セットの選択方法を同定し，BO における検索性能への影響を確認した。

4.2 データセット作成方法

本研究にて提案する手法，DFT 計算にて得られた化合物の構造情報から複数の記述子群およびデータセットを準備し，GPR モデルを構築して BO を実施する手順は以下のとおりである(図 19)。

1. 対象となる m 個の化合物の分子構造(mol ファイル等)を準備する。
2. 様々な基底関数・汎関数の組み合わせ(n 通り)に対して Gaussian[40]を用いて DFT 計算を実施する。
3. DFT 計算で得られた構造・電子状態をもとに Mordred[21,22,39]，Codessa[41]等の記述子計算方法を用いて 0 次元，1 次元，2 次元，3 次元記述子を計算する。
4. DFT 計算で得られた k 個の記述子群を平均化し，連続値の説明変数 **x** や目的変数 **y** を組み合わせて 1 つのデータセットを作成する。記述子群の数 k は化合物の数，用いる基底関数・汎関数，組み合わせ方などにより変化する。
5. 選択・作成されたデータセットを用いて GPR モデルを構築し，BO にて獲得関数 EI が高いサンプルを次の実験条件として選択する。

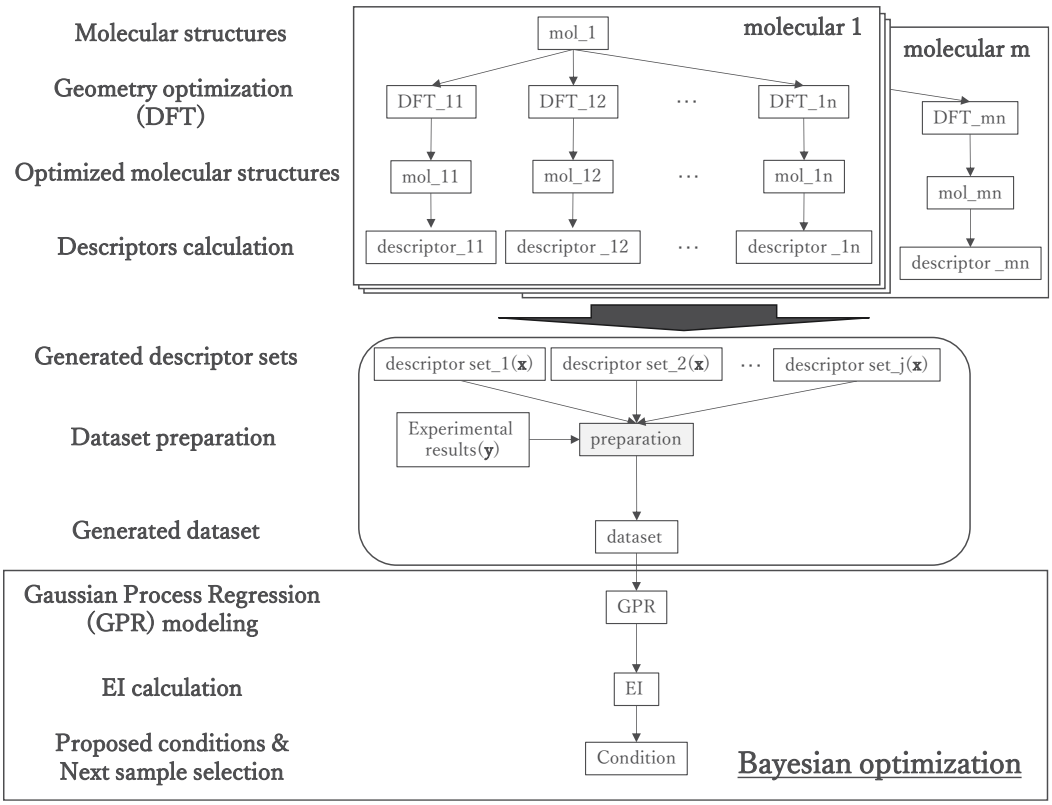
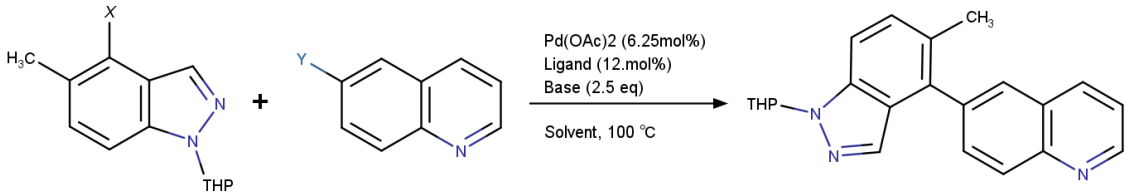


図 19　BO のためのデータセット作成手順

4.3 データセット

本検討では Shields ら[17]が報告したパラジウム触媒による直接的アリール化(スキーム 1: 反応 A)と鈴木-宮浦カップリング(スキーム 2：反応 B)における BO の探索性能が検証された。反応 A は第 3 章で用いたものと同様であり，実験条件のうち反応基質，触媒と配位子の当量は固定し，反応温度 3 種類，基質濃度 3 種類，配位子 12 種類，溶媒 4 種類，塩基 3 種類の組み合わせとなる全 1,728 通りの実験が行われた(表 11，表 12)。目的変数である反応収率が 95%以上の条件は 10 通りで全体の 0.58%，収率が 98%以上の条件は 7 通りで全体の 0.41%であった。配位子，溶媒，塩基は化合物であり，カテゴリカル変数であるため，オープンソースの分子記述子計算ソフトウェア MORDRED[21,22]にて 20 種類の化合物の分子構造を 2 次元記述子に変換して計算に用いた。また，量子科学計算ソフトウェア Gaussian[40]で DFT 計算を実施した分子構造に対して，分子記述子計算ソフトウェア Codessa[41]にて，3 次元記述子に変換して計算に用いた。DFT 計算では基底関数は STO-3G, 3-21G, 3-21Gd, 6-31G, 6-31Gd, 6-31Gdp, 6-31G+d, 6-311G, 6-311Gd, 6-311Gdp の 10 種類，汎関数は有機合成分野で良く採用される B3LYP のみを用いた。得られた記述子は主成分分析(PCA)にて 20 変数に低次元化した。その際の累積寄与率は 99%以上であった。反応 B は実験条件のうち反応基質，塩基，溶媒，触媒，配位子の当量は固定し，求電子

剤 4 種類，求核剤 3 種類，配位子 11 種類，塩基 7 種類，溶媒 4 種類の組み合わせとなる全 3,696 通りの実験が行われた(表 20)。目的変数である反応収率が 95%以上の条件は 71 通りで全体の 1.92%，収率が 98%以上の条件は 10 通りで全体の 0.27%であった。求電子剤，求核剤，配位子，塩基，溶媒は化合物であり，カテゴリカル変数であるため，記述子に変換して計算に用いた。記述子計算方法および後処理は反応 A の場合と同様である。

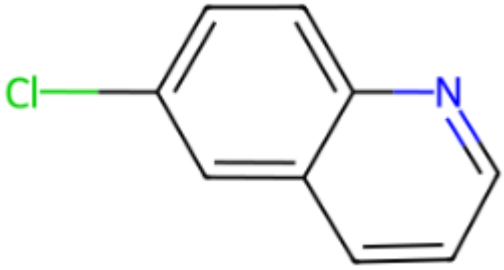
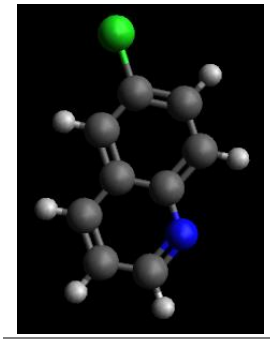
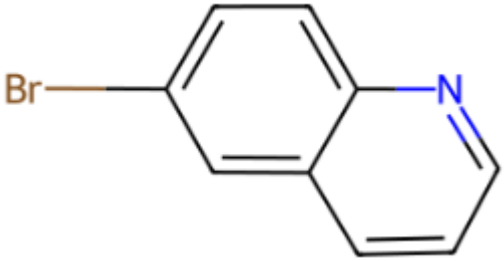
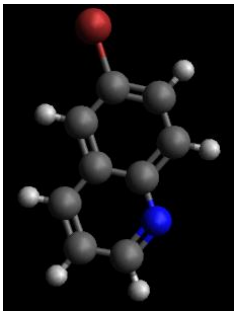


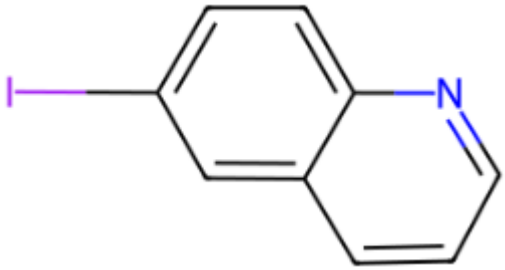
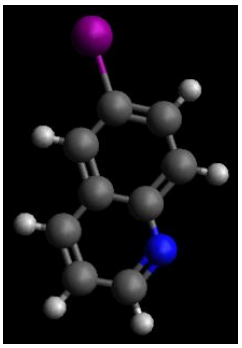
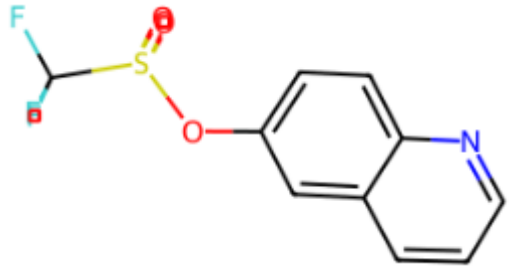
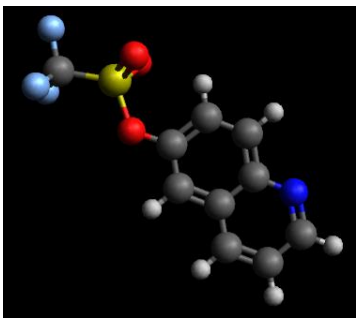
スキーム 2. 反応 B: 鈴木–宮浦カップリング

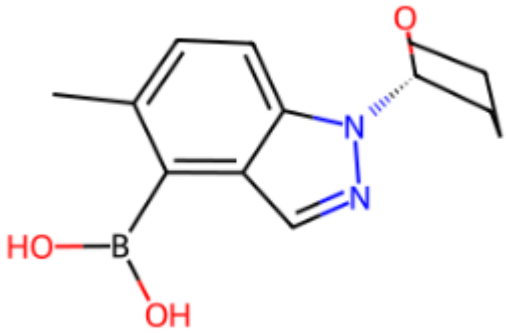
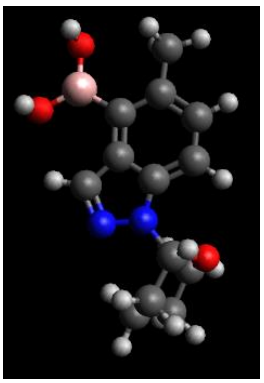
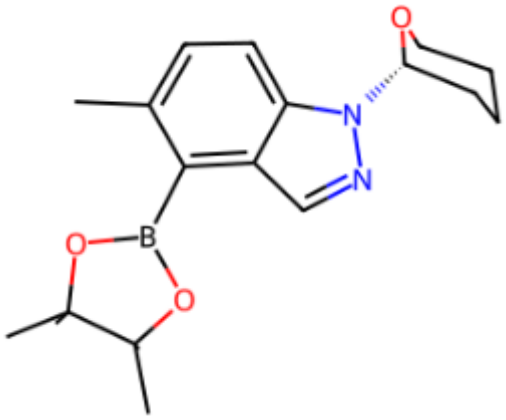
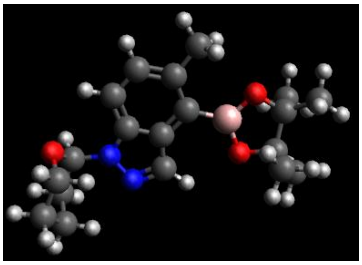
表 20　実験条件(反応 B: 鈴木–宮浦カップリング)

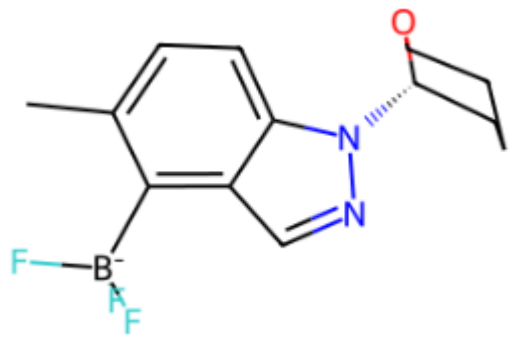
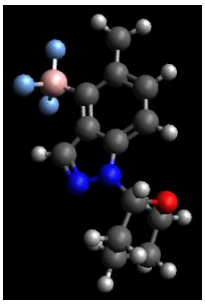
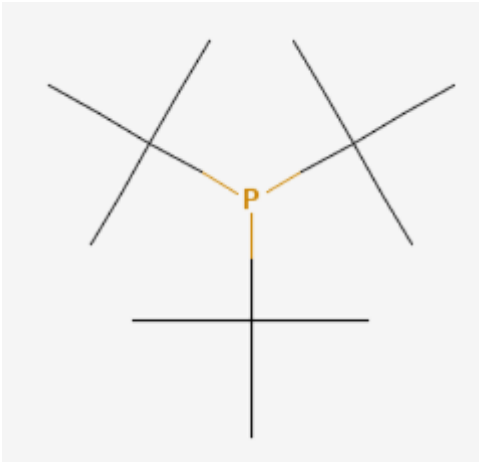
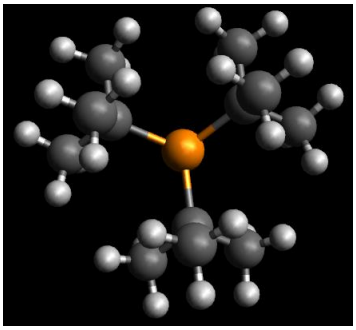
求電子剤(X)	求核剤(Y)	配位子	塩基	溶媒
Cl	B(OH)2	PtBu3	NaOH	Acetonitrile/H2O
Br	Bpin	PPh3	NaHCO3	Tetrahydrofuran/H2O
I	BF3K	AmPhos	CsF	N,N-dimethylformamide/H2O
OTf		PCy3	K3PO4	Methanol/H2O
		PoTol3	KOH	
		cataCXium A	LiOtBu	
		SPhos	Triethylamine	
		dtbpf		
		XPhos		
		dppf		
		Xantphos		

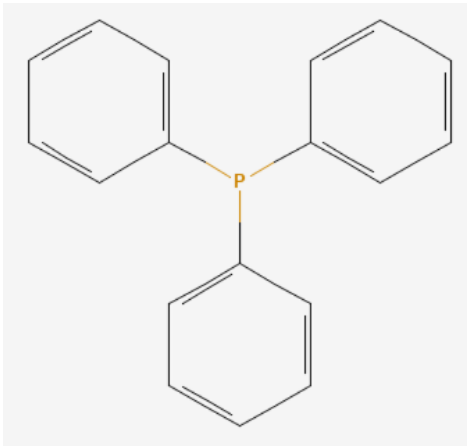
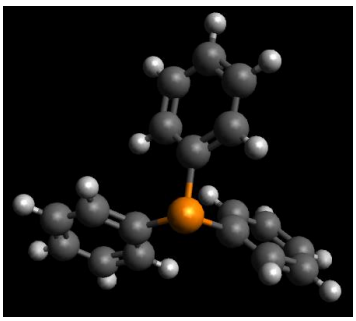
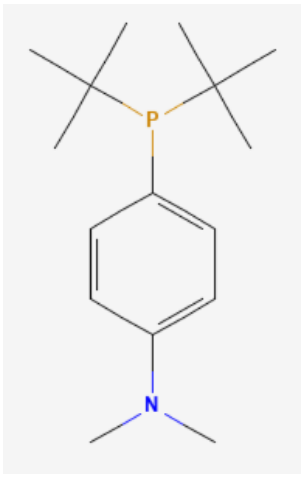
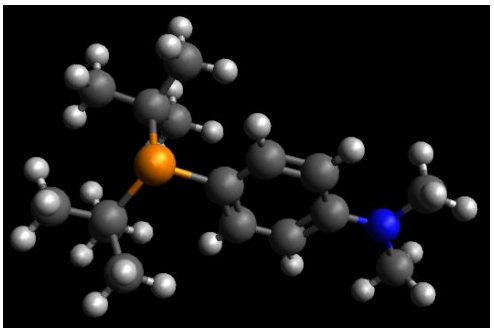
表 21　反応試剤情報(反応 B: 鈴木–宮浦カップリング)

Name	SMILES
求電子剤（X=Cl）  	[H]C1=NC2=C([H])C([H])=C(Cl)C([H])=C2C([H])=C1[H]
求電子剤(X=Br)  	[H]C1=NC2=C([H])C([H])=C(Br)C([H])=C2C([H])=C1[H]

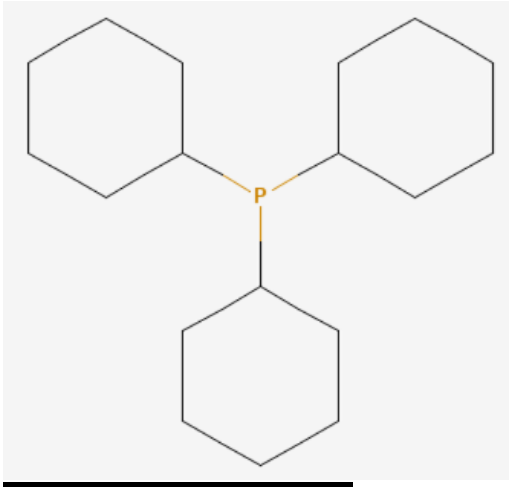
求電子剤(X=I)  	[H]C1=NC2=C([H])C([H])=C(I)C([H])=C2C([H])=C1[H]
求電子剤(X=OTf)  	[H]C1=NC2=C([H])C([H])=C(OS(=O)(=O)C(F)(F)F)C([H])=C2C([H])=C1[H]

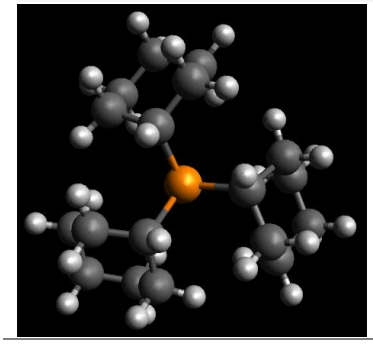
求核剂（Y=B(OH)2）  	[H]OB(O[H])C1=C2C([H])=NN(C2=C([H])C([H])=C1C([H])([H])[H])[C@]1([H])OC([H])([H])C([H])([H])C([H])([H])C1([H])[H]
求核剂(Y=Bpin)  	[H]C1=NN(C2=C([H])C([H])=C(C(B3OC(C([H])([H])[H])(C([H])([H])[H])C(O3)(C([H])([H])[H])C([H])([H])[H])=C12)C([H])([H])[H])[C@]1([H])OC([H])([H])C([H])([H])C([H])([H])C1([H])[H]

求核剂(Y=BF3K)  	[H]C1=NN(C2=C([H])C([H])=C(C(=C12)[B-](F)(F)F)C([H])([H])[H])[C@]1([H])OC([H])([H])C([H])([H])C([H])([H])C1([H])[H]
PtBu3  	[H]C([H])([H])C(P(C(C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H])C(C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H]

PPh3  	[H]C1=C([H])C([H])=C(C([H])=C1[H])P(C1=C([H])C([H])=C([H])C([H])=C1[H])C1=C([H])C([H])=C([H])C([H])=C1[H]
AmPhos  	[H]C1=C([H])C(=C([H])C([H])=C1N(C([H])([H])[H])C([H])([H])[H])P(C(C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H])C(C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H]

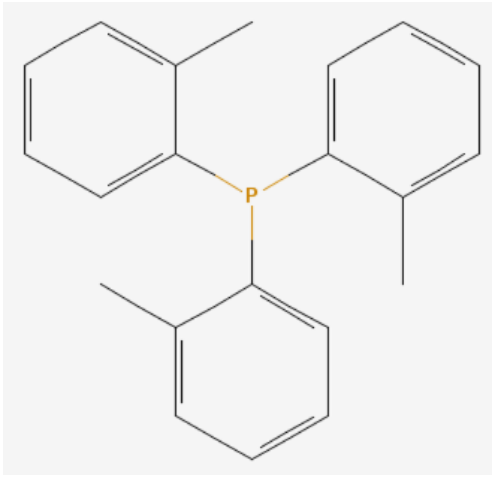
PCy3

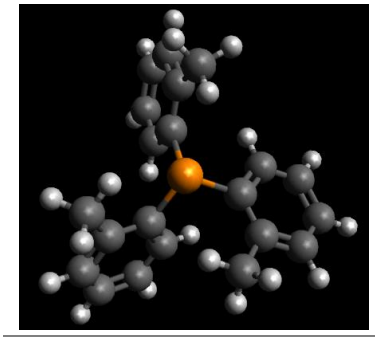




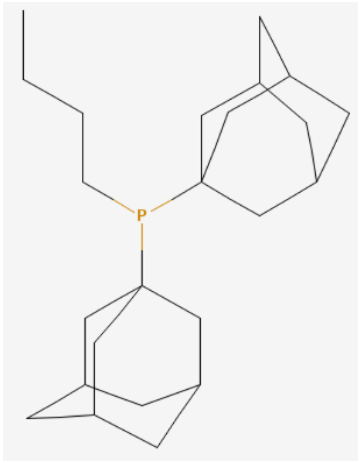
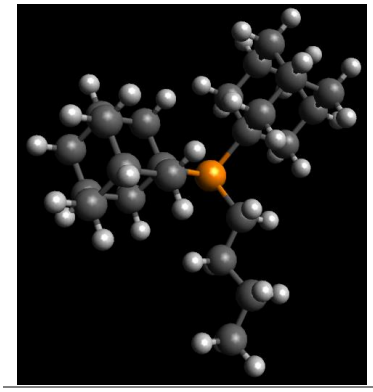
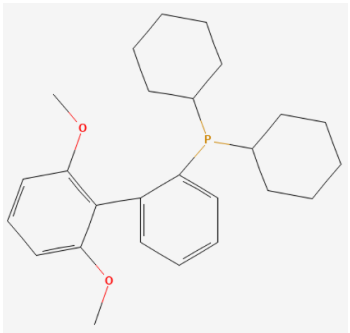
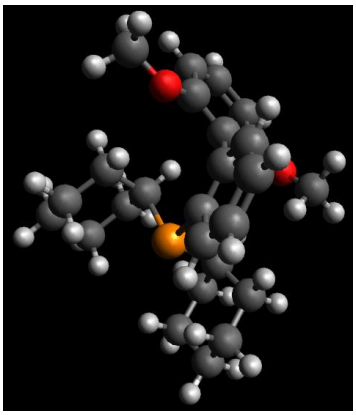
[H]C1([H])C([H])([H])C([H])([H])C([H])([H])P(C2([H])C([H])([H])C([H])([H])C([H])([H])C2([H])[H])C2([H])C([H])([H])C([H])([H])C([H])([H])C([H])([H])C2([H])[H])C([H])([H])C1([H])[H]

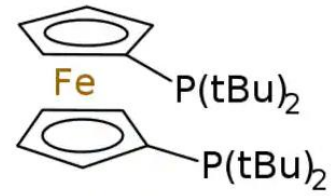
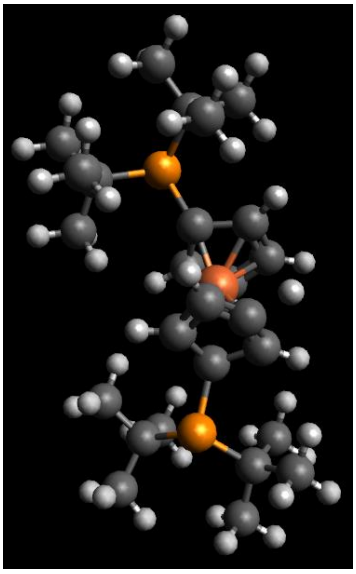
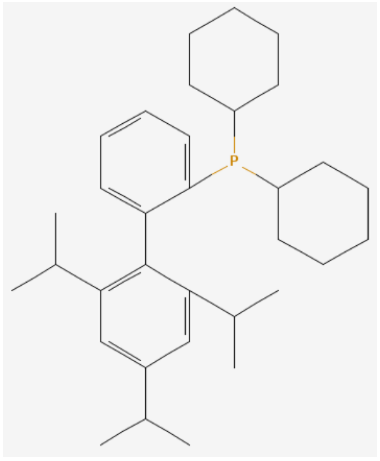
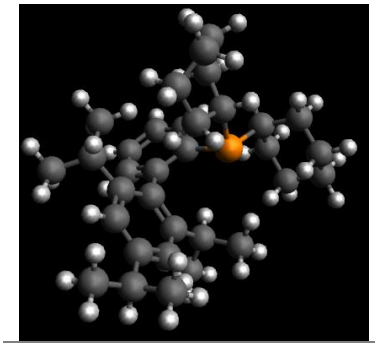
PoTol3

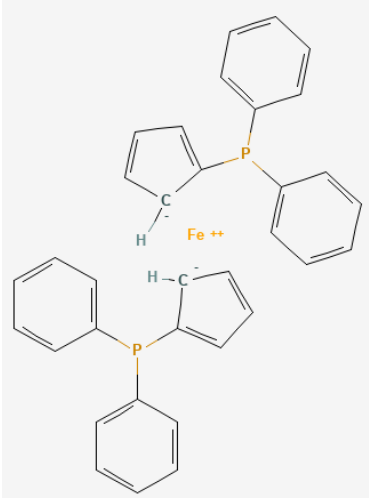
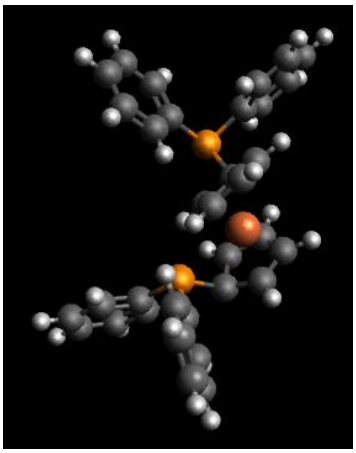
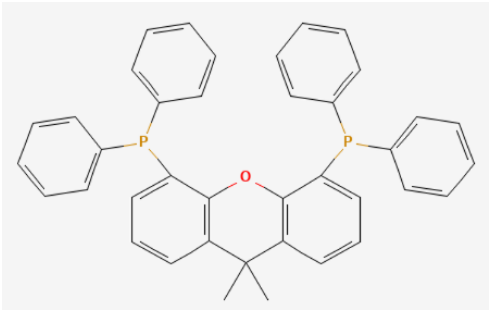
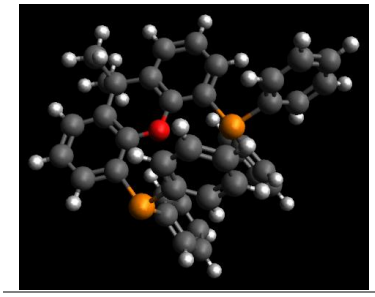


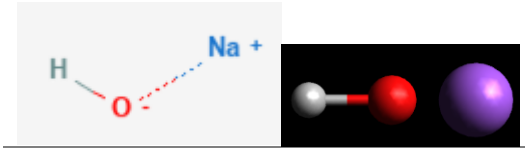
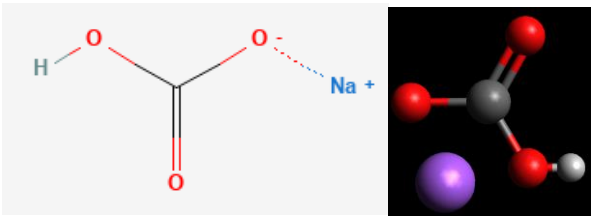
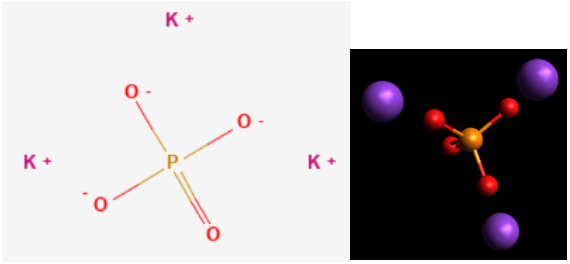
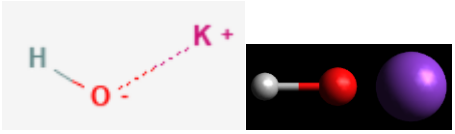
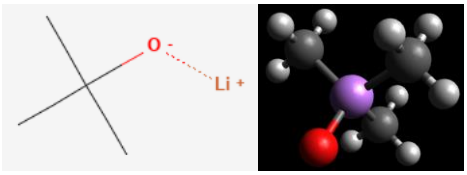
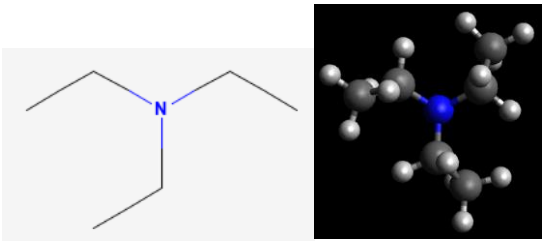


[H]C1=C([H])C([H])=C(C(=C1[H])P(C1=C([H])C([H])=C([H])C([H])=C1C([H])([H])[H])C1=C([H])C([H])=C([H])C([H])=C1C([H])([H])[H])C([H])([H])[H]

cataCXium A  	[H]C([H])([H])C([H])([H])C([H])([H])C([H])([H])P([C@]12C([H])([H])[C@@]3([H])C([H])([H])[C@]([H])(C([H])([H])[C@]([H])(C3([H])[H])C1([H])[H])C2([H])[H])[C@]12C([H])([H])[C@@]3([H])C([H])([H])[C@]([H])(C([H])([H])[C@]([H])(C3([H])[H])C1([H])[H])C2([H])[H]
Sphos  	[H]C1=C([H])C([H])=C(C(=C1[H])P(C1([H])C([H])([H])C([H])([H])C([H])([H])C([H])([H])C1([H])[H])C1([H])C([H])([H])C([H])([H])C([H])([H])C([H])([H])C([H])([H])C1([H])[H])C1=C(OC([H])([H])[H])C([H])=C([H])C([H])=C1OC([H])([H])[H]

Dtbpf  	[Fe].[H]C1=C([H])[C-](C([H])=C1[H])P(C(C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H])C(C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H].[H]C1=C([H])[C-](C([H])=C1[H])P(C(C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H])C(C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H]
Xphos  	[H]C1=C([H])C([H])=C(C(=C1[H])P(C1([H])C([H])([H])C([H])([H])C([H])([H])C([H])([H])C1([H])[H])C1([H])C([H])([H])C([H])([H])C([H])([H])C([H])([H])C([H])([H])C1([H])[H])C1=C(C([H])=C(C([H])=C1C([H])(C([H])([H])[H])C([H])([H])[H])C([H])(C([H])([H])[H])C([H])([H])[H])C([H])(C([H])([H])[H])C([H])([H])[H]

dppf  	[Fe].[H][C-]1[C-]([H])[C-]([H])[C-]([C-]1[H])P(C1=C([H])C([H])=C([H])C([H])=C1[H])C1=C([H])C([H])=C([H])C([H])=C1[H].[H][C-]1[C-]([H])[C-]([H])[C-]([C-]1[H])P(C1=C([H])C([H])=C([H])C([H])=C1[H])C1=C([H])C([H])=C([H])C([H])=C1[H]
Xantphos  	[H]C1=C([H])C([H])=C(C([H])=C1[H])P(C1=C([H])C([H])=C([H])C([H])=C1[H])C1=C([H])C([H])=C([H])C2=C1OC1=C(C([H])=C([H])C([H])=C1P(C1=C([H])C([H])=C([H])C([H])=C1[H])C1=C([H])C([H])=C([H])C([H])=C1[H])C2(C([H])([H])[H])C([H])([H])[H]

NaOH 	[Na+].[O-][H]
NaHCO3 	[Na+].[H]OC([O-])=O
CsF 	F[Cs]
K3PO4 	[K+].[K+].[K+].[O-]P([O-])([O-])=O
KOH 	[K+].[O-][H]
LiOtBu 	[Li+].[H]C([H])([H])C([O-])(C([H])([H])[H])C([H])([H])[H]
Triethylamine 	[H]C([H])([H])C([H])([H])N(C([H])([H])C([H])([H])[H])C([H])([H])C([H])([H])[H]

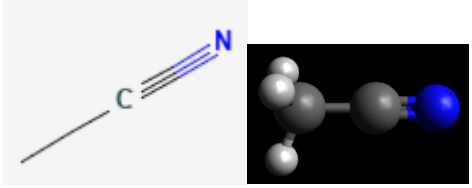
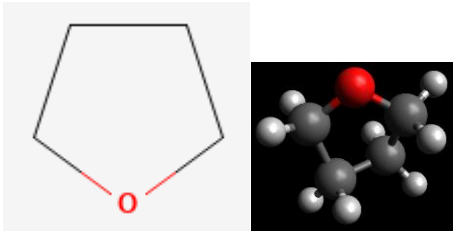
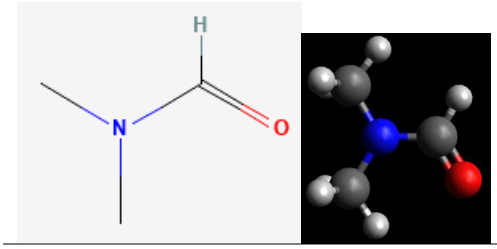
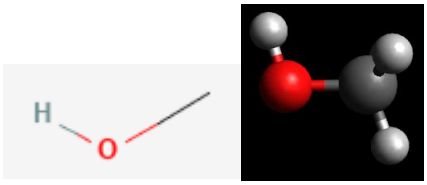
Acetonitrile 	[H]C([H])([H])C#N
Tetrahydrofuran 	[H]C1([H])OC([H])([H])C([H])([H])C1([H])[H]
N,N-dimethylformamide 	[H]C(=O)N(C([H])([H])[H])C([H])([H])[H]
Methanol 	[H]OC([H])([H])[H]

表 22　収率上位 10 条件(反応 B: 鈴木–宮浦カップリング)

No.	求電子剤	求核剤	配位子	溶媒	塩基	収率 [%]
1	I	Bpin	AmPhos	LiOtBu	MeOH	100
2	I	Bpin	PoTol3	LiOtBu	MeOH	100
3	I	Bpin	SPhos	LiOtBu	MeOH	100
4	I	Bpin	AmPhos	LiOtBu	MeCN	99.15
5	I	Bpin	AmPhos	KOH	MeCN	99.05
6	I	BOH2	PPh3	LiOtBu	MeOH	98.69
7	OTf	Bpin	cataCXium	CsF	MeCN	98.61
8	OTf	BOH2	PPh3	NaHCO3	MeCN	98.53
9	I	Bpin	PtBu3	LiOtBu	MeCN	98.42
10	I	BOH2	PPh3	LiOtBu	MeCN	98.20

4.4 ベンチマーク

前述の反応 A,B を対象として DFT 計算における基底関数が異なるデータセットを準備し，BO 探索性能の確認を実施した。それぞれのケースにおける反応，記述子群の数，終了判定条件，実験提案数(1 ラウンドあたりの実験数)は表 23 の通りである。Case1, 2 および 4 では 10 個すべての基底関数にて DFT 計算で得られた分子情報から計算された記述子群を用いた。Case3 ではカテゴリカル変数である配位子, 溶媒, 塩基に対して, 基底関数 STO-3G，3-21G および 6-31G にて DFT 計算で得られた分子情報から計算された記述子群を用いた。Case5, 6 ではカテゴリカル変数である求電子剤，求核剤，配位子，塩基，溶媒に対して，基底関数 STO-3G，6-31G にて DFT 計算で得られた分子情報から計算された記述子群を用いた。初期条件はランダムで決定し，乱数の影響を排除するためにすべてのケースにおいて 10,000 回以上の計算を実施した。終了判定条件に到達するまでに要したラウンド数の平均及び標準偏差を確認したが，基底関数の選択方法と探索性能の間に，より厳密な DFT 計算を行った方が探索性能はよくなるというような類の関係性は見いだせなかった(表 28, 表 29，表 30，表 31，表 32，表 33)。本結果は単一の記述子群を用いた際の探索性能として，データセット作成方法の評価の際のベンチマークに用いた。

表 23　様々な実験条件における BO における探索性能比較

Case No.	1	2	3	4	5	6
Reaction	A	A	A	B	B	B
Number of descriptor sets	10	10	27	10	32	32
Target yield	98	95	98	95	98	95
Number of proposed experiments	5	5	5	5	10	5

4.5 比較手法

BO の探索性能に対する記述子の影響を確認するために，記述子群をすべて一つのデータセットとして利用する方法と，記述子群の中から BO で利用する記述子群を選択してデータセットを作成する 5 種類の方法(ランダム選択, GPR モデルの予測性能, BO の獲得関数, 記述子群の類似性: D 最適基準，重心からの平均距離)を用いて探索性能の比較を行った。いずれもの方法もデータセットの選択方法以外の計算手順は Dataset preparation method で提案した方法を同様である。記述子群を選択する方法では BO で利用する記述子群を計算ごとに選択しなおしてデータセットを作成した。記述子群を複数選択する(アンサンブルを行う)際には，それぞれの評価指標が高い/低い順に記述子群を選択した。5 種類の方法の詳細を以下に示す。

I.　Random selection of descriptor set

ランダムに記述子群を選択する方法では，一様乱数を用いて BO で利用する記述子群をランダムで選択した。

II & III.　Descriptor set selection based on diversity

記述子群の多様性をもとに選択する方法では，BO の探索性能が悪い理由は局所最適解から抜け出すことができないためと考え，可能な限り類似していない記述子群を選択することで多様な条件が提案されるようにした。まず，記述子群の類似性を評価するために記述子群を一次元ベクトル化したのちに，それらを一つのデータセットとして結合した。構築した記述子群のデータセットに対して，各記述子群の重心からの平均距離や D 最適基準を求めて類似度の評価を行った。選択された複数の記述子群において，重心からの平均距離が大きい場合，または D 最適基準が大きい場合に，記述子群が類似していないと判断をした。

IV.　Descriptor set selection based on prediction performance of GPR model

GPR モデルの予測性能を用いる方法では，GPR の汎化性能が高いほど BO の探索性能も高くなるだろうという考えのもと，BO にて GPR モデルを構築するたびにモデルのクロスバリデーションを行い，MSE(Mean Squared Error)を指標として汎化性能を評価し，最も MSE が小さなモデル(記述子群)を選択して条件探索に利用した。本手法はすべての GPR モデルに対してバリデーションを行う必要があり，計算負荷が非常に大きいため，クロスバリデーションのフォールド数は提案実験数(1 ラウンドあたりの実験数)とした。提案実験数が 1 の場合はクロスバリデーションが実施できないため，提案実験数は 5 以上程度とすることが望ましい。

V.　Descriptor set selection based on expected improvement

BO の獲得関数の値を用いて選択する方法では，同じ化合物から得られた記述子を用いて計算された獲得関数であれば，改善幅の期待値(EI)が高い条件を選択すれば BO の探索性能も高くなるだろうという考えのもと，BO にて獲得関数を計算した後に，獲得関数が大きな値を持つモデル(記述子群)を選択して BO に利用した。記述子群から作成されたデータセットすべてに対して GPR モデルを構築して EI を計算する必要があるため計算負荷は比較的大きい。複数の BO で計算される獲得関数(EI)の値をすべて記憶しておき，最終的に EI の値が高いサンプルおよびモデルを採用して次の実験条件の提案を行った。

4.6　計算結果と考察

表 23 で示した 6 種類のケースに対して，表 24 に示す 12 種類の方法でデータセットを作成して探索性能の違いを確認した。それぞれのケースにおける基底関数，汎関数，データ

セット数，終了判定条件，実験提案数(1 ラウンドあたりの実験数)は表 23 の通りである。初期条件はランダムで決定し，初期条件や乱数の影響を可能な限り小さくするためにそれぞれ 10,000 回以上の計算を実施した。また，BO に用いる記述子群を選択する方法を用いる際には，評価指標ごとに単一の記述子群を用いる場合，複数(5 つ)の記述子群(モデル)を用いてアンサンブルを行う場合の 2 種類について評価を実施した。方法名の後に記載している数字は選択した記述子群の数を示す。

表 24　データセット作成方法一覧

Method name	Number of selected descriptor sets	Explanation
Rand1	1	Random selection of the descriptor set
Rand5	5	
dopt1	1	Descriptor set selection based on diversity (D-optimality)
dopt5	5	
dist1	1	Descriptor set selection based on diversity (average distance from the center of gravity)
dist5	5	
CV1	1	Descriptor set selection based on the prediction performance of the GPR model
CV5	5	
EI1	1	Descriptor set selection based on expected improvement
EI5	5	
all	1	All the descriptor sets combined into a single descriptor set
ave	1	Average of all the descriptor sets used as a single descriptor set (proposed method)

それぞれのデータセット作成方法を用いて BO を実施した際の探索結果を図 20~図 25 に示す。横軸は終了判定条件に到達するために要したラウンド数の平均，縦軸は標準偏差を示す。その他の凡例中の表記はデータ作成方法の名称であり，Datasets は前節でベンチマークとして計算された単一の記述子群を用いて BO を実施した際の探索結果を示す。DFT 計算で得られたすべての記述子の平均値を用いる方法は，すべてのケースにおいて最も高い探索性能を示し，記述子群を単一で用いた場合よりも平均および標準偏差ともに小さくなる傾向がみられた。特に Case 4 のように探索性能が著しく低下する記述子群が存在する場合であっても，外れ記述子群の影響を大きく受けることはなかった。記述子群を選択する方法では，選択した記述子群の数が単一，複数の場合いずれ場合であっても探索性能が大きく向上することはなかった。今回記述子群を選択するために利用した 5 つの指標はいずれも BO における探索性能との相関が高くなかったためと考えられる。また，多様性が高くなるよう

に記述子群を選択する方法においても探索性能が向上することはなかった。データセットの多様性と BO の探索性能の間にはあまり関係性がないと思われる。すべての記述子群を用いる方法は記述子群を選択する方法よりは良好な探索性能を示すケースが多かったものの，今回提案した平均化を行う手法には及ばなかった。終了条件に到達するまでに要するラウンド数の累積相対度数を，提案手法(ave)の累積相対度数で規格化したものを図 26~図 31 に示す。横軸は各 Trial において終了条件に到達するまでに要したラウンド数，縦軸は規格化された累積相対度数を表す。乱数の影響により 1 ラウンド目の結果にばらつきが生じているが，2 ラウンド目以降はほぼすべてのケースで平均化を行う手法(ave)の探索性能が相対的に高いことが確認できる。

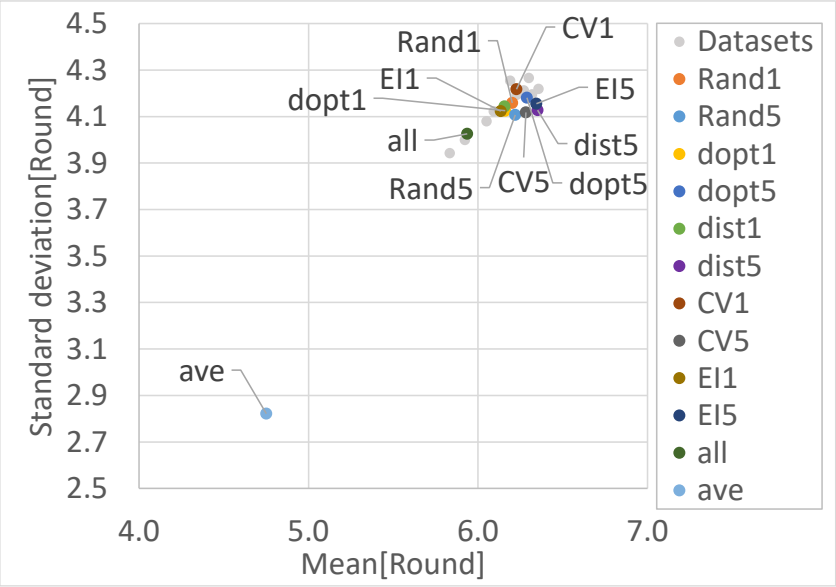


図 20　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差(Case 1)

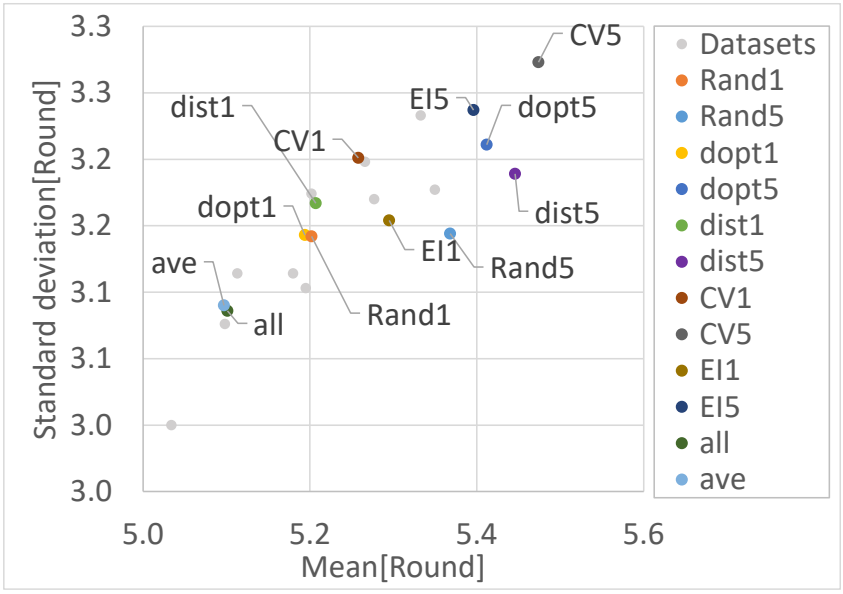


図 21　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差(Case 2)

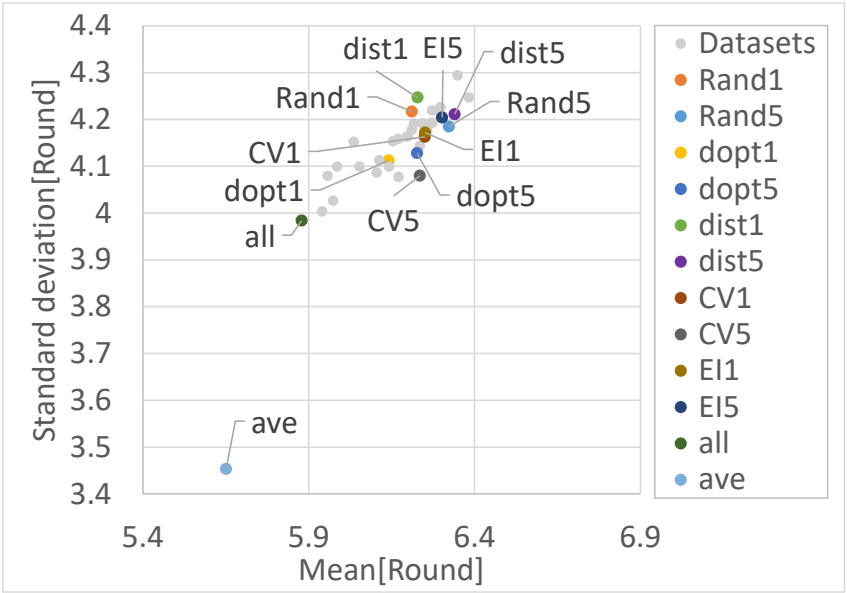


図 22　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差(Case 3)

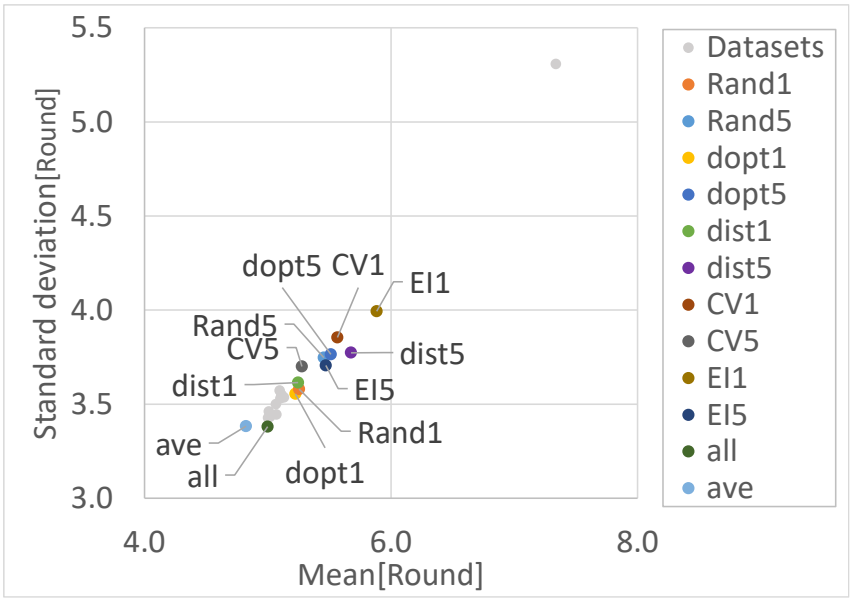


図 23　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差(Case 4)

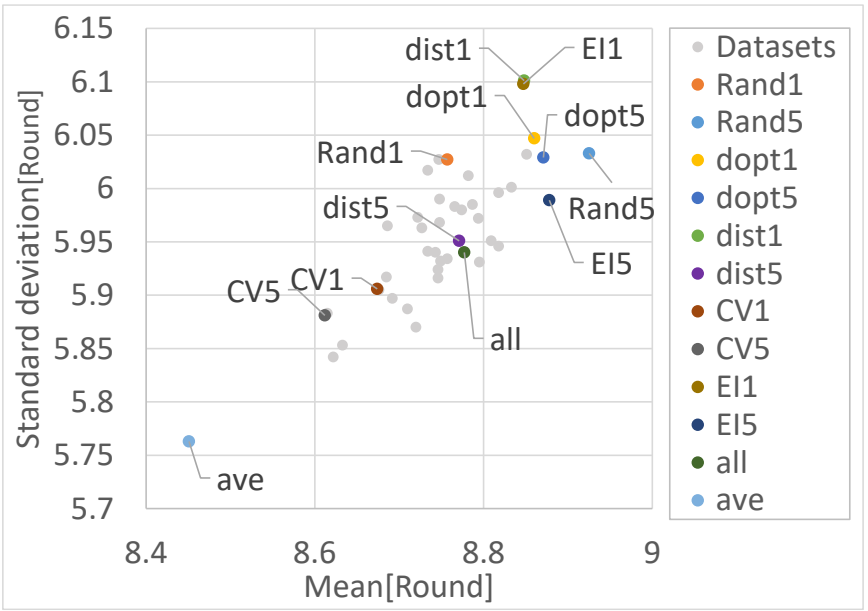


図 24　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差(Case 5)

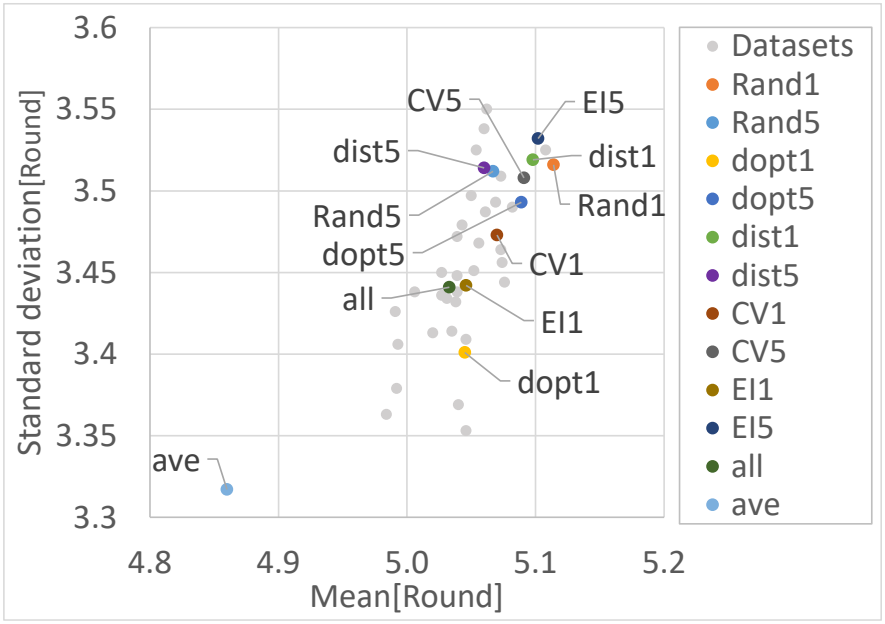


図 25　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差(Case 6)

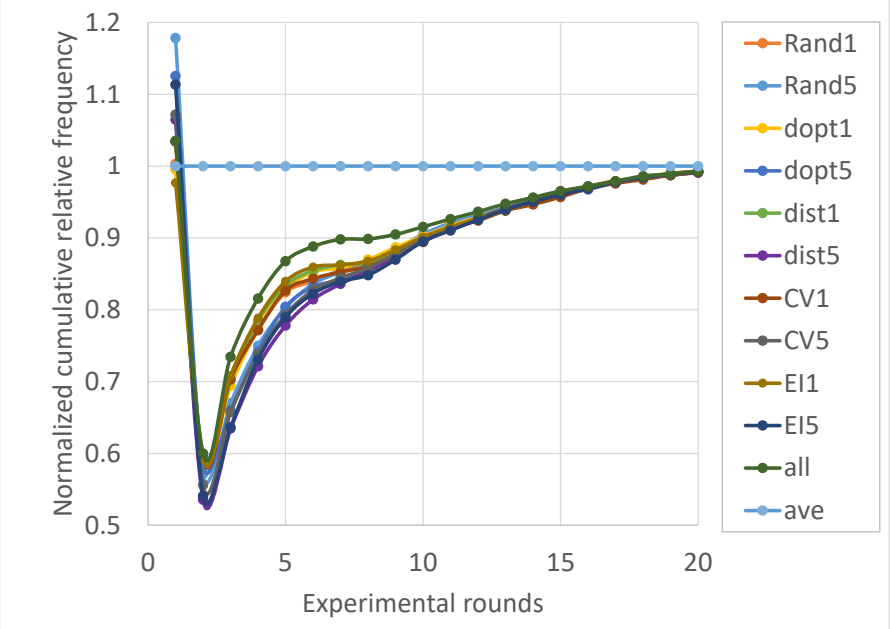


図 26　実験ラウンド数に対する正規化された累積相対度数(Case 1)

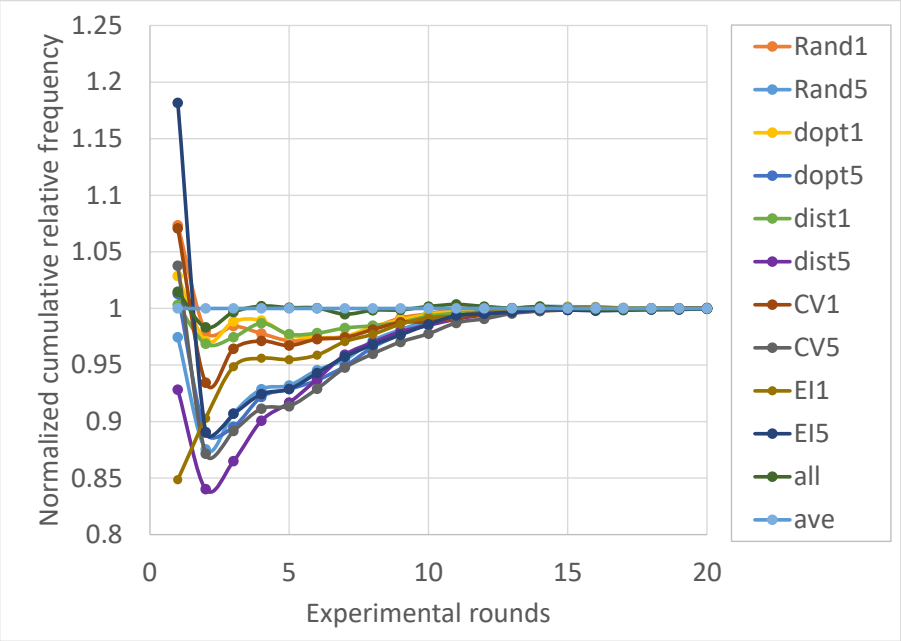


図 27　実験ラウンド数に対する正規化された累積相対度数(Case 2)

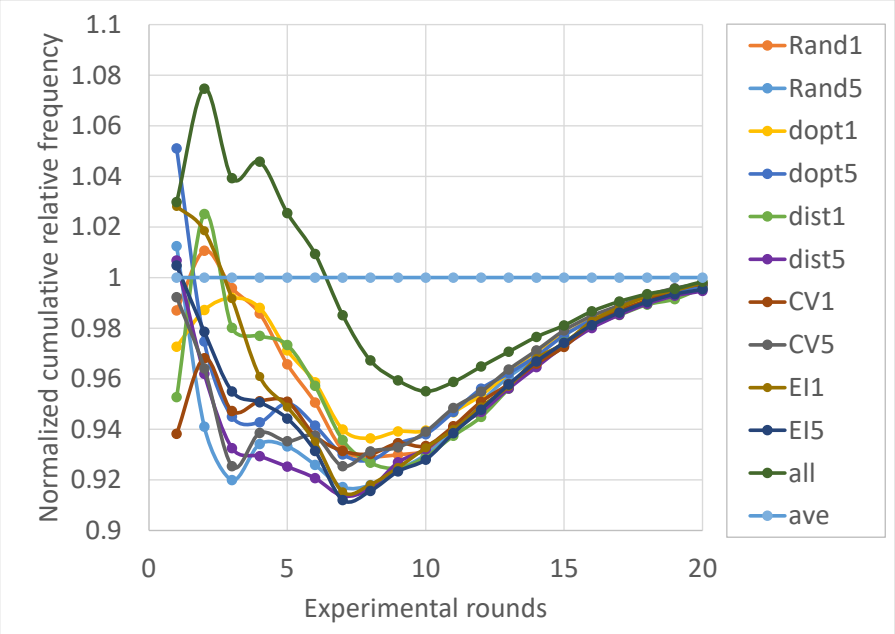


図 28　実験ラウンド数に対する正規化された累積相対度数(Case 3)

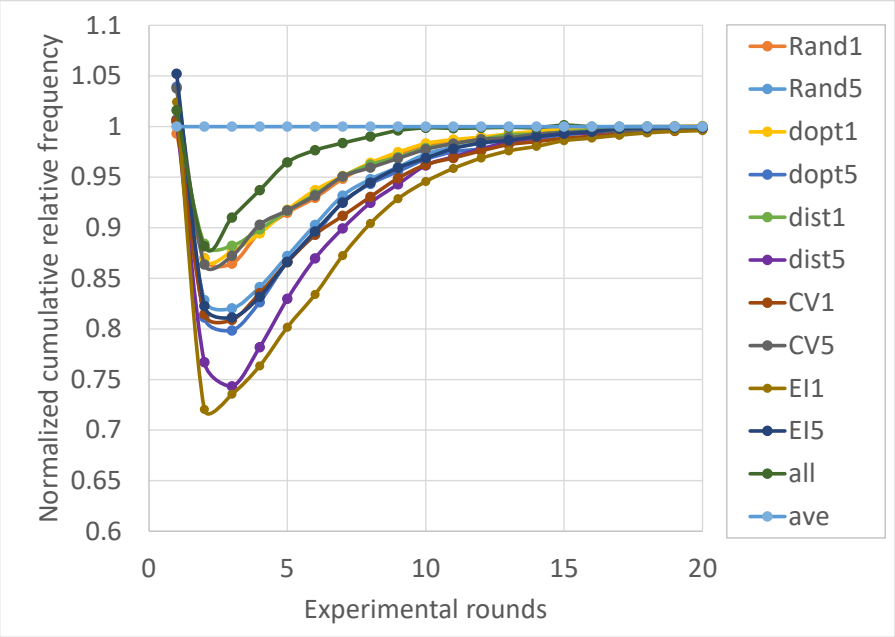


図 29　実験ラウンド数に対する正規化された累積相対度数(Case 4)

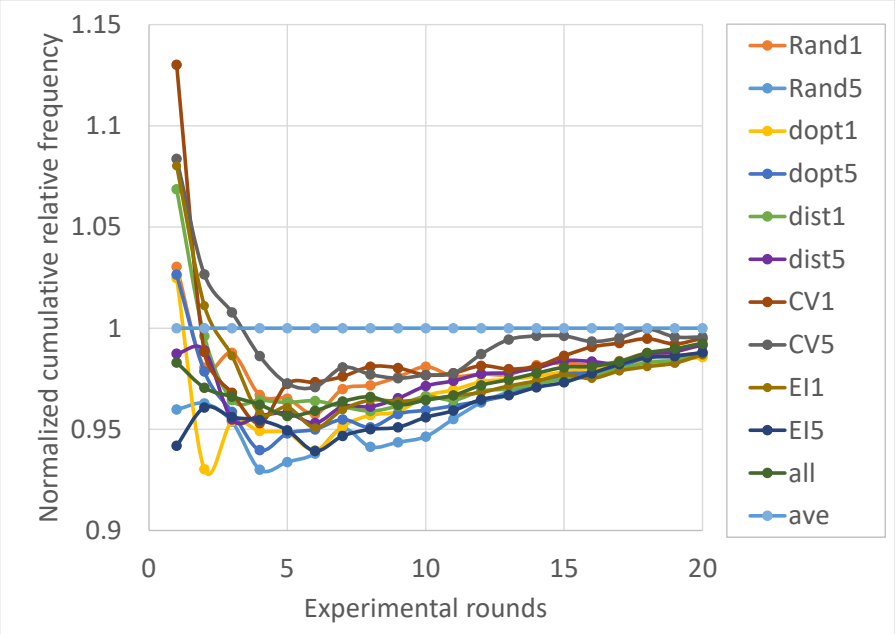


図 30　実験ラウンド数に対する正規化された累積相対度数(Case 5)

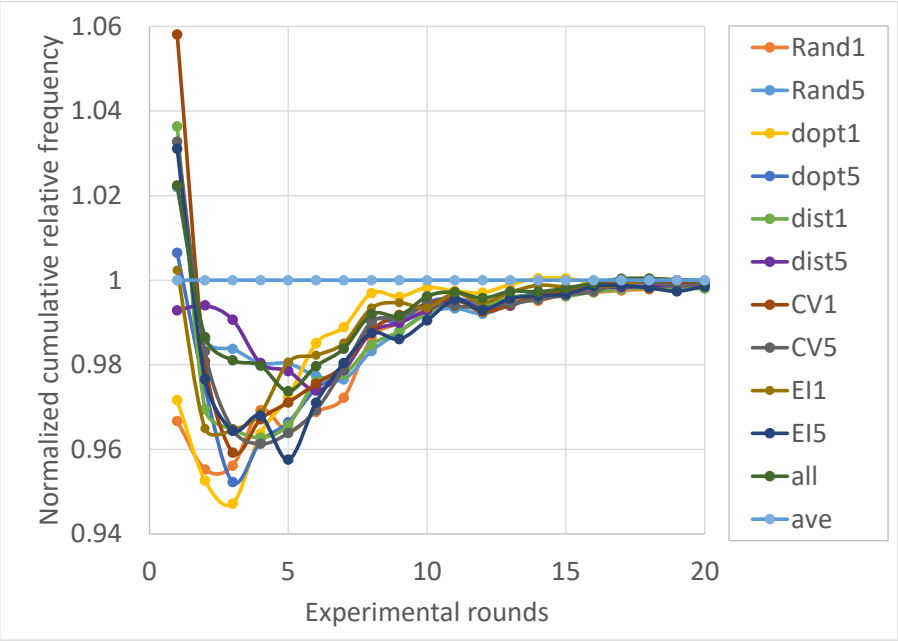


図 31　実験ラウンド数に対する正規化された累積相対度数(Case 6)

記述子群を選択する方法では探索性能が大きく向上することはなかった。Case 1 (反応 A, データセット数 10，終了条件 95%，提案数 5)を対象として，複数の GPR および BO でアンサンブルを行った場合の探索性能の変化を確認した。初期条件はランダムで決定し，初期条件や乱数の影響を可能な限り小さくするためにそれぞれ 10,000 回以上の計算を実施した。10 種類の記述子群から 2 つの記述子群を選択した場合，それぞれ単一の記述子群から作成したデータセットで探索を行った際に終了条件に到達するまでに要するラウンド数の平均値と，BO でアンサンブルを実施した際の終了条件に到達するまでに要するラウンド数の平均値に比較的高い相関(R=0.77)がみられた(図 32)。これは BO でアンサンブルを行うと，単一の記述子群から作成したデータセットを利用して探索を実施した場合の平均的な探索性能になることを意味する。つまり，探索性能が高い記述子群を選択できれば良い結果となり，探索性能が低い記述子群を選択すれば悪い結果になるということである。一方で，実験を行う前もしくは実験中に探索性能が高い記述子群だけを選択することは難しく，今回の検証ではそのような指標を見つけることはできなかった。

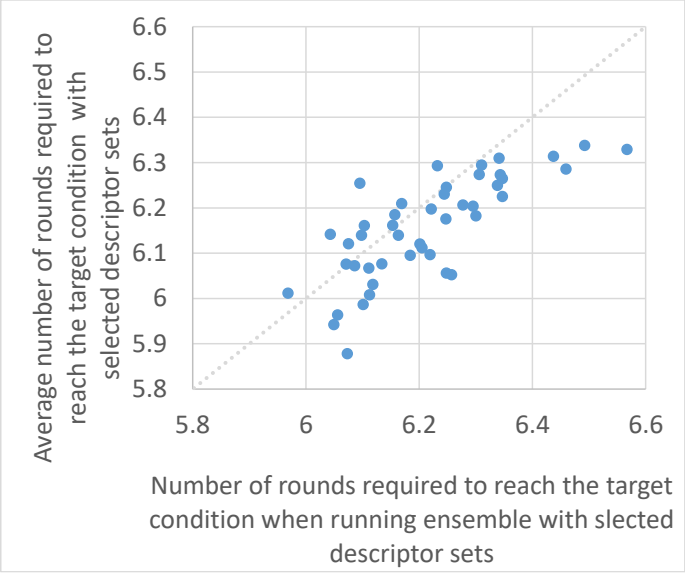


図 32　終了条件に到達するまでに要した実験ラウンド数の比較

平均化処理を行い新たな記述子群を作成する方法では，多くのケースで単体の記述子群から作成されたデータセットで BO を行った場合よりも探索性能が向上した。Case 1 (反応 A，データセット数 10，終了条件 95%，提案数 5)を対象として，平均化に用いる記述子群の数が 2, 3, 4, 5, 8, 10 の場合の探索性能をすべての記述子群の組み合わせに対して確認した(図 33, 図 34, 図 35, 図 36)。DFT から得られた記述子群を平均化して作成されたデータセットを用いた場合，それぞれ単一の記述子群から作成されたデータセットを用いた場合よりも終了条件に到達するまでに要するラウンド数の平均および標準偏差が小さくなる傾向がみられた。また，平均化に用いる記述子群が多くなればなるほど，複数の記述子群の平均化を実施した方が単一の記述子群を用いた場合の平均より探索性能が向上する割合が増加する傾向がみられた(表 25)。

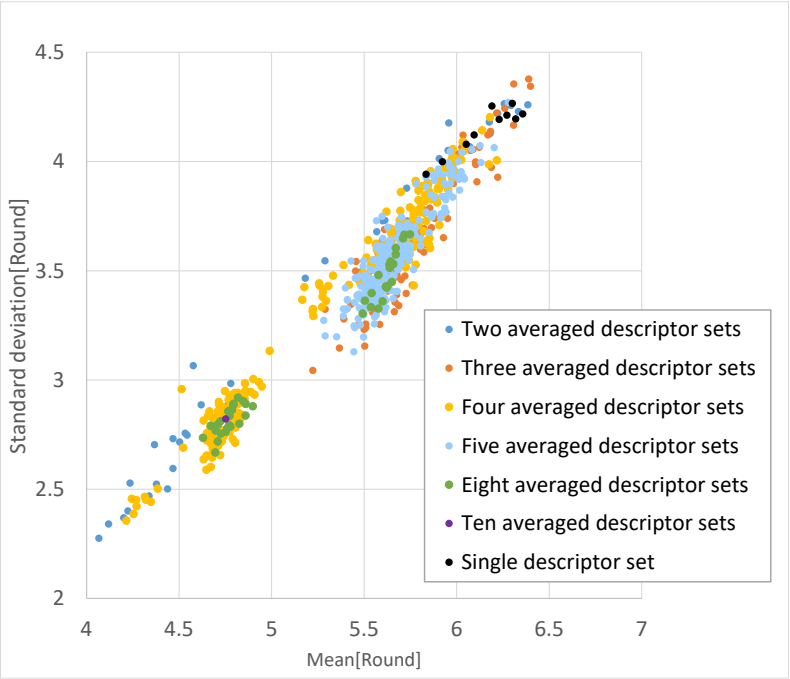


図 33　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差

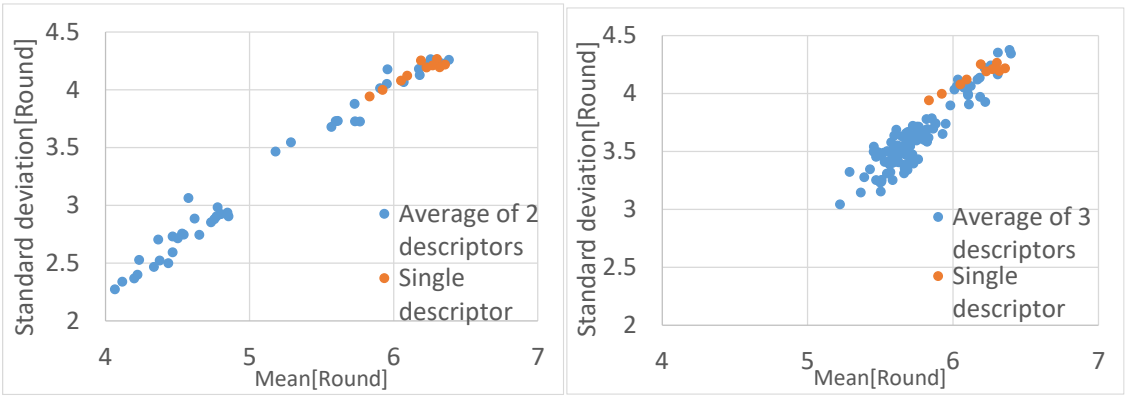


図 34　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差
(左: 記述子群の数 2，右: 記述子群の数 3)

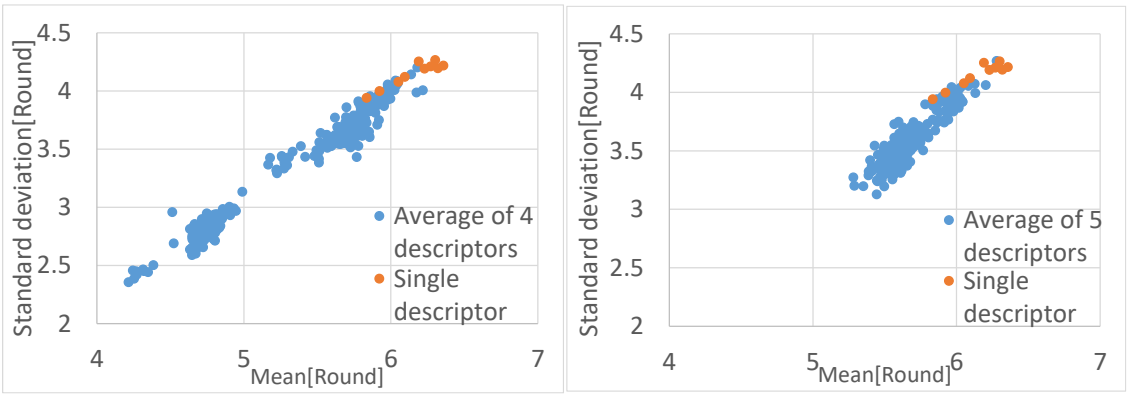


図 35　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差
(左: 記述子群の数 4，右: 記述子群の数 5)

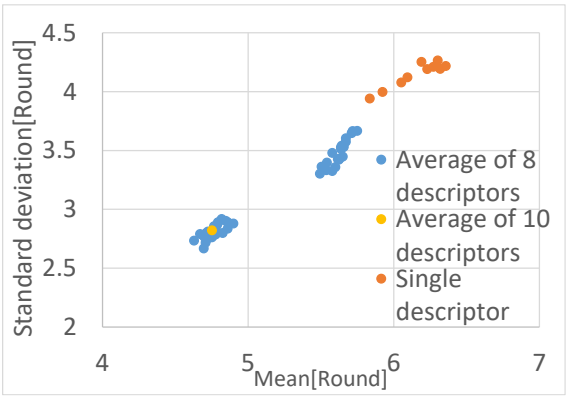


図 36　終了条件に到達するまでに要した実験ラウンド数の平均及び標準偏差
(記述子群の数 8, 10)

表 25　複数データセットの平均を用いた場合の単一データセットを用いた場合対する終了条件に到達するまでに要する実験ラウンド数の改善率（A: 複数データセット, B: 単一データセット）

記述子群の数	組み合わせ総数	平均(B)に対する平均(A)の改善率[%]	標準偏差(B)に対する標準偏差(A)の改善率[%]
2	45	86.7	86.7
3	120	90.8	95.8
4	210	98.1	99.5
5	252	98.6	99.5
8	45	100	100
10	1	100	100

表 26　終了条件に到達するまでに要した実験ラウンド数平均の最大，平均，最小

記述子群の数	組み合わせ総数	最小	平均	最大
2	45	3.706	5.070	6.384
3	120	5.222	5.737	6.399
4	210	3.990	5.315	6.217
5	252	5.281	5.663	6.279
8	45	4.628	5.147	5.748
10	1	4.751	4.751	4.751

表 27　終了条件に到達するまでに要した実験ラウンド数標準偏差の最大，平均，最小

記述子群の数	組み合わせ総数	最小	平均	最大
2	45	2.044	3.208	4.265
3	120	3.044	3.615	4.378
4	210	2.195	3.334	4.203
5	252	3.129	3.568	4.270
8	45	2.668	3.111	3.667
10	1	2.822	2.822	2.822

表 28　終了条件に到達するまでに要した実験ラウンド数の平均と標準偏差(Case 1)

No.	基底関数	汎関数	平均	標準偏差
1	STO-3G	B3LYP	5.923	3.999
2	3-21G	B3LYP	6.094	4.122
3	3-21Gd	B3LYP	6.301	4.266
4	6-31G	B3LYP	6.229	4.193
5	6-31Gd	B3LYP	6.271	4.212
6	6-31Gdp	B3LYP	6.051	4.079
7	6-31G+d	B3LYP	6.190	4.254
8	6-311G	B3LYP	6.357	4.218
9	6-311Gd	B3LYP	6.319	4.195
10	6-311Gdp	B3LYP	5.834	3.942

表 29　終了条件に到達するまでに要した実験ラウンド数の平均と標準偏差(Case 2)

No.	基底関数	汎関数	平均	標準偏差
1	STO-3G	B3LYP	5.113	3.114
2	3-21G	B3LYP	5.195	3.103
3	3-21Gd	B3LYP	5.277	3.170
4	6-31G	B3LYP	5.180	3.114
5	6-31Gd	B3LYP	5.266	3.198
6	6-31Gdp	B3LYP	5.098	3.076
7	6-31G+d	B3LYP	5.202	3.174
8	6-311G	B3LYP	5.350	3.177
9	6-311Gd	B3LYP	5.333	3.233
10	6-311Gdp	B3LYP	5.034	3.000

表 30　終了条件に到達するまでに要した実験ラウンド数の平均と標準偏差(Case 3)

No.	基底関数			汎関数	平均	標準偏差
	配位子	塩基	溶媒			
1	STO-3G	STO-3G	STO-3G	B3LYP	5.958	4.084
2	STO-3G	STO-3G	3-21G	B3LYP	6.229	4.124
3	STO-3G	STO-3G	6-31G	B3LYP	6.369	4.243
4	STO-3G	3-21G	STO-3G	B3LYP	5.961	4.017
5	STO-3G	3-21G	3-21G	B3LYP	6.253	4.203
6	STO-3G	3-21G	6-31G	B3LYP	6.311	4.227
7	STO-3G	6-31G	STO-3G	B3LYP	5.958	4.057
8	STO-3G	6-31G	3-21G	B3LYP	6.276	4.186
9	STO-3G	6-31G	6-31G	B3LYP	6.354	4.279
10	3-21G	STO-3G	STO-3G	B3LYP	6.063	4.108
11	3-21G	STO-3G	3-21G	B3LYP	6.103	4.104
12	3-21G	STO-3G	6-31G	B3LYP	6.253	4.176
13	3-21G	3-21G	STO-3G	B3LYP	6.032	4.144
14	3-21G	3-21G	3-21G	B3LYP	6.183	4.165
15	3-21G	3-21G	6-31G	B3LYP	6.17	4.072
16	3-21G	6-31G	STO-3G	B3LYP	5.978	4.031
17	3-21G	6-31G	3-21G	B3LYP	6.153	4.096
18	3-21G	6-31G	6-31G	B3LYP	6.195	4.166
19	6-31G	STO-3G	STO-3G	B3LYP	6.316	4.24
20	6-31G	STO-3G	3-21G	B3LYP	6.255	4.223
21	6-31G	STO-3G	6-31G	B3LYP	6.123	4.128
22	6-31G	3-21G	STO-3G	B3LYP	6.225	4.191
23	6-31G	3-21G	3-21G	B3LYP	6.264	4.213
24	6-31G	3-21G	6-31G	B3LYP	6.181	4.149
25	6-31G	6-31G	STO-3G	B3LYP	6.205	4.184
26	6-31G	6-31G	3-21G	B3LYP	6.314	4.235
27	6-31G	6-31G	6-31G	B3LYP	6.108	4.075

表 31　終了条件に到達するまでに要した実験ラウンド数の平均と標準偏差(Case 4)

No.	基底関数	汎関数	平均	標準偏差
1	STO-3G	B3LYP	5.004	3.427
2	3-21G	B3LYP	7.339	5.308
3	3-21Gd	B3LYP	5.020	3.433
4	6-31G	B3LYP	5.098	3.572
5	6-31Gd	B3LYP	5.066	3.500
6	6-31Gdp	B3LYP	5.104	3.531
7	6-31G+d	B3LYP	5.132	3.536
8	6-311G	B3LYP	5.010	3.461
9	6-311Gd	B3LYP	5.112	3.550
10	6-311Gdp	B3LYP	5.071	3.445

表 32　終了条件に到達するまでに要した実験ラウンド数の平均と標準偏差(Case 5)

No.	基底関数					汎関数	平均	標準偏差
	求電子剤	求核剤	配位子	塩基	溶媒			
1	STO-3G	STO-3G	STO-3G	STO-3G	STO-3G	B3LYP	8.746	5.938
2	STO-3G	STO-3G	STO-3G	STO-3G	6-31G	B3LYP	8.680	5.923
3	STO-3G	STO-3G	STO-3G	6-31G	STO-3G	B3LYP	8.685	5.924
4	STO-3G	STO-3G	STO-3G	6-31G	6-31G	B3LYP	8.736	5.923
5	STO-3G	STO-3G	6-31G	STO-3G	STO-3G	B3LYP	8.734	6.025
6	STO-3G	STO-3G	6-31G	STO-3G	6-31G	B3LYP	8.740	5.948
7	STO-3G	STO-3G	6-31G	6-31G	STO-3G	B3LYP	8.787	6.003
8	STO-3G	STO-3G	6-31G	6-31G	6-31G	B3LYP	8.738	6.000
9	STO-3G	6-31G	STO-3G	STO-3G	STO-3G	B3LYP	8.730	5.989
10	STO-3G	6-31G	STO-3G	STO-3G	6-31G	B3LYP	8.730	5.888
11	STO-3G	6-31G	STO-3G	6-31G	STO-3G	B3LYP	8.785	5.976
12	STO-3G	6-31G	STO-3G	6-31G	6-31G	B3LYP	8.770	5.999
13	STO-3G	6-31G	6-31G	STO-3G	STO-3G	B3LYP	8.723	6.013
14	STO-3G	6-31G	6-31G	STO-3G	6-31G	B3LYP	8.714	5.980
15	STO-3G	6-31G	6-31G	6-31G	STO-3G	B3LYP	8.634	5.866
16	STO-3G	6-31G	6-31G	6-31G	6-31G	B3LYP	8.813	5.964
17	6-31G	STO-3G	STO-3G	STO-3G	STO-3G	B3LYP	8.787	5.985
18	6-31G	STO-3G	STO-3G	STO-3G	6-31G	B3LYP	8.710	5.887
19	6-31G	STO-3G	STO-3G	6-31G	STO-3G	B3LYP	8.757	5.934
20	6-31G	STO-3G	STO-3G	6-31G	6-31G	B3LYP	8.694	5.893
21	6-31G	STO-3G	6-31G	STO-3G	STO-3G	B3LYP	8.833	6.001
22	6-31G	STO-3G	6-31G	STO-3G	6-31G	B3LYP	8.818	5.946
23	6-31G	STO-3G	6-31G	6-31G	STO-3G	B3LYP	8.818	5.996
24	6-31G	STO-3G	6-31G	6-31G	6-31G	B3LYP	8.622	5.842
25	6-31G	6-31G	STO-3G	STO-3G	STO-3G	B3LYP	8.727	5.963
26	6-31G	6-31G	STO-3G	STO-3G	6-31G	B3LYP	8.692	5.897
27	6-31G	6-31G	STO-3G	6-31G	STO-3G	B3LYP	8.795	5.931
28	6-31G	6-31G	STO-3G	6-31G	6-31G	B3LYP	8.615	5.883
29	6-31G	6-31G	6-31G	STO-3G	STO-3G	B3LYP	8.782	6.012
30	6-31G	6-31G	6-31G	STO-3G	6-31G	B3LYP	8.746	5.916
31	6-31G	6-31G	6-31G	6-31G	STO-3G	B3LYP	8.749	5.932
32	6-31G	6-31G	6-31G	6-31G	6-31G	B3LYP	8.851	6.032

表 33　終了条件に到達するまでに要した実験ラウンド数の平均と標準偏差(Case 6)

No.	基底関数					汎関数	平均	標準偏差
	求電子剤	求核剤	配位子	塩基	溶媒			
1	STO-3G	STO-3G	STO-3G	STO-3G	STO-3G	B3LYP	5.026	3.434
2	STO-3G	STO-3G	STO-3G	STO-3G	6-31G	B3LYP	5.120	3.538
3	STO-3G	STO-3G	STO-3G	6-31G	STO-3G	B3LYP	5.039	3.48
4	STO-3G	STO-3G	STO-3G	6-31G	6-31G	B3LYP	5.054	3.476
5	STO-3G	STO-3G	6-31G	STO-3G	STO-3G	B3LYP	5.017	3.421
6	STO-3G	STO-3G	6-31G	STO-3G	6-31G	B3LYP	5.038	3.436
7	STO-3G	STO-3G	6-31G	6-31G	STO-3G	B3LYP	5.058	3.544
8	STO-3G	STO-3G	6-31G	6-31G	6-31G	B3LYP	5.014	3.465
9	STO-3G	6-31G	STO-3G	STO-3G	STO-3G	B3LYP	5.043	3.370
10	STO-3G	6-31G	STO-3G	STO-3G	6-31G	B3LYP	4.980	3.392
11	STO-3G	6-31G	STO-3G	6-31G	STO-3G	B3LYP	4.981	3.364
12	STO-3G	6-31G	STO-3G	6-31G	6-31G	B3LYP	5.148	3.589
13	STO-3G	6-31G	6-31G	STO-3G	STO-3G	B3LYP	5.012	3.461
14	STO-3G	6-31G	6-31G	STO-3G	6-31G	B3LYP	5.065	3.482
15	STO-3G	6-31G	6-31G	6-31G	STO-3G	B3LYP	5.050	3.522
16	STO-3G	6-31G	6-31G	6-31G	6-31G	B3LYP	4.999	3.386
17	6-31G	STO-3G	STO-3G	STO-3G	STO-3G	B3LYP	5.096	3.471
18	6-31G	STO-3G	STO-3G	STO-3G	6-31G	B3LYP	5.071	3.427
19	6-31G	STO-3G	STO-3G	6-31G	STO-3G	B3LYP	5.014	3.413
20	6-31G	STO-3G	STO-3G	6-31G	6-31G	B3LYP	5.014	3.388
21	6-31G	STO-3G	6-31G	STO-3G	STO-3G	B3LYP	5.051	3.526
22	6-31G	STO-3G	6-31G	STO-3G	6-31G	B3LYP	5.091	3.467
23	6-31G	STO-3G	6-31G	6-31G	STO-3G	B3LYP	5.07	3.462
24	6-31G	STO-3G	6-31G	6-31G	6-31G	B3LYP	4.996	3.382
25	6-31G	6-31G	STO-3G	STO-3G	STO-3G	B3LYP	5.088	3.494
26	6-31G	6-31G	STO-3G	STO-3G	6-31G	B3LYP	5.046	3.504
27	6-31G	6-31G	STO-3G	6-31G	STO-3G	B3LYP	5.029	3.424
28	6-31G	6-31G	STO-3G	6-31G	6-31G	B3LYP	5.052	3.451
29	6-31G	6-31G	6-31G	STO-3G	STO-3G	B3LYP	5.051	3.372
30	6-31G	6-31G	6-31G	STO-3G	6-31G	B3LYP	5.065	3.464
31	6-31G	6-31G	6-31G	6-31G	STO-3G	B3LYP	5.065	3.576
32	6-31G	6-31G	6-31G	6-31G	6-31G	B3LYP	5.036	3.474

DFT 計算で選択する基底関数・汎関数の種類によって BO の探索性能に変化が生じる理由は，構造最適化を実施した分子およびその構造から計算される記述子に誤差が含まれるからと考えられる。DFT 計算時に与えるパラメータとして基底関数，汎関数を変化させているが，厳密には分子構造の収束状態はその他の計算条件の影響もうける。例えば初期構造が悪く局所最適になっている場合，設定パラメータが悪く収束しきれていない場合などがあり，必ずしも対象分子を正しく表現できていない可能性がある。一般的に分析機器などの測定器からデータを取得する際，測定される値は測定者の手技，機器間差，設定パラメータなど，測定者や測定器に大きく依存し，誤差が含まれる値が観測される。DFT 計算を用いて立体構造の最適化を行う際も同様であり，与える分子の初期構造や収束条件，基底関数，汎関数などの設定パラメータに大きく依存し，誤差が含まれる値が観測される。いずれの場合も真値は不明であり，誤差が含まれる値が観測される点は共通している。一般的に分析機器から得られる値の信頼性を上げるためには，測定数を増やす，測定機器を変えるなどして得られた複数の観測値の平均値を利用するといった対策が取られる。DFT 計算においても同様であり，DFT 計算を測定器と考えれば，複数の DFT 計算で得られた構造情報および記述子の平均値を用いることにより，より真の状態を表現できるようになる。その結果，BO における探索性能が向上したのではないかと考えている。一方で平均化することにより真の状態に近づくとしても，記述子で表現される空間が最適化を行いたい目的変数の挙動を表現できていなければならない。今回対象とした反応で用いた記述子は上記条件に適合したためこのような効果が表れたのではないかと推察している。記述子群の平均化は非常にシンプルで目新しさはないが非常に大きな効果を得ることができた。計算負荷も小さく(図 37, 図 38)，ドメイン知識があまりない研究者でも簡単に利用できる点も非常に優れている。また，複数の反応，記述子の組み合わせで提案手法の有効性が確認されたことから，本手法は合成反応全般に適用できる可能性があると考えている。今回対象とした直接的アリール化およびカップリング以外の反応や反応条件最適化以外の対象にも適用可能かどうか，どのような基底関数・汎関数の組み合わせでも再現性があるかどうか今後検証が必要である。

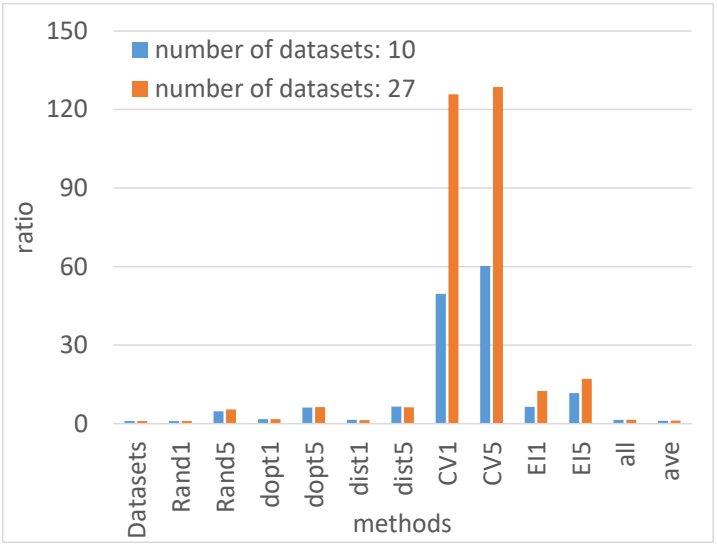


図 37　単一データセットを用いた場合を 1 とした場合の計算時間の比較(反応 A: パラジウム触媒を用いた直接的アリール化)

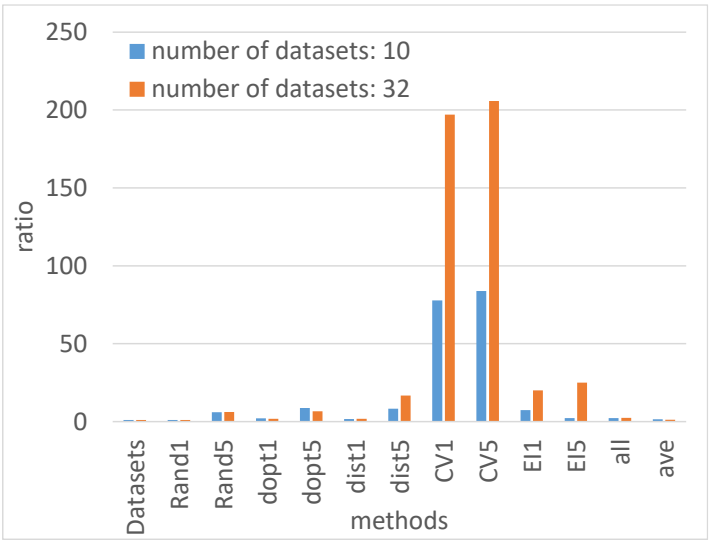


図 38　単一データセットを用いた場合を 1 とした場合の計算時間の比較(反応 B :鈴木–宮浦カップリング)

4.7 まとめ

BO で化合物を取り扱う場合，分子構造情報から計算された記述子を説明変数として利用することができる。そこで様々な組み合わせの基底関数・汎関数で DFT 計算された記述子を活用して，BO の探索性能を向上させる方法を開発した。直接的アリール化および鈴木-宮浦カップリング反応に対して，いくつかの基底関数・汎関数の組み合わせで記述子群を作成し，それらから作成したデータセットを用いて探索性能の検証を実施した。複数の記述子群を平均化して作成したデータセットを BO に用いた場合，単一の記述子群から作成したデータセットを用いる場合よりも探索性能が向上した。また，平均化に用いる記述子群が多くなればなるほど，探索性能が向上した割合が増加する傾向があった。複数の記述子群を平均化する方法は計算負荷も小さく，量子科学計算に関する知識があまりない研究者でも簡単に利用可能である。今後，今回提案した方法がほかの合成反応や反応だけでなく化合物を扱う場合，基底関数・汎関数が異なる場合にも一般的に利用可能かどうかを確認する必要がある。また，カテゴリカル変数となる因子の数や基底関数・汎関数の数が増えると組み合わせ数が膨大となるという課題が残されているため，計算負荷を下げるための工夫も必要である。

4.8 参考文献

37. P.Honarmandi, Accelerated materials design using batch Bayesian optimization: A case study for solving the inverse problem from materials microstructure to process specification, Computational Materials Science, Volume 210, July 2022
38. Kyohei Hanaoka, Bayesian optimization for goal-oriented multi-objective inverse material design, iScience, VOLUME 24, ISSUE 7, 102781, JULY 23, 2021
39. https://mordred-descriptor.github.io/documentation/master/descriptors.html (accessed 2023-20th-June)
40. https://gaussian.com/ (accessed 2023-20th-June).
41. http://www.codessa-pro.com (accessed 2023-20th-June).

第5章　結論

本論文では，医薬品原薬製造プロセス開発への機械学習の活用における課題の把握と対策の実行を行い，製造プロセス開発の効率向上による検討期間の短縮による上市までのスピードアップや開発コスト削減，安定生産へとつなげることを目的として，機械学習を用いた予測モデルの開発とベイズ最適化(BO)を用いた適応的実験計画法の活用を検討した。

1 つ目の研究では，有機過酸化物の構造式から自己促進分解温度(SADT)を推算するモデルの構築を試みた。化合物の構造式から記述子を計算し，偏最小二乗（PLS: Partial Least Squares）回帰，サポートベクター回帰(SVR: Support Vector Regression)を適用した結果，比較的精度良く SADT を予測できるモデルを構築することができた。

2 つ目の研究では，クラスタリングを用いたベイズ最適化の初期条件の決定法を検討した。BO で最適解を効率的に探索するには，GPR モデルを構築する際に適切な初期サンプルを提供する必要がある。本研究では，目的変数に大きな影響を与える因子をカバーするクラスタリング情報に基づく初期サンプル選択手法を提案し，BO とのカップリング反応条件の最適化に適用した。その結果，クラスタを適切に形成し，各クラスタから初期サンプルを選択した場合，提案手法はランダムサンプリングや D 最適基準に基づくサンプリングよりも少ない実験回数で最適解に到達することを確認した。

3 つ目の研究では，DFT を用いた場合にベイズ最適化の探索性能を向上させるための方法を検討した。BO で化合物を取り扱う場合，分子構造情報から計算された記述子を説明変数として利用することができる。そこで様々な組み合わせの基底関数・汎関数で DFT 計算された記述子を活用して，BO の探索性能を向上させることができる方法を開発した。複数の記述子群を平均化する方法は計算負荷も小さく，量子科学計算に関する知識があまりない研究者でも簡単に利用可能な方法である。

本研究では，原薬プロセス開発への機械学習の適用を検討した。予測モデルおよびそれを用いた最適化はプロセス開発において様々な場面で活用できることがわかった。また，SADT 予測モデルの予測精度や計算負荷，ベイズ最適化の初期サンプル選択や分子記述子の利用など，いくつかの課題に対して考察を行い対策の提案を実施した。しかし，依然として様々な課題も残されている。例えば，ベイズ最適化において目的変数が複数になった場合の計算負荷が大きい点，実験や分析などに多くの時間を要しベイズ最適化の効果を最大限活用できない場合がある点，研究者が実験空間を定義しなければならない点などである。今後，ベイズ最適化のアルゴリズム改良や計算機能力の向上，ロボット等を活用したオートメーション，生成 AI の活用などにより，これらの課題はそう遠くない未来に解決できるのではないかと推察する。一方で，研究開発の現場では，研究者の勘と経験による試行錯誤で実験条件が決定されることも多く，機械学習の活用に対する心理的なハードルはいまだ高い。特に，合成原薬開発は比較的成熟した分野であるにもかかわらず，昔ながらのスタイルで実験が行われることも多く，機械学習を活用するインフラが整っていないと感じる場面が散見

される。機械学習のさらなる活用のためには，適用対象の自動化やデジタル化を進めることも必要だと考える。今後は予測モデルやベイズ最適化のような機械学習の普及および活用を進め，医薬品原薬プロセス開発の最適化だけでなく，様々な最適化に広く適用できるように検討を継続していきたい。

謝辞

本論文は著者が明治大学大学院理工学研究科博士後期課程において，データ化学工学研究室にて行った研究成果をまとめたものです。本研究を行うにあたり，多くの方々にお世話になりました。この場をお借りしてお礼を述べたいと思います。

本論文をまとめるにあたり，研究の機会を与えて頂き，その遂行が円滑に行われるために，適切なご指導とご鞭撻を頂きました金子宏昌准教授に心より感謝致します。研究方針への助言や論文執筆，学会発表など，様々な場面で助けていただきました。また，副査をお引き受け頂いた深澤倫子教授，土本晃久教授にも感謝申し上げます。お忙しい中論文をご精読頂き大変有益なご助言を頂きました。明治大学理工学部データ化学工学研究室ならびにアステラス製薬原薬研究所の皆様には在学中に様々な面で助けていただきました。ご協力，ご支援ありがとうございました。最後に，私の博士後期課程への進学を快諾し，3 年間の学生生活を辛抱強く支えてくれた妻，子供たちに深く感謝致します。