Imputation of Sensory Properties Using Deep Learning

20th March 2022

Dmitriy Chekmarev (IFF) and Samar Mahmoud (Optibrium)

Overview

- Introduction to sensory properties
  - Data used in this project
- The Alchemite™ method for deep learning imputation
- Results
- Conclusions

Journal of Computer-Aided Molecular Design (2021) 35, pp 1125–1140
doi.org/10.1007/s10822-021-00424-3; https://bit.ly/SensoryImputation

Introduction to Sensory Properties and Data Set

Introduction to Sensory Properties (Olfactive properties)

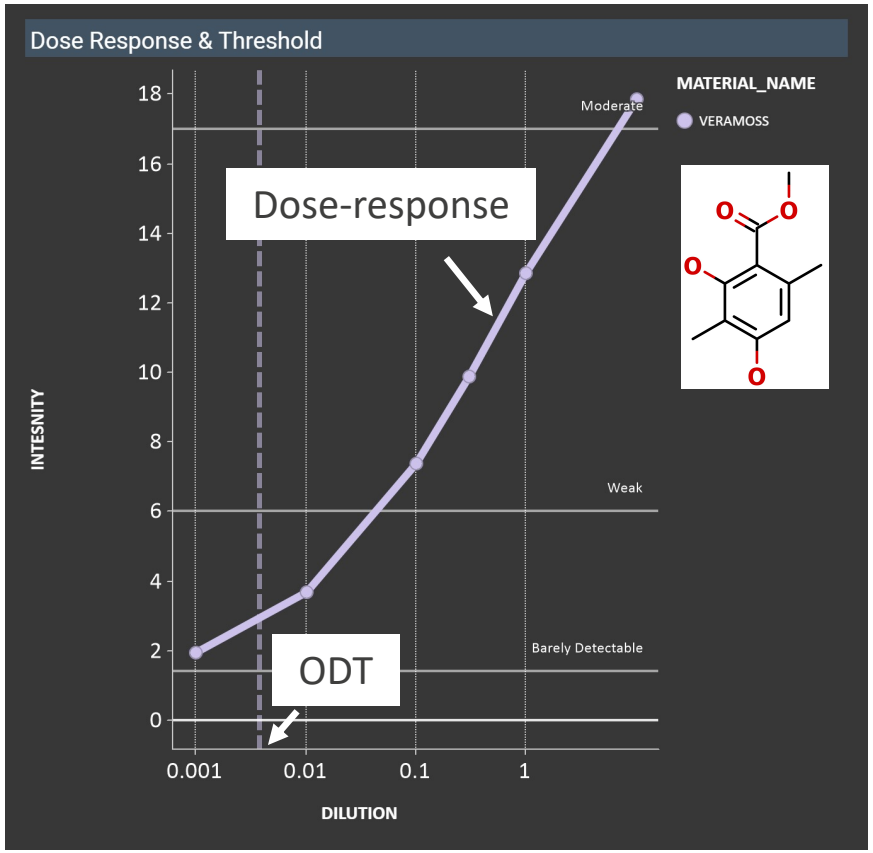
- Odor Intensity and Odor Detection Threshold are the key olfactive properties which define the performance of fragrance ingredients in various applications

- Odor Intensity dose-response curve
  - A series of evaluations of odor intensity performed at different concentrations by a panel of trained testers
  - Helps perfumers to create fragrance formulas

- Odor Detection Threshold
  - The lowest concentration that is detectable 50% of the time
  - Helps to prioritize high-value ingredients

- Challenges of predicting sensory properties
  - Noisy – subjective assessment by human subjects
  - Inherent variability among human subjects
  - Expensive and time-consuming

Veramoss (IFF) CAS: 4707-47-5



Introduction to IFF Project Objectives

- Alchemite deep learning can simultaneously predict several properties for multiple compounds from very sparse data

1. Assess the ability of Alchemite to impute missing structure-property data, physical chemistry and sensory, of fragrance molecules using molecular descriptors and limited experimental data (**Imputation model**)

2. Understand the ability of Alchemite to predict physical chemistry and sensory properties of fragrance molecules based ONLY on molecular descriptors (**Virtual model**)

3. Compare with conventional QSAR machine learning models

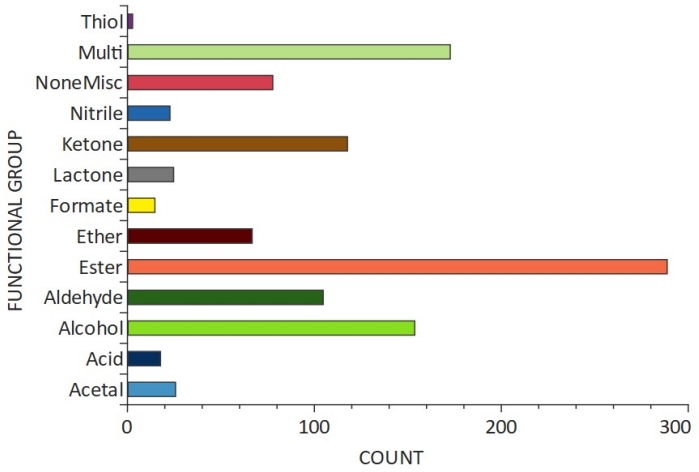
Introduction to Data Set

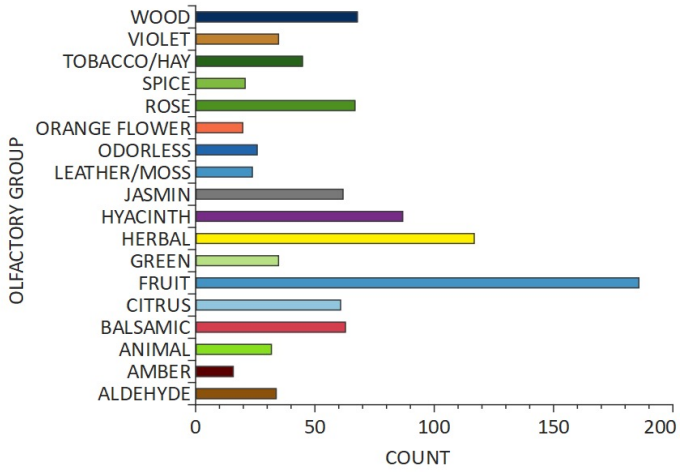
- 1094 molecules from IFF proprietary catalog with at least one measured property: vapor pressure, water solubility, odor detection thresholds, dose-response odor intensities

- Varying degree of sparsity across properties

	VP	WS	ODT	I1	I2	I3	I4
Sparsity (% of missing data)	15%	18%	56%	37%	37%	37%	37%

- Distribution of data set compounds across chemical classes and odor categories

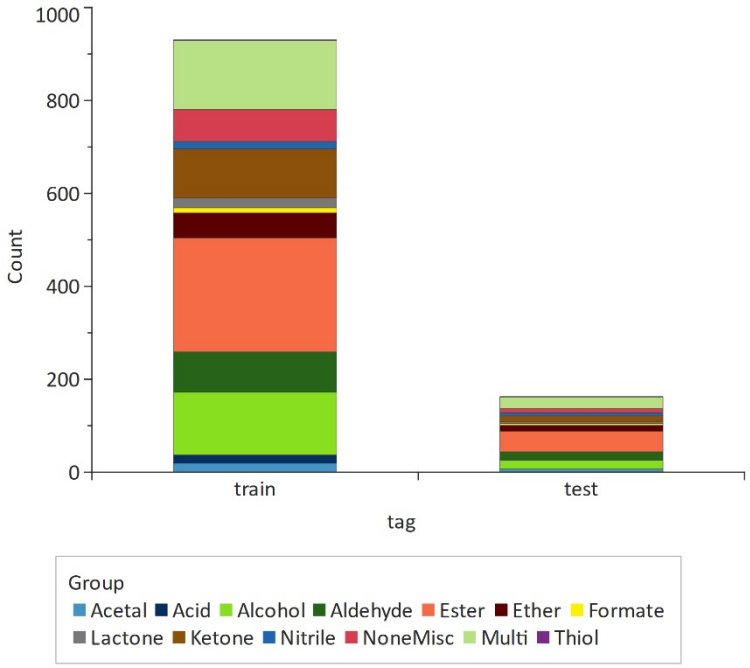




Introduction to Data Set (cont.)

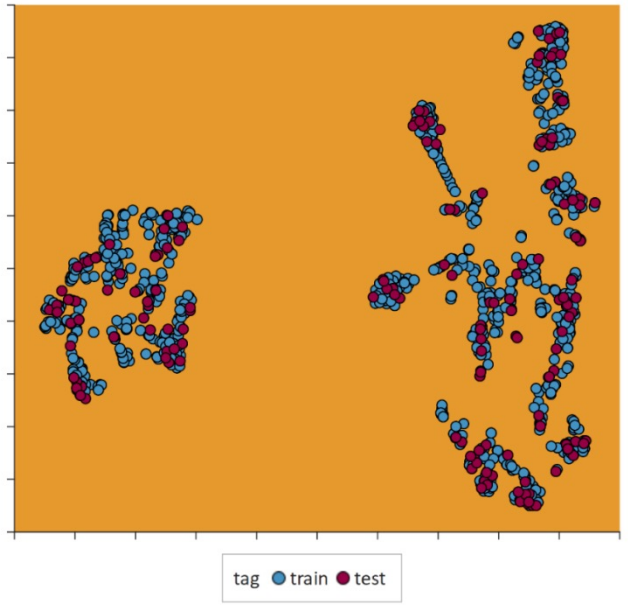
- Training/Test Split: 931/163 molecules (85%/15 %)
  - Molecular structures were provided to Optibrium, data was blinded
  - IFF initially held back the test set for model validation

Chemical composition of training and test sets



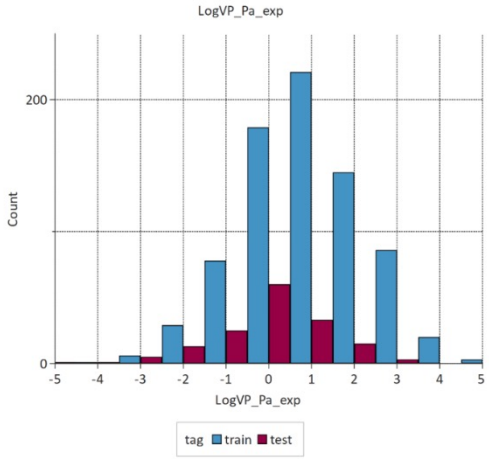
Diverse chemical space sampling

virtual clustering based on chemical structure using t-SNE

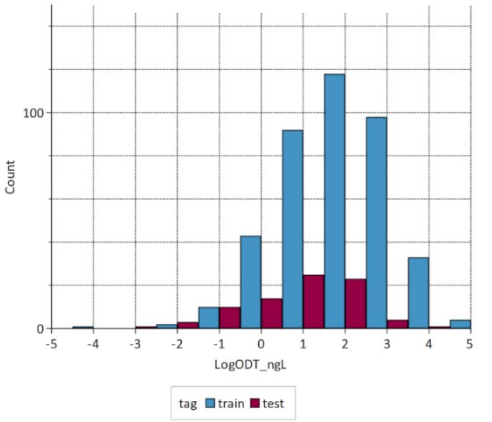


Representative property distribution

LogVP



LogODT

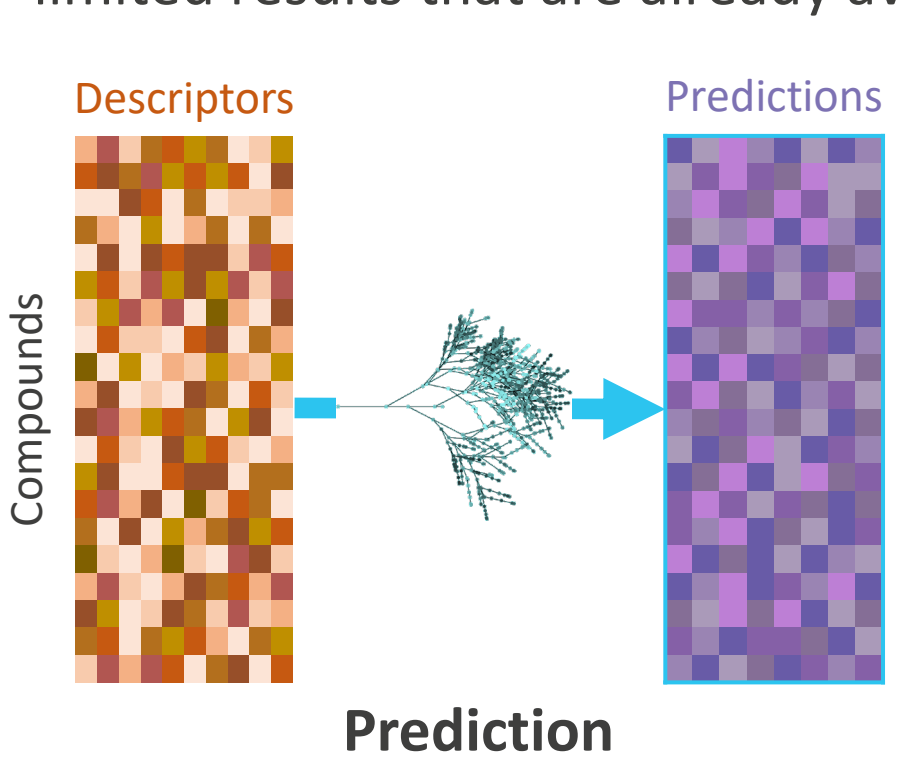


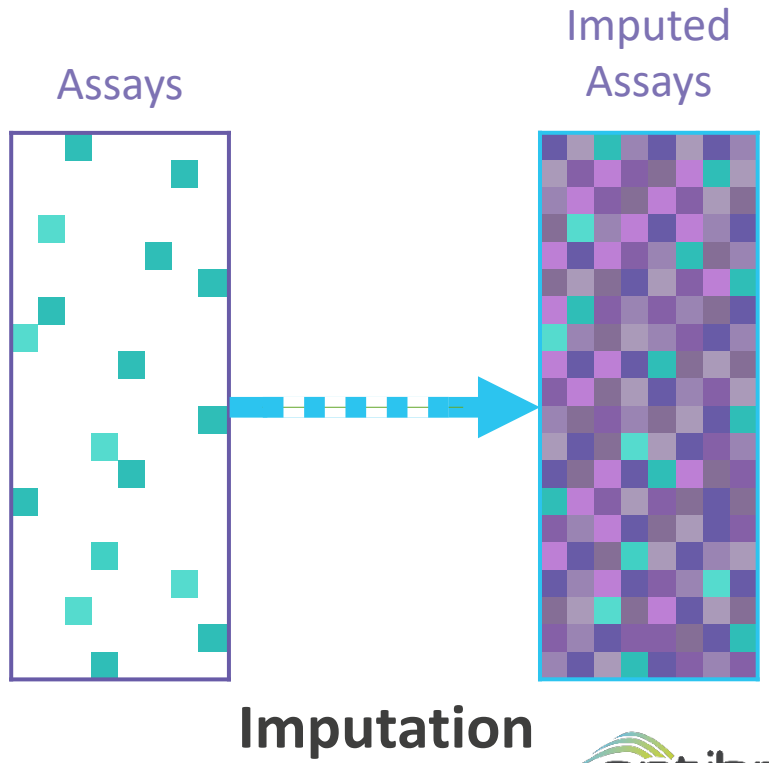
The Alchemite Method for Deep Learning Imputation

Prediction vs. Imputation

- Prediction uses input ‘features’ to predict one or more property values for a compound, e.g. QSAR models

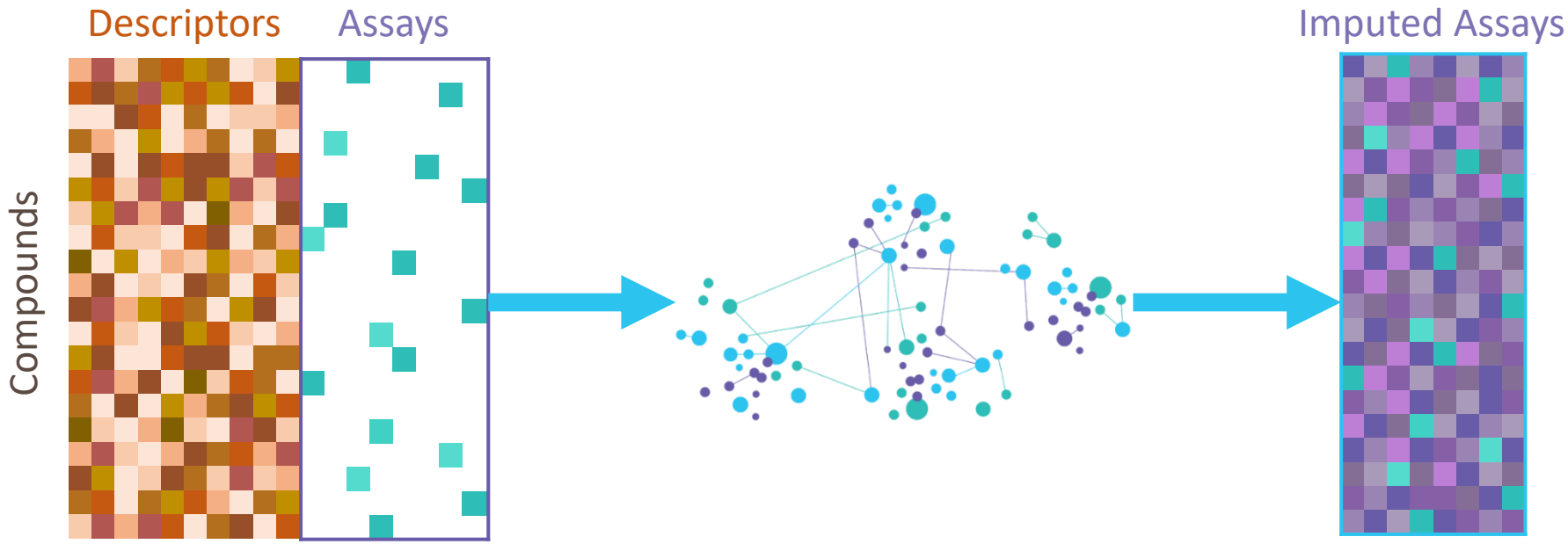
- Imputation is the process of filling in the gaps in sparse experimental data using the limited results that are already available





Alchemite™ Deep Learning Imputation
Optibrium's exclusive partnership with Intellegens

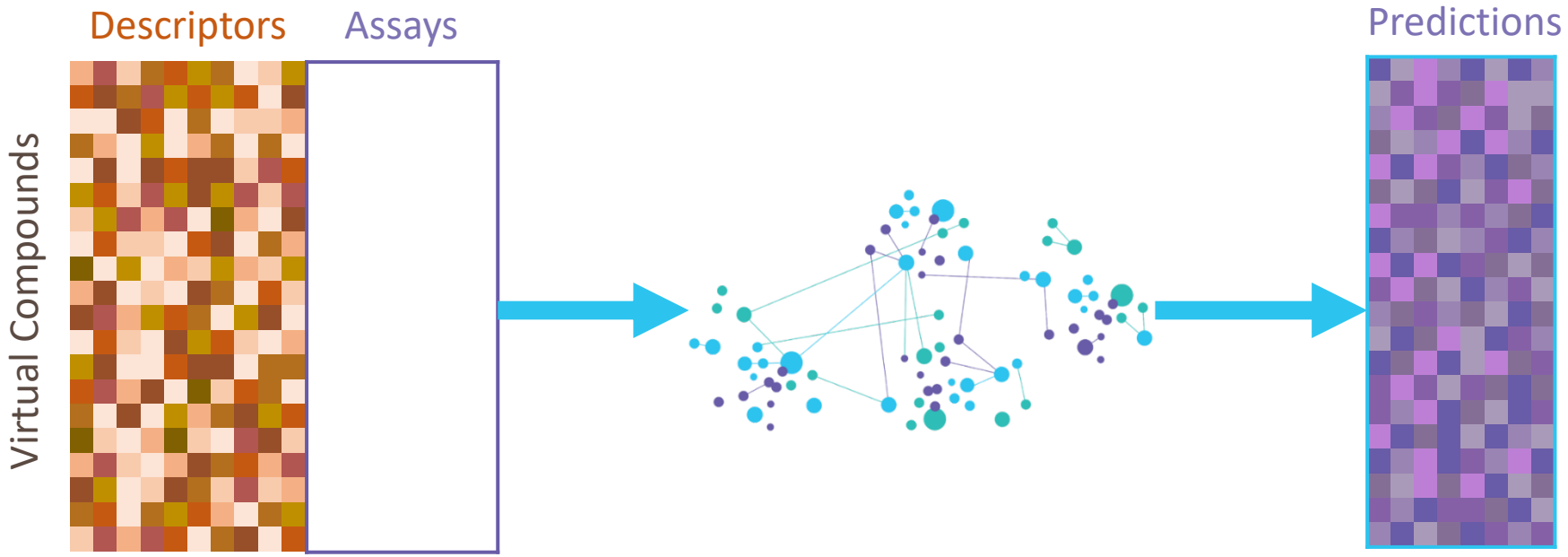
- Learns directly from relationships between experimental endpoints as well as SAR
  - Makes better use of sparse and noisy experimental data than conventional QSAR models
- 'Fills in' the gaps in your data and makes predictions for 'virtual' compounds
  - Generates more accurate predictions to target high-quality compounds



Whitehead *et al.* J. Chem Inf. Model. (2019) **59**(3) pp. 1197-1204, B. Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857

Alchemite™ Deep Learning Imputation
Optibrium's exclusive partnership with Intellegens

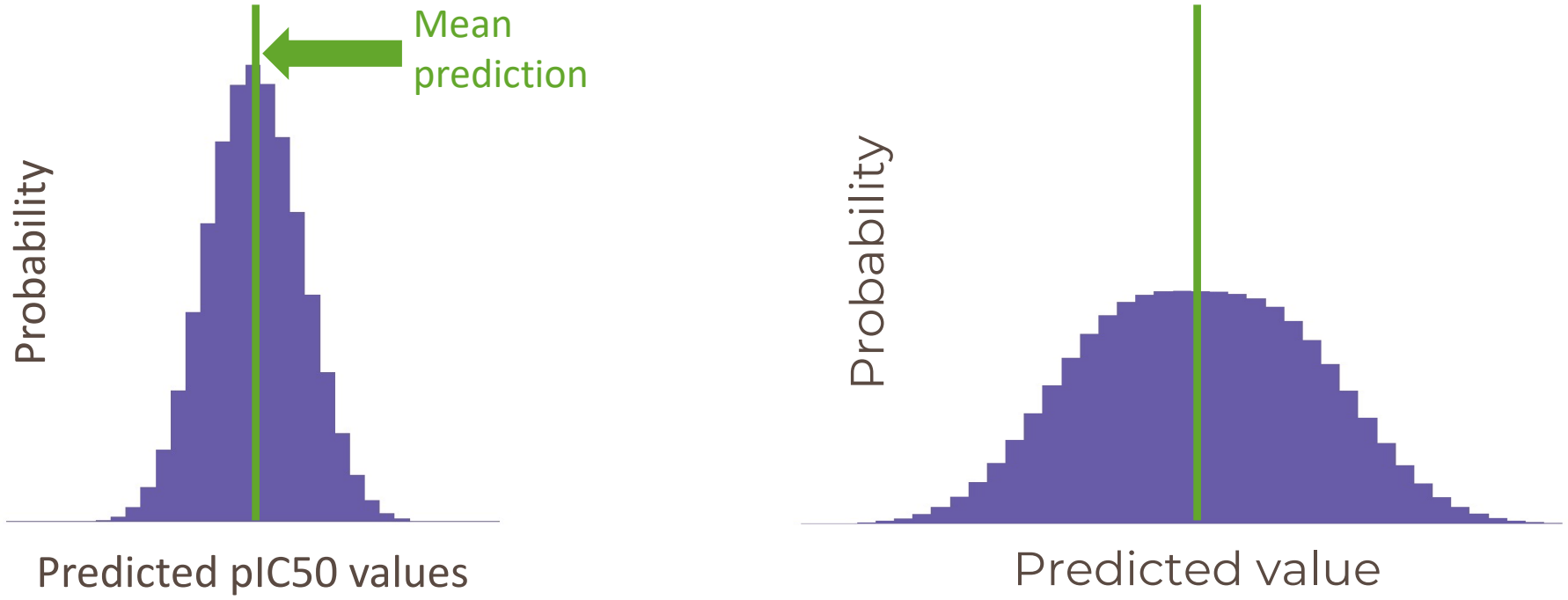
- Learns directly from relationships between experimental endpoints as well as SAR
  - Makes better use of sparse and noisy experimental data than conventional QSAR models
- 'Fills in' the gaps in your data and makes predictions for 'virtual' compounds
  - Generates more accurate predictions to target high-quality compounds



Whitehead *et al.* J. Chem Inf. Model. (2019) **59**(3) pp. 1197-1204, B. Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857

Alchemite™ Deep Learning Imputation
Optibrium's exclusive partnership with Intellegens

- Estimates uncertainty in each individual prediction
  - Highlights the most accurate predictions on which to base decisions
- Confidently targets high-quality compounds and prioritise experimental resources



Whitehead *et al.* J. Chem Inf. Model. (2019) **59**(3) pp. 1197-1204, B. Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857

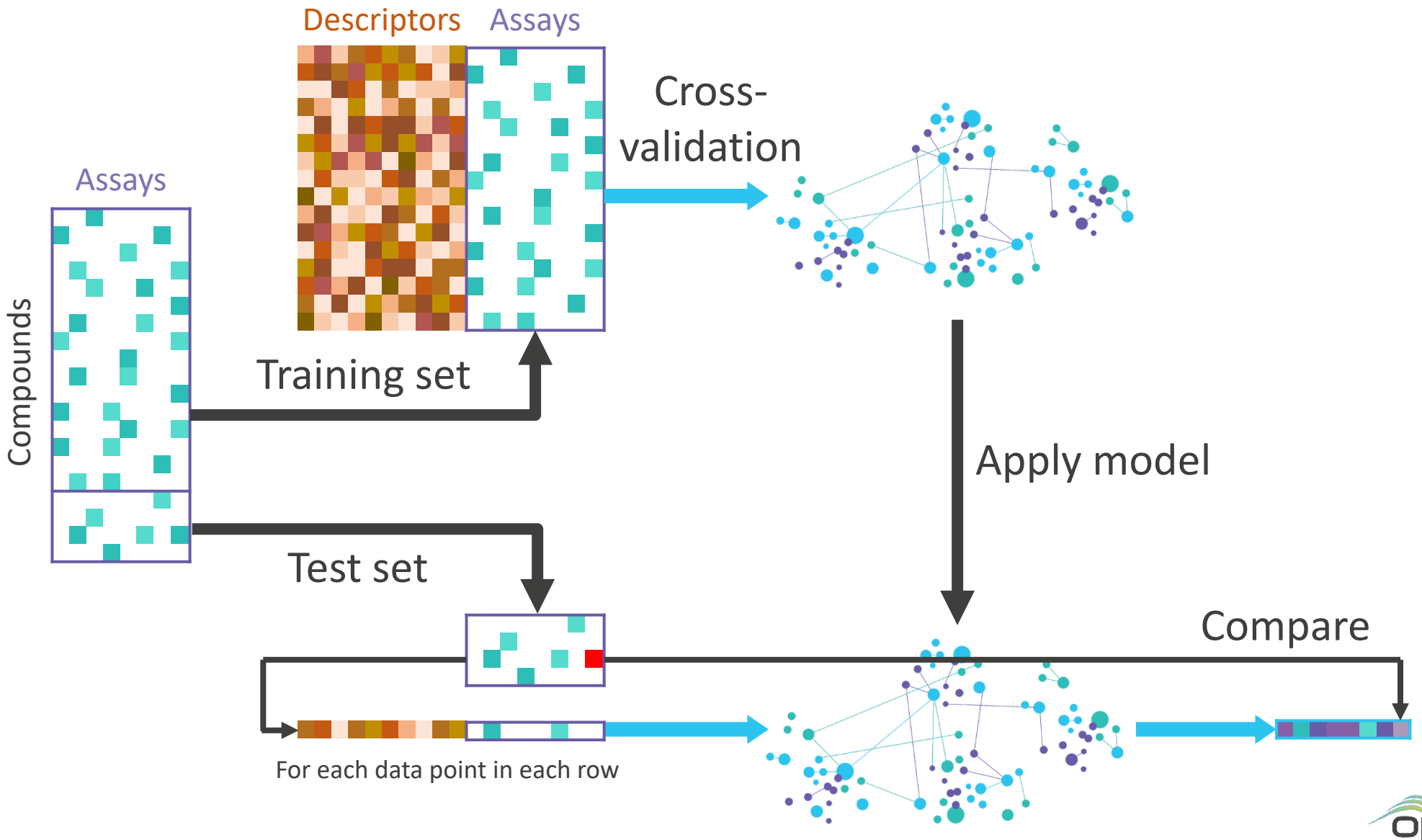
Project Results

Imputation vs Virtual Models

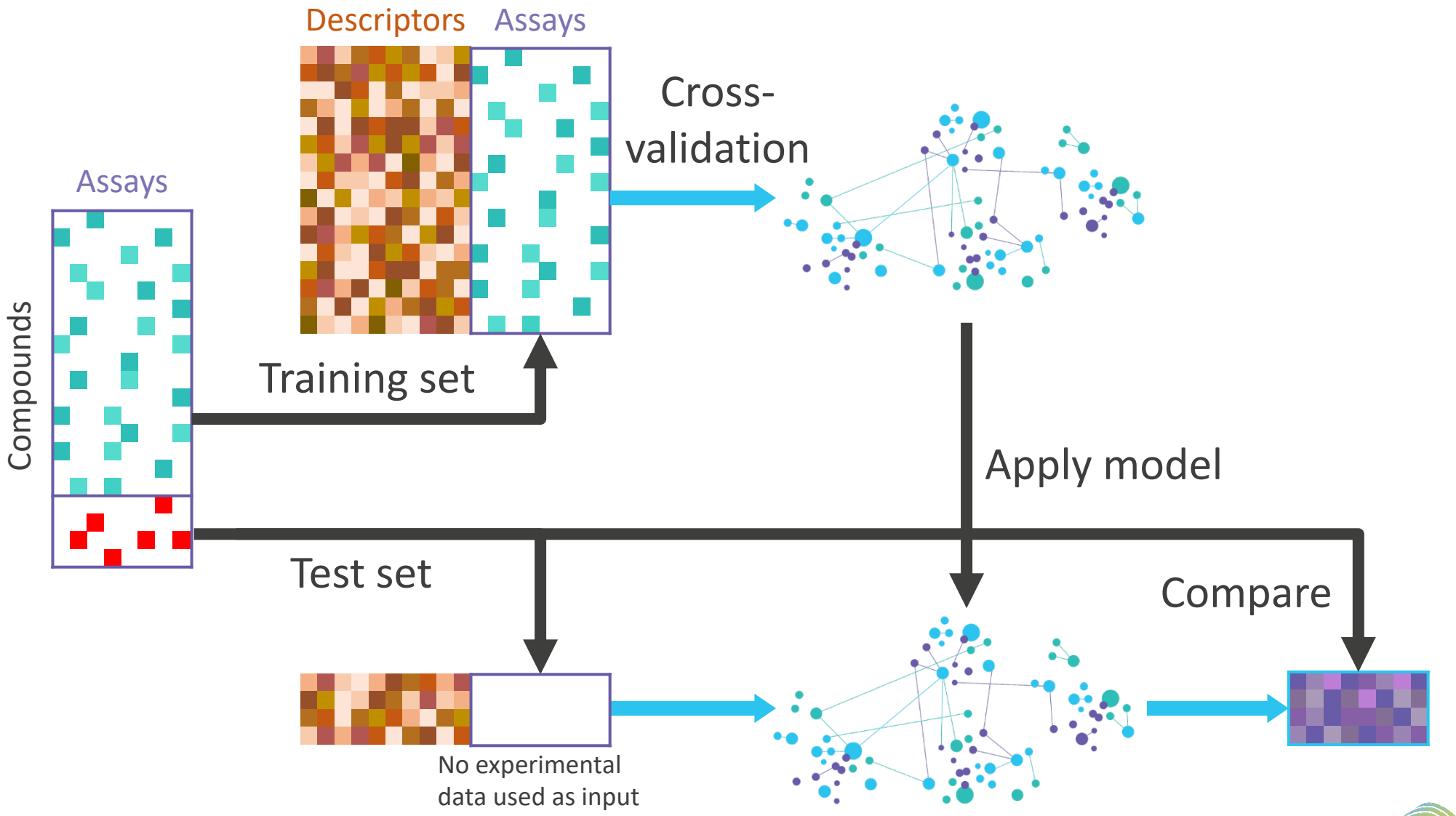
- **Imputation**: These models generate predictions for the test compounds using sparse experimental data as input, in addition to molecular descriptors
  - These models ‘fill in the gaps’ in the experimental data for compounds that have been synthesised and tested in some assays

- **Virtual**: These models are built to expect only molecular descriptors as input
  - These models make predictions based only on compound structure, i.e., for a compound that has not yet been synthesised or tested

Build/Test Process for Imputation Model

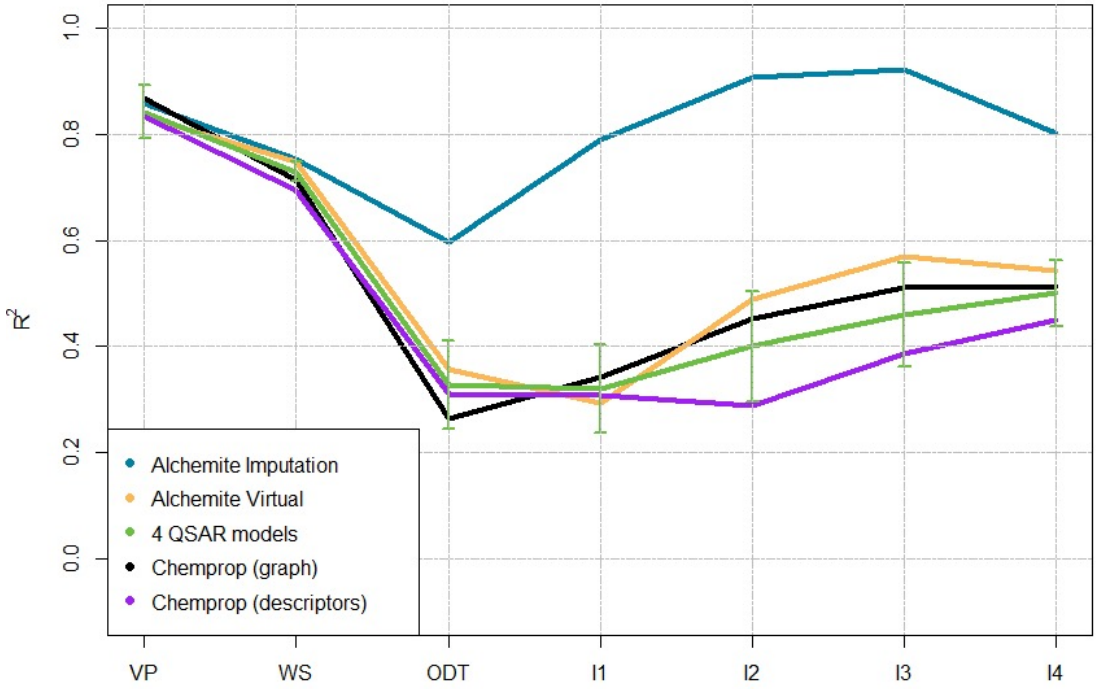


Build/Test Process for Virtual Model



Internal Validation Results

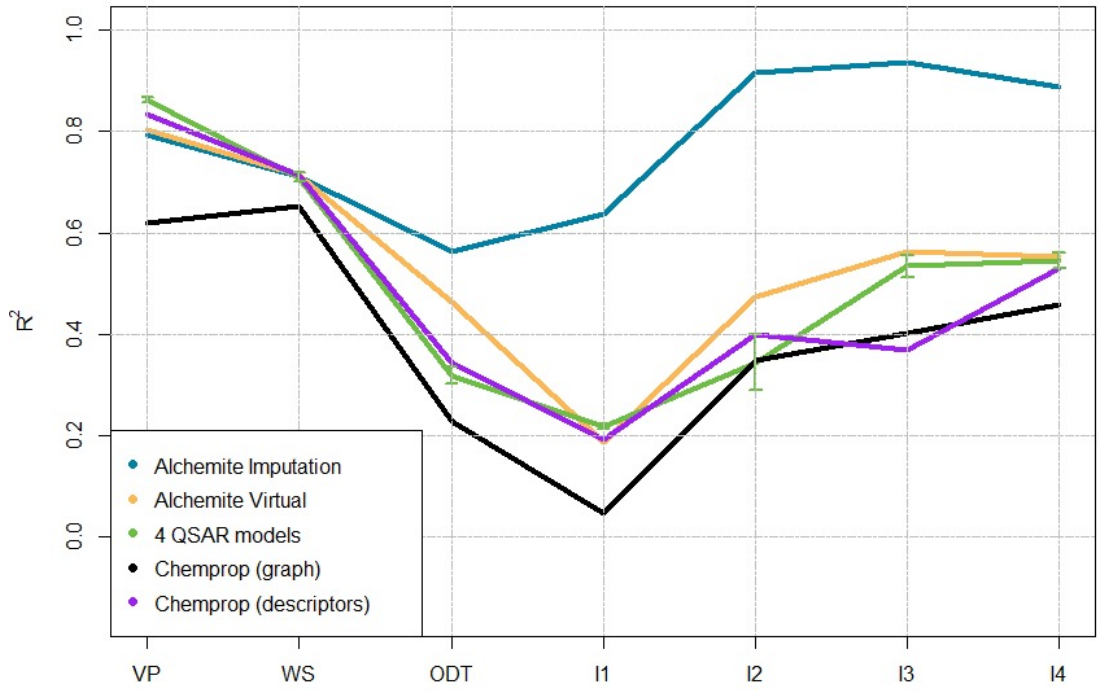
- Clear advantage for **Alchemite Imputation** over all virtual models for sensory properties
- All models perform equivalently for physicochemical properties (VP and WS)
- Virtual models perform similarly for sensory properties
  - We observe a small advantage for the Alchemite Virtual model



Coefficient of Determination (R^2)							
	VP	WS	ODT	I1	I2	I3	I4
Alchemite Imputation	0.85	0.75	0.60	0.79	0.90	0.92	0.8
Alchemite Virtual	0.83	0.75	0.36	0.29	0.49	0.57	0.54
Chemprop (graph)	0.86	0.71	0.26	0.34	0.45	0.51	0.51
Chemprop (descriptors)	0.83	0.69	0.31	0.31	0.29	0.38	0.45
4 QSAR models (average)	0.84±0.05	0.73±0.02	0.33±0.08	0.32±0.08	0.40±0.10	0.46±0.10	0.50±0.06

Independent Test Set Results

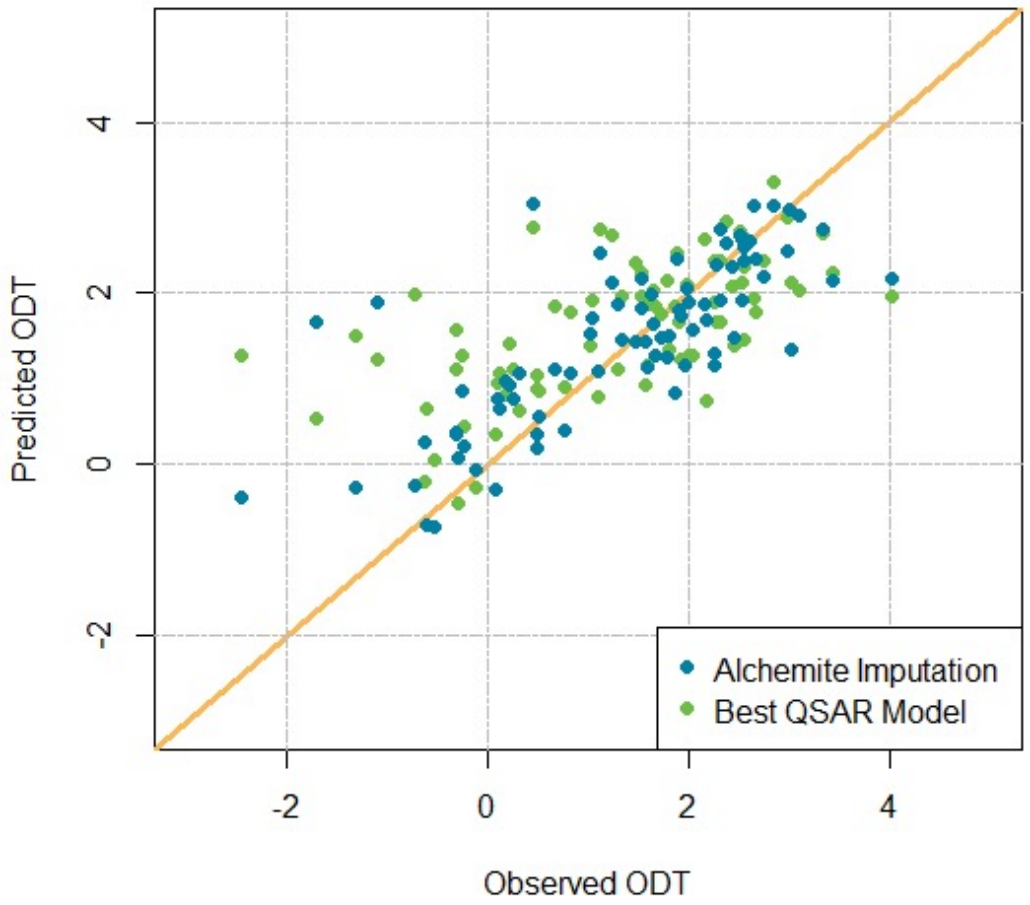
- The test set results are generally consistent with internal validation
  - These confirm the substantial benefit conferred by the use of **Alchemite Imputation**
- R^2 values for I1 are lower for all methods than the independent test set
  - Due to greater variability between test subjects at the lowest concentration
- There is a notable reduction in the performance of Chemprop (graph) relative to the internal validation
  - The graph representations learned from the training set may not capture the SAR of the test compounds



Coefficient of Determination (R^2)							
	VP	WS	ODT	I1	I2	I3	I4
Alchemite Imputation	0.79	0.71	0.56	0.63	0.92	0.94	0.89
Alchemite Virtual	0.81	0.71	0.38	0.19	0.48	0.55	0.53
Chemprop (graph)	0.62	0.65	0.23	0.05	0.35	0.40	0.46
Chemprop (descriptors)	0.83	0.71	0.34	0.19	0.40	0.37	0.53
4 QSAR models (average)	0.80±0.07	0.70±0.02	0.29±0.04	0.2±0.03	0.32±0.08	0.52±0.02	0.53±0.02

ODT Prediction – Alchemite Imputation vs QSAR

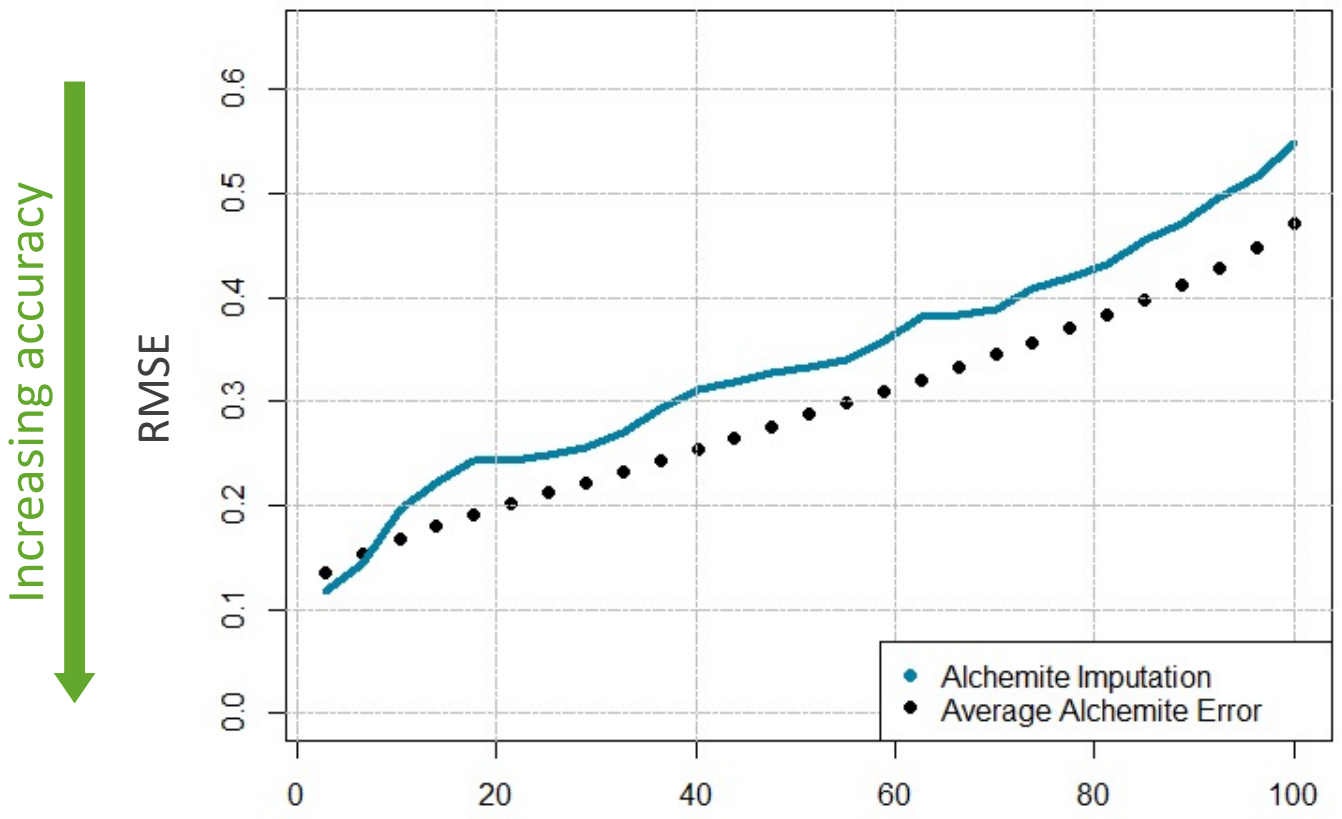
Independent test set



Model	R^2
Alchemite Imputation	0.56
Best QSAR Model	0.34

- Greater scatter in predictions for the best QSAR model than for Alchemite Imputation illustrates the difference in R^2

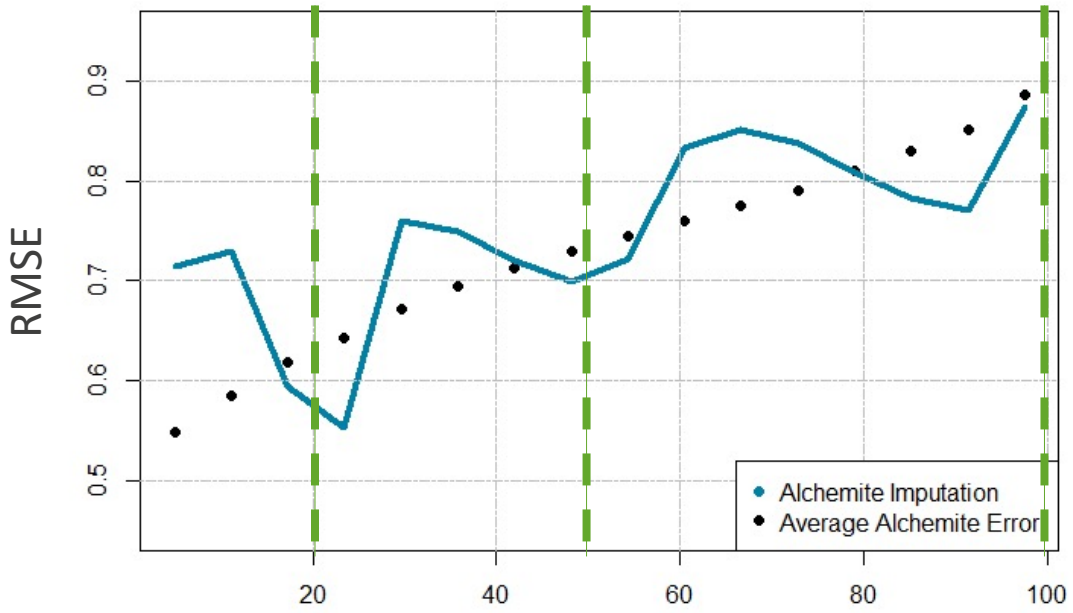
Focussing on the Most Confident Results



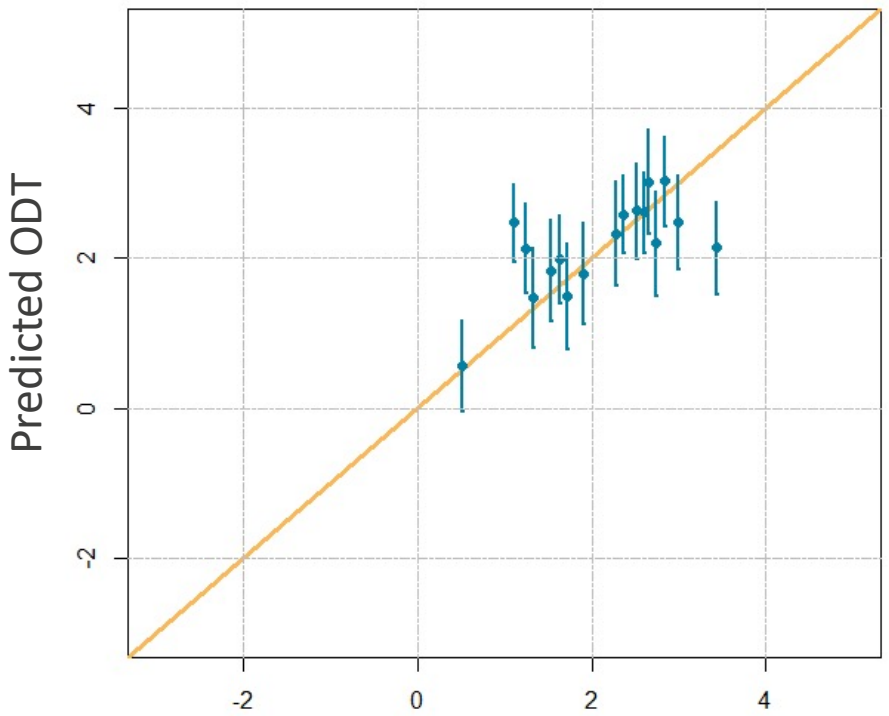
RMSE: Root-Mean-Square Error

Focusing on the Most Confident Results

ODT Endpoint



Most confidently predicted percentage of test set



Observed ODT

- Excellent correlation between model confidence (error bars) and observed accuracy
- The model can reliably identify the most accurate predictions

Conclusions

- Sensory property data are noisy, due to inter-individual variability between test subjects, and present a particular challenge for predictive modelling

- **Alchemite Imputation** offers a substantial advantage over conventional QSAR and multi-target GCNN models
  - Despite the sparsity of the experimental data, Alchemite can extract significant additional information
  - This also confers advantages for **extrapolation in chemical space** and **detection of activity cliffs**

- Alchemite **uncertainty estimates** can be used reliably to identify the most accurate predictions
  - Make decisions based on the most confident results
  - Avoid missed opportunities caused by rejecting compounds based on inaccurate data

Download paper: https://bit.ly/SensoryImputation, doi.org/10.1007/s10822-021-00424-3

optibrium™

Acknowledgements

IFF
Shyam Vyas
Jeff Kattas
Jack Bikker

Intellegens
Thomas Whitehead
Gareth Conduit

Optibrium
Tamsin Mansley
Ben Irwin
Matt Segall

Download JCAMD paper: https://bit.ly/SensoryImputation
Journal of Computer-Aided Molecular Design (2021) 35, pp 1125–1140
doi.org/10.1007/s10822-021-00424-3