サイエンティフィック・システム研究会 科学技術計算分科会 2017年度会合

マテリアルズインフォマティクスの最前線

吉田 亮 [1,2,3,4] *yoshidar@ism.ac.jp*

[1] 情報・システム研究機構 統計数理研究所 モデリング研究系 准教授
[2] 情報・システム研究機構 統計数理研究所 ものづくりデータ科学研究センター センター長
[3] 総合研究大学院大学 複合科学研究科 統計科学専攻 准教授
[4] 物質・材料研究機構 特別研究員

Mi²i "Materials research by Information Integration" Initiative
情報統合型物質・材料開発イニシアティブ

“外挿的予測と発見” のデータ科学

Data Science towards Extrapolative Prediction and Discovery

- データサイエンスの予測は “内挿的” (材料開発者の経験や勘も同様に)
- “外挿的予測と発見” を実現するシナリオ？

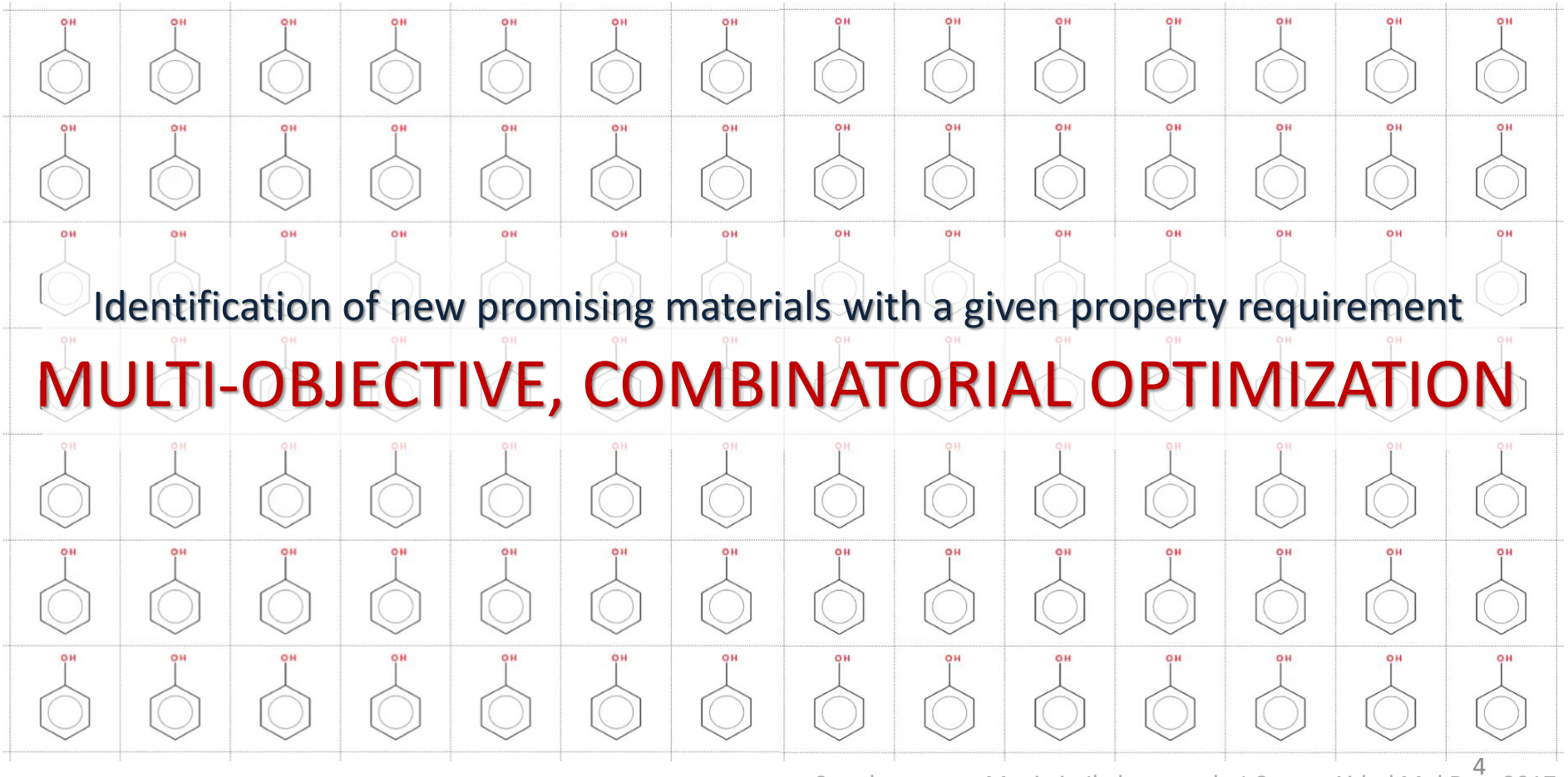
Bayesian Inverse Materials Design

ベイズ推論で新物質を発見する

MACHINE LEARNING

SAILING THE BIG MATERIAL SPACE

e.g. 10^{60} small-sized organic molecules (GO game: 250^{150})



Identification of new promising materials with a given property requirement

MULTI-OBJECTIVE, COMBINATORIAL OPTIMIZATION

Supplementary Movie in Ikebata et al., J Compt Aided Mol Des, 2017

BAYESIAN INFERENCE

Transforming forward models into the backward prediction

Given data

Y

FORWARD PREDICTION

BACKWARD PREDICTION

Unknown

S

Posterior - backward prediction

Likeliness of *S* with respect to observed *Y*

Likelihood - forward prediction

Goodness-of-fit of *S* with respect to *Y*

Prior distribution

Prior knowledge on *S*

Backward Forward Prior

$$S \sim p(S|Y) \propto p(Y|S) \times p(S)$$

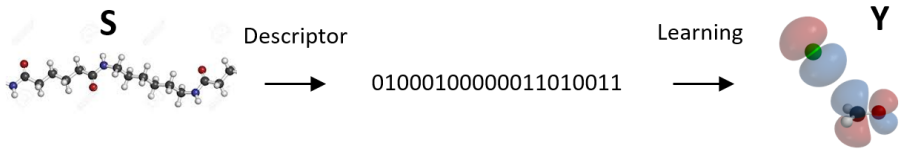
Inversion

BAYESIAN INVERSE MATERIALS DESIGN

Ikebata *et al.*, *J Compt Aided Mol Des*, 2017

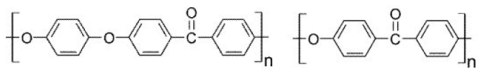
FORWARD PREDICTION - p(Y|S)

Prediction of properties $Y = (Y_1, \ldots, Y_m)$ of input material S
Machine learning with structure-property data



BACKWARD PREDICTION - p(S|Y∈U)

New structures S with properties falling within a desired region $Y \in U$
Generation of new 'materials'

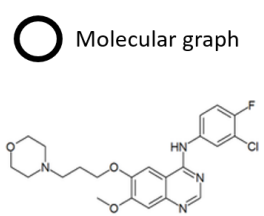


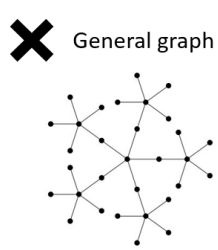
backward Y → S forward S → Y prior

$$S \sim p\left(S \middle| Y \in U\right) \propto p\left(Y \in U \middle| S\right) \times p\left(S\right)$$

PRIOR DISTRIBUTION - p(S)

Regularization to generate "*chemically favorable structures*"
Natural Language Processing techniques: "*chemical language model*"





Geometry
Exclusion of chemically unfavorable bonds
Chemical stability
Drug-likeliness
Synthesizability

R package iqspr v2.3 @CRAN
https://cran.r-project.org/web/packages/iqspr/index.htm

R package iqspr v2.3 @GitHub
https://github.com/GLambard/inverse-molecular-design

高分子材料ビッグデータに基づく材料探索を中心に

ベイズ推論×PolyInfo

"Materials research by Information Integration" Initiative
情報統合型物質・材料開発イニシアティブ

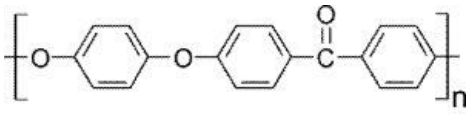
PolyInfo (@NIMS) PROJECT

Morikawa (Tokyo Tech)
Xu, Kuwajima, Kondo (NIMS)
Hongo (JAIST)
Yamada, Lambard (ISM)
Wu S
(ISM)

WORLD LARGEST POLYMER DATABASE × STATE-OF-THE-ART MACHINE LEARNING

Target: higher thermal conductivity polymers

Structure (S) → **Descriptor (ϕ)** → **Properties (Y)**



0100010000011010011

λ: thermal conductivity
T_g: glass transition temp.
T_m: melting points
and so on

S = molecular structure of repeat unit (monomer)
(and/or process/other parameters)

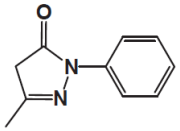
Quantitative Structure-Property Relationship

順方向の物性予測モデルを作る

FORWARD PREDICTION

Quantitative structure-property relationship analysis (QSPR)

Structure (S)



Descriptor (ϕ)

01000100000011010011

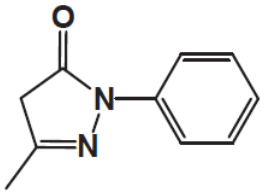
Molecular fingerprint
PubChem
MDL MACCS keys
ECFP6
Kolekota-Roth
etc

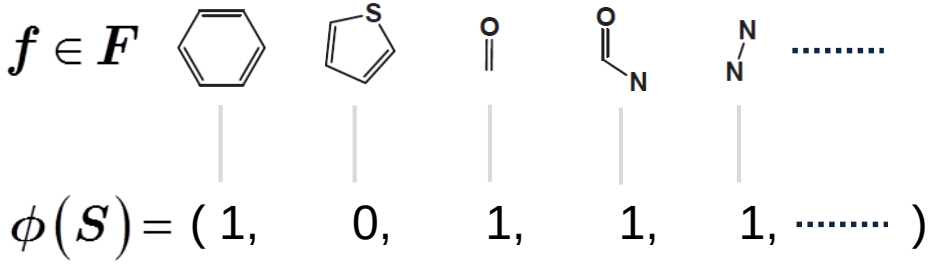
Properties (Y)

Physiological activity
Toxicity
Band gap
Energy

Deep Learning
Elastic Net
Random Forest
Gradient Boosting
Support Vector Machine
etc

e.g. **molecular fingerprint**





R PACKAGES: DESCRIPTOR & ML ALGORITHM

Regression and classification

Method	Package
Elastic-net	glmnet
Deep Learning	h2o, mxnet
Support Vector Machine	kernlab
Gaussian Process regression	GPfit, laGP
Boosting	Boost
Bagging	Ipred
Random Forest	randomForest, ranger, Rborist
Gradient Boosting	xgboost

Hyperparameter optimization

Method	Package
Bayesian Optimization	rBayesianOptimization

Descriptors in the *rcdk* package

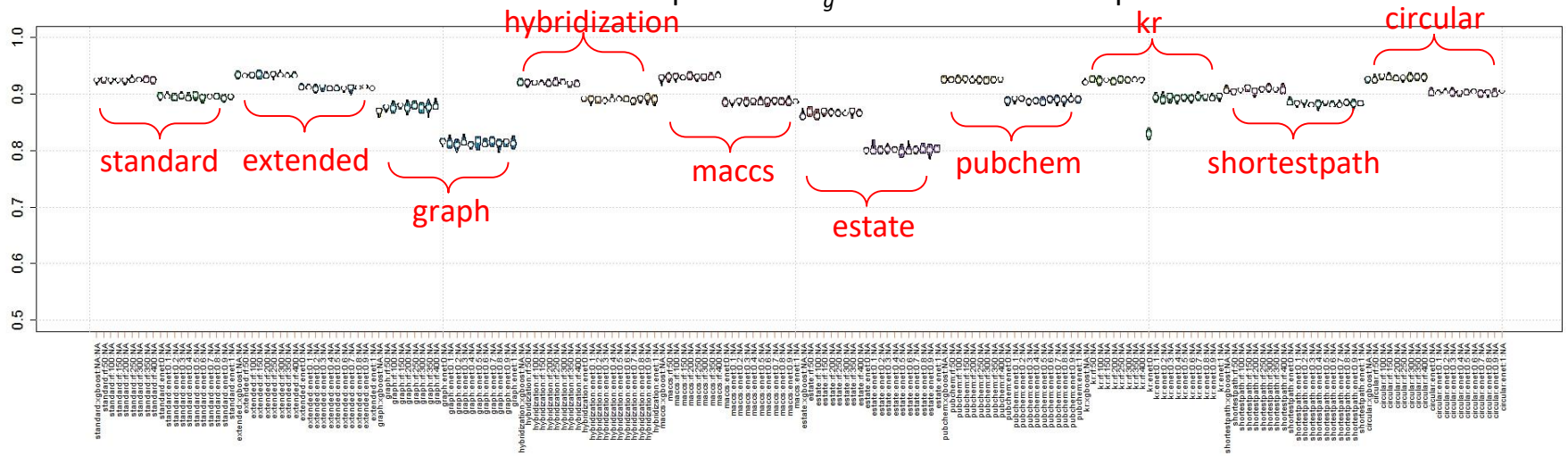
Name	Description
standard	Paths of a given length
extended	standard + ring + atomic property
maccs	166 bits MDL MACCS keys
circular	ECFP6
pubchem	881 bits PubChem fingerprint
graph	standard + connectivity
kr	4860 bits Kolekota-Roth
hybridization	standard + Info. on hybridization
shortestpath	shortest paths between atoms
signature	count type of fingerprint

+ more than 200 physical and chemical descriptors

POLYMER PROPERTY

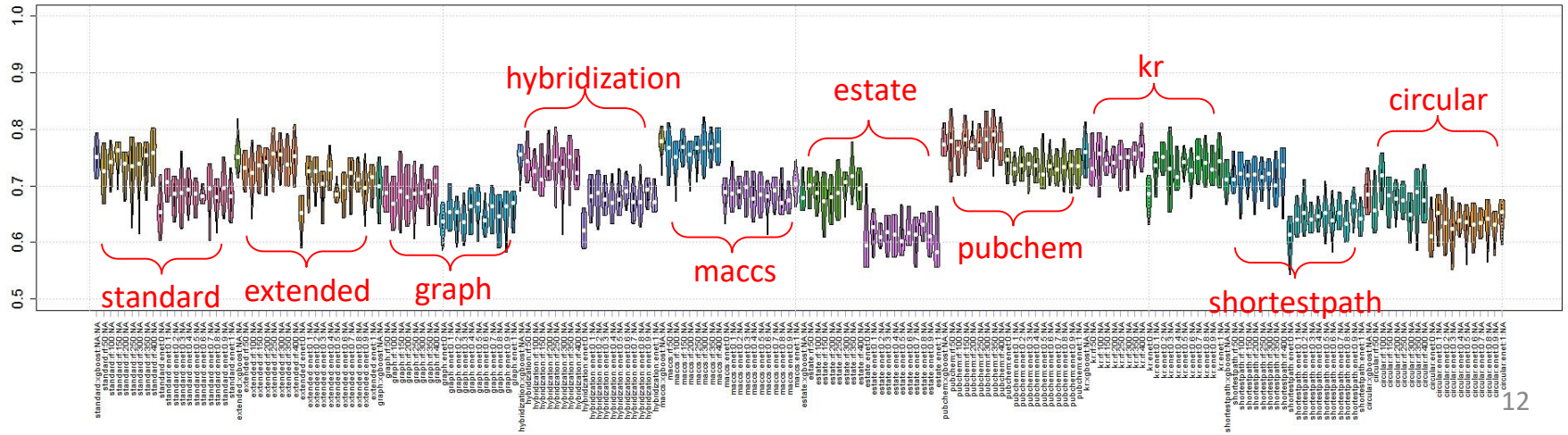
Glass transition temperature

Correlation between observed and predicted T_g with different descriptors and ML models



Density

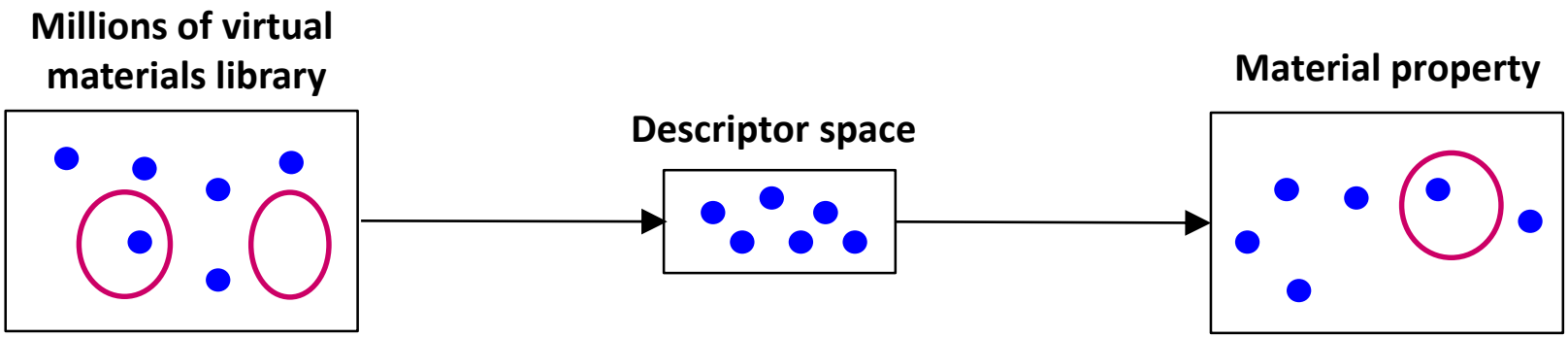
Correlation between observed and predicted D_s with different descriptors and ML models



One of the grand challenges in Materials Informatics

HIGH-THROUGHPUT SCREENING

高品質・大量の仮想物質を作製・保有することが，莫大な価値を生み出す



Machine Learning based HTS (w/wo Bayesian Optimization) for materials property prediction

Gómez-Bombarelli et al., *Nat Mater*, 15, 1120-1127, 2016

Seko et al., *Phys Rev B*, 89, 054303, 2014

Seko et al., *Phys Rev Lett*, 115, 205901, 2015

Balachandran et al., *Sci Rep*, 6, 19660, 2016

Kiyohara et al., *Sci Adv*, 2, e1600746, 2016

Rupp et al., *Phys Rev Lett*, 108, 058301, 2012

Pilania et al., *Sci Rep*, 3, 2810, 2013

Deep Learning for solving Schrödinger equation

Mills et al., Deep learning and the Schrödinger equation, *arXiv:1702.01361*, 2017

Carleo et al., Solving the quantum many-body problem with artificial neural networks, *Science*, 355(6325), 602-606, 2017

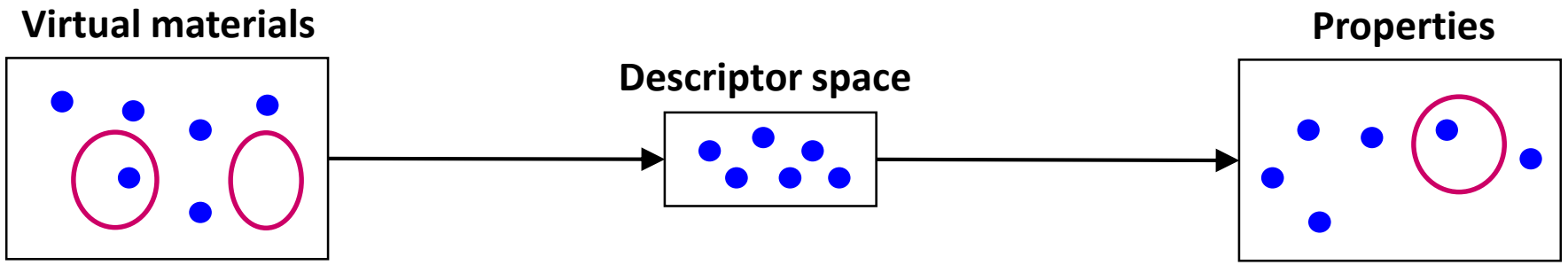
入力はポテンシャル画像，出力は波動関数

Inverse Quantitative Structure-Property Relationship

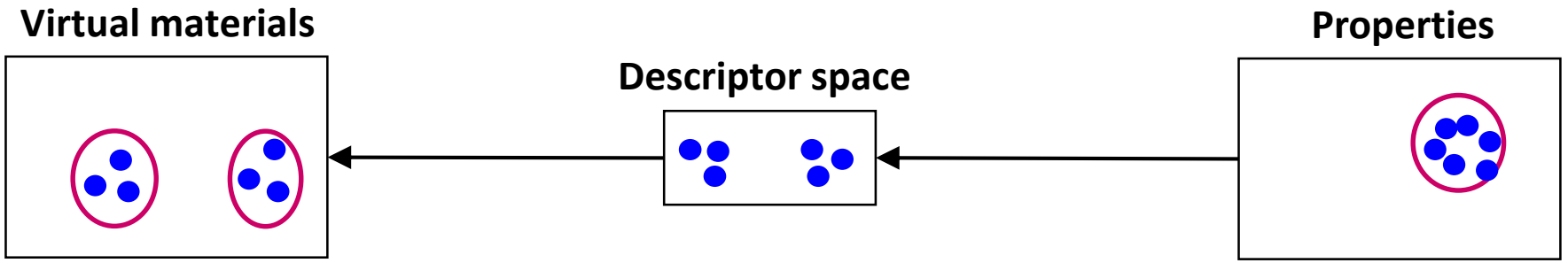
逆方向の計算で埋蔵物質を発掘

BACKWARD PREDICTION

$P(Y|S)$ Limitation = Low hit rate - when the input space is exceedingly high-dimensional, the exhaustive search is impossible to perform



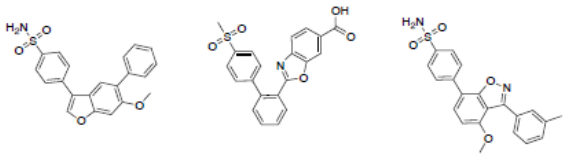
$P(S|Y) \propto P(Y|S)P(S)$ Explore promising materials intensively around high probability regions of the posterior



BACKWARD PREDICTION

MONTE CARLO STRUCTURE GENERATOR

Creation of promising hypothetical materials (*e.g.* molecules) with the posterior probability distribution



Backward prediction　　Forward prediction　　Prior

$$\sim p\left(S \middle| Y \in U\right) \propto \prod_{g=1}^{m} p\left(Y_g \in U_g \middle| S\right) \times p\left(S\right)$$

※ Prior: “ポリマーらしさ” の学習済モデル

J Comput Aided Mol Des (2017) 31:379–391
DOI 10.1007/s10822-016-0008-z

CrossMark

Bayesian molecular design with a chemical language model

Hisaki Ikebata[1] · Kenta Hongo[2,3,4] · Tetsu Isomura[5] · Ryo Maezono[2] · Ryo Yoshida[1,3,6]

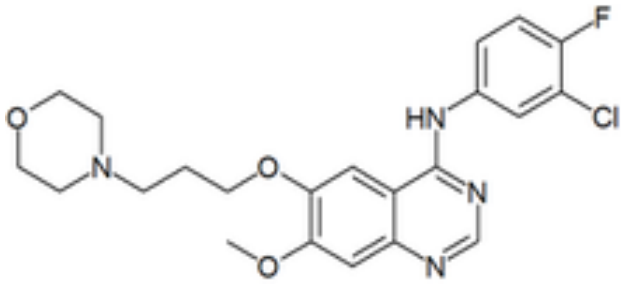
Received: 25 October 2016 / Accepted: 31 December 2016 / Published online: 9 March 2017
© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract The aim of computational molecular design is the identification of promising hypothetical molecules with a predefined set of desired properties. We address the issue of accelerating the material discovery with state-of-the-art machine learning techniques. The method involves two dif-

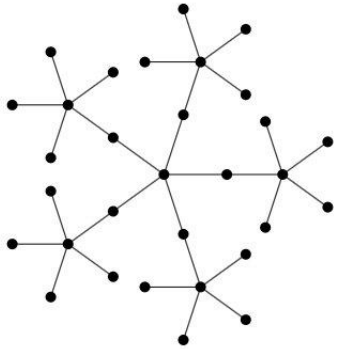
this issue, we derive a chemical language model that acquires commonly occurring patterns of chemical fragments through natural language processing of ASCII strings of existing compounds, which follow the SMILES chemical language notation. In the backward prediction

CHEMICALLY FAVORABLE STRUCTURE?

Molecular graph



General graph



p(S)

prior knowledge

Geometry

Exclusion of chemically unfavorable bonds, *e.g.* F-N, C=O=C

Chemical stability

Drug-likeliness

Synthesizability

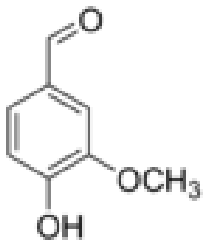
Liquid crystalline likeliness

p(S) CHEMICAL LANGUAGE MODEL

Ikebata *et al.*, *J Compt Aided Mol Des*, 2017

■ 実在の化合物セットから ‘化合物らしさ’ を学習

- **化学構造を文字列で表現 (SMILES)**
- **言語モデルを用いて文字のパターンや文法を学習**



vanillin $C_8H_8O_3$

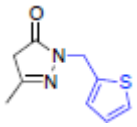
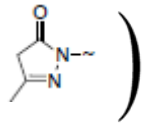
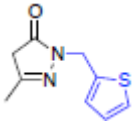
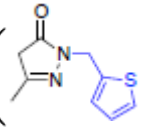
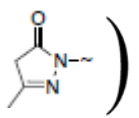
O=Cc**1**ccc**(**O**)**c**(**OC**)**c**1**

環の始点と終点は同じ数字

側鎖は括弧

■ 訓練された言語モデル：‘化合物らしい構造’ の生成器として使用

- Prior dist.
- Proposal distribution in SMC or MCMC

$$p\left(\text{  } \middle| Y \in U, \text{  }\right) \propto p\left(Y \in U \middle| \text{  }\right) p\left(\text{  } \middle| \text{  }\right)$$

TWO TECHNICAL ISSUES

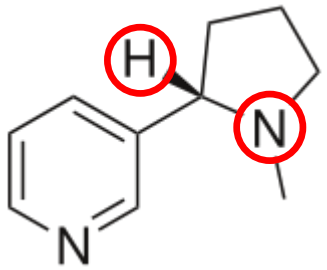
- **Grammar error:** Avoid the occurrence of unclosed ring or side-chain indicators

`S[1:i-1] = CCCCCC(CCCC(OC(O)C`

Two unclosed parentheses '('

The Markov probability should be biased such that two closing letters ')' appear somewhere in the rightward

- **Long term dependency:** 'Neighbors' in a molecular graph are not neighboring in the string representation

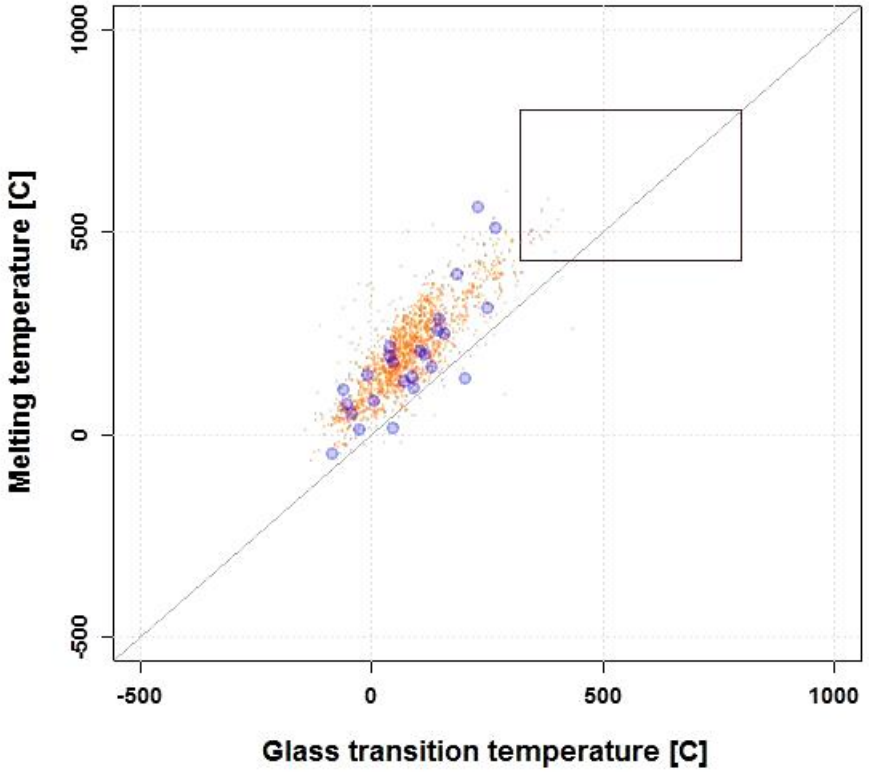


`CN1CCC[C@H]1c2cccnc2`

NEW FUNCTIONAL POLYMERS

Higher glass transition temperatures and higher melting points

Property refinement process



Structure modification

Orange: existing polymers with experimentally measured T_g and T_m

Blue: polymers computationally created by *iqspr* (circle size: the magnitude of likelihood)

NEW FUNCTIONAL POLYMERS

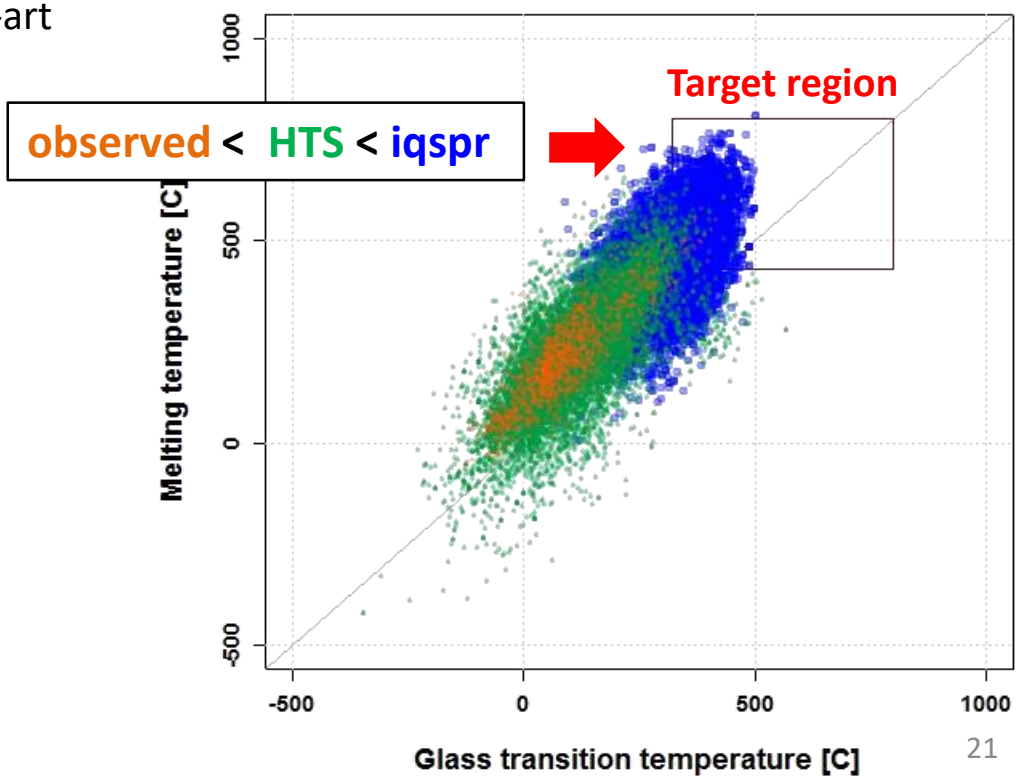
Higher glass transition temperatures and higher melting points

Further expansion of the materials landscape (T_g and T_m)

Orange: polymers with observed T_g and T_m
Green: HTS of polymers with unobserved T_g and T_m
Blue: newly created polymers of iqspr

Result & Conclusion

Our algorithm has predicted the existence of yet-undiscovered many (9,685) polymers with T_g and T_m much higher than those of the state-of-the-art



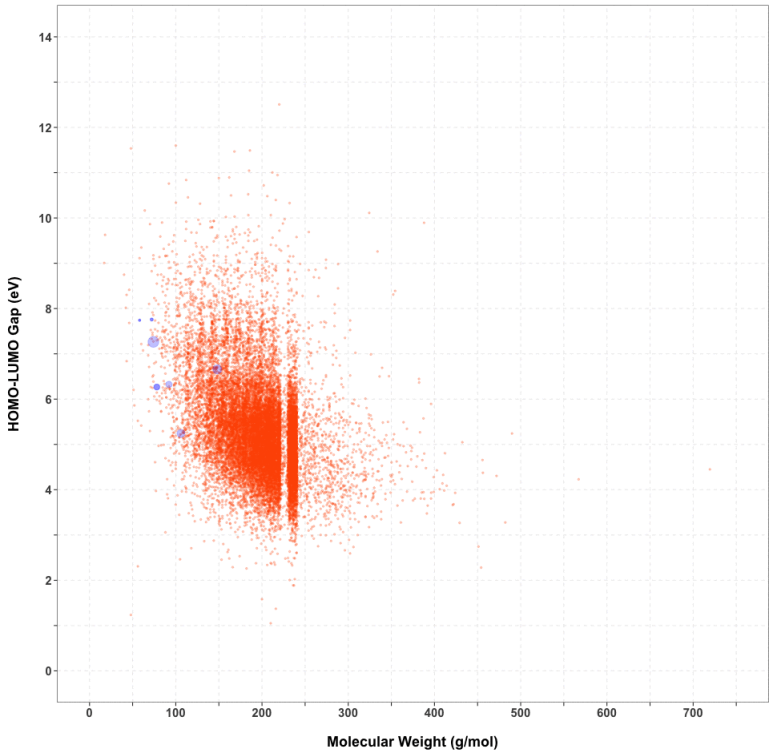
LOW BAND GAP MOLECULES

Application of iqspr on novel monomer discovery

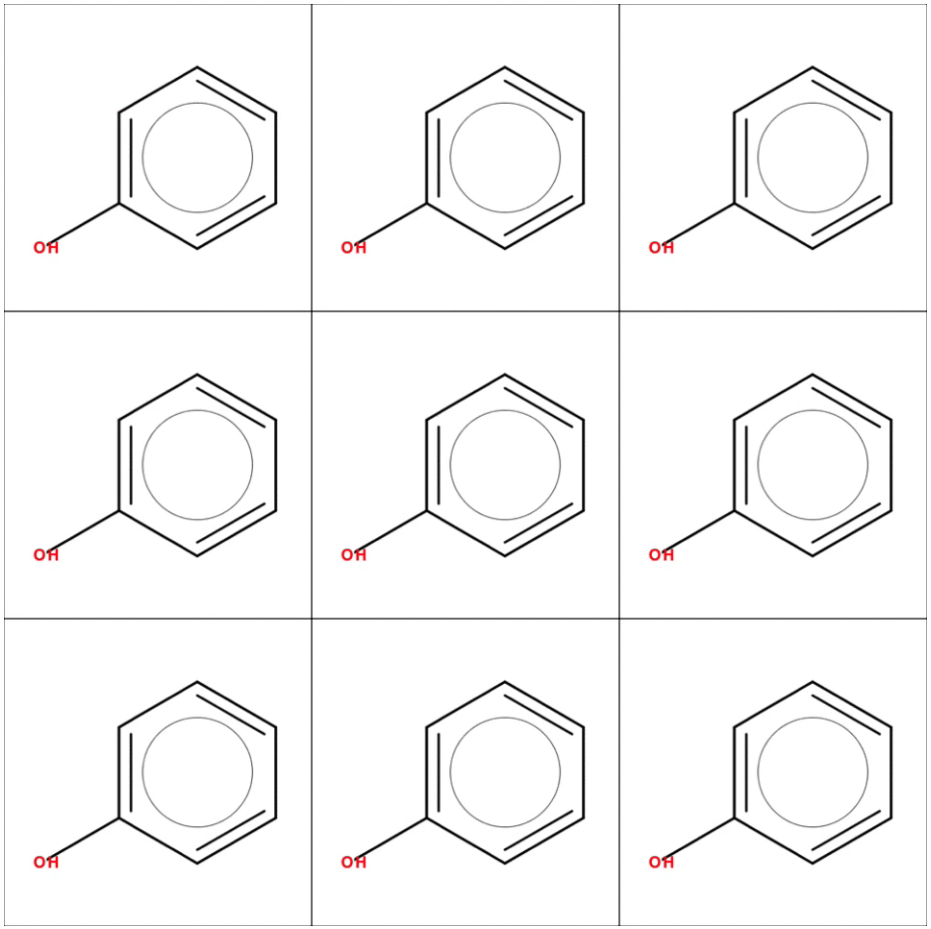
Property requirements

Band gap in [0, 3] eV
MW unconstrained

Band gap vs molecular weight
(vertical, horizontal axes)



Orange: existing molecules (PubChem)
Blue: molecules created by *iqspr*



R PACKAGE iqspr v2.3

CRAN: https://cran.r-project.org/web/packages/iqspr/index.htm
GitHub: https://github.com/GLambard/inverse-molecular-design

Lambard, G
(ISM)

- Version 1 released: 2016/05/22
- Number of downloads: 2,805 (2017/08/31)
- **Version 2.3 released: 2017/07/27**

Tutorial for the package iqspr v2.1

Guillaume Lambard

2017-05-10

Introduction

The structure of chemical species can be uniquely encoded in a single string of standard text characters called SMILES (Simplified Molecular Input Line Entry Specification). A very nice presentation of the SMILES notation can be found here. If one knows the SMILES of a chemical compound, its 2D structure can be univoquely re-constructed. One of the aspect of the SMILES format is that it is particularly useful in the prediction of properties of compounds. The link that exists between the structure of a compound and its properties is generally called a QSPR (Quantitative Structure-Properties Relationship), and it has been widely used in cheminformatics for the design of new drugs or dyes. Generally, compounds structures are primarily investigated by chemists following a trial-and-error construction controlled by their existing knowledge of the chemistry and their intuition. The properties of the investigated compounds are then checked by direct experiments and/or driven by a QSPR analysis. In this kind of analysis, numerous descriptors can be build from the SMILES format. These descriptors can be represented a set of binary and/or continuous properties based on the existence of certain fragments in a molecule, or on the ability of its bonds to rotate for example. An introduction and overview concerning the molecular descriptors can be found here. Then, the descriptors are parsed as input features for a given regression model to predict output properties for a list of novel compounds. This kind of reconstruction of the properties of compounds from descriptors is called a forward prediction.

This package is entirely devoted to the inverse problem which is the backward prediction. **The generation of entirely novel SMILES in output, and consequently chemical compounds, from input targeted properties initially constrained by a user**. Thanks to the inverse-QSPR model(via), this is now possible. This package has the ambition to become a useful tool, in the innovation of novel materials, in a field that is now widely referred as Materials Informatics. It is important to note that as SMILES is the basis format for this package, only organic molecular non-crystalline compounds can be generated.

Forward prediction methods

- Deep Learning (mxnet)
- Gradient Boosting (xgboost)

⋮

- **User-plugin models**
- **van Krevelen models**

Other additional features

- Synthetic Accessibility scoring
- A set of molecular filters
- Richer descriptor set

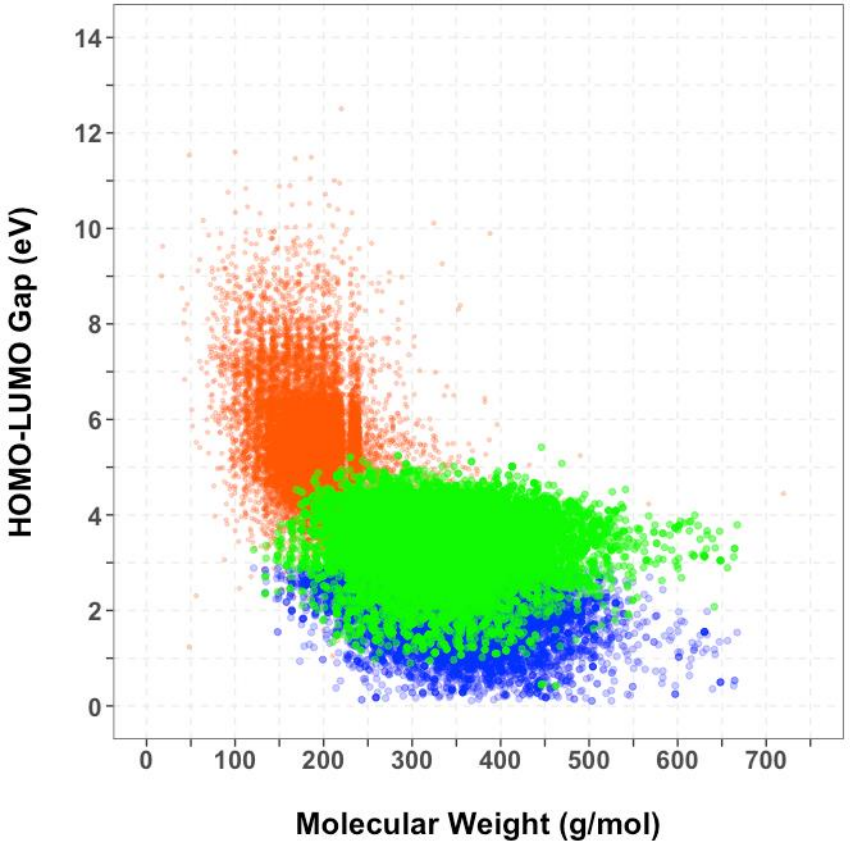
GO BEYOND INTERPOLATIVE PREDICTION

内挿的予測の限界を超える

計算機実験による検証

FIRST PRINCIPLES CALCULATION (DFT)

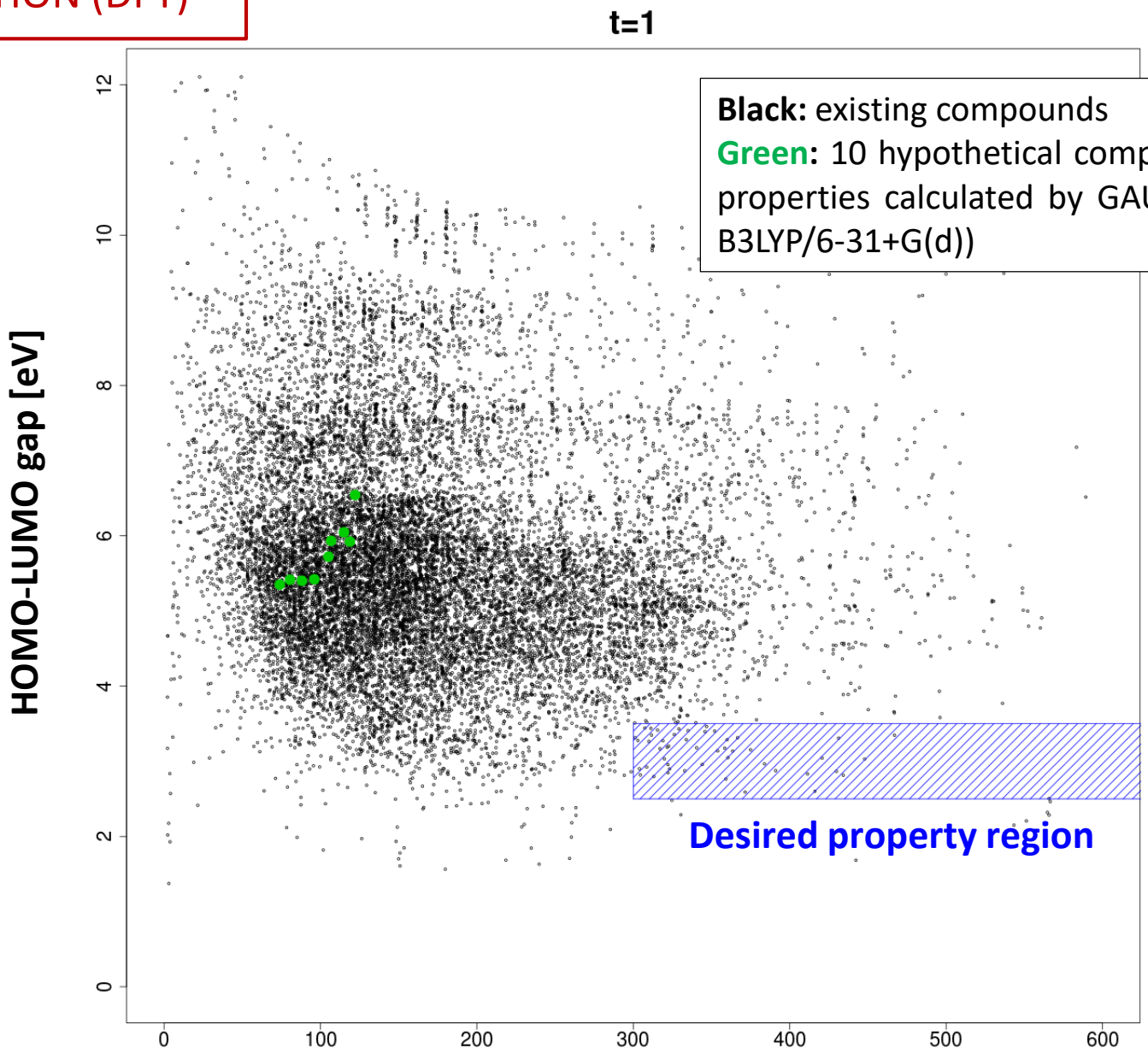
- Identification of promising materials based on DFT calculation (filtering)
- Poor prediction of the machine learning models in regions with few data points



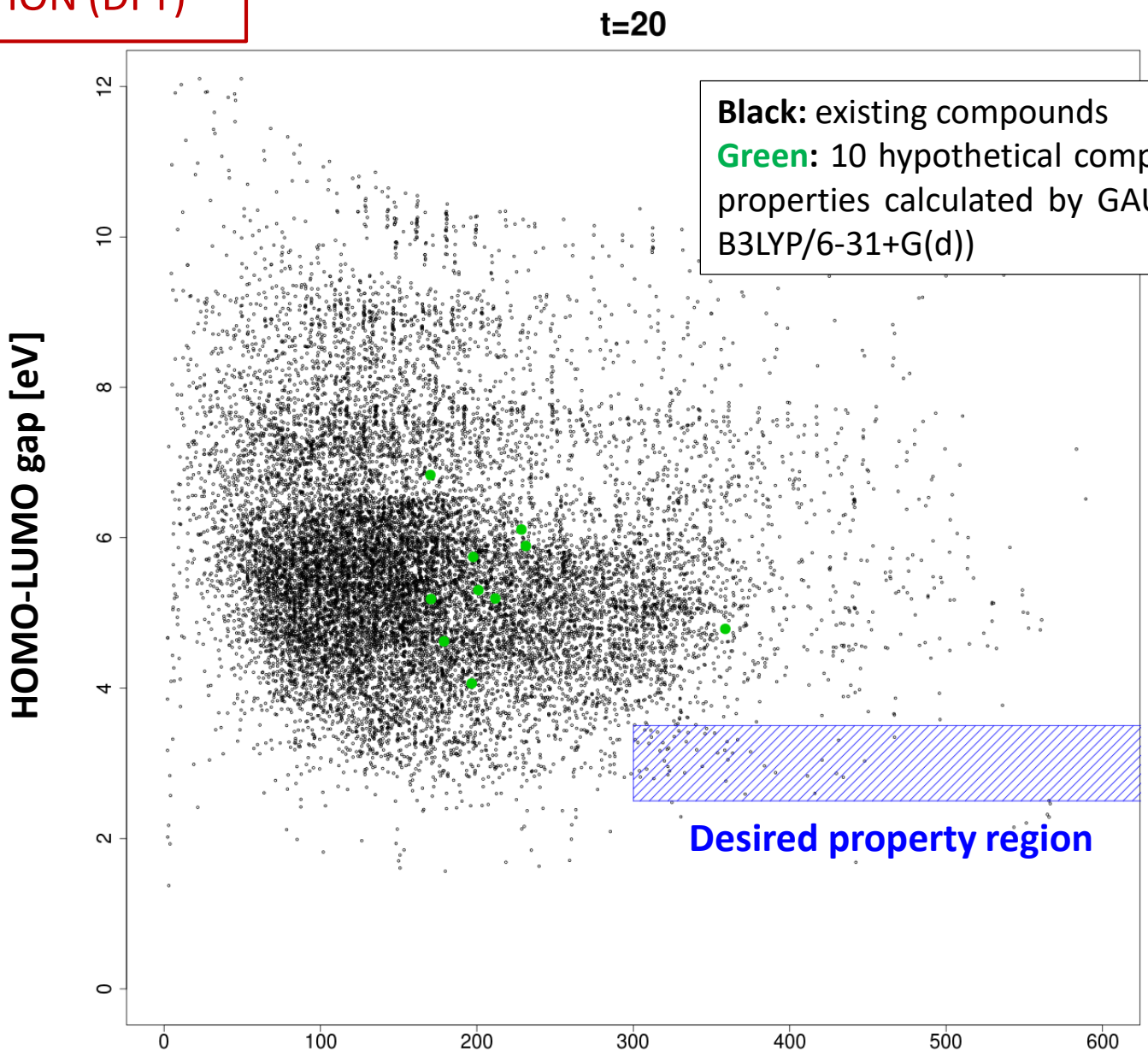
DFT scheme
B3LYP functional, 6-31G basis set

Orange: existing molecules (DFT)
Blue: hypothetical (ML prediction)
Green: hypothetical (DFT)

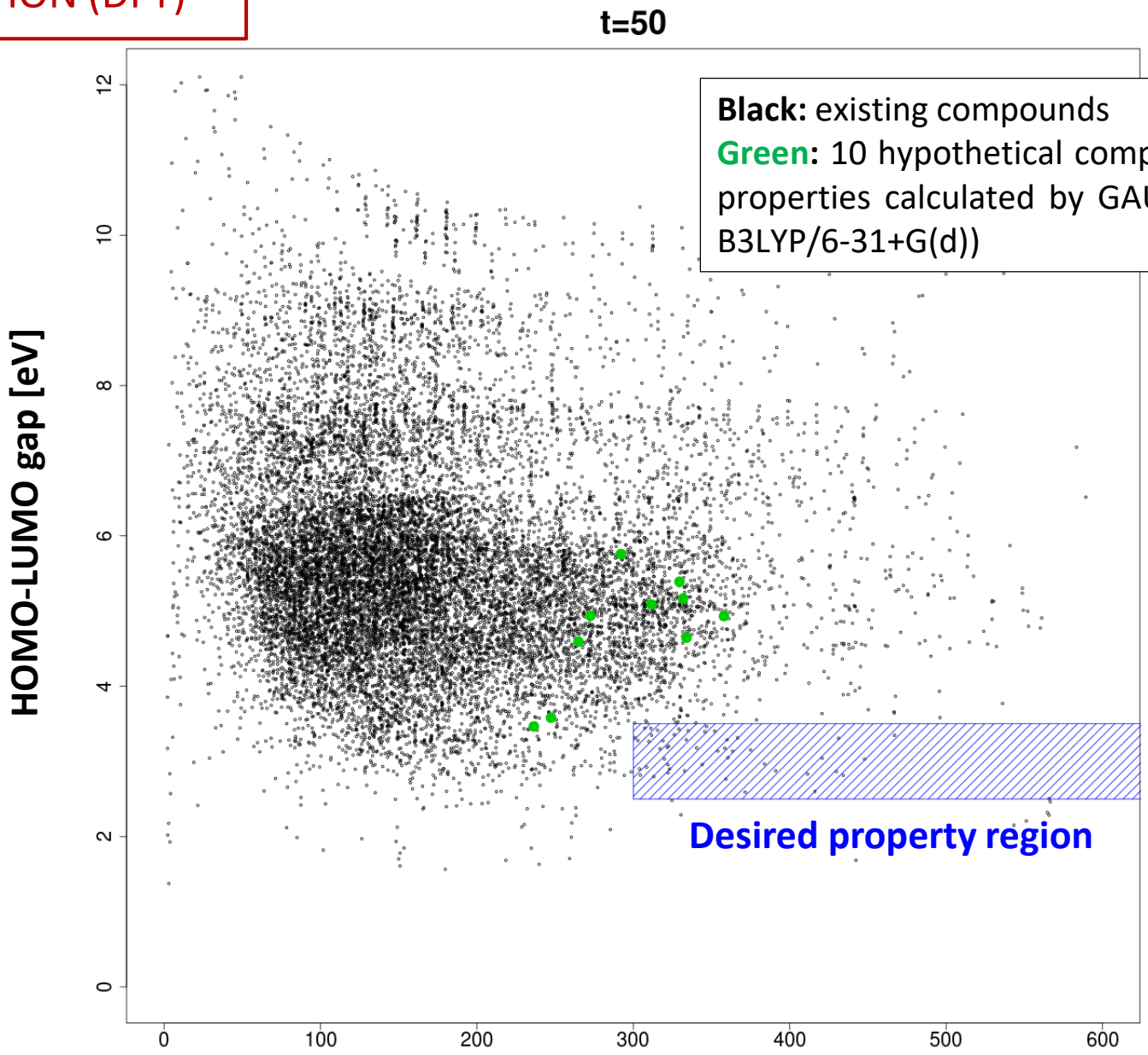
VALIDATION (DFT)



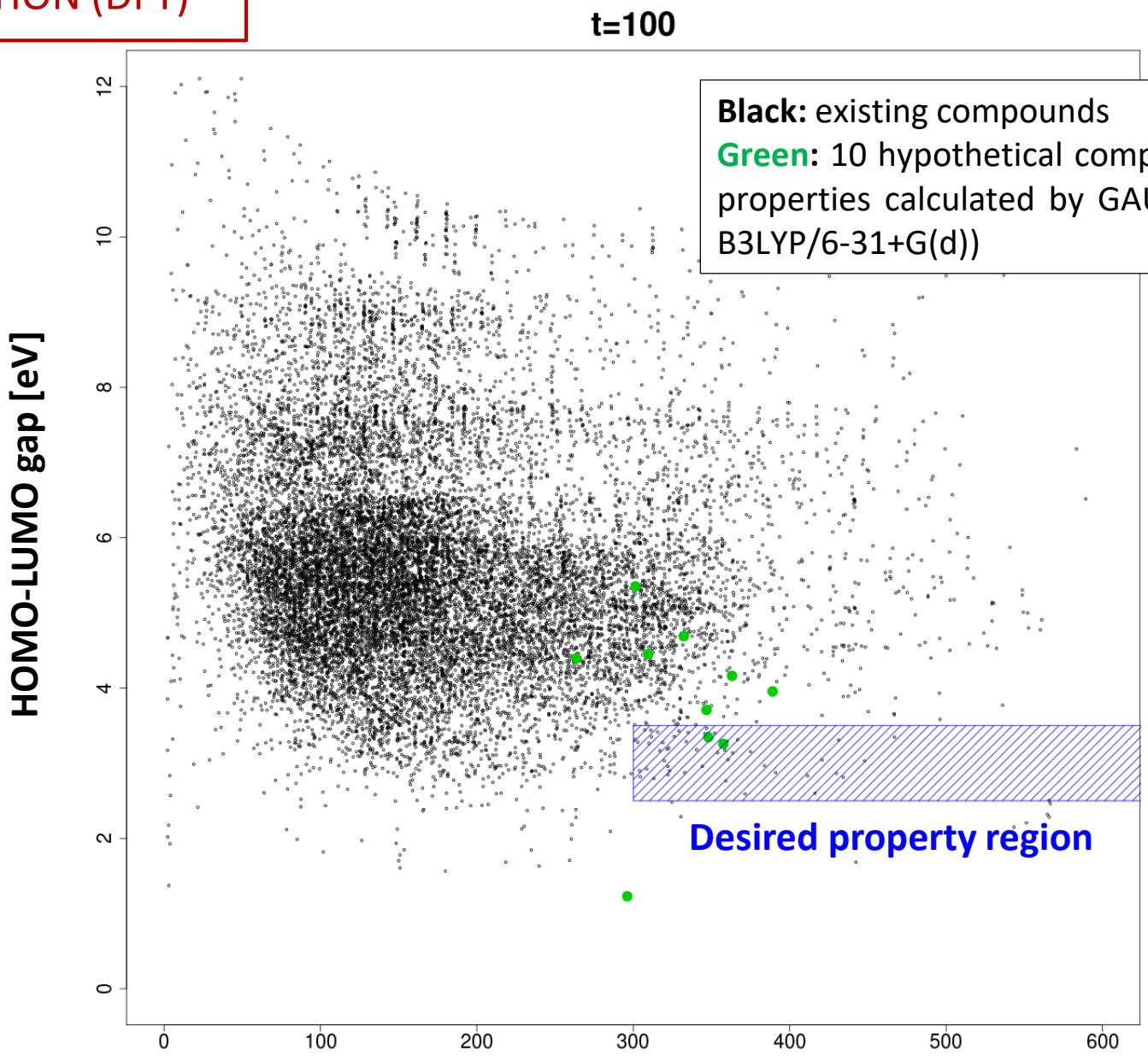
VALIDATION (DFT)



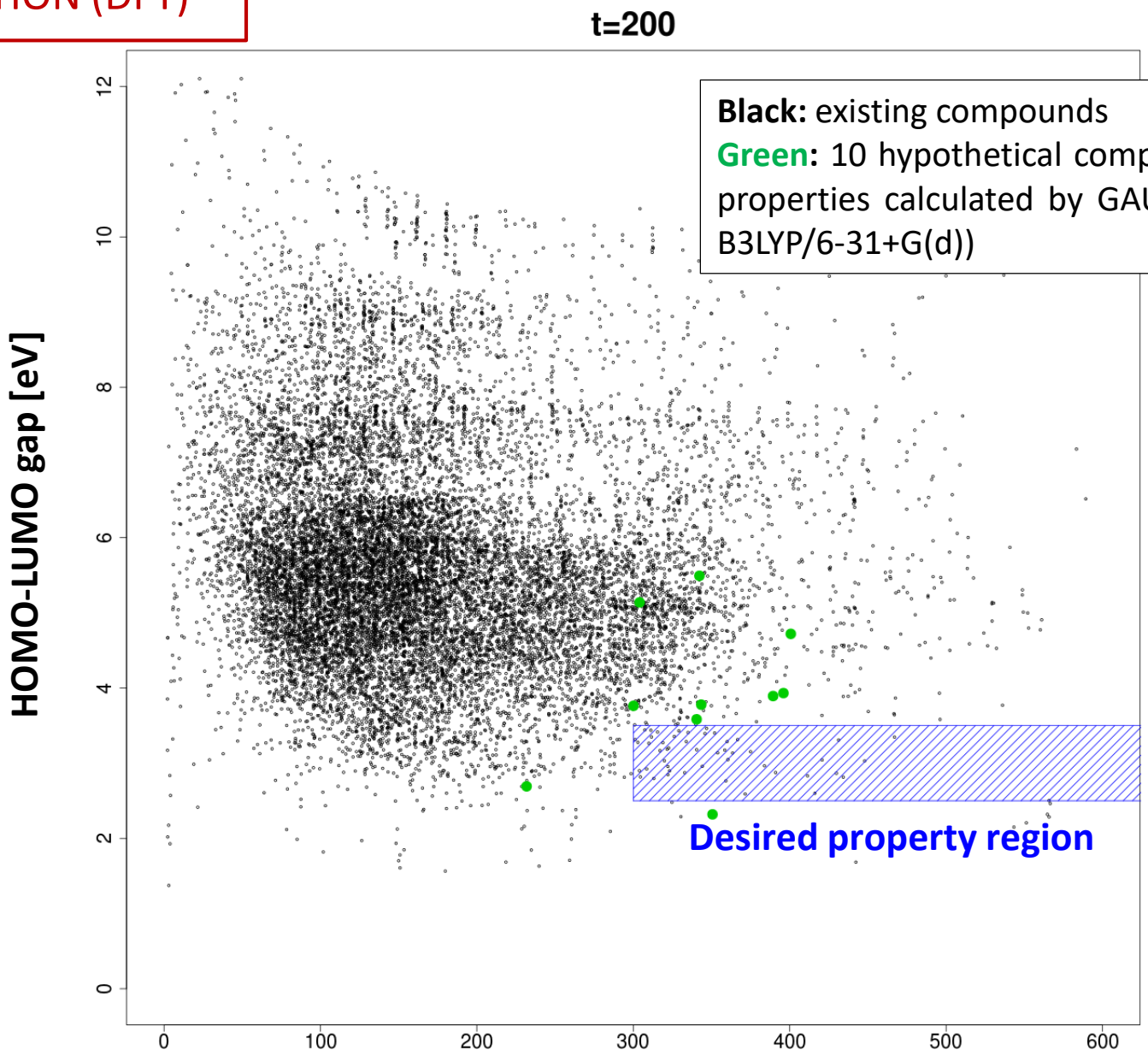
VALIDATION (DFT)



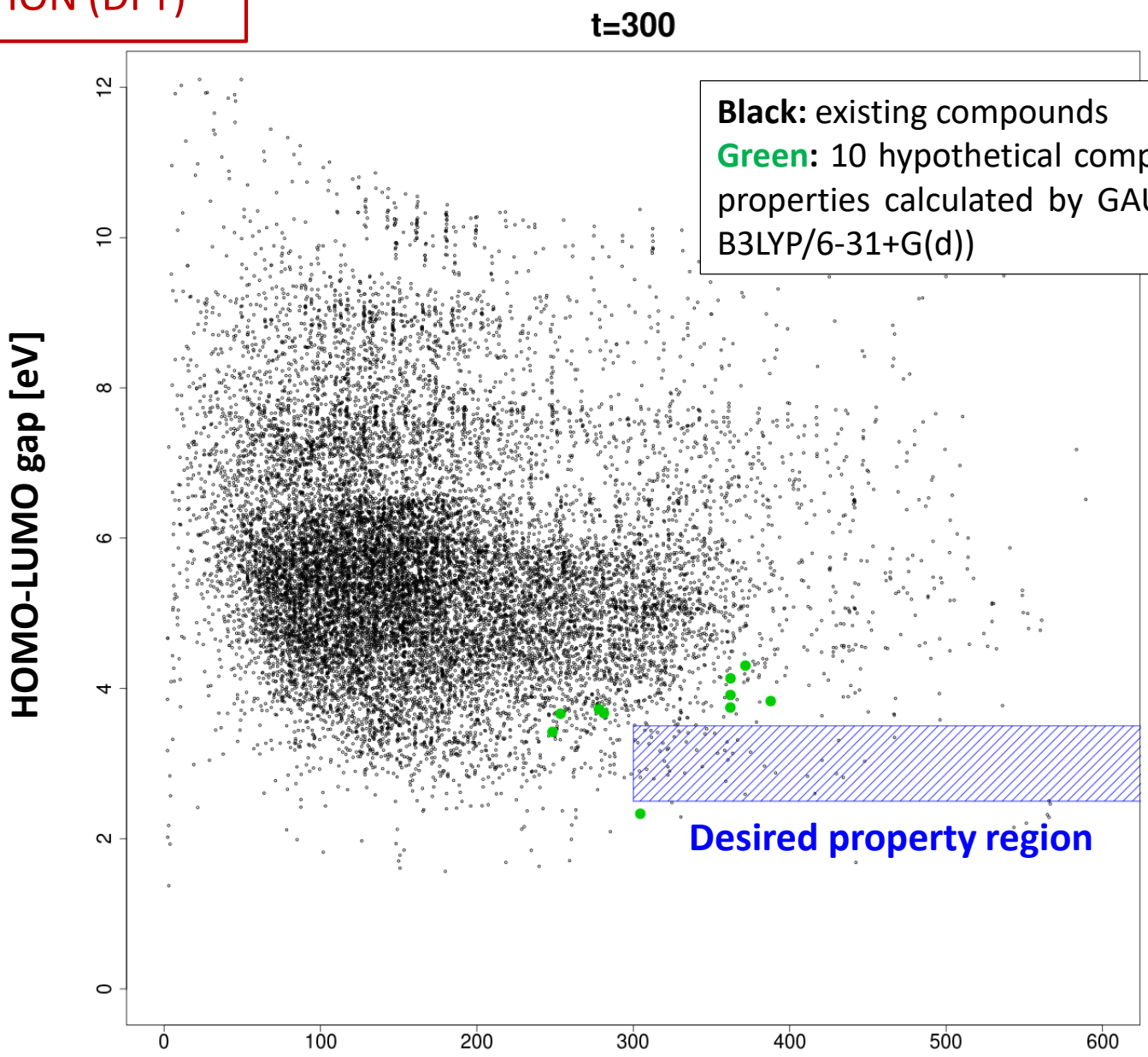
VALIDATION (DFT)



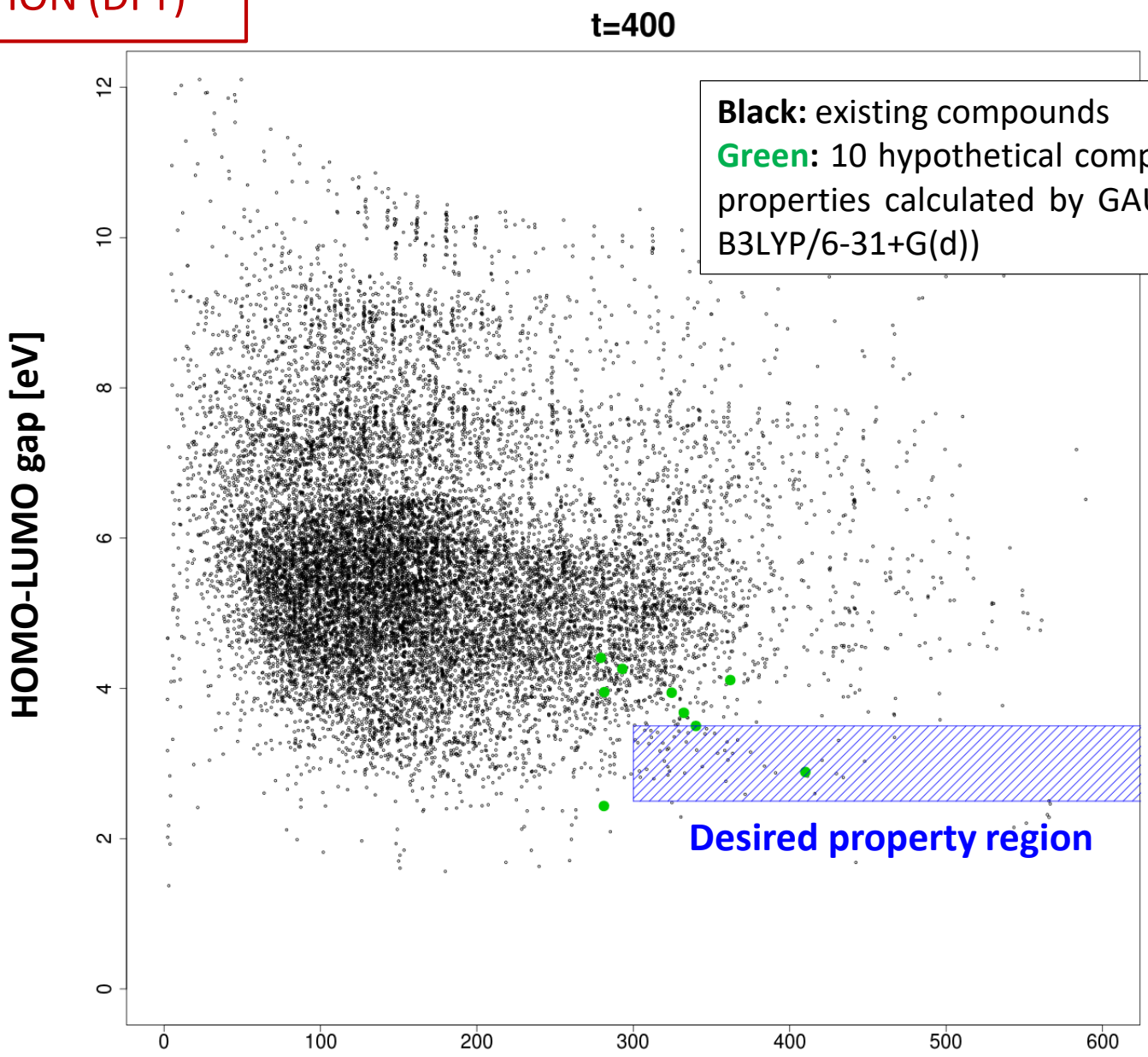
VALIDATION (DFT)



VALIDATION (DFT)

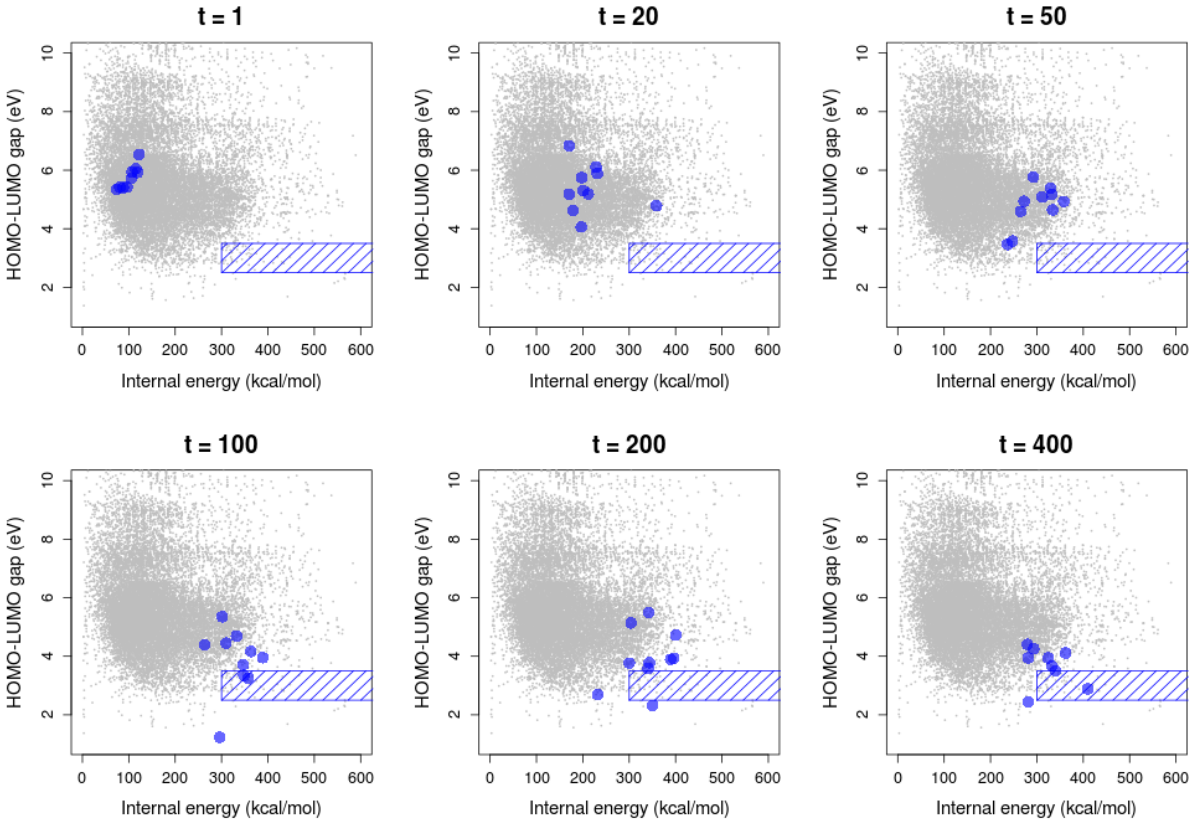


VALIDATION (DFT)



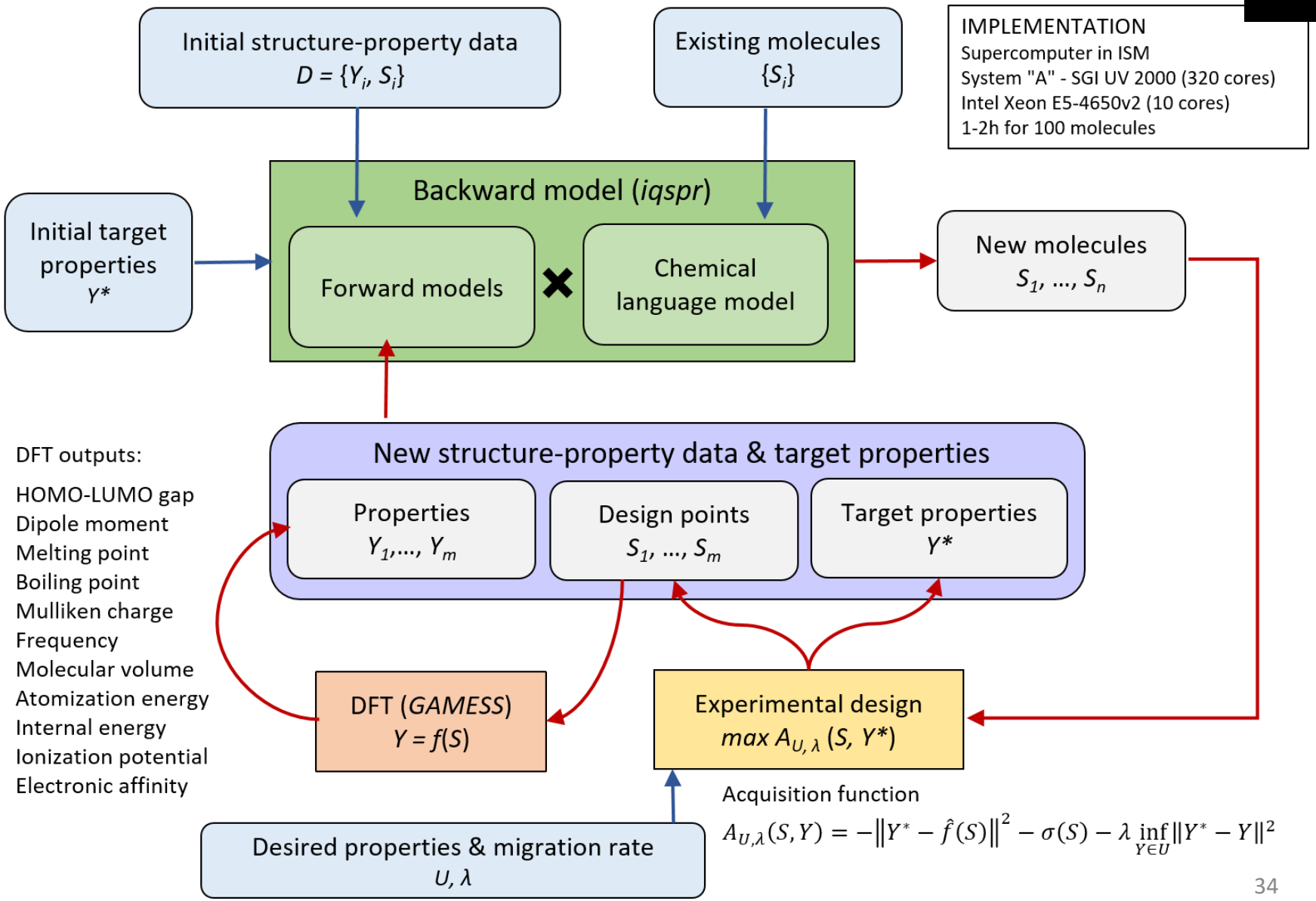
GO BEYOND INTERPOLATIVE PREDICTION

- Limitation of *iqspr*: unreachable to regions where data are sparsely populated (*i.e.*, interpolative prediction)
- Extrapolative prediction by the integration of *iqspr* and DFT (computer experiments)



Vertical axis: HOMO-LUMO gap
Horizontal axis: internal energy
Gray dot: existing molecules
Black dot: newly created molecules

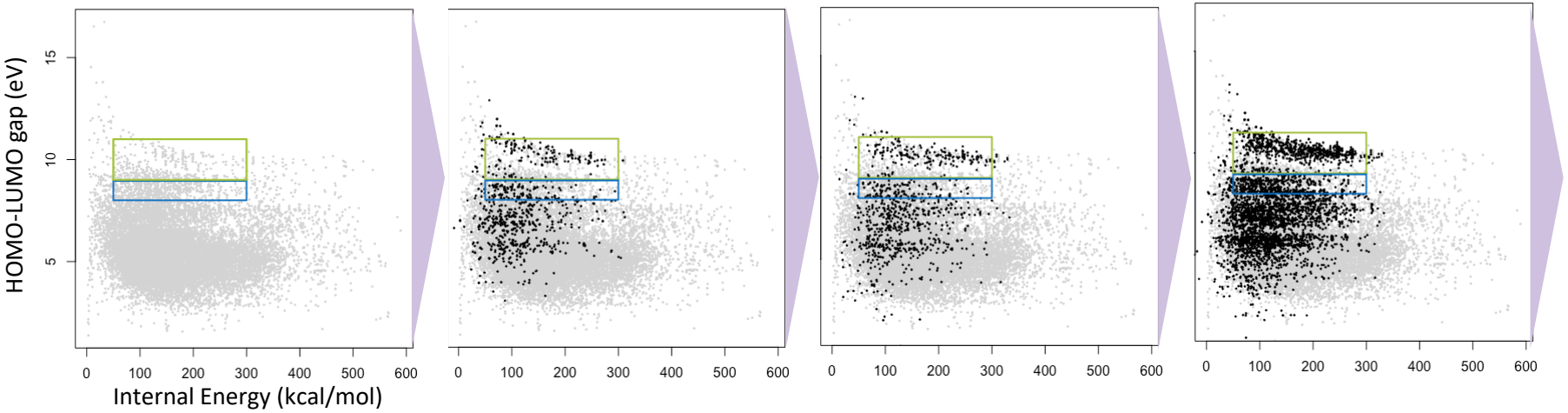
SPACIER Go Beyond Interpolative Prediction



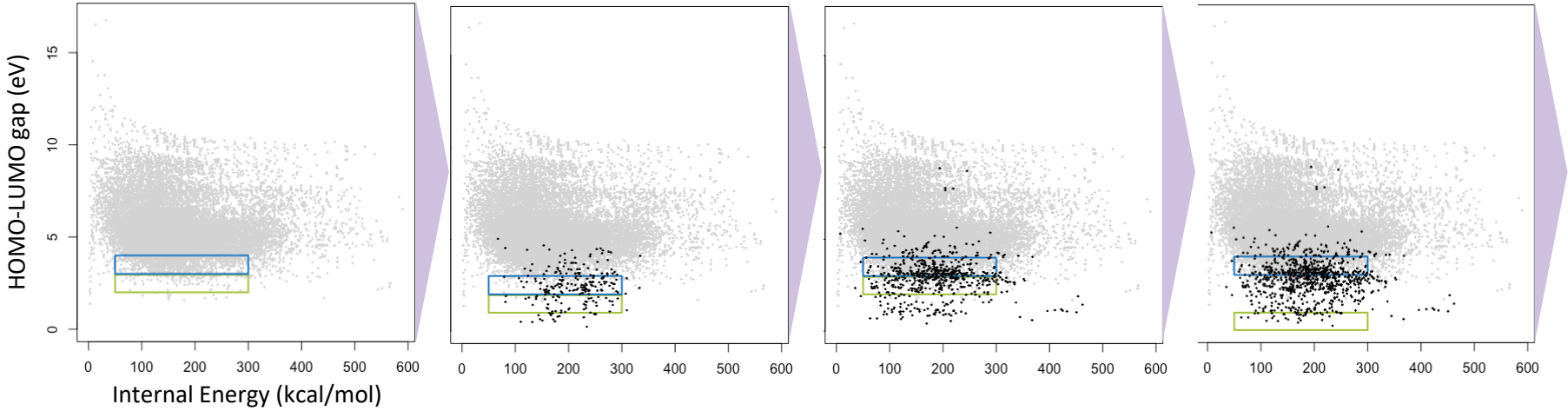
SPACIER Go Beyond Interpolative Prediction

Vertical axis: HOMO-LUMO gap
Horizontal axis: internal energy
Gray dot: existing molecules
Black dot: newly created molecules

Higher HOMO-LUMO gap

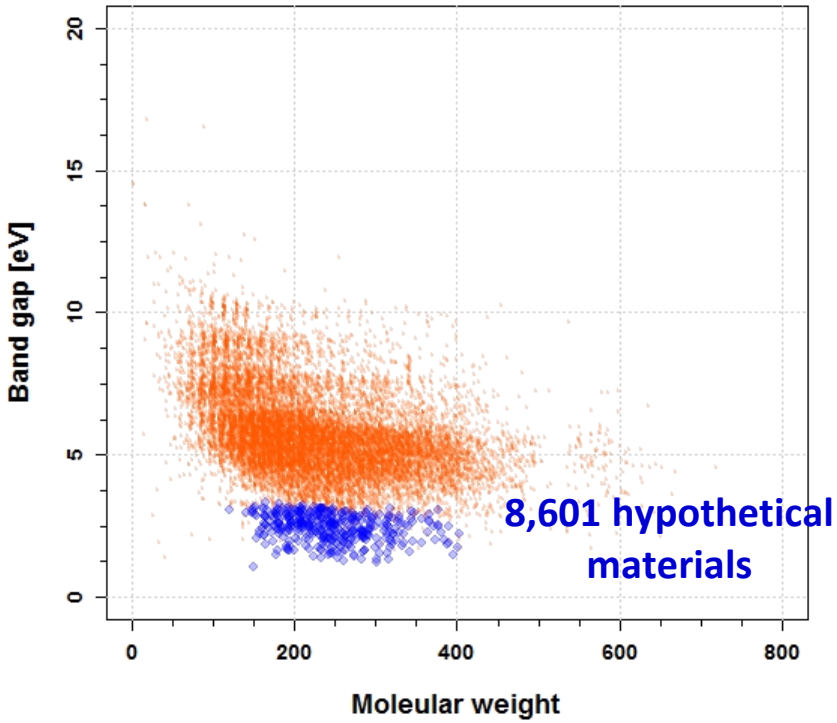


Lower HOMO-LUMO gap



SPACIER 低バンドギャップモノマーの発掘

学術的意義：物質・材料科学の事前知識を全く活用しない完全なデータ科学駆動型手法でここまで到達できた.



データが存在しない低バンドギャップ領域に埋蔵する大量の新物質が発掘された.

Orange: DFT properties of existing materials
Blue: newly discovered materials (SPACIER)

Synthesis
Experimental validation
in the next 1-2 years

材料研究と実験計画法

- **材料研究では，実験（実空間・仮想空間）とデータ科学を連動させることが極めて重要**
- **目的：① 実験コストの抑制，②外挿こそが究極の目標**

実験が済んだあとに統計学者に意見を求めても、大抵の場合、それは単に終わったことを議論しようと言っているようなものだ。もしかしたら統計学者は、失敗の原因を語り出すかもしれない。

Fisher R.A. (1938) Presidential Address, *Sankhyā,* 4:14-17.

R. A. フィッシャー (1890-1962)

現代の統計学，集団遺伝学の創始者．ロザムステッド農事試験場における穀物量の変動に関する研究を通して，実験計画法や分散分析等，数々の統計理論を確立．

物性計算（第一原理計算など）の試行回数を削減

物質・材料科学におけるベイズ最適化の活用

Seko et al., *Phys Rev B*, 2014
Property: melting point
Input: crystal structure

Seko et al., *Phys Rev Lett*, 2015
Property: LTC (lattice thermal conductivity)
Input: crystal structure

Balachandran et al., *Sci Rep*, 2016
Property: elastic properties [bulk, shear, and Young's modulus]
Input: crystal structure

Kiyohara et al., *Sci Adv*, 2016
Property: grain boundary energy
Input: interface structure (grain boundary)

Ju et al., *arXiv*, arXiv:1609.04972
Property: interfacial thermal conductivity
Input: interface structure

[SOFTWARE] Ueno et al., COMBO: an efficient Bayesian Optimization library for materials science, *Mater Dis*, 2016

ベイズ最適化に基づく実空間の合成・機能計測の循環

4×9回の実験サイクル 新規形状記憶合金を同定

ARTICLE

Received 2 Sep 2015 | Accepted 4 Mar 2016 | Published 15 Apr 2016

DOI: 10.1038/ncomms11241 OPEN

Accelerated search for materials with targeted properties by adaptive design

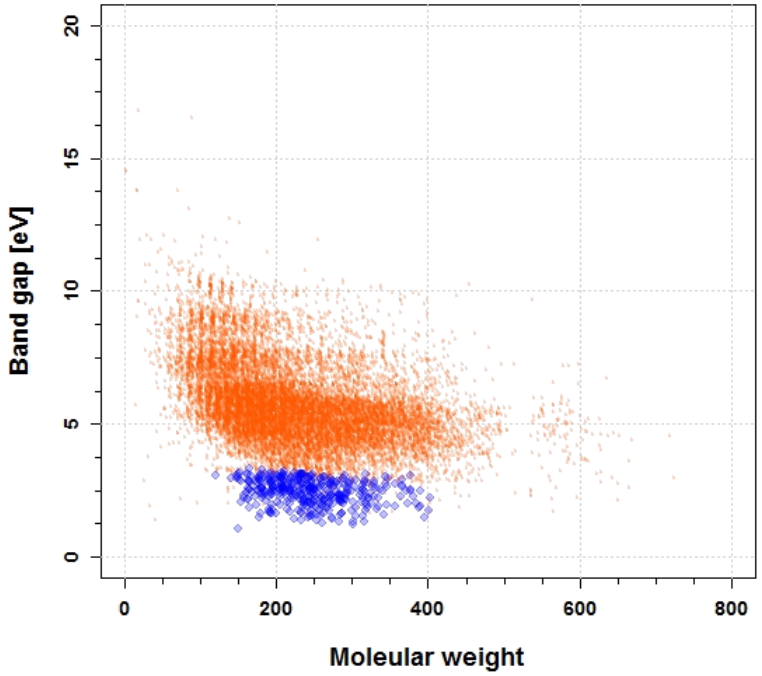
Dezhen Xue[1,2], Prasanna V. Balachandran[1], John Hogden[3], James Theiler[4], Deqing Xue[2] & Turab Lookman[1]

Finding new materials with targeted properties has traditionally been guided by intuition, and trial and error. With increasing chemical complexity, the combinatorial possibilities are too large for an Edisonian approach to be practical. Here we show how an adaptive design strategy, tightly coupled with experiments, can accelerate the discovery process by sequentially identifying the next experiments or calculations, to effectively navigate the complex search space. Our strategy uses inference and global optimization to balance the trade-off between exploitation and exploration of the search space. We demonstrate this by finding very low thermal hysteresis (ΔT) NiTi-based shape memory alloys, with $Ti_{50.0}Ni_{46.7}Cu_{0.8}Fe_{2.3}Pd_{0.2}$ possessing the smallest ΔT (1.84 K). We synthesize and characterize 36 predicted compositions (9 feedback loops) from a potential space of ~800,000 compositions. Of these, 14 had smaller ΔT than any of the 22 in the original data set.

外挿アルゴリズムの到来：何が変わるか

著作権・知財の在り方

- アルゴリズム（& データ）さえあれば，誰もが新材料を発掘できる時代？
- 遥かに高い生産性を有する（弱い）AI による著作物の数が爆発的に増加
- 独占の不可：全候補物質の合成は不可能．同じ計算をすれば，第三者も他の物質を発掘できる．データやアルゴリズムのオープン化は危険かもしれない．
- 特許の網目を抜ける技術としても（*e.g.* A社とB社の特許化合物のハイブリッド）



例：SPACIERによる低バンドギャップポリマーの探索．
データが存在しない低バンドギャップ領域に埋蔵する大量の新物質が発掘された．

Orange: DFT properties of existing materials
Blue: newly discovered materials (SPACIER)