Data-driven prediction of materials properties in an automated fashion

Caroline M. Krauter[a], H. Shaun Kwak[b], Thomas J.L. Mustard[c], Alexander Goldberg[d], Steve L. Dixon[e], and Mathew D. Halls[d]

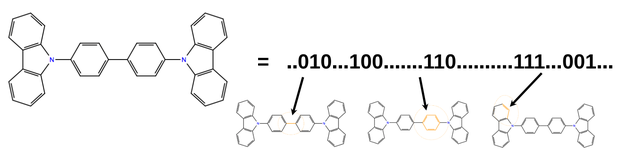
[a] Schrödinger GmbH, Mannheim, Germany. [b] Schrödinger Inc., Cambridge, MA, US. [c] Schrödinger Inc., Portland, OR, US. [d] Schrödinger Inc., San Diego, CA, US. [e] Schrödinger Inc., New York, NY, US.

Abstract. There is pressing need for the use of a rapid and reliable data-driven prediction scheme for materials development and optimization. It can drastically speed up the process of assessing key control variables for materials properties, avoiding the needs of scanning the entire design space with costly experimental measurements and computationally intensive simulations. However, complexity of data generation, model building, and validation procedures for the learned-model approaches could pose a major obstacle, making them less accessible from the materials science and engineering community.

In this work, we showcase the latest Schrödinger developments in computerized algorithms for automated generation and ranking of predictive regression models, which is readily available for the design of new chemistry in molecular space. The methodology is demonstrated with large-scale virtual screening of a design space for thermally activated delayed fluorescence (TADF) materials, catalytic activity prediction for Ziegler-Natta catalysts, and selectivity prediction for the Tsuji reaction. The automated data-driven predictive scheme provides unbiased measures to quickly assess the key design rules for a wide variety of applications, which could significantly lower the barrier towards large-scale virtual screening for developing novel materials solutions.

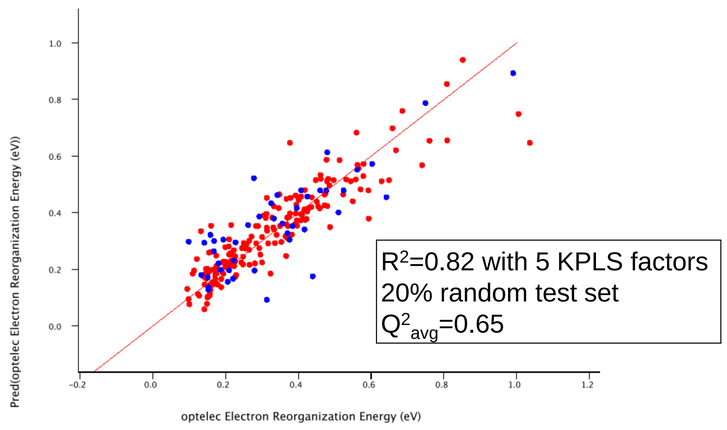
Fingerprint-based QSPR for interactive designs

Binary fingerprints as predictive descriptors [1]
- Assembled binary strings solely based on 2D descriptors
- Dendritic fingerprints: fingerprinting scheme widely tested and validated for small molecule space
- Topology and chemistry zipped in binary strings

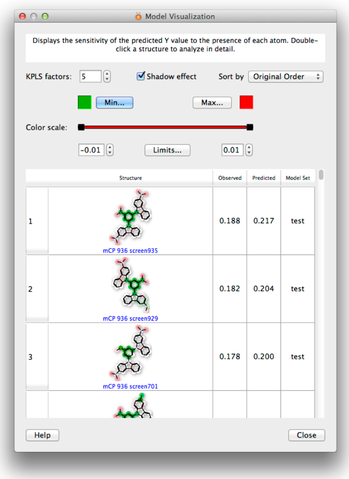


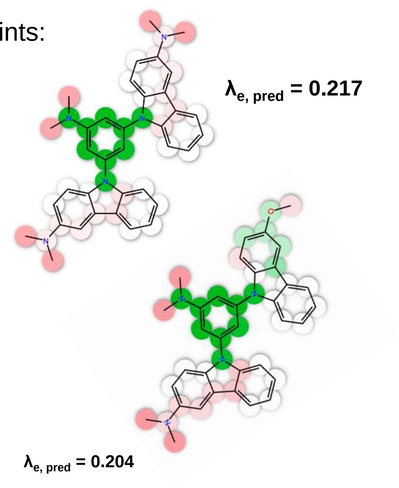
Kernel-based partial least square (KPLS) regression to push QSAR beyond black-box model [2]

Example: QSAR model for optoelectronic properties
KPLS regression model for 200+ compounds with dendritic fingerprint



Model Visualization with fingerprints:





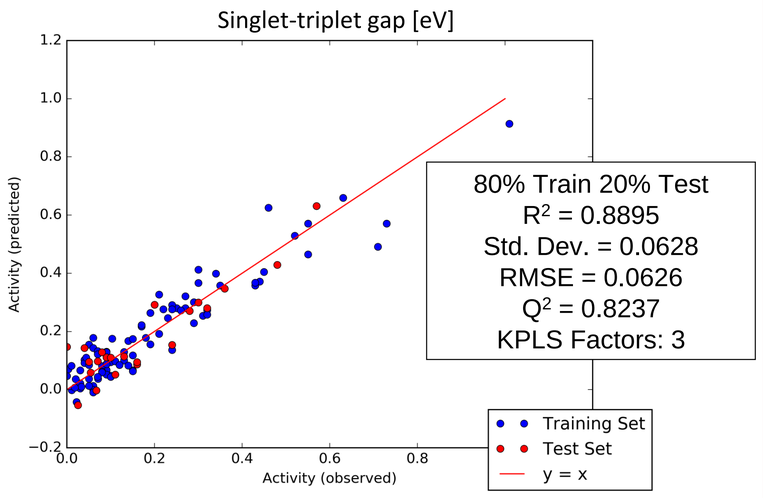
Automated QSPR

Follows best-practices QSPR methods for model building [3]:

Automating the building, validation and deployment of single or consensus models:
- Descriptor generation (from 497 topological descriptors and 4 binary fingerprints)
- Feature selection
- Model generation
- Cross-validation over multiple test sets
- Scoring to rank models with respect to accuracy

Example: TADF discovery

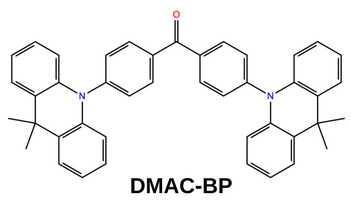
AutoQSAR model built using 113 TADF molecules with known ΔE_{ST}
- 80% Train 20% Test
- The 113 TADF "parents" broken into fragments and fragments were recombine via chemically viable enumeration scheme
- 58,110 "children" were produced
- Best QSAR model was used on children
- Structures with predicted $\Delta E_{ST} \leq 0.05$ eV were subjected to QM (B3LYP/6-31G* OPT and single-points averaged over B3LYP/6-31G* TDDFT and M06-2X/6-31G* TDDFT)



➢ ~4k of 58K children predicted to have a $\Delta E_{ST} \leq 0.05$ eV

TDDFT-predicted $\Delta E_{ST} \leq 0.10$ eV:
B3LYP/6-31G* **75%** (>2,900)
M06-2X/6-31G* **5%** (>150)
Average **8%** (>300)

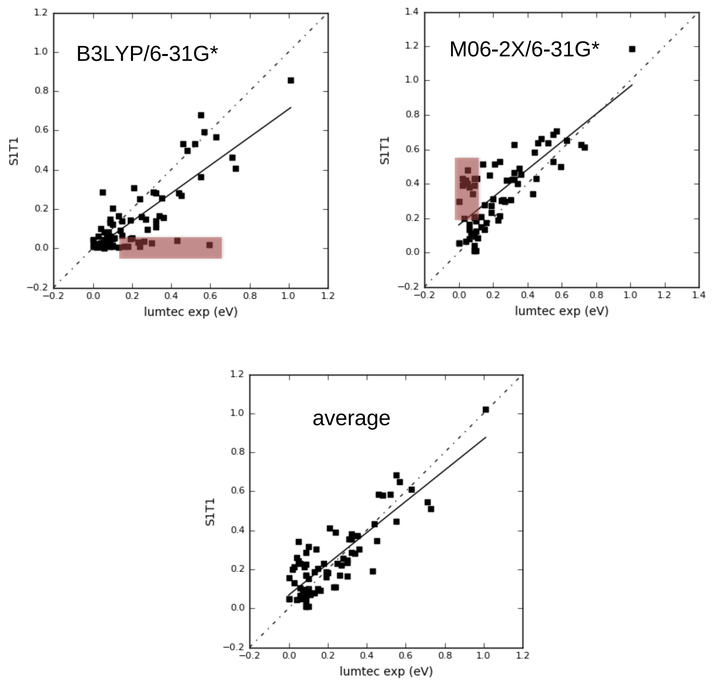
➢ 18 structures are known TADF molecules



DMAC-BP

	ΔE_{ST}	λ_{EL}
Experimental	0.07 eV	508 nm
Computed	0.008 eV	519 nm

Motivation for averaging over B3LYP and M06-2X:
- B3LYP tends to underestimate ΔE_{ST}
- M06-2X tends to overestimate ΔE_{ST}
- Averaging: less false negative or false positive results

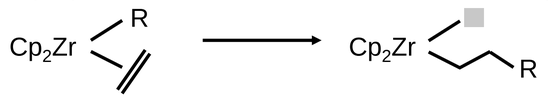


Examples: homogeneous catalysis

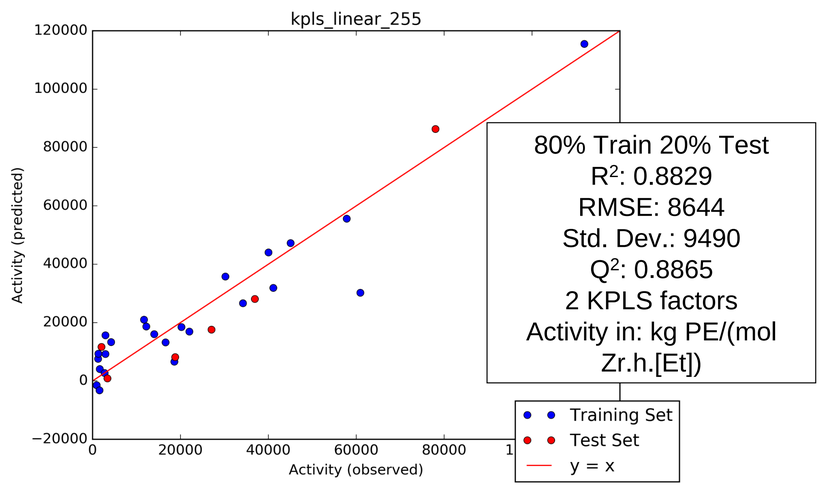
AutoQSAR was used to determine the applicability of machine learning techniques to the design of transition metal catalyst
Two systems studied:

Metallocene-catalyzed olefin polymerization: turnover frequency prediction

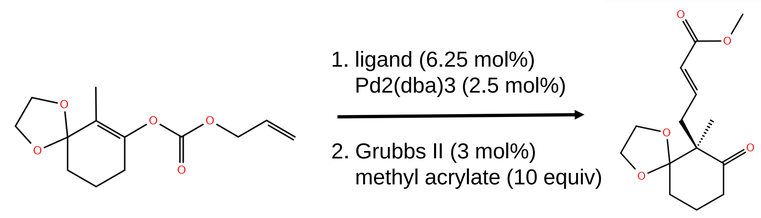
The catalyst activity is inversely proportional to the reaction barrier (and proportional to all side/deactivation reactions):



AutoQSAR model built using 30 experimentally known catalysts [4]

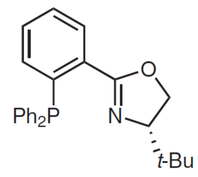


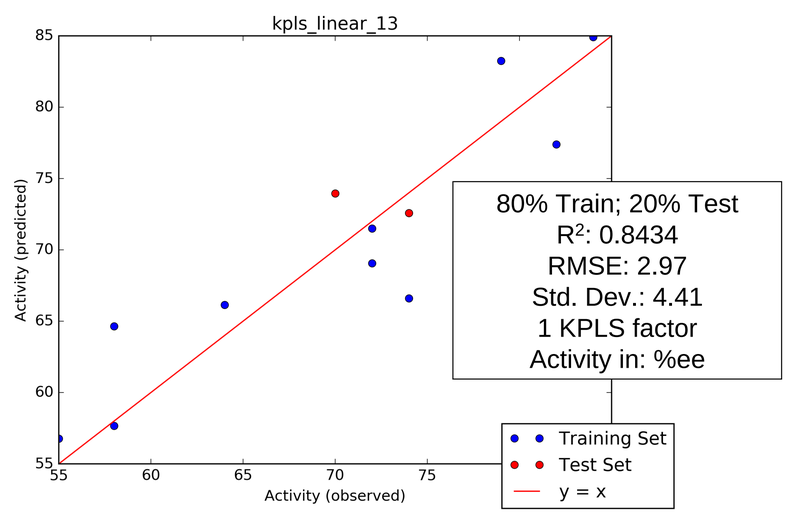
Palladium-catalyzed asymmetric decarboxylative alkylation reaction: selectivity of the Tsuji reaction



➢ Selectivity is determined by the ligand on Pd

AutoQSAR model built using experimentally known catalysts derived from the following ligand [5]:





References

[1] a) J. Duan, S.L. Dixon, J.F. Lowrie, W. Sherman, *J. Molec. Graph. Model.* **29**, 157 (2010). b) M. Sastry, J.F. Lowrie, S.L. Dixon, W. Sherman, *J. Chem. Inf. Model.* **50**, 771 (2010).
[2] Y. An, W. Sherman, and S.L. Dixon, *J. Chem. Inf. Model.* **53**, 2312 (2013).
[3] S.L. Dixon, J. Duan, E. Smith, C.D. Von Bargen, W. Sherman, and M.P. Repasky, Future Med. Chem. **8**, 1825 (2016).
[4] A. J. van Reenen, Recent Advances in Metallocene Catalyzed Polymerization of Olefins and other Monomers, lecture prepared for the 2nd annual UNESCO training school, March 29-31, 1999.
[5] N.T. McDougal, S.C. Virgil, and B.M. Stoltz, SYNLETT 2010, 11, 1712–1716.