

Methods for Simulating Combinatorial Discovery:

Novel QSAR Methods for Exploring Chemical
Space

Prof. Frank Burden
Chemistry Department
Monash University

Dr. Dave Winkler,
CSIRO Molecular Science



Why work on QSAR?

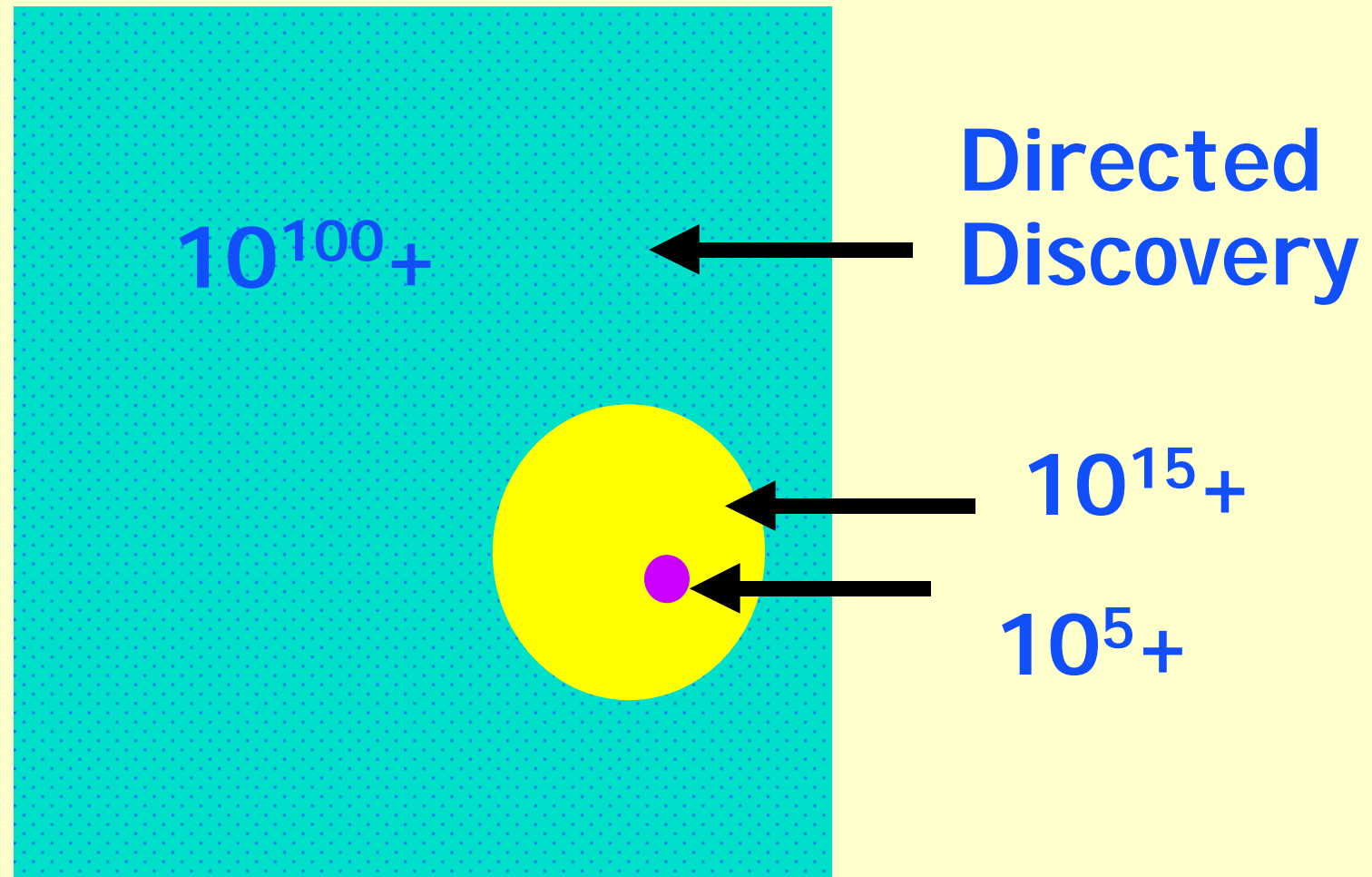
- **Extract information from very large SAR data sets?**
- **Explore the combinatorial 'universe' to find novel leads for a nominated activity**



- Combinatorial chemistry & HTS is revolutionizing molecular discovery.
- But, combinatorial 'universe' is vast ($10^{100}+$ = 70! or a Googol) and only a tiny fraction is accessible to combichem.
- Huge (10^{12}) virtual libraries now available (eg ChemSpace®).



Combinatorial Universe



Strengths of QSAR

- Computationally cheap.
- Proven, over 30 years, many successes in optimization/design
- Handles complexity
- Suitable for in vivo predictions



Apparent weaknesses of QSAR

- Overfitting, chance correlations
- 'Ill-posed' problem - instability
- Interpretation of models
- Subjective non-linearity treatment



Our Research

- Novel information-rich, efficient molecular descriptors
- A new robust method for mapping structure to activity
- Elimination of model optimization, **validation and testing**?
- Applications to dataset mining, simulation of combinatorial discovery, toxicity prediction



QSAR Components

- **Molecular representations**
- SAR mapping methodologies
- Validation, testing and model optimization



Desirable properties of descriptors

- independence of **orientation**
- **easy** to generate quickly
- information-rich
- applicable to **any** structure



Improved Descriptors

- **Molecular fingerprints (1D)**
- **Molecular eigenvalues (2D)**
- **Molecular multipole moments (3D)**



Molecular Fingerprints (1D)

- Simple structural indices
- Functional group
- Holograms



Atomistic descriptors

Count atoms of various types and connectivity. e.g. **NCCH₂CH₂OH**

5 hydrogens	5 x H1
1 carbon with two connections	1 x C2
2 carbons with four connections	2 x C4
1 oxygen with two connections	1 x O2
1 nitrogen with one connection	1 x N1



Extended Atomistic descriptors

Count

N (=O) ; C (=O)

(-C-) -O- (-C-) ; O (-H)

C(ar) N(ar) etc



Atomistic descriptors

Used HQSAR unit fragment length molecular holograms to derive objectively the atomistic descriptors using Andrews binding data on 200 drug-like receptor ligands. Found 26 atomistic descriptors of which 21 contributed to binding significantly.

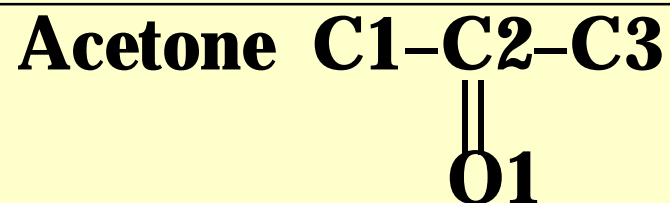
Most positive: N_+ , N_{ar} , S_{ring} , N_3 , C_{ar} , P_{phosp} ,
 C_{ar} , C_4 , C_3 , C_2 , hal , O_3 , O_2

Andrews *et al*, J. Med. Chem., 1984, 27, 1648-57



Molecular Eigenvalues (2D)

Eigenvalues of a modified adjacency matrix



Gives
results
similar to
Randic
indices

	C1	C2	C3	O1
C1	0	1	0	0
C2	1	0	1	$\sqrt{2}$
C3	0	1	0	0
O1	0	$\sqrt{2}$	0	1.3



Molecular Eigenvalues (2D)

These were adapted by Pearlman into BCUT diversity measures. We are exploring use as QSAR descriptors

	a1	a2	a3	an
a1	p11	1	0		0
a2	1	p22	1		$\sqrt{2}$
a3	0	1	p33		0
....					
an	0	$\sqrt{2}$	0		pnn

- scaled connectivity
- steric, electronic, lipophilic properties



Molecular multipole moments

<u>Order</u>	<u>mass</u>	<u>charge</u>	<u>lipophilicity</u>
0	Σm_i	Σq_i	Σf_i
1	0 (cog)	μ	lipophilic dipole moment
2	moment of inertia	electrostatic quadrupole moment	lipophilic quadrupole moment

No molecular alignments required

Silverman *et al* , J. Med. Chem., 1996, 39, 2129-40,
Platt *et al* , J. Computat. Chem., 1996, 17, 358-66



Other Descriptors under development

- Rings
- Functional groups
- Enhanced atomistic
- Binned Charges
- Moments



QSAR Components

- Molecular representations
- **SAR mapping methodologies**
- Validation, testing and model optimization



Neural Networks Advantages

- **Self Learning**
- **Generalising Ability**
- **Subjective decisions on SAR form not needed**
- **Can recognise patterns**
- **Can deal with fuzzy, noisy or missing data**
- **Accounts well for non-linear SAR and generally gives a better model than linear methods**



Frustrations with Validation

- What is optimum number of parameters?
- What is optimum architecture (ANN)?
- Cross validation scales badly
- Which model from test set is optimum?
- What statistical test should be used to find optimum model?



Bayesian inference

- Bayesian methods are optimal methods for solving learning problems
- **Any other method not approximating them will not do as well on average**
- Orthodox statistics provide several models with different criteria for choosing best model - **Bayesian statistics only offers one optimum model.**



Bayes' Theorem

$$P(\mathbf{w} | D, H_i) = \frac{P(D | \mathbf{w}, H_i) P(\mathbf{w} | H_i)}{P(D | H_i)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$



Bayesian Regularised Artificial Neural Networks (BRANNs)

The Bayesian neural network is a back-propagation network with a regulariser added which is related to the prior (distribution of weights).

Reference: David MacKay
Neural Computation, 1992, 4, 448-472



In Practice

$$P(\mathbf{w} | H_i) = \left(\frac{\mathbf{a}}{2p} \right)^{K/2} \exp\left(-\frac{\mathbf{a}}{2} \|\mathbf{w}\|^2\right)$$

Which coupled with the output noise model gives

$$S(\mathbf{w}) = \frac{\mathbf{b}}{2} \sum_{n=1}^N \left(f(\mathbf{x})^{(n)} - y^{(n)} \right)^2 + \frac{\mathbf{a}}{2} \sum_{k=1}^K w_k^2$$



Bayesian Regularised Artificial Neural Networks (BRANNs)

- Evaluates **effective number of parameters** in regression
- **Eliminates need for a validation set**
- Trains to the **same** model
- **Is insensitive to number of hidden nodes beyond the minimum**

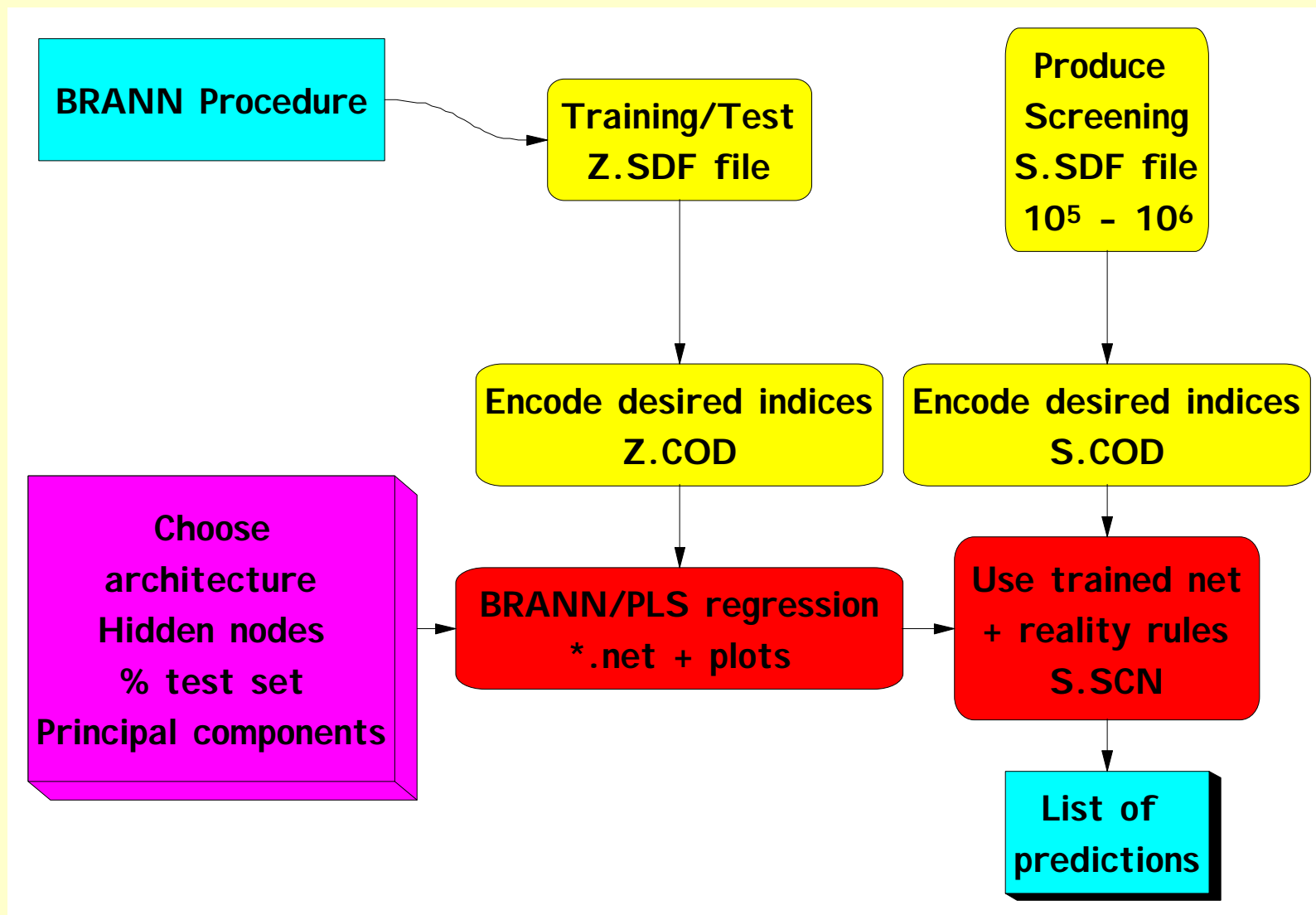


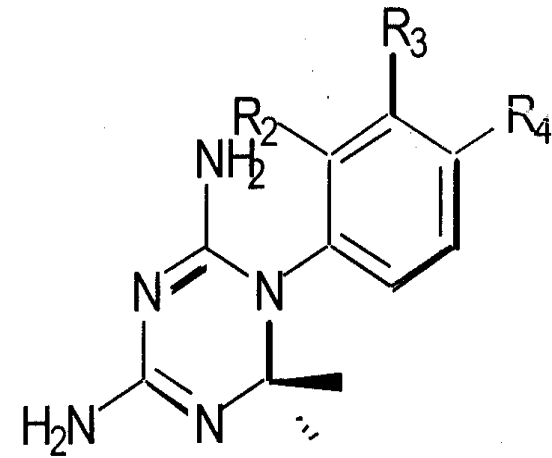
QSAR Components

- Molecular representations
- SAR mapping methodologies
- **Validation, testing and model optimization**



Methodology

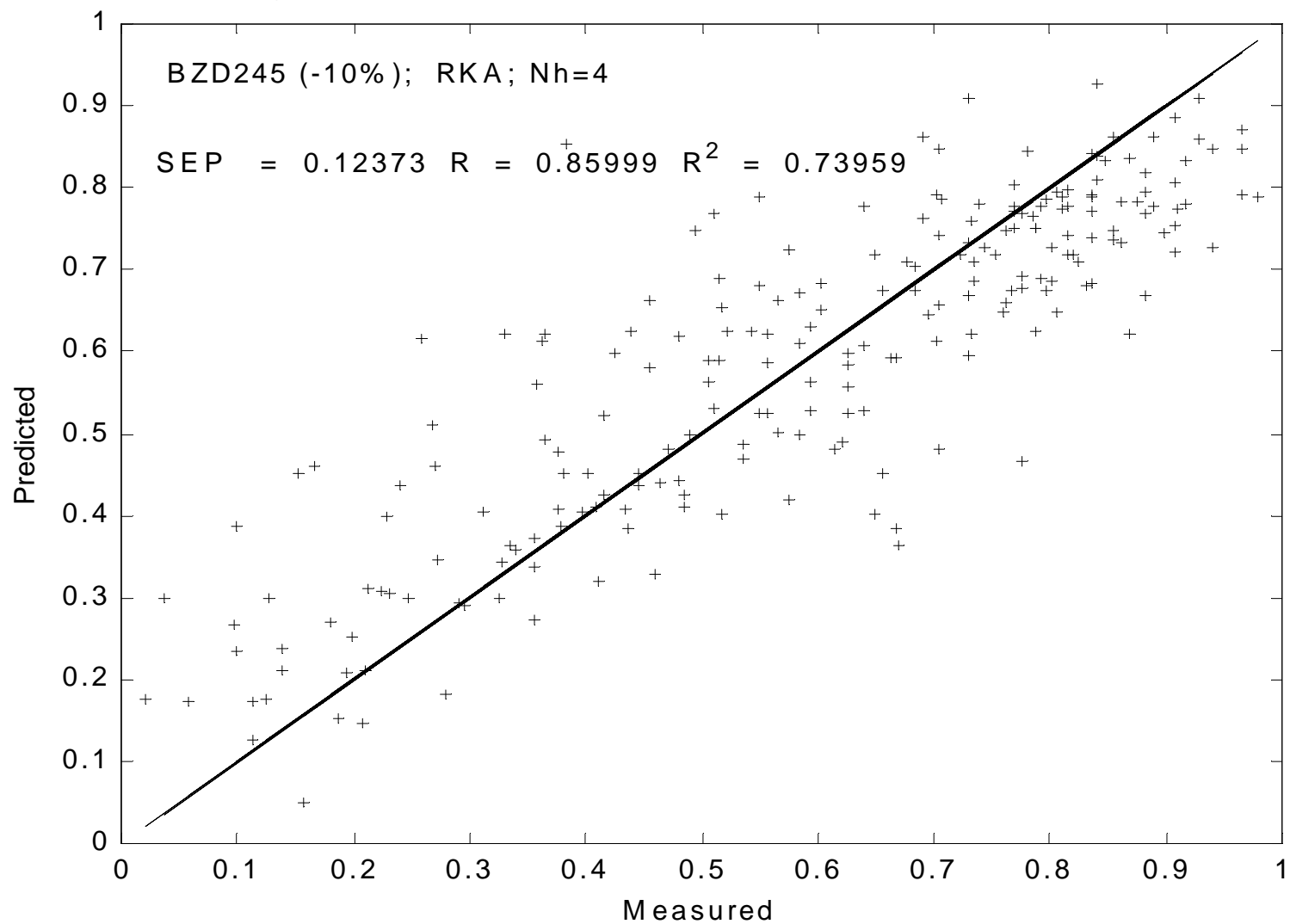




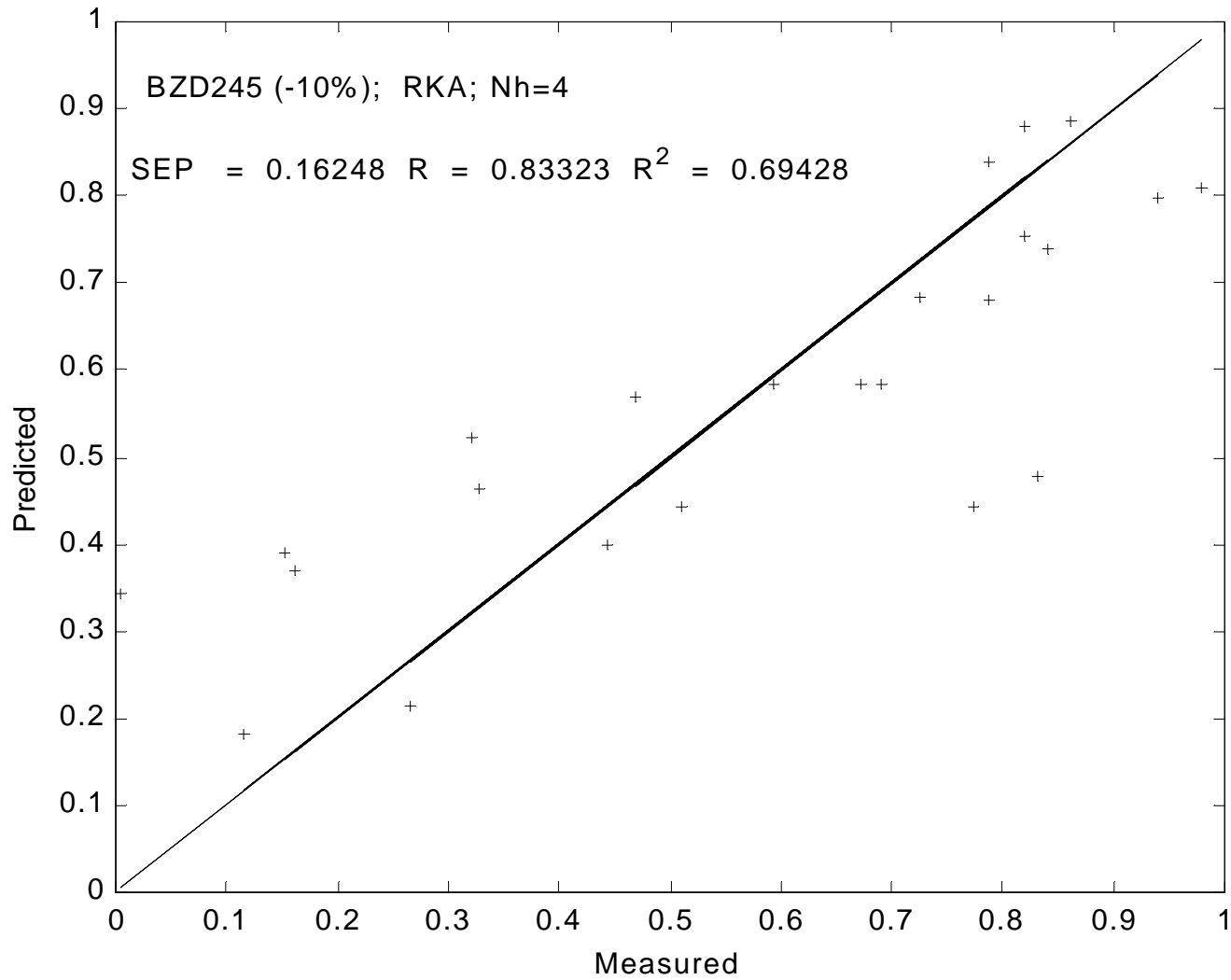
Benzodiazepine/ GABA_A Virtual Receptor



Training BRANN: Y1 with 14 PC Components 14-Sep-1999 22:43:15



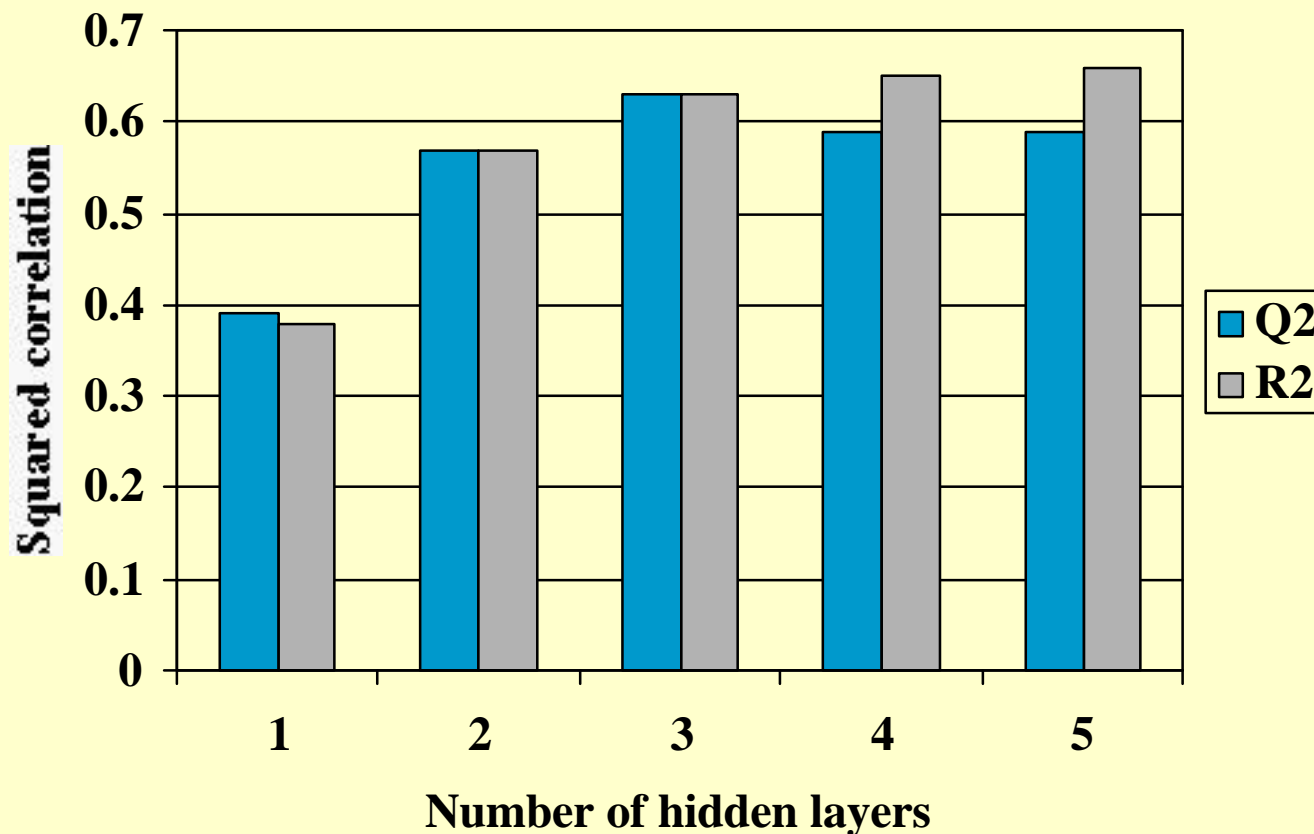
Prediction BRANN: Y1 with 14 PC Components 14-Sep-1999 22:43:16



No	NPC	Nh1	Nh2	SEE	R ²	SEP	Q ²	Npar	SE(All)
1.0000	14.0000	4.0000	0	0.1237	0.7396	0.1625	0.6943	55.6453	0.1280
2.0000	14.0000	4.0000	0	0.1247	0.7350	0.2050	0.5122	55.0465	0.1347
3.0000	14.0000	4.0000	0	0.1341	0.6935	0.1695	0.6708	52.0162	0.1380
4.0000	14.0000	4.0000	0	0.1343	0.6922	0.1878	0.6214	53.4172	0.1404
5.0000	14.0000	4.0000	0	0.1241	0.7365	0.1654	0.6860	56.0884	0.1288
6.0000	14.0000	4.0000	0	0.1240	0.7373	0.1871	0.5719	55.7540	0.1316
7.0000	14.0000	4.0000	0	0.1316	0.7062	0.1875	0.5687	53.5425	0.1381
8.0000	14.0000	4.0000	0	0.1336	0.6955	0.1907	0.6070	53.8444	0.1402
9.0000	14.0000	4.0000	0	0.1311	0.7072	0.1986	0.5194	53.3952	0.1391
10.0000	14.0000	4.0000	0	0.1260	0.7298	0.2165	0.4654	55.0955	0.1375



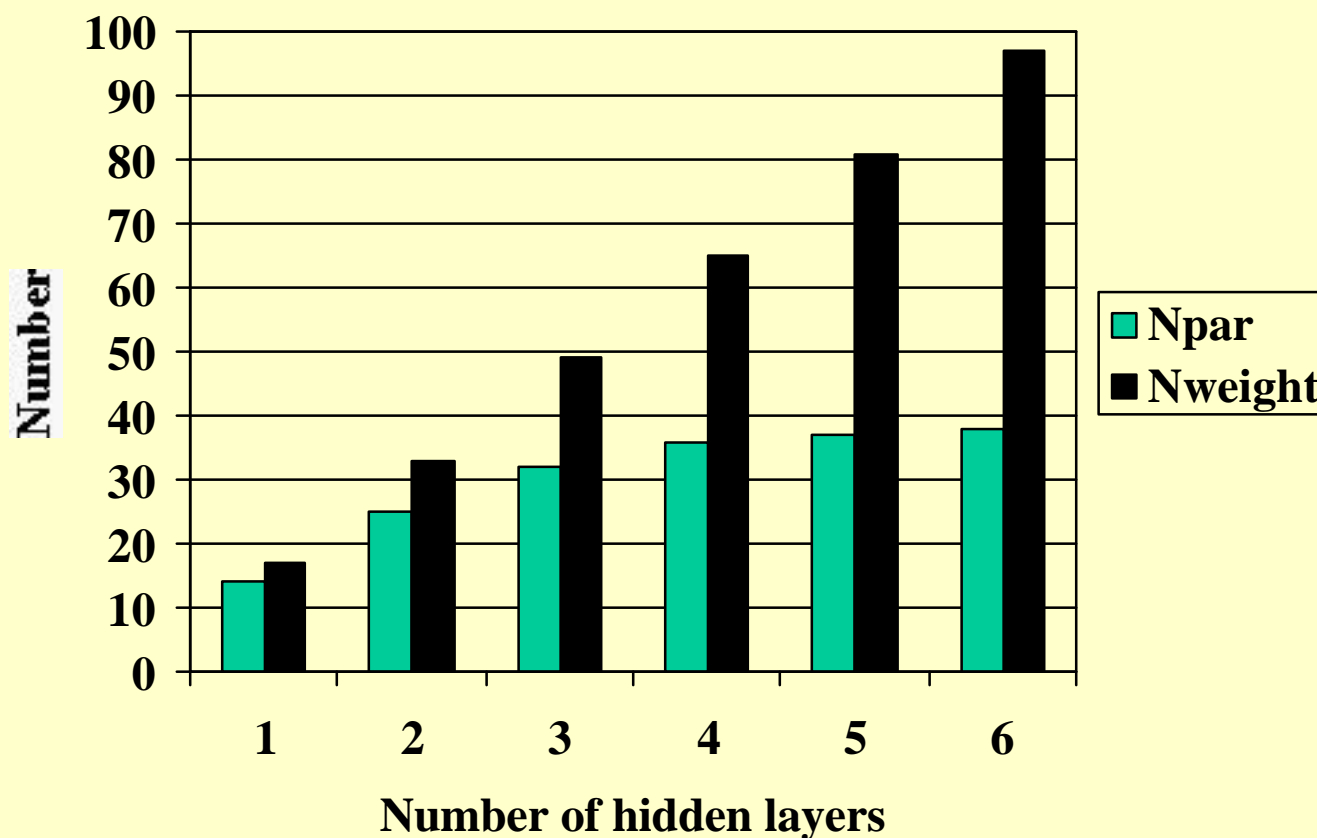
Net architecture optimization



192 mixed structure M1 muscarinic agonists, RKA indices



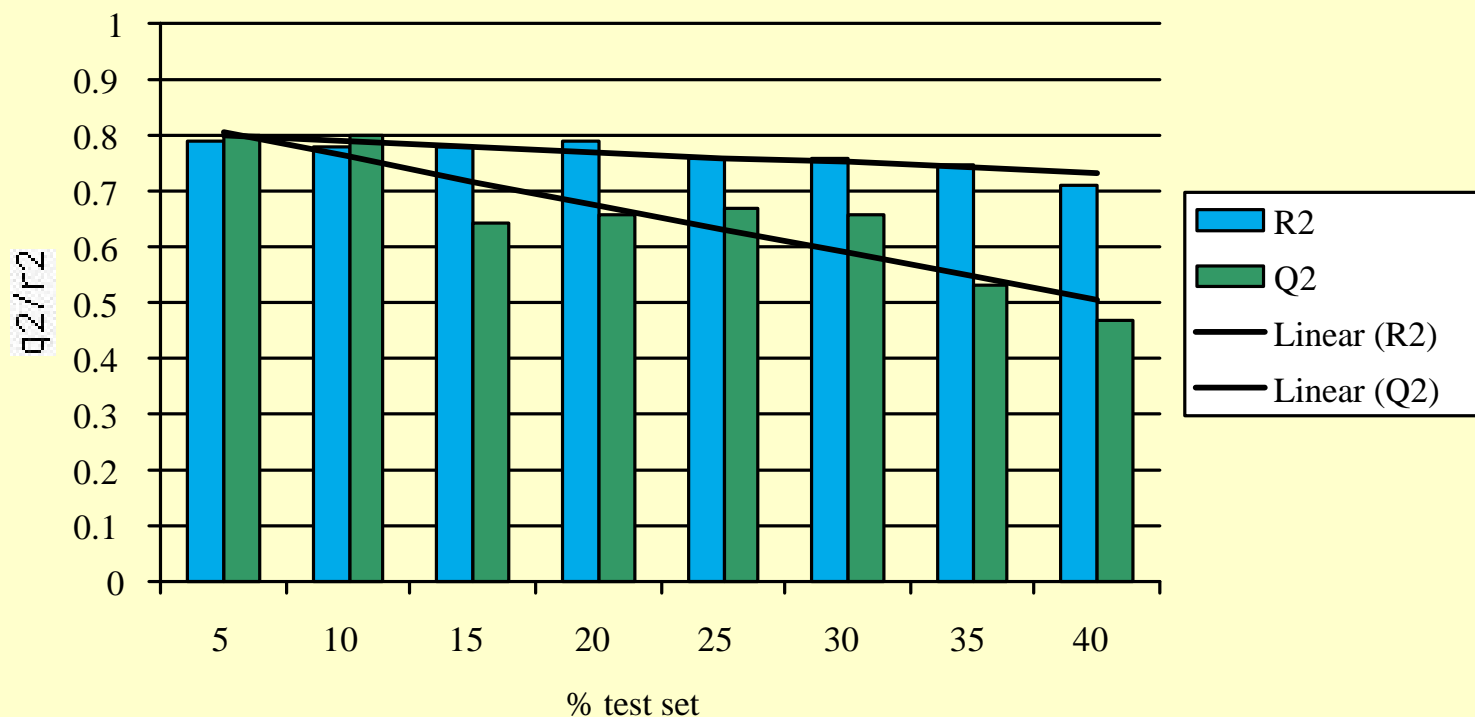
Number of effective parameters



192 mixed structure M1 muscarinic agonists, RKA indices



Do Bayesian nets need test sets?



245 mixed structure benzodiazepine agonists, RKA indices



Towards Virtual Combinatorial Screening

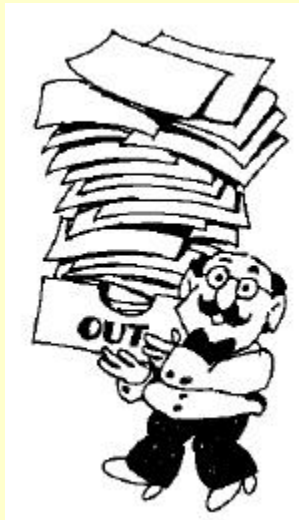


Virtual Combinatorial Approaches

- **Virtual receptors** which allow rapid screening of large virtual combinatorial libraries.
- **Genetically-optimized virtual libraries** which can explore any promising area of structure space

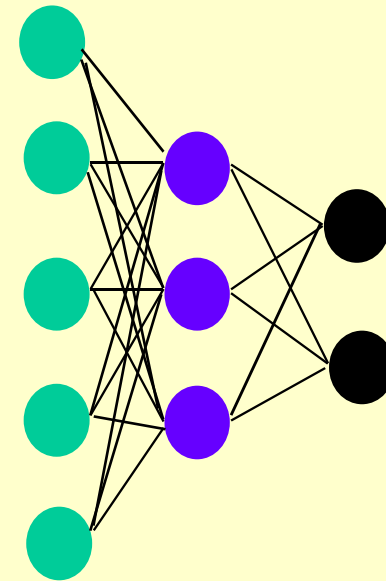
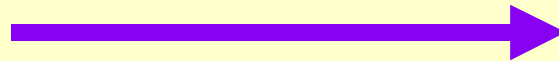


Virtual Receptor Generation



**literature
data**

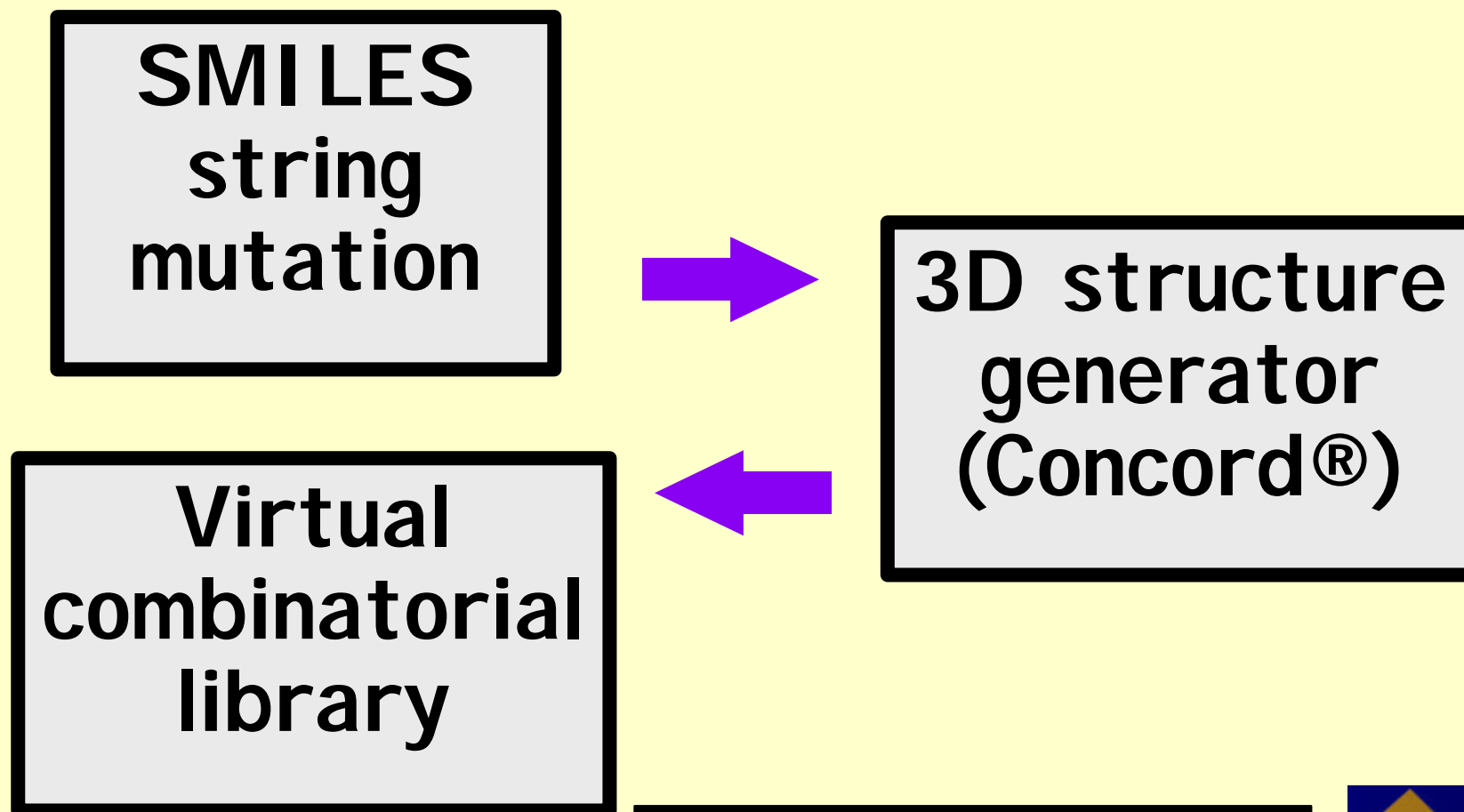
representation



receptor model

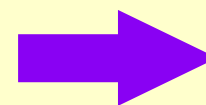
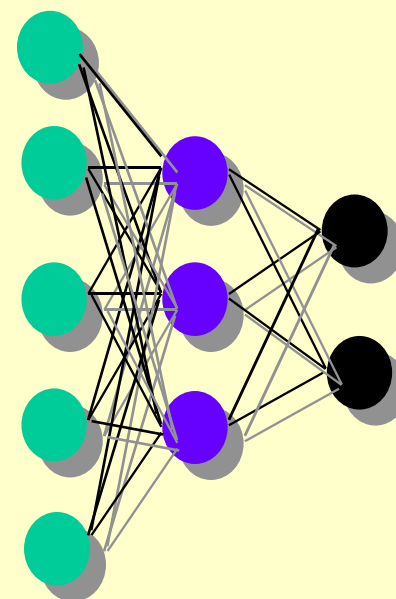
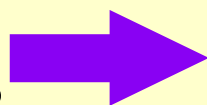


Virtual Library Generation



Virtual Library Screening

virtual
database
 $10^{11} - 10^{12}$
compounds



novel
drug
leads

receptor
model



Considerations

- **Descriptor coverage** of molecular properties relevant to receptor binding
 - **Training set coverage** of sufficient diversity to fully map receptor requirements
- However**, we are attempting to produce an enriched library of hits not a precise QSAR model



Other papers

See web site:

<http://positronium.chem.monash.edu.au/burden/>

or

<http://130.194.164.85/burden/>



Acknowledgements

David Winkler, CSIRO, Co-author

Dan Platt, IBM Watson Labs (Molecular moments)

Glen Kellogg, Virginia Commonwealth Uni (Hydropoles)

Graham Richards, Oxford Univ. (Representations)

David Livingston, Portsmouth Univ. (Validation)

David Manallack, Chiroscience (Neural networks)

David Mackay, Cambridge Univ. (Bayesian inference)

