ADDOPT

ADVANCED DIGITAL DESIGN OF PHARMACEUTICAL THERAPEUTICS

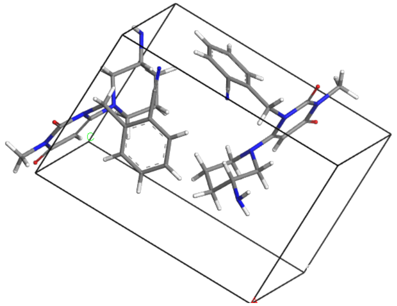
# De-risking early stage drug development : A big data approach to address lattice energy prediction challenges associated with a diverse chemical space

ADDoPT Dissemination Event 28th March 2019

**Rebecca Mackenzie,** Dawn Geatches, Chris Morris (Hartree), Richard Marchese Robinson (Leeds), Andy Maloney (CCDC), Bob Docherty, Klimentina Pencheva, Ernest Chow (Pfizer), Colin Edge (GSK)

# Why lattice energy is important for the pharma industry?

The lattice energy of the selected physical form of the drug dictates physico-chemical properties (e.g. stability, solubility, process-ability), and is important for the understanding of the thermodynamic relationships.



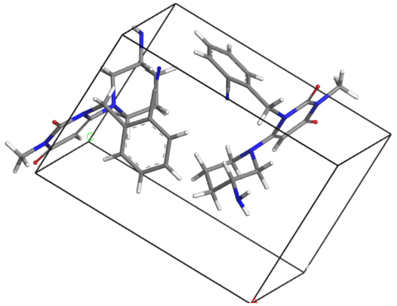
There are different ways to quantify this:

- Sublimation enthalpy ($\Delta H_{sub}$) - Energy required to break the packing lattice (tend to be used by experimental scientists)
- Lattice energy ($E_{latt}$) – Energy released upon formation of crystal packing arrangement (tend to be used by computational scientists)
- The approximate relationship between the two is expressed by:

$$\Delta H_{sub} \approx -E_{latt} - 2RT$$

# Why lattice energy is important for the pharma industry?

The lattice energy of the selected physical form of the drug dictates physico-chemical properties (e.g. stability, solubility, process-ability), and is important for the understanding of the thermodynamic relationships.



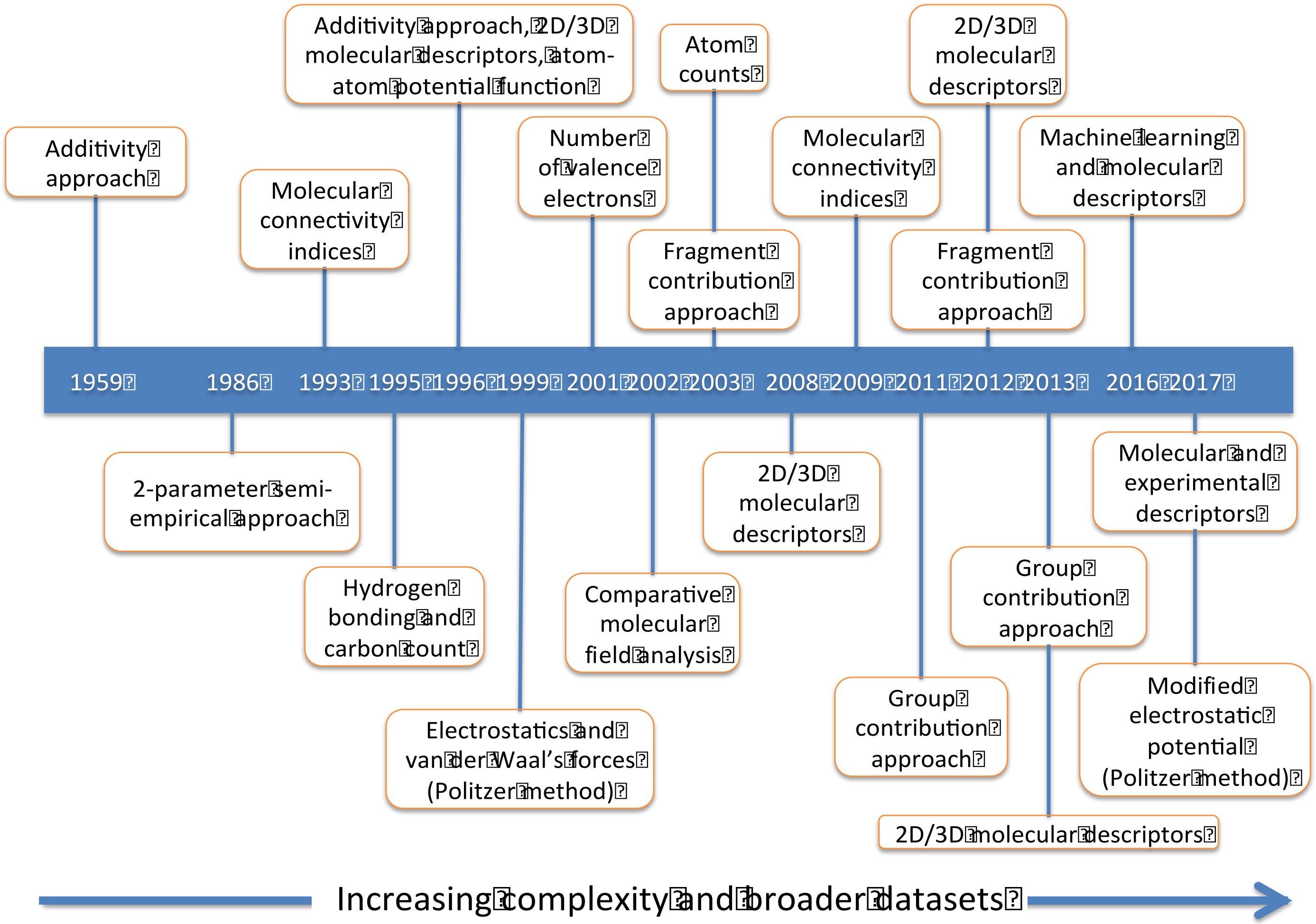
There are different ways to quantify this:

- Sublimation enthalpy ($\Delta H_{sub}$) - Energy required to break the packing lattice (tend to be used by experimental scientists)
- Lattice energy ($E_{latt}$) – Energy released upon formation of crystal packing arrangement (tend to be used by computational scientists)
- The approximate relationship between the two is expressed by:

$$\Delta H_{sub} \approx -E_{latt} - 2RT$$

**In this work we demonstrate the power of big data and machine learning driven by cross community efforts.**

# Timeline of different approaches used to build packing energy prediction models



Additivity approach

Molecular connectivity indices

Additivity approach, 2D/3D molecular descriptors, atom-atom potential function

Number of valence electrons

Atom counts

Fragment contribution approach

Molecular connectivity indices

2D/3D molecular descriptors

Fragment contribution approach

Machine learning and molecular descriptors

1959 1986 1993 1995 1996 1999 2001 2002 2003 2008 2009 2011 2012 2013 2016 2017

2-parameter semi-empirical approach

Hydrogen bonding and carbon count

Electrostatics and van der Waal's forces (Politzer method)

Comparative molecular field analysis

2D/3D molecular descriptors

Group contribution approach

Group contribution approach

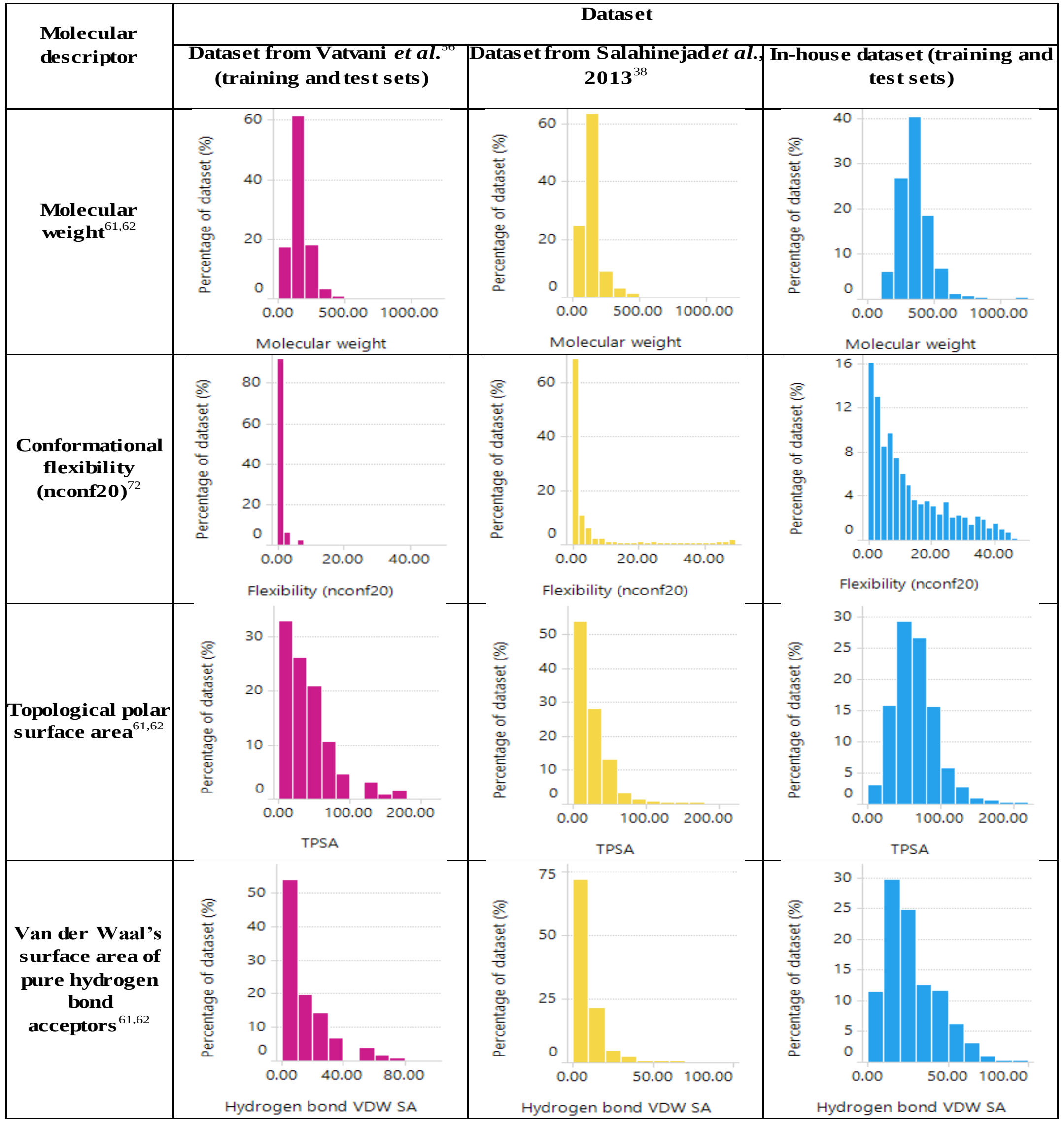
2D/3D molecular descriptors

Molecular and experimental descriptors

Modified electrostatic potential (Politzer method)

Increasing complexity and broader datasets

# Comparison of selected descriptors cross data-sets



| Molecular descriptor | Dataset | | |
|---|---|---|---|
| | Dataset from Vatvani *et al.*[36] (training and test sets) | Dataset from Salahinejad *et al.*, 2013[38] | In-house dataset (training and test sets) |
| Molecular weight[61,62] | Molecular weight | Molecular weight | Molecular weight |
| Conformational flexibility (nconf20)[72] | Flexibility (nconf20) | Flexibility (nconf20) | Flexibility (nconf20) |
| Topological polar surface area[61,62] | TPSA | TPSA | TPSA |
| Van der Waal's surface area of pure hydrogen bond acceptors[61,62] | Hydrogen bond VDW SA | Hydrogen bond VDW SA | Hydrogen bond VDW SA |

- Problem: data-sets do not provide enough coverage of strong packing crystals which are typical for drug molecules.

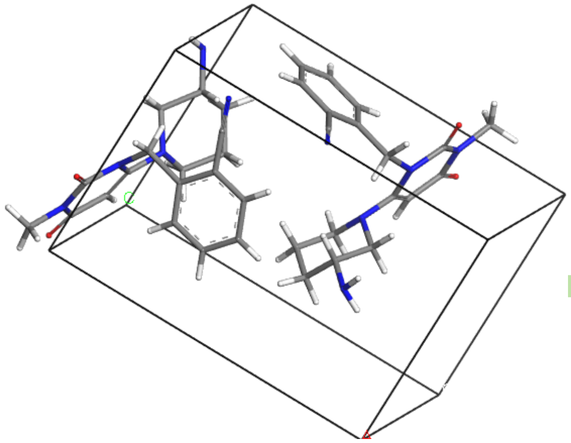
- Limits the predictability and therefore applicability of the model.

- Histograms highlight the differences in key molecular descriptors for selected data-sets.

# Objectives

Utilise big data and statistical approaches to predict lattice energy using 2D molecular information only

Crystal structure database



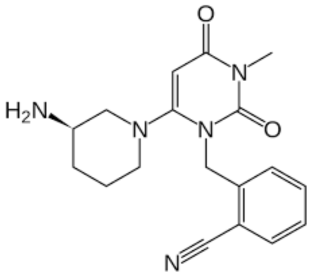
Lattice energy calculation

| Mol | Energy |
|---|---|
| A | 100 |
| B | 200 |
| C | 300 |
| D | 200 |

Regression model

Key: Atomistic Modelling QSPR workflow

2D structure



Molecular descriptor

Descriptors associated with molecular structure

| Mol | MW | PSA |
|---|---|---|
| A | 300 | 110 |
| B | 350 | 120 |
| C | 400 | 100 |
| D | 350 | 150 |

# Data collection

*For atomistic modelling*

Enthalpies of sublimation

$$\Delta H_{sub} \approx -E_{latt} - 2RT$$

- Literature paper sources
- National institute of standards and technology (NIST) database
- 428 sublimation enthalpies at a known temperature, linked to 256 Cambridge Structural Database (CSD) entries – each corresponding to a unique molecule.

*For QSPR*

Calculated lattice energies

- From single crystal structures (generic and industrial)
- 60,000 organic molecules crystal structures from CSD
- 1,500 Pfizer internal data-set

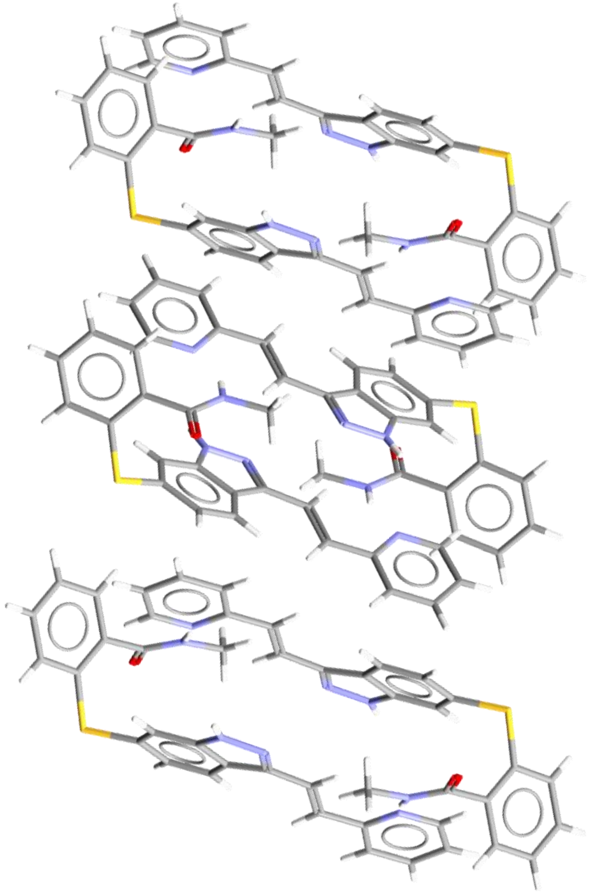
# Data collection

*For atomistic modelling*

Enthalpies of sublimation

*For QSPR*

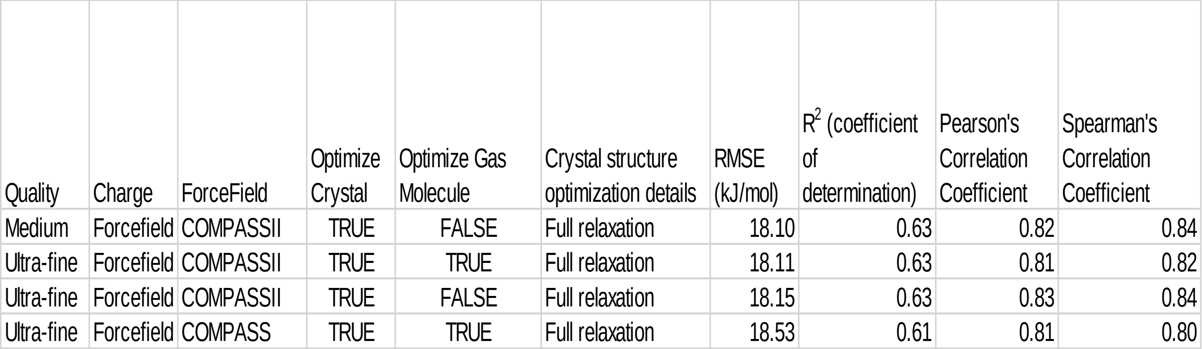
Calculated lattice energies



**Highlights the power of uniting resources at the Hartree Centre from across the community (industry, academics, subject matter experts)**

# Atomistic Modelling

1. Using the crystal structure database together with enthalpies of sublimations (N = 256)
2. Applied a variety of atomistic models to calculate lattice energy (force-fields including COMPASS II force-field; Density Functional Theory etc.)
3. Benchmarked these methods (324) to obtain best performing atomistic model.



| Quality | Charge | ForceField | Optimize Crystal | Optimize Gas Molecule | Crystal structure optimization details | RMSE (kJ/mol) | $R^2$ (coefficient of determination) | Pearson's Correlation Coefficient | Spearman's Correlation Coefficient |
|---|---|---|---|---|---|---|---|---|---|
| Medium | Forcefield | COMPASSII | TRUE | FALSE | Full relaxation | 18.10 | 0.63 | 0.82 | 0.84 |
| Ultra-fine | Forcefield | COMPASSII | TRUE | TRUE | Full relaxation | 18.11 | 0.63 | 0.81 | 0.82 |
| Ultra-fine | Forcefield | COMPASSII | TRUE | FALSE | Full relaxation | 18.15 | 0.63 | 0.83 | 0.84 |
| Ultra-fine | Forcefield | COMPASS | TRUE | TRUE | Full relaxation | 18.53 | 0.61 | 0.81 | 0.80 |

# QSPR – Machine Learning

| Mol | MW | PSA | Energy |
|---|---|---|---|
| A | 300 | 110 | 100 |
| B | 350 | 120 | 200 |
| C | 400 | 100 | 300 |
| D | 350 | 150 | 200 |

Use calculated lattice energies dataset (N = 60,000) together with descriptors

→

| Mol | Predicted Energy |
|---|---|
| A | 100 |
| B | 200 |
| C | 300 |
| D | 200 |

Create and compare a series of statistical regression models to predict calculated lattice energy.

# Machine learning methods

Linear Regression

Nearest Neighbour

AdaBoosted Tree

Decision Tree

Random Forest

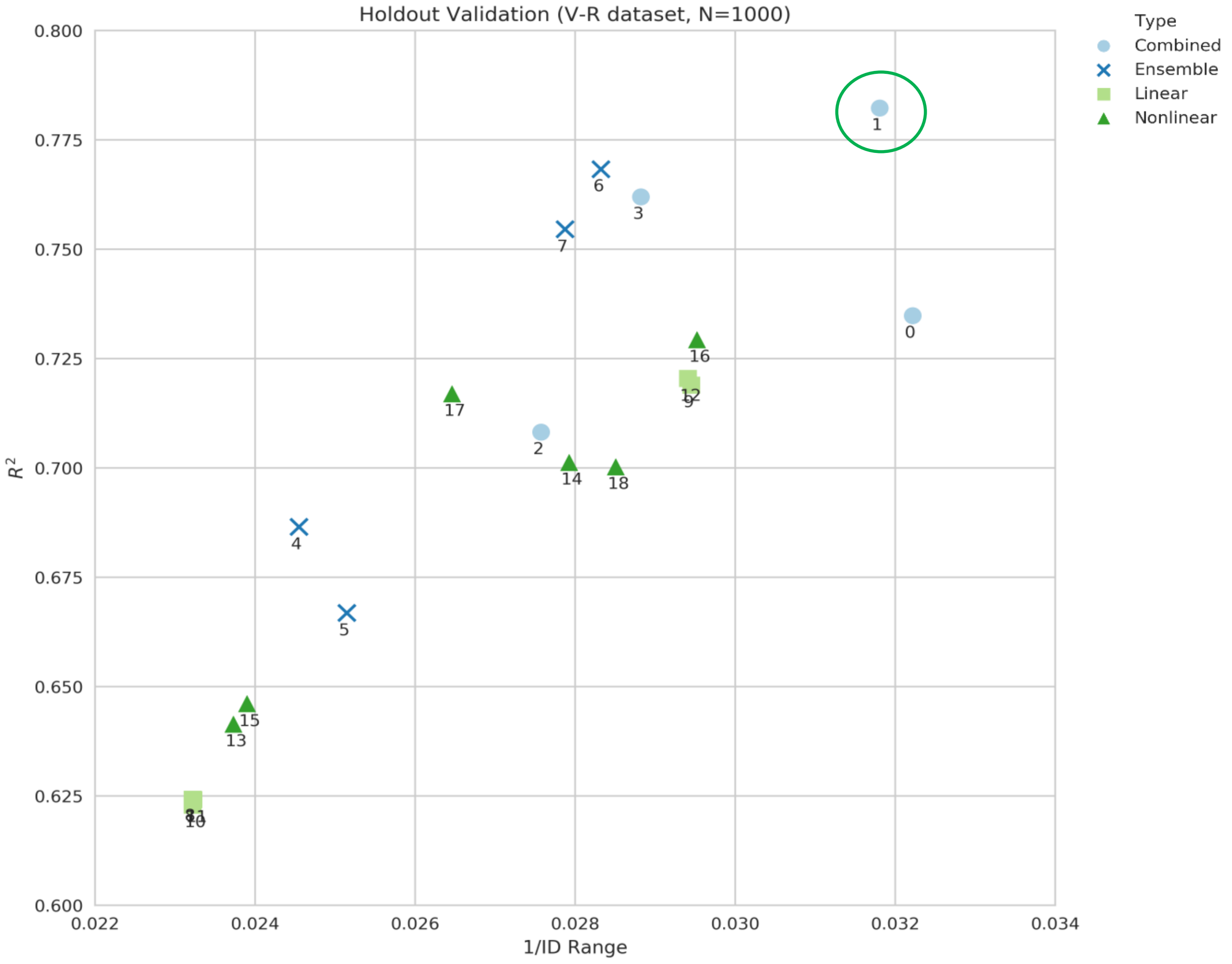
Robust Regression

Kernel – using fingerprints

Gradient Boosted Regression Tree

Combined models

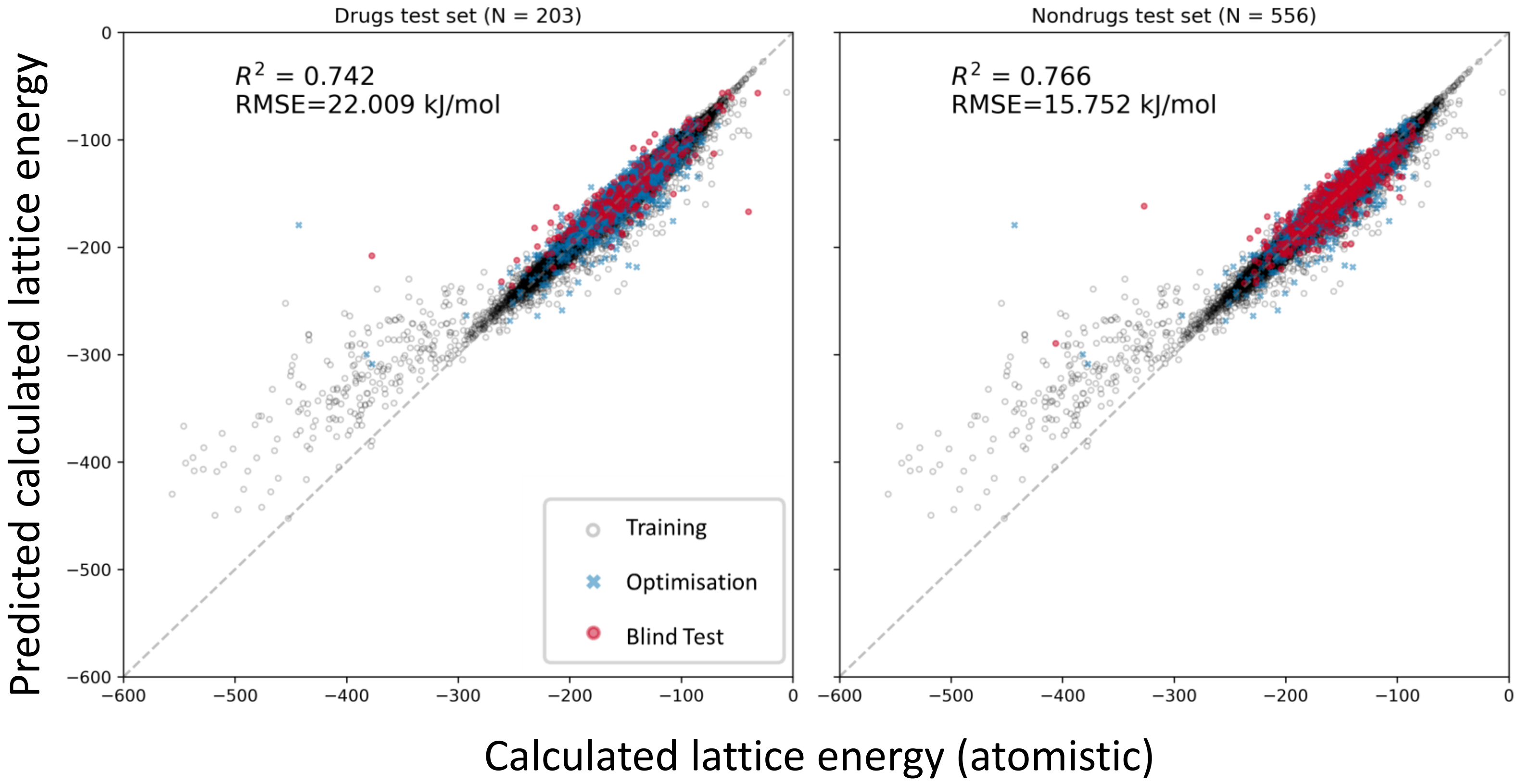
- Best performing model: $R^2 = 0.782$; RMSE = 16.545 kJ mol$^{-1}$



1-18: Model #

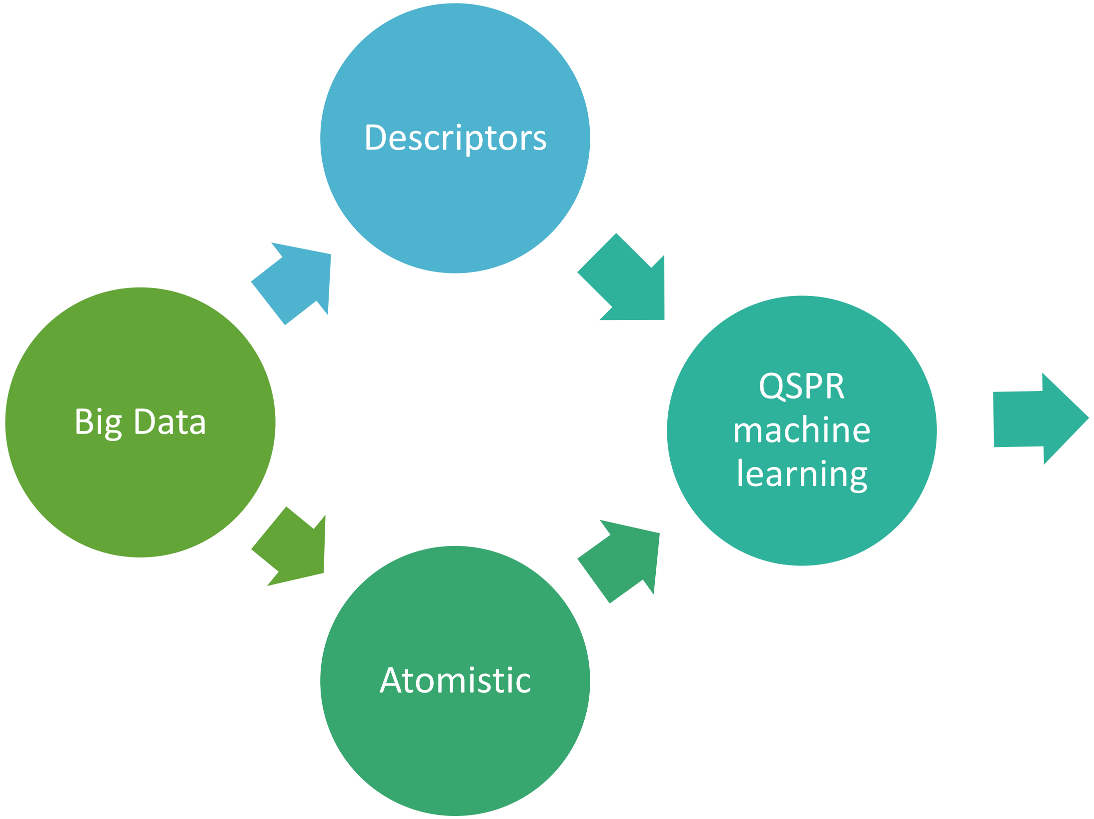
# ML model results

The best model is able to predict lattice energies for current known drug (from DrugBank ID) and nondrug molecules.



Model appears to deviate for low lattice energy, however this is only apparent for training. There is no test data for this at present.

# Summary



**Industrial Impact**

Predictive workflow for understanding process-ability of the potential drug product during the development stage

# Future work

The application of big data and predictive sciences to streamline pharmaceutical development as exemplified by these efforts confirms the growing momentum in this area.

There are a number of options available cross-industry for further development of this work:

- Enrich the training set with a greater quantity and diversity of molecules would be valuable to improve the accuracy and range of applicability of our ADDoPT model.
- Translate the workflow method and apply to industry specific properties.

**Thank you for listening.**